STRUCTURAL SPATIO-TEMPORAL TRANSFORM FOR ROBUST VISUAL TRACKING

Yazhe Tang[1,2]*, Mingjie Lao*[1]*, Feng Lin*[1] *and Denglu Wu*[1]

[1]Temasek Laboratories, National University of Singapore, Singapore
[2]Department of Precision Mechanical Engineering, Shanghai University, Shanghai

ABSTRACT

This paper presents a tracking method which decouples tracking process into translation and scale estimation steps. A coarse estimation step in translation is implemented with particle filter first. Then, a two layers of correlation filters is proposed to robustly estimate object variation in translation and scale accurately. The appearance of object is divided into several local blocks. Each block is a basic unit for data updating and it is capable of accurately locating the subcontext of target based on the trained block filters. A local weight vector is developed to structurally and flexibly formulate spatial-temporal transform feature map with online learning framework. The block-updated filters are assembled to a final tracker for the accurate translation estimation. To handle the adaptive scale variation, a sample pyramid based tracker is built to estimate the scale accurately. Experiments on the public benchmark demonstrate the advantage of proposed algorithm over the state-of-the-art approaches.

Index Terms— Visual tracking, Fourier transform, Computer vision

1. INTRODUCTION

Visual tracking is one of the essential tasks for intelligent visual analysis since it has been widely used in visual surveillance, human computer interaction, motion understanding and intelligent transportation system, etc. A fruitful literature has been reported with promising results in tracking area[1, 2, 3, 4]. However, although much progress has been made in recent years, the problems are still remained to develop a robust tracking algorithm in complex and dynamic scenes. Challenges include appearance changes of the object, pos and illumination variations, occlusions and significant viewpoint changes, etc. To handle difficult tracking scenarios, lots of online updating tracking methods have been developed over the last few years.

The existing tracking algorithms can be roughly categorized into two classes: generative models and discriminative ones. Generative models formulate the tracking problem as feature matching by searching for the image region that best matches a template or an appearance model [5, 6, 7, 8]. These methods perform well when there is no dramatic appearance change and when the background is not complex. Discriminative approaches formulate visual tracking as a binary classification problem which aims to design a classifier for distinguishing the target from the background. From the framework of classifier level, early discriminative methods most rely on the offline trained classifier for online tracking applications [9]. Since the classifier is trained offline, these kinds of methods are hard to be extended to more complex scenarios with appearance variations and occlusion. To accommodate the appearance variations of target, online trained classifiers are introduced to model the varied objects [10, 11].

This paper proposes a novel tracker with two layers of correlation filters, which divides tracking process in structural appearance coding based translation and sample pyramid based scale estimation for a robust visual tracking. The paper begins by reviewing the relevant work in the Section 2. Section 3 presents the details of correlation filter foundation, structural spatio-temporal transform framework and sample pyramid for scale estimation. Section 4 gives the experiments and performance evaluation of our proposed algorithm. Finally, we conclude this paper in Section 5.

2. RELATED WORK

Correlation filters have been widely used in signal processing domain. They have been successfully extended to computer vision for visual detection and tracking recently due to their promising performance with computational efficiency [12, 13]. They take advantage of the benefit that convolution of images in spatial domain is equivalent to an element-wise product in the Fourier domain. Some researches related to correlation visual application has been reported recently. Heriques et al. present to use correlation filters in a kernel space with the CSK method [14] which achieves the highest speed in a recent benchmark [15]. The CSK method builds on illumination intensity features and is further enhanced by adopting HOG features in the KCF algorithm [16]. Zhang et al. [17] make use of context information into filter learning

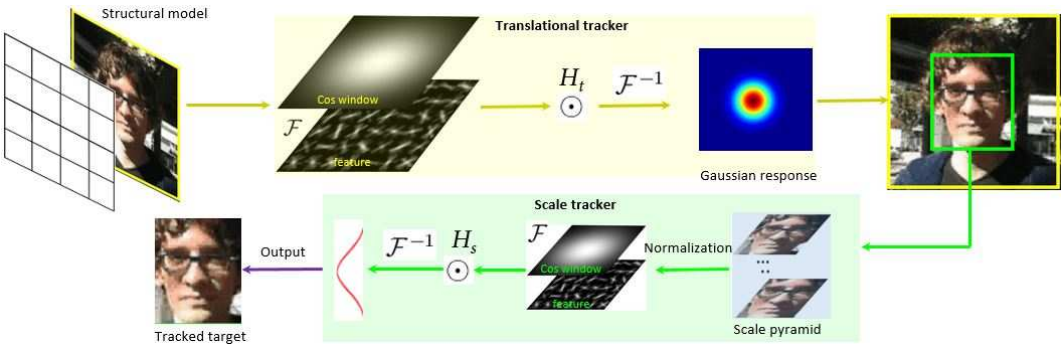

Fig. 1: The schematic diagram of structural correlational tracker.

and model the scale change based on consecutive correlation response. These correlation trackers are based on the holistic model. They are susceptible to drifting and less effective to handle long-term occlusion and out-of-view problems.

Structural framework aims to integrate the spatial information of appearance of object for a better tracking results. In [18], a local sparse representation scheme is employed to model the target appearance and then represent the basis distribution of target with the sparse coding histogram. Because of the structural appearance representation, this method performs well in handling the partial occlusion. Hare et al. [19] adopt a kernelled structured output support vector machine for adaptive online learning based object tracking. More recently, Yan et al. [20] present a structured partial least squares based appearance model for object tracking which is not only able to discriminate the target from the background but also able to tolerate the appearance variations due to the structured system updating framework.

The main contributions of this paper are listed as follows. First, a structural spatio-temporal appearance transform framework is presented for adaptive online learning based visual tracking. Coupling of spatial structure information, local patches corruption will not affect the entire appearance model response. Second, a coarse to fine translation estimation integrated with two layers of tracking framework is formulated for the accurate translation and scale estimation. Third, a sample pyramid is proposed for effective scale estimation. Experiments verify that our proposed algorithm achieved the satisfactory performance.

3. DISCRIMINATIVE TRACKER

The tracking framework of this paper is implementing particle filter for coarse translation estimation firstly. Then, two layers of correlation filter is applied for accurate translation and scale estimation (Fig.1).

3.1. Multi-dimensional Correlation Filter

Correlation filters model the appearance of a target using a filter h trained on a number of grayscale image patches $f_1, f_2, ..., f_t$ with $M \times N$ pixels centered around the target. The tracker considers all cyclic shift $x_{m,n}$, $(m,n) \in 0, ..., M-1 \times 0, ..., N-1$ as the training examples for the classifier. These are labelled with the desired Gaussian functions $y_1, y_2, ..., y_t$, so that $y(m,n)$ is the label for $x(m,n)$. Multi-dimensional feature map is considered for a signal representation. Then, the filter can be expressed in the spatial domain as solving the ridge regression problem:

$$E(h) = \frac{1}{2}\sum_{i=1}^{t} \|y_i - \sum_{l=1}^{k} h^l \star f_i^l\|_2^2 + \frac{\lambda}{2}\sum_{l=1}^{k} \|h^l\|_2^2 \qquad (1)$$

where f^l and h^l refers to the lth channel of the vectorized image and filter respectively. The star $\star$ denotes circular operator and λ is the regularization parameter. Solving this multi-channel form in the spatial domain is even more intractable. To reduce the problem complexity, we transform the spatial convolution to the element-wise production in frequency domain and solve this equation as

$$H = \frac{\sum_{l=1}^{k} F^l \odot Y^{l*}}{\lambda + \sum_{l=1}^{k} F^{l*} \odot F^l} \qquad (2)$$

where Capital letters denote the discrete Fourier transforms of the corresponding functions. The dot $\odot$ denotes element-wise production and $*$ is complex conjugate. The tracking task is carried out on an image patch z in the new frame with the search window size $M \times N$ by computing the response

$$\gamma = \mathcal{F}^{-1}(F \odot H^*) \qquad (3)$$

3.2. Structural Correlation Filter

Figure 2 presents the structural coding template. Although the candidate is partially occluded by a book, it is still the best candidate sample which should be considered as the tracking

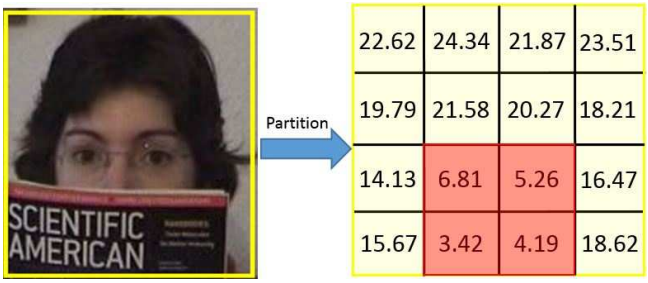

Fig. 2: Structural updating model of correlational tracker.

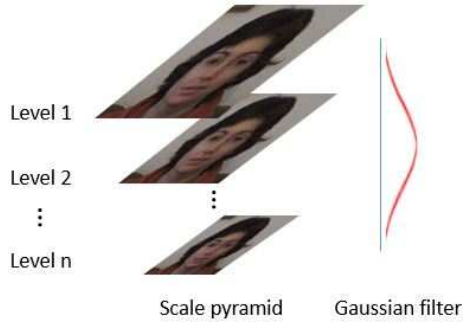

Fig. 3: Structural scale pyramid sampling framework.

results since the upper part of target is obviously observable. The structural pattern is divided into p blocks. Each block is a basic coding unit and adaptively implements online updating. The final filter is made up of all the weighted block-filters.

The Peak-to-Sidelobe Ratio(PSR) measures the strength of a correlation peak. The PSR is defined as $\frac{y_{max}-\mu}{\sigma}$, where y_{max} is the peak values and μ and σ are the means and standard deviations of the sidelobe. This paper uses PSR as a metric to qualify the similarity between the target and the candidate.

We define a weight vector $W = \{w_1, ..., w_p\}$ to help the appearance block maintain the underlying structure information of target sample. The input sample is normalized into a patch with predefined width and height firstly. Then, we partition the normalized sample into p local blocks. A weight is attached to each block. The weight of block at time t is calculated based on the block's correlation response at time $t-1$. Then, weight w_b is defined as

$$w(b) = \begin{cases} 1 & \text{if } PSR_b \geq \mathcal{T}_t \\ \frac{1}{p} & \text{otherwise} \end{cases} \tag{4}$$

where b is the index of structural block, $b \in (0, 1, ..., p)$. To maintain the model stability, we adopt a pre-defined threshold $\mathcal{T}_t$. Refer to Eq. (4), weight w_b will be set to 1 and filter will be fully updated if $PSR_b > \mathcal{T}_t$. Otherwise, the update extent of block will be suppressed as $\frac{1}{p}$ to reduce the corruption of interference.

The numerator and denominator of H in Eq. (2) is abbreviated as A and B, respectively. We update the A_b and B_b of block b at time t separately as

$$\begin{cases} A_{t,b} = (1-\eta)A_{t-1,b} + w_b\eta\sum_{l=1}^{k} Y_{t-1,b}^{l*} \odot F_{t-1,b}^{l} \\ B_{t,b} = (1-\eta)B_{t-1,b} + w_b\eta\sum_{l=1}^{k} F_{t-1,b}^{l*} \odot F_{t-1,b}^{l} \end{cases} \tag{5}$$

where η is a learning rate parameter.

Then, the final correlation response γ of translational filter at a rectangular region z are computed as Eq. (7). The new target state is found by maximizing the score γ

$$\gamma_t = \mathcal{F}^{-1}\{\frac{\sum_{l=1}^{k} A_{t-1}^{*} \odot Z_t^l}{B_{t-1} + \lambda}\} \tag{6}$$

3.3. Scale Pyramid

The proposed algorithm specially uses a correlation filter for scale estimation with the computational efficiency in fourier domain. Different from the translational estimation, we only extract the target patch not including of padding area for computational efficiency. We construct a scale pyramid around the estimated location from translational tracker for scale estimation (Fig. 3). Let $V \times U$ denote the target size in a frame and N is the number of scale level $n \in \{\lfloor -\frac{N-1}{2} \rfloor, \lfloor -\frac{N-2}{2} \rfloor, ..., \lfloor -\frac{N-1}{2} \rfloor\}$. The tracker extracts an image patch I_n with size of $a^n V \times a^n U$ centered around the estimated location of translational tracker. a is a scale factor. We uniformly resize all patches in pyramid with size $V \times U$ and use HOG features to construct the feature pyramid. The training sample is then set to a rectangular cuboid of the feature pyramid. The cuboid is of size $V \times U \times S$ and is centred at the target's estimated location and scale. We update scale filter holistically to capture the scale variation of whole target. The scale tracking filter is learned using formula (7)

$$\begin{cases} A_t = (1-\eta)A_{t-1} + \eta\sum_{l=1}^{k} Y_{t-1}^{l*} \odot F_{t-1}^{l} \\ B_t = (1-\eta)B_{t-1} + \eta\sum_{l=1}^{k} F_{t-1}^{l*} \odot F_{t-1}^{l} \end{cases} \tag{7}$$

4. EXPERIMENTS

This section presents experiments to demonstrate the performance of our proposed algorithm. We evaluate our method on a benchmark dataset [15] and select six representative videos with comparisons to state-of-the-art methods. All tracking methods are evaluated using distance precision and overlap precision as shown in Table 1 and Table 2, respectively. The first and second highest values are highlighted by bold and underline. Distance precision is computed as the relative number of frames in the sequence where the centre location error is within the given threshold (20 pixels) of the ground truth. Overlap precision is defined as the percentage of frames where the bounding box overlap surpasses a threshold (0.5).

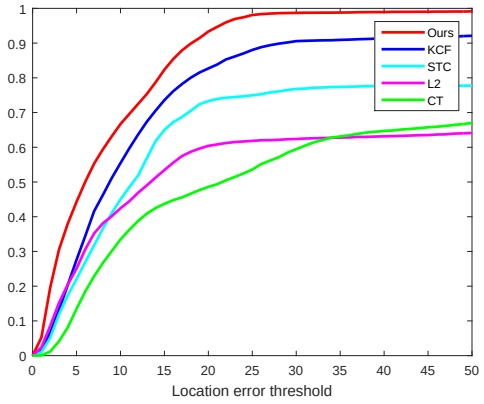

Fig. 4: The average distance precision comparisons.

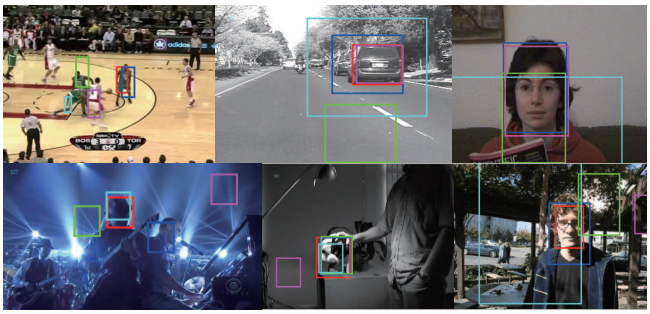

Fig. 5: The tracking samples of the selected datasets. They are $Basketball$, $Car4$,$FaceOccl1$,$Shaking$,$Sylvester$ and $Trellis$ from left to right.

We used 50 particles for the coarse local sampling. The regularization parameter is set as $\lambda = 0.005$. The size of padding window for translation estimation is set to be 1.2 times of the target size. The learning rate η in Eq. (5) is set to be 0.02. We use N=20 as number of scales for scale pyramid with a scale factor of $a = 1.02$. The threshold of $\mathcal{T}_t$ and $\mathcal{T}_s$ are set as 0.25 and 0.3. All experiments are conducted using MATLAB implementation on a Intel I7 3.4GHz machine with 16GB RAM. We use F-HOG for image representation and the first dimension of feature is image intensity (graylevel) value. Our algorithm performs well at around $12fps$. Additionally, we also test the L2 tracker [21], CT tracker [1], STC tracker [17], KCF tracker [16] for comparative purpose.

Tables I and II show that our algorithm performs favorably against the state-of-the-art methods in distance precision (DP) and overlap precision (OP). Our algorithm achieved the best performance in $Car4$, $Shaking$, $Sylvester$ and $Trellis$, and top performance in $Basketball$ and $FaceOccl1$ from the perspectives of DP and OP evaluations.

Figure 4 presents an average DP for comprehensive evaluation of the selected six datasets. The x axis denotes the threshold in DP and y axis is the performance score normalized from 0 to 1. Higher score in y axis denotes a better performance. We set the threshold range from 0 to 50. As shown in Fig.4, our algorithm achieved the best score performance among the other methods. It demonstrates that our algorithm is robust to handle illumination change, occlusion, and pose variation, etc.. KCF obtained the second high performance. KCF performed good in the most datasets except of $Shaking$ and it lost the target in early stage of $Shaking$ due to serious illumination change coupled with posture variation. STC have the medium performance in this comparison. STC performed unstable in $Basketball$, $FaceOccl1$ and $Trellis$ at the component of scale estimation, which causes tracking failure. Figure 4 shows that L2 and CT cannot obtain the satisfactory performance. The representative tracking samples are presented in Fig. 5.

Table 1: Comparison results of distance precision.

Video Clip	L2	CT	STC	KCF	Ours
Basketball	60.7	26.8	56	**92.3**	87.6
Car4	**100**	34.6	96.7	95	**100**
FaceOccl1	96.7	48	25	72.8	**98**
Shaking	1.2	15.6	94.8	2.5	**97.9**
Sylvester	37.4	93.7	94.1	94.5	**96.4**
Trellis	44.1	31.3	73.8	98.2	**100**

Table 2: Comparison results of overlap precision.

Video Clip	L2	CT	STC	KCF	Ours
Basketball	0.55	0.25	0.23	0.90	**0.91**
Car4	**1**	0.24	0.21	0.25	**1**
FaceOccl1	**0.99**	0.95	0.25	**0.99**	0.98
Shaking	0.01	0.16	0.82	0.02	**0.95**
Sylvester	0.33	0.74	0.56	0.81	**0.84**
Trellis	41	0.09	0.51	0.82	**0.96**

5. CONCLUSIONS

This paper have presented a novel tracker to decouple the tracking process with translation and scale estimation steps. A coarse to fine method has been proposed to integrate particle filter with structural correlation filter for effective target locating in spatial domain. Structural updating in spatial domain can make a reasonable online learning, which may effectively prevent drift and handle occlusion for translational tracker. A sample pyramid has been introduced for a robust scale estimation. Experiments on publicly available benchmark video sequences show the superiority of our proposed method over the representatives of state-of-the-art trackers.

6. REFERENCES

[1] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang, "Fast compressive tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 10, pp. 2002–2015, 2014.

[2] Tianxiang Bai, You-Fu Li, and Xiaolong Zhou, "Learning local appearances with sparse representation for robust and fast visual tracking," *Cybernetics, IEEE Transactions on*, vol. 45, no. 4, pp. 663–675, 2015.

[3] Wei Zhong, Huchuan Lu, and Ming-Hsuan Yang, "Robust object tracking via sparse collaborative appearance model," *Image Processing, IEEE Transactions on*, vol. 23, no. 5, pp. 2356–2368, 2014.

[4] Yazhe Tang, Youfu Li, and Jun Luo, "Parametric distortion-adaptive neighborhood for omnidirectional camera," *Applied optics*, vol. 54, no. 23, pp. 6969–6978, 2015.

[5] Allan D Jepson, David J Fleet, and Thomas F El-Maraghi, "Robust online appearance models for visual tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 10, pp. 1296–1311, 2003.

[6] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[7] Yazhe Tang, YF Li, Shuzhi Sam Ge, Jun Luo, and Hongliang Ren, "Distortion invariant joint-feature for visual tracking in catadioptric omnidirectional vision," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2418–2423.

[8] Xue Mei and Haibin Ling, "Robust visual tracking and vehicle classification via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 11, pp. 2259–2272, 2011.

[9] Shai Avidan, "Support vector tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 8, pp. 1064–1072, 2004.

[10] Helmut Grabner, Christian Leistner, and Horst Bischof, "Semi-supervised on-line boosting for robust tracking," in *Computer Vision–ECCV 2008*, pp. 234–247. Springer, 2008.

[11] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas, "Tracking-learning-detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1409–1422, 2012.

[12] David S Bolme, J Ross Beveridge, Bruce Draper, Yui Man Lui, et al., "Visual object tracking using adaptive correlation filters," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2544–2550.

[13] David S Bolme, Bruce Draper, J Ross Beveridge, et al., "Average of synthetic exact filters," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2105–2112.

[14] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Computer Vision–ECCV 2012*, pp. 702–715. Springer, 2012.

[15] Yaowu Wu, Jungyoul Lim, and Ming-Hsuan Yang, "Object tracking benchmark," 2015.

[16] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista, "High-speed tracking with kernelized correlation filters," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 3, pp. 583–596, 2015.

[17] Kaihua Zhang, Lei Zhang, Qingshan Liu, David Zhang, and Ming-Hsuan Yang, "Fast visual tracking via dense spatio-temporal context learning," in *Computer Vision–ECCV 2014*, pp. 127–141. Springer, 2014.

[18] Baiyang Liu, Junzhou Huang, Lin Yang, and Casimir Kulikowsk, "Robust tracking using local sparse appearance model and k-selection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1313–1320.

[19] Sam Hare, Amir Saffari, and Philip HS Torr, "Struck: Structured output tracking with kernels," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 263–270.

[20] Jia Yan, Xi Chen, Dexiang Deng, and Qiuping Zhu, "Structured partial least squares based appearance model for visual tracking," *Neurocomputing*, vol. 144, pp. 581–595, 2014.

[21] Ziyang Xiao, Huchuan Lu, and Dong Wang, "L2-rls-based object tracking," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 8, pp. 1301–1309, 2014.