SPATIAL-SEQUENTIAL-SPECTRAL CONTEXT AWARENESS TRACKING

Jianwu Fang [†‡], *Zheng Li* [†] *and Jianru Xue* [†]

[†] Xi'an Jiaotong University;[‡] Chang'an University

ABSTRACT

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. Supplemented by the infrared perception, this work exploits the role of visual context in tracking in a spatial-sequential-spectral view, by which to excavate dominance of different contexts in various scenarios. Specifically, we infer it in the Fourier domain with a real-time speed, and incorporate a fully-occlusion handling and scale adaptation with a trajectory regression filter and object contour closure, respectively. Extensive experiments on 50 video clips simultaneously containing registered RGB and thermal bands demonstrate that our tracker shows a state-of-the-art performance.

***Index Terms*—** Visual tracking, context understanding, trajectory regression, scale adaptation

1. INTRODUCTION

The task of tracking shows many promising strengths in various applications, such as robotics, surveillance, augmented reality, etc. However, because of the unpredicted dynamic disturbance of target, the robust tracking is still not solved.

For the sake of tracking success, visual target context is a particular and effective attempt which keeps a watchful eye on the locally surrounding information of the target with various dimension view, i.e., spatial, temporal, and so forth. The existing contextual trackers mainly concentrate on either the local spatial context consisting of a target object and its immediate surrounding background within a determined region, or the local temporal context in consecutive frames where the motion consistency of spatial regions or points are preferred [1]. The feasibility of this kind of context sometimes relies on the video quality. That is to say that some extremely terrible weather or light conditions captured by visual light videos weaken the feasibility of the spatial-temporal context.

Supplemented by the infrared perception, this paper proposes *3SContext* (select *S*patial-*S*equential-*S*pectral *Context* to track), a tracker which selects Spatial-Sequential-Spectral context to approximate the most discriminative power for target-background separation. Compared to other trackers fusing multi-band [7, 8], we generate a real-time inference in a FFT correlation computing framework, to select the best spectral band by exploiting both the intra-frame and inter-frame context. 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. Fig.1 demonstrates the superiority of our tracker in some extreme conditions. 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 in [7] while running at a real-time speed.

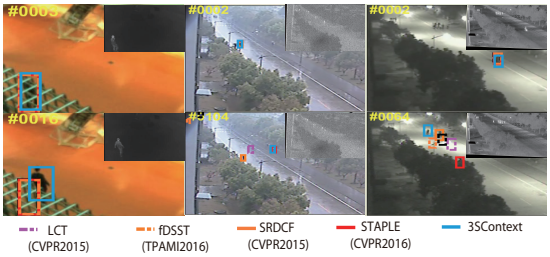

Fig. 1. The tracking results on some typically difficult frames in visible light videos tracked by LCT [2], RPT [3], fDSST [4], SRDCF [5] , STAPLE [6] and our 3SContext. The corresponding thermal infrared band is shown in the right-top corner in each frame.

2. RELATED WORKS

Context understanding: Context understanding means to comprehend the relationship of the target with its local surrounding situation. One kind of methods [9, 10] selected many supporters with motion pattern similar to the target, and distracters with dissimilar motion pattern. Spatial-temporal context (STC) tracker [11] computed the dense spatial correlations between target and its local surrounding region in a scene by solving a deconvolution problem, and formulated the tracking problem into computing a confidence map as a convolution problem that integrates the spatio-temporal context

This work is supported by the National Key R&D Program Project under Grant 2016YFB1001004, and also supported by the Natural Science Foundation of China under Grant 61603057, China Postdoctoral Science Foundation under Grant 2017M613152, and is also partially supported by Collaborative Research with MSRA.

information. Saliency prior context [12] was also exploited. We differ from these approaches in that we extend the context into *S*patial-*S*equential-*S*pectral dimension, and make a truly reflection of the spectra discrimination in a scene.

Appearance model enhancement: To make the appearance model be stronger, many studies exploit the combination of multiple features and online model learning schemes [13]. As for the *multi-feature combination*, Yuan *et al.* [8] proposed a robust superpixel based tracking via depth fusion. STAPLE [6] designed a simple combination of a HOG feature based Correlation Filters and a global colour histogram. Nam *et al.* [14] proposed a Tree-structured Convolutional Neural Network (TCNN) tracker to preserve model consistency and handle appearance multi-modality. For *online learning*, STRUCK [15] localized the target by minimising the structured output objective. Multiple-Instance Learning (MIL) was used to train with bags of positive examples [16].

Efficiency and long-term demand: For *efficiency demand*, Correlation Filters (CF) [17] are currently the most efficient framework. They replace the time-consuming convolution operation in the time domain by pointwise multiplication in the frequency domain. The typical Kernelized Correlation Filter (KCF) [18] extended CF by exploiting the circulant structure of the tracking object into Fourier analysis. Based on KCF, many new trackers are proposed, such as Discriminative Scale-Space Tracking (DSST) [4] for a powerful scale adaptation, Spatially Regularised Correlation Filters (SRDCF) [5] with a spatially regularisation of the filter coefficients. In terms of *long-term demand* [19, 2, 20], Multi-Expert Entropy Minimisation (MEEM) [21] and Long-term Correlation Tracker (LCT) [2] trained classifiers to re-detect the target. Multi-Store Tracker (MUSTer) [22] maintained a long-term memory of SIFT keypoints for object and background.

3. METHODOLOGY

3.1. 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:

$$r_t = \arg\max_{r_t \in I_t} f(\mathbf{M}(I_t, r); \mathcal{C}_t), \tag{1}$$

where $\mathcal{C}_t = \{\kappa_t^c, \varsigma_t^c, \rho_t^c\}$ is the context space containing the spatial context κ^c, sequential context ς^c and spectral context ρ^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. We model the the joint distribution of these contextual cues as distribution $\mathbf{P}(\mathcal{C}_t)$, the context items in $\mathcal{C}_t$ are not independent. Inspired by conditional probability:

$$\begin{aligned} &\mathbf{P}(\mathcal{C}_t) = \mathbf{P}(\kappa_t^c, \varsigma_t^c, \rho_t^c) \\ &= \mathbf{P}(\kappa_t^c)\mathbf{P}(\varsigma_t^c|\kappa_t^c)\mathbf{P}(\rho_t^c|\kappa_t^c, \varsigma_t^c). \end{aligned} \tag{2}$$

Eq. 2 denotes their conditional dependency. Intuitively speaking, if we want to select the best spectra in each time, we need to consider its discrimination ability for target in spatial and sequential dimension. Similarly, sequential context is constructed by gathering the separability of target/background in spatial domain. Therefore, to solve Eq. 1, we infer it from following procedure.

3.2. Tracking inference

By κ_t^c: Spatial context models the relationship between the object location with its local surrounding region. Suppose the surrounding region is R_t whose size is twice as r_t. Assume that the object location $\mathbf{o}$ is the center of r_t. In this paper, the spatial context is modeled to compute the distance between the locations $\mathbf{s}$ in R_t with $\mathbf{o}$. To achieve this, a radial function $g_t(\mathbf{o} - \mathbf{s})$ similar to that in [11] is used, which obtains a dense measure, and is claimed that $g_t(\mathbf{o} - \mathbf{s})$ is not radially symmetric to avoid ambiguities when similar objects appear in close proximity. In order to reduce the frequency effect of image boundary, a cosine window $w_t(\mathbf{s} - \widehat{\mathbf{o}})$ is utilized, where $\widehat{\mathbf{o}}$ is the predicted object location at previous time. In time domain, two more similar observation can get a confidence map with higher peak value after convolution, i.e., for the spatial context $g_t(\mathbf{o} - \mathbf{s})$ and weighted surrounding region $I_t(\mathbf{s})w_t(\mathbf{s} - \widehat{\mathbf{o}})$, tracking can be inferred by:

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

where $\otimes$ is the convolution operator. By observation, Eq.3 can be computed by Fast Fourier Transform (FFT), i.e., $\mathcal{F}(f(\mathbf{M}(I_t, r); \kappa^c{}_t)) = \mathcal{F}(g_t(\mathbf{o})) \odot \mathcal{F}(I_t(\mathbf{o})w(\mathbf{o} - \widehat{\mathbf{o}}))$. For the first frame, $\mathcal{F}(f(\mathbf{M}(I_1, r); \kappa^c{}_1))$ is modeled by a *gaussian kernel* treated as a desired confidence map. Hence Eq.1 stimulated by κ_t^c can be estimated by $\mathcal{F}^{-1}(\mathcal{F}(f(\mathbf{M}(I_t, r); \kappa^c{}_t)))$. $\mathcal{F}^{-1}$ is the inverse-FFT operation. Denote $\mathcal{F}(g_t(\mathbf{o}))$ as the Fourier domain of spatial context at time t.

By $\kappa_t^c, \varsigma_t^c$: 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. Thus, spatial-sequential context $\mathcal{G}_t(\mathbf{o})$ is obtained by an online adaptivity rate ξ (denoted as 0.3 in this work), simply as:

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

Based on the statistic of FFT, Eq. 4 is equivalent to low passing $g(\mathbf{o})$ [11]. Hence, it can avoid the image noise caused by disturbance. Following Eq. 3, infer $f(\mathbf{M}(I_t, r); \kappa^c{}_t, \varsigma_t^c)$ as:

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

By $\kappa_t^c, \varsigma_t^c, \rho_t^c$: In terms of the spatial-sequential-spectral context, our goal is to select the best spectral band according to the discriminative ability of $\kappa_t^c, \varsigma_t^c$ in each spectra.

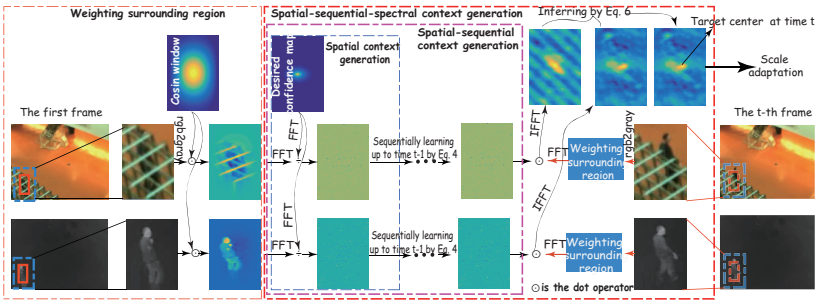

Fig. 2. The flowchart of tracking inference for RGB and thermal infrared bands. Note that each spectral band is gray-scale in inferring.

Among the visual effect of different spectra, the band with clearer object-background boundary has stronger discriminative ability. To approach this, we measure the correlation between the searching region in the t^{th} frame with the spatial-sequential context in each spectral band learned up to the $(t\text{-}1)^{th}$ frame. The underlying meaning is that the spatial-sequential context modeled by the video frames with clearer object-background boundary is more discriminative than others. Assume that there are K spectral bands. Denote S_t^k=$\mathcal{F}^{-1}(\mathcal{F}(f(\mathbf{M}(I_t^k, r); (\kappa_t^c)^k, (\varsigma_t^c)^k)))$ as the confidence map of the t^{th} frame in k^{th} spectral band, $k = 1, 2, ..., K$. Different from many related methods [7, 8], we compute the object score as $(\rho_t^c)^k$=$max(S_t^k)$-$mean(S_t^k)$. Eq. 1 is inferred as:

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

This spectral selection is simple but effective, as verified in our experiments. To maintain the reliability of $(\kappa_t^c)^k$, $(\varsigma_t^c)^k$ in each band, Eq. 4 is updated only when $(\rho_t^c)^k > \mathcal{T}$, where $\mathcal{T}$ is a predefined threshold to restrain the noise spatial context (denoted as 0.002). The flowchart of the tracking inference for RGB and thermal infrared bands is shown in Fig. 2.

3.3. Fully-occlusion handling

We introduce the fast least trimmed squares (FAST-LTS) model [23] to regress the object trajectory with the tracked target centroids collected from a short previous time, and makes a trajectory prediction for upcoming frames. Based on the moving tendency, we can provide credible target candidates when the target undergos fully occlusion. Given m collected object centroids $\mathcal{P} = \{p_1; p_2; ...; p_m\}$, where $p_i = (x_i, y_i)$ is the coordinate of the object centroid. Denote $\mathcal{P}_x$=$\{x_i\}_{i=1}^m$ as the x-axis set of $\mathcal{P}$, and $\mathcal{P}_y$ as the y-axis of $\mathcal{P}$. FAST-LTS regression is:

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

where $\mathbf{u}$ is the parameter vector with the regression order j by fitting a linear model. For the first order regression, it is the fitting slope and offset. FAST-LTS regression can get rid of the noise point [23], which is adequate for our tracking, see ref. [23] for details. In this paper, we set j as 1 to make a first-order prediction, and collect $m = 30$ frames to regress. The re-detection mechanism will start only when $(\rho_t^c)^k < \mathcal{T}$, and the candidate target samples are selected from the forward extension line with the constraint of $\mathbf{u}$ from the object centroid in current time to the image boundary. Then, the re-detection is conducted by Eq. 5 in the selected spectral band.

3.4. Scale adaptation

This work introduces the contour closure to estimate the target scale. That is because the object in each frame has manifest closed contour in comparison to other image regions. Assume the target location r_t (a rectangle region) by inferring Eq. 1 is obtained. We crop an image region LR_t whose size is 2.5 times the size as r_t in the selected spectral band. Then EdgeBox is adopted on LR_t to generate several bounding boxes with probable closed contour. In particular, EdgeBox is used on color image. When the thermal infrared band is selected, we pseudo-colorize the selected frame. In addition, we find that EdgeBox sometimes cannot assign the largest score to the true target. Therefore, this work defines the following criterion to select the object bounding box:

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

where $\widehat{B}_{t-1}$ is the estimated object bounding box in previous time ($\widehat{B}_1$ is the ground-truth in the first frame), $\lambda(B_t{}^p)$ the object score of $B_t{}^p$ assigned by EdgeBox. Here, the scale of $\widehat{B}_t$ is adaptively updated as $\widehat{B}_t = (1-\eta)\widehat{B}_{t-1} + \eta B_t$ with a learning rate η (set as 0.3 in this work).

4. EXPERIMENTS

4.1. Evaluation details

Dataset: In this section, we apply the proposed tracker on 50 video sequences in [7], simultaneously containing registered RGB and thermal infrared bands. The resolution of each band

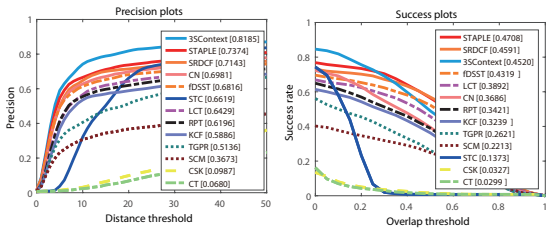

Fig. 3. The overall performance on all the sequences.

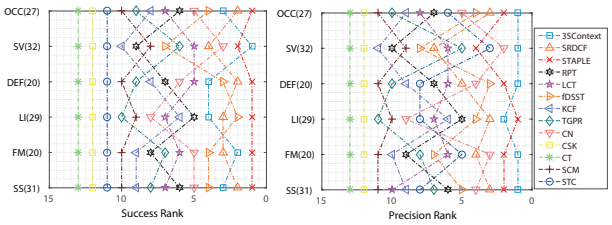

Fig. 4. The performance rank of trackers corresponding to success rate and precision on each challenging attribute. Rank-1 represents the best tracker. The number in the bracket specifies the number of sequences belonging to each kind of challenging attribute.

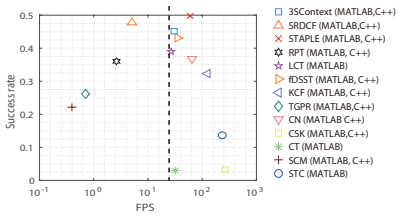

Fig. 5. Success rate vs. running speed. The x-axis is shown with a semilog display. The dashed vertical line denotes the real-time performance threshold (25fps used in this work).

is 384×288. Because the gray-scale image is utilized for each band, the number of spectral band K is 2. The ground-truth of each video frame is labeled separately in RGB and infrared band. In this work, we use the intersection of the bounding box of the RGB and thermal infrared band in the first frame to initialize the target. The challenging attributes contain occlusion (OCC), fast motion (FM), scale variation (SV), deformation (DEF), low illumination (LI), and small size (SS).

Metrics: Two widely used metrics, precision rate and success rate, are utilized to evaluate the tracking performance. Note that all the precision and success rate of other trackers are the best one among two modality. For precision rate, the Euclidean distance between the center location of the tracked object and the ground truth is used. Precision score calculates the percentage of frames whose outputted Euclidean distance is less than a predefined distance threshold. Success rate computes the bounding box overlap between the tracked object B_T and the ground-truth B_G. The detailed operation is $|\frac{B_T \cap B_G}{B_T \cup B_G}|$, where $\cap$ and $\cup$ denote the intersection and union operators, and $|\cdot|$ counts the number of pixels.

Competitors: In this work, twelve popular trackers are selected as the competitors. They are SRDCF [5] , STAPLE [6], LCT [2], RPT [3], fDSST [4], KCF [18], TGPR [24], CN [25], SCM [26], CT [27], CSK [28] and STC [11]. All the trackers are re-run in three times in this work. Fig.3 demonstrates the overall performance, and Fig. 4 shows the rank evaluation of the trackers on different challenging attributes.

4.2. Result analysis

Tracking evaluation: From Fig. 3 and Fig. 4, it can be observed that our 3SContext, STAPLE [6] and SRDCF [5] are the top three trackers. Among them, our 3SContext outperforms the others in precision score and STAPLE shows the best in the success evaluation. The main reason is that, 3SContext can select the best spectral band for each frame while the other ones drift away when encountering the situation that the target has similar appearance or temperature with other objects or background. In addition, 3SContext is a gray-scale tracker, the performance can be improved in future by fusing more informative clues, such as color and texture. Besides, 3SContext generates the best scale adaptation among the competitors, reported by the rank evaluation in Fig. 4. We find that our spatial-sequential-spectral context can boost the precision and success rate respectively at 20% and 30% rate of STC [11] only with spatial-temporal context exploitation.

Efficiency evaluation: The running efficiency is one important factor for application realization. Hence, we report the success rate vs. running speed in Fig. 5. The corresponding compiler is also denoted. The statistical result is generated on a platform with an i7 CPU running of 2.7GHz and 8GB RAM. From the success rate vs. running speed analysis, 3SContext, STAPLE, and fDSST are the top three trackers.

From the above analysis, we can conclude that our 3SContext demonstrates a state-of-the-art performance, especially for the precision and scale adaptation.

5. CONCLUSION

This paper proposed a robust real-time tracker via spatial-sequential-spectral context (3SContext) understanding incorporating with fully-occlusion handling and scale adaptation. Extensive experiments on 50 video sequences simultaneously containing registered RGB and thermal infrared bands demonstrated that the proposed tracker 1) generated a state-of-the-art performance, 2) outperformed other trackers in scale adaptation. We will improve our tracker by exploiting more informative visual clues, as well as a more comprehend evaluation with VOT16 measures [29].

6. REFERENCES

[1] M. Yang, Y. Wu, and G. Hua, "Context-aware visual tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1195–209, 2009.

[2] C. Ma, X. Yang, C. Zhang, and M. Yang, "Long-term correlation tracking," in *CVPR*, 2015, pp. 5388–5396.

[3] Yang Li, Jianke Zhu, and Steven C. H. Hoi, "Reliable patch trackers: Robust visual tracking by exploiting reliable patches," in *CVPR*, 2015, pp. 353–361.

[4] M. Danelljan, G. Hager, F. Khan, and M. Felsberg, "Discriminative scale space tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[5] M. Danelljan, G. Hager, F. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *ICCV*, 2015, pp. 4310–4318.

[6] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. Torr, "Staple: Complementary learners for real-time tracking," in *CVPR*, 2016, pp. 1401–1409.

[7] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," vol. PP, no. 99, pp. 1–1, 2016.

[8] Y. Yuan, J. Fang, and Q. Wang, "Robust superpixel tracking via depth fusion," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 15–26, 2014.

[9] T. Dinh, N. Vo, and G. Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *CVPR*, 2011, pp. 1177–1184.

[10] G. Zhu, J. Wang, C. Zhao, and H. Lu, "Weighted part context learning for visual tracking.," *IEEE Trans. Image Processing*, vol. 24, no. 12, pp. 5140–5151, 2015.

[11] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M. Yang, "Fast visual tracking via dense spatio-temporal context learning," 2014, pp. 127–141, ECCV.

[12] C. Ma, Z. Miao, and X. Zhang, "Saliency prior context model for visual tracking," in *ICIP*, 2016, pp. 1724–1728.

[13] Jianwu Fang, Hongke Xu, Qi Wang, and Tianjun Wu, "Online hash tracking with spatio-temporal saliency auxiliary," *Computer Vision and Image Understanding, Accepted*, 2017.

[14] H. Nam, M. Baek, and B. Han, "Modeling and propagating cnns in a tree structure for visual tracking," *CoRR*, 2016.

[15] S. Hare, A. Saffari, and P. Torr, "Struck: Structured output tracking with kernels," in *ICCV*, 2011, pp. 263–270.

[16] B. Babenko, M. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 983–990, 2009.

[17] D. S Bolme, J. R Beveridge, B. A Draper, and Yui Man Lui, "Visual object tracking using adaptive correlation filters," in *CVPR*, 2010, pp. 2544–2550.

[18] J. Henriques, C. Rui, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[19] J. S Supancic and D Ramanan, "Self-paced learning for long-term tracking," in *CVPR*, 2013, pp. 2379–2386.

[20] Z. Kalal, K. Mikolajczyk, and J. Matas, "Tracking-learning-detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, 2012.

[21] J. Zhang, S. Ma, and S. Sclaroff, "Meem: Robust tracking via multiple experts using entropy minimization," in *ECCV*, 2014, pp. 188–203.

[22] Z. Hong, Z. Chen, C. Wang, and X. Mei, "Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking," in *CVPR*, 2015, pp. 749–758.

[23] 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.

[24] J. Gao, H. Ling, W. Hu, and J. Xing, "Transfer learning based visual tracking with gaussian processes regression," in *ECCV*, 2014, pp. 188–203.

[25] M. Danelljan, F. Khan, M. Felsberg, and J. Weijer, "Adaptive color attributes for real-time visual tracking," in *CVPR*, 2014, pp. 1090–1097.

[26] W. Zhong, H. Lu, and M. Yang, "Robust object tracking via sparsity-based collaborative model," in *CVPR*, 2012, pp. 1838–1845.

[27] K. Zhang, L. Zhang, and Ming H. Yang, "Real-time compressive tracking," in *ECCV*, 2012, pp. 864–877.

[28] J. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *ECCV*, 2012, pp. 702–715.

[29] M. Felsberg *et al.*, "The thermal infrared visual object tracking vot-tir2016 challenge results," in *ECCV Workshop*, 2016.