SELECTIVE OBJECT AND CONTEXT TRACKING

Ce Zhou[1], *Qing Guo*[1,†], *Liang Wan*[2], *Wei Feng*[1]

[1] School of Computer Science and Technology, Tianjin University, Tianjin, 300072, China
[2] School of Computer Software, Tianjin University, Tianjin, 300072, China

ABSTRACT

Robust appearance model is significantly important to state-of-the-art trackers. However, such trackers highly rely on the reliability of foreground appearance model. When the foreground is seriously occluded or the scene contains multiple objects with similar appearance, such foundation is destroyed. To extend the ability of trackers to handle these difficulties, we propose selective object and context tracking to locate the target according to the reliability of the foreground appearance model which is determined by two measures about whether the target is occluded or surrounded by similar objects. Extensive experiments show that our method achieves better performance than state-of-the-art trackers on VOT TIR-2015 dataset and is able to track the target even when the foreground appearance is completely unreliable.

Index Terms— Visual tracking, unreliable foreground appearance model, selective tracking, correlation filters

1. INTRODUCTION

Visual tracking is still a challenging problem in computer vision. Recently, a lot of trackers [1–8] are proposed and achieve significant improvements on public datasets [9, 10] by making the best of the foreground appearance model. For example, the correlation filter based trackers, e.g. kernelized correlation filters (KCF) [11], spatio-temporal context learning (STC) [12] and spatially regularized correlation filters (SRDCF) [13], can get high tracking accuracy and real-time performance. As is shown in the first column of Fig. 1, with foreground appearance model, we get a discriminative confidence map with a very high score at object location, which helps to track the object accurately. At this time, the foreground appearance model is defined as *reliable*.

However, the foreground appearance model becomes *unreliable* when the confidence map is not prominent at target location. As is shown in the second and third column of Fig. 1, when the target is seriously occluded or the scene contains multiple similar objects, the confidence map calculated by foreground appearance model is less discriminative, which makes the tracker lose the target easily. Note that, this problem cannot be solved by simply using a more discriminative feature or a better classifier updating strategy, because the target may be totally occluded.

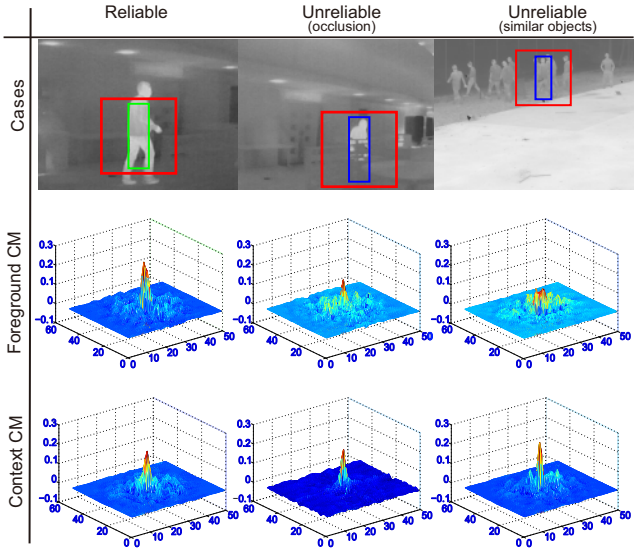

Fig. 1. Each column represents a case with two confidence maps, i.e. Foreground CM and Context CM, generated by foreground appearance model and context appearance model, respectively. Green and blue bounding boxes are the target locations generated by Foreground CM and Context CM, respectively. The red bounding boxes excluding the green or blue boxes represent the range of context.

In addition to using foreground appearance model, we show that context appearance model can also help to locate the target when foreground appearance model becomes unreliable. As is shown in Fig. 1, the confidence map (CM) calculated by context appearance model is much more discriminative than that generated by foreground appearance model.

In this paper, we propose selective object and context tracking to locate the target. Specifically, we learn two appearance models by selectively regularizing the foreground and the context. Then, we propose two measures to judge the reliability of foreground appearance model w.r.t whether the target is occluded or surrounded by multiple similar objects. If the foreground appearance model is unreliable, we use the context appearance model to track the object. Our method achieves the best performance on VOT TIR-2015 dataset [14] comparing with several other state-of-the-art trackers.

† is corresponding author. Email: tsingqguo@tju.edu.cn.

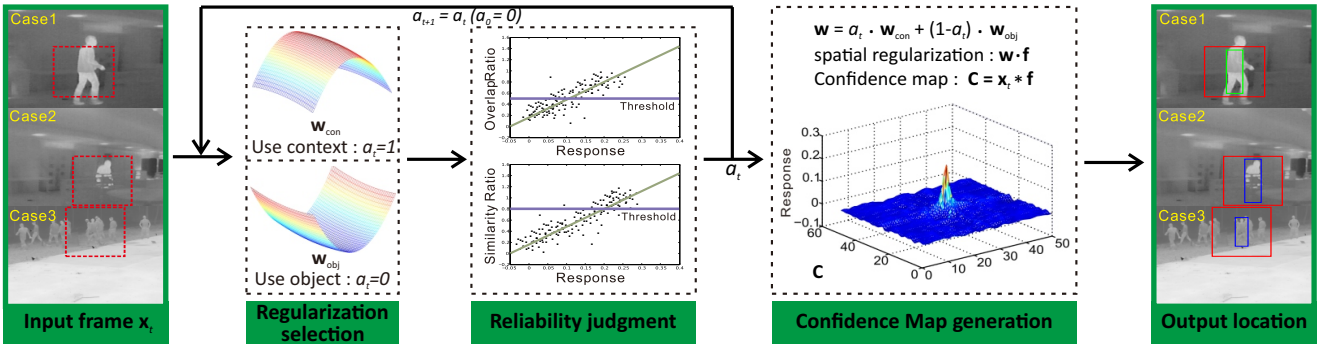

Fig. 2. Algorithm flow of the proposed tracker. Case1: reliable appearance, case2: object occlusion, case3: multiple similar objects. Given input frame t with a red dotted bounding box as candidate search region, we obtain a confidence map by selectively using object and context models, i.e. $\mathbf{w}_{\mathrm{obj}}$ and $\mathbf{w}_{\mathrm{con}}$. The selection is determined by a process of reliability judgment w.r.t. the confidence map generated by $\mathbf{w}_{\mathrm{obj}}$.

2. RELATED WORK

Context based trackers. Several works have proposed to use context information to improve tracking performance [12,15–17]. Spatio-temporal context learning based tracker (STC) [12] models the relationship between the target location and its context and is actually an improved correlation filter based tracker, whose performance is mainly based on the foreground appearance model and cannot handle serious occlusion or separate multiple similar objects. Other context based trackers [15–17] use objects in the context to handle occlusion. However, these methods have to firstly select objects near the target as the helpers to locate the target. They are more complex than our method and cannot handle the situation where there exist multiple similar objects in a scene.

Correlation filter based trackers. Correlation filter based trackers utilize foreground appearance model to train a filter from a set of training samples $\{\mathbf{x}_k, \mathbf{y}_k\}_{k=1}^{t}$. They utilize circular correlation and perform most computations in the Fourier Domain through Fast Fourier Transform (FFT) with high efficiency. For example, Bolme et al. proposed Minimum Output Sum of Squared Error (MOSSE) in [18]. Afterwards, due to the unwanted boundaries produced by periodic assumption in MOSSE, Danelljan et al. introduces SRDCF in [13], in the training stage for frame $\mathbf{x}_k$, the goal is to find a function that minimizes the squared error between sample $\mathbf{x}_k$ and regression label $\mathbf{y}_k$,

$$\min_{\mathbf{f}_{\mathrm{obj}}} \sum_{k=1}^{t} \alpha_k || \sum_{l=1}^{d} \mathbf{x}_k^l * \mathbf{f}_{\mathrm{obj}}^l - \mathbf{y}_k ||^2 + \sum_{l=1}^{d} ||\mathbf{w}_{\mathrm{obj}} \cdot \mathbf{f}_{\mathrm{obj}}^l||^2, \quad (1)$$

here, the weights α_k determine the impact of $\mathbf{x}_k$, $*$ and $\cdot$ denote circular convolution and element-wise production, respectively, $\mathbf{x}_k^l$ and $\mathbf{f}_{\mathrm{obj}}^l$ denote the l-th channel of sample $\mathbf{x}_k$ and correlation filter $\mathbf{f}_{\mathrm{obj}}$, respectively. The former part of Eq. 1 indicates the total squared error; the later part of Eq. 1 introduces the spatial regularization function $\mathbf{w}_{\mathrm{obj}}$ to penalize the region residing in the background. In detection stage, by applying the updated filter in image patch, a confidence map $\mathbf{C}$ is then computed and maximized to estimate the state of the target. Afterwards, a d-dimensional feature map is then extracted around the target, and eventually a new sample $(\mathbf{x}_k, \mathbf{y}_k)$ is added into the training set.

These correlation trackers can indeed deal with tracking problems with high efficiency. However, they are strongly dependent on foreground appearance model. Therefore, They are much more likely to fail in complicated situations whose foreground appearance model is unreliable.

3. THE PROPOSED METHOD

3.1. Overview

We propose selective object and context tracking (SOCT) for visual tracking. SRDCF [13] is utilized as our baseline tracker. The proposed tracking framework contains two parts: detection and updating stage. Updating stage is the same with [13], thus we mainly focus on detection stage as illustrated in Fig. 2. In the first frame, by applying the newly initialized filter to the image patch we obtain the confidence map of the target region, by which we train the Ridge Regression model (details are shown in section 3.3). Parameter a_0 is initialized as 0 indicating that we use object spatial regularization. In frame $\mathbf{x}_t$, we use the regression model to judge the reliability of foreground appearance model. If the model is reliable, then object confidence map will predict the target state; if not, a_t is then transferred to 1 in order to use context spatial regularization for tracking (details are shown in 3.2), and then the derived confidence map of the context is used to predict the target state. In Fig. 2, confidence map $\mathbf{C}$ is a selected between object CM and context CM. Note that, in frame $\mathbf{x}_t$, we set $a_{t+1} = a_t$ to instruct model selection of frame $\mathbf{x}_{t+1}$.

3.2. Context Regularization

Discriminative trackers are strongly dependent on foreground appearance model, thus when it is unreliable, these trackers often fail. In such situation, we enable context tracking.

The SRDCF tracker introduces a spatial weight function $\mathbf{w}_{\mathrm{obj}}$ to penalize the magnitude of its filter coefficients in the learning according to spatial locations, which is shown in Eq. 1. When the foreground appearance model is unreliable, we propose to track the context by penalizing the magnitude on the opposite way. Specifically, for the region in the center, we penalize it by assigning higher weights and vice versa. When utilizing the context model, the resulting optimization problem can be expressed as,

$$\min_{\mathbf{f}_{\mathrm{con}}} \sum_{k=1}^{t} \alpha_k || \sum_{l=1}^{d} \mathbf{x}_k^l * \mathbf{f}_{\mathrm{con}}^l - \mathbf{y}_k||^2 + \sum_{l=1}^{d} ||\mathbf{w}_{\mathrm{con}} \cdot \mathbf{f}_{\mathrm{con}}^l||^2, \quad (2)$$

here, $\mathbf{w}_{\mathrm{con}}$ and $\mathbf{f}_{\mathrm{con}}$ denote weight function and filter in context tracking, respectively. The generation of context spatial regularization $\mathbf{w}_{\mathrm{con}}$ is as follows. Firstly, calculate the maximum and minimum value of weight $\mathbf{w}_{\mathrm{obj}}$: $\max(\mathbf{w}_{\mathrm{obj}})$ and $\min(\mathbf{w}_{\mathrm{obj}})$, then

$$\mathbf{w}_{\mathrm{con}}(m,n) = \max(\mathbf{w}_{\mathrm{obj}}) + \min(\mathbf{w}_{\mathrm{obj}}) - \mathbf{w}_{\mathrm{obj}}(m,n), \quad (3)$$

here, (m,n) denotes a position in image patch. When the model is converted to context tracking, context filter $\mathbf{f}_{\mathrm{con}}$ is initialized. The spatial weight function $\mathbf{w}_{\mathrm{con}}$ mainly focuses on the region residing in the background. Additionally, when in context tracking, the original object tracking filter $\mathbf{f}_{\mathrm{con}}$ is set aside, which contains the object state and waits until reliable foreground appearance model occurs again.

3.3. Selective Tracking

The difficulty of our method is how to scale the reliability into a certain range for judging the reliability of foreground appearance model. Here, we consider two unreliable situations: object occlusion and multiple similar objects.

Object occlusion. In the first frame, given bounding box $\mathbf{F}$, we apply the newly initialized filter to the image patch and derive the response $\mathbf{R}$, i.e. the confidence map in which each value $\mathbf{R}_{ij}$ corresponds to a bounding box $\mathbf{B}_{ij}$ with the same size as the target. The overlap ratio o_{ij} between the bounding box $\mathbf{B}_{ij}$ and $\mathbf{F}$ can be calculated as,

$$o_{ij} = \frac{\mathbf{B}_{ij} \bigcap \mathbf{F}}{\mathbf{B}_{ij} \bigcup \mathbf{F}}, \quad (4)$$

Then we consider the Ridge Regression model to train samples $\{\mathbf{O}, \mathbf{R}\}$ in which matrix $\mathbf{O}$ indicates a collection of o_{ij}. The goal of training the new generated samples is to find a function $f(\mathbf{z}) = \theta^T \mathbf{z}$ that minimizes the squared error between samples $\mathbf{z}_i$ and regression label $\mathbf{g}_i$,

$$\min_{\theta} \sum_{i} (f(\mathbf{z}_i) - \mathbf{g}_i)^2 + \eta ||\theta||^2, \quad (5)$$

Here, η denotes the regularization parameter which controls over fitting, and its close-form solution is given by,

$$\theta = (\mathbf{Z}^T \mathbf{Z} + \eta \mathbf{I})^{-1} \mathbf{Z}^T \mathbf{g}, \quad (6)$$

here, the data matrix $\mathbf{Z}$ is composed of $\mathbf{z}_i$, and each component of $\mathbf{g}$ is a regression label $\mathbf{g}_i$, $\mathbf{I}$ is an identity matrix. We eventually derive the desired function $f(\mathbf{z}) = \theta^T \mathbf{z}$. In each frame, after utilizing the maximum response into regression model, the overlap_ratio is derived. If the overlap_ratio is low enough, the target is considered to be severely occluded.

Multiple similar objects. In real life, it's common to see there are multiple similar objects in a scene, such as lots of people running in marathon. It's at times too difficult to divide one person from others in such a big scene. Therefore, when meeting this situation, we consider tracking the context with a larger patch than that of the target and then use context tracking to instruct the original object tracking. By this way, more information can be extracted from the context patch, and thus increasing accuracy and robustness. The SRDCF tracker utilizes spatial regularization to penalize the region residing in the background, thus if we want to judge whether their exist similar objects, we should do our work without using spatial regularization. Afterwards, we calculate the average response in object and context region respectively. Then, in each frame, the similar ratio s can be derived as,

$$s = \frac{\mathrm{ave}_{\mathrm{bg}}}{\mathrm{ave}_{\mathrm{fg}}}, \quad (7)$$

here, $\mathrm{ave}_{\mathrm{bg}}$ and $\mathrm{ave}_{\mathrm{fg}}$ denote the average response of the context region and the image region, respectively. If the similar_ratio is high enough, the target is considered to be surrounded with multiple similar objects.

4. EXPERIMENTS

4.1. Setup

In this section, we evaluate our tracker on benchmark VOT TIR-2015 with 20 challenging TIR video sequences. Challenging factors include motion change, camera change, dynamics change, occlusion and size change. TIR itself is challenging due to lack of enough appearance information and it is more likely for TIR data meeting unreliable appearance. Therefore, we use TIR to evaluate our tracker. Weight function $\mathbf{w}_{\mathrm{con}}$ is generated by function $\mathbf{w}_{\mathrm{con}}(m,n) = \mu + \eta(m/R)^2 + \eta(n/C)^2$ in which $R \times C$ denotes target size, μ and η is set 0.1 and 3 respectively. Thresholds of overlap_ratio and similar_ratio are set to 0.55 and 0.6, respectively. Other parameters are the same as the baseline tracker, readers can refer to [13]. We compare our method with 11 state-of-the-art tracking methods: SRDCF [13], NSMAF [19], scalable kernel correlation filter with sparse feature integration (sKCF) [20], multi-channel Multiple-Instance-Learning tracker (CMIL) [21], restore point guided kernelized correlation filter (KCFv2) [21], spatio-temporal context tracker (STC) [12], SumShift tracker [22], geometric structure Hyper-Graph based tracker (G2T) [21], flock of trackers (FoT) [23], point-based Kanade Lukas Tomasi colorFilter (PKLTF) [24], ASMS [25].

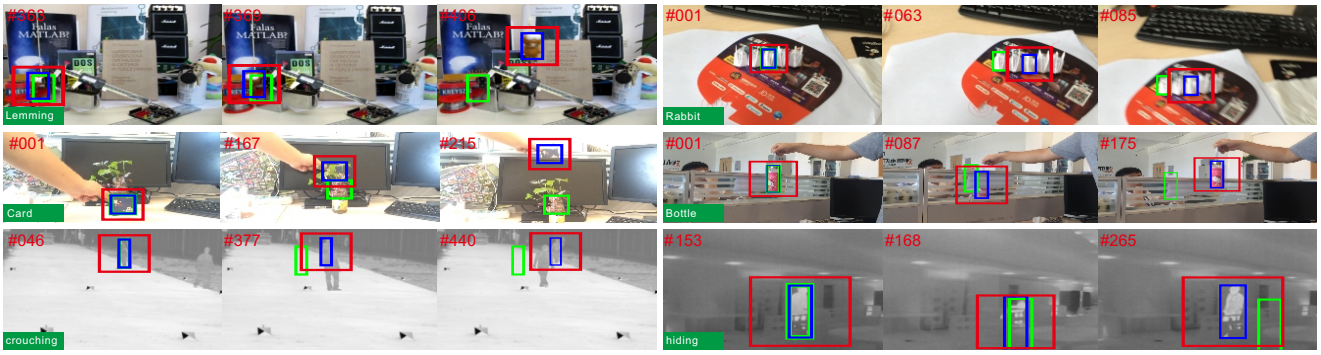

Fig. 3. Some tracking results comparison between SOCT (blue box) and SRDCF (green box).

Table 1. AR rank and overlap of 11 trackers on VOT TIR-2015 dataset. Smaller rank and bigger overlap are better.

	ASMS	PKLTF	FoT	G2T	SumShift	STC	KCFv2	CMIL	sKCF	NSAMF	SRDCF	SOCT
Accuracy Rank	4.35	5.80	6.60	5.05	3.40	3.85	5.05	3.85	3.70	3.45	3.05	**2.65**
Robustness Rank	5.80	6.00	8.85	6.05	3.25	6.70	5.35	5.00	4.20	3.45	2.75	**2.20**
Overlap	0.1179	0.1369	0.1486	0.1548	0.1688	0.1950	0.1998	0.2190	0.2264	0.2382	0.2502	**0.2616**

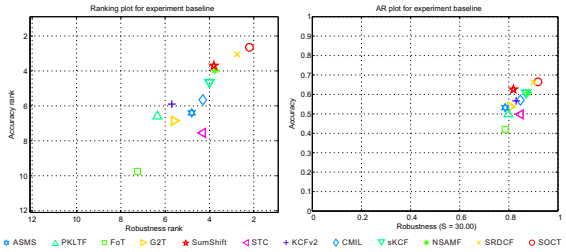

Fig. 4. The AR plots by sequence pooling. It considers the results from all sequences and thus derives a single rank list. Trackers in top right are considered to be better.

We use two measures to evaluate the performance of the above methods: accuracy and robustness. The accuracy measurement measures the overlap ratio between the estimated bounding box and the ground truth; while the robustness measurement measures how many times the tracker fails to track the object. Specifically, robustness is scaled by calculating the probability of tracker failing after S frames.

4.2. Advantages of context spatial regularization

Current datasets contain few related videos, thus we collect some videos to evaluate our tracker. These videos mostly contain occlusion or multiple similar objects. The comparison results are shown in Fig. 3. In sequence Rabbit, both trackers can track the third rabbit in the beginning. However, when the camera motion is fast enough, SRDCF tracks the left first rabbit instead of the third one while our tracker can still track it. In sequence Lemming, Card, bottle and hiding, when the targets are occluded, SRDCF keeps staying in the occlusion position while our tracker is able to track them when the objects show up again. Our approach outperforms the SRDCF tracker with performance improvement, especially for videos with occlusion or multiple similar objects.

4.3. Comparison to state-of-the-arts

We compare our approach with 11 methods and fix the parameters in VOT TIR-2015. The results are shown in Fig. 4 and Tab. 1. In Fig. 4, SOCT keeps in the most top-right indicating the best performance. In Tab. 1, in accuracy and robustness rank we get the best: 2.65 and 2.20, and in overlap we get the highest 0.2616, followed by SRDCF 0.2502. These results show that our method achieves better performance than state-of-the-arts by introducing context model. Besides, the speed of SRDCF and SOCT is about 3fps and 2fps, respectively, showing that our tracker can still operate in real time.

5. CONCLUSION

In this paper, we have proposed selective object and context tracking to handle unreliable foreground appearance model. Our method achieves the best performance on the VOT TIR-2015 dataset comparing with several state-of-the-art trackers. In the first place, we use the proposed two measures to evaluate the reliability of the foreground appearance model based on its confidence map. Then, the context appearance model is selectively learnt and used to track the object according to the reliability of foreground appearance model. In our future work, we will extend our method to adapt other trackers and make it be a general strategy to achieve more robust performance. Besides, we will construct a classifier to replace the two measures to realize more accurate judgment.

Acknowledgements This work is supported by the National Natural Science Foundation of China (NSFC 61671325, 61572354).

6. REFERENCES

[1] K. Zhang, Q. Liu, Y. Wu, and M. H. Yang, "Robust visual tracking via convolutional networks without training," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1779–1792, 2016.

[2] Q. Guo, W. Feng, C. Zhou, and B. Wu, "Structure-regularized compressive tracking," in *ICME*, 2016.

[3] J. Xiao, R. Stolkin, and A. Leonardis, "Single target tracking using adaptive clustered decision trees and dynamic multi-level appearance models," in *CVPR*, 2015.

[4] T. Zhang, S. Liu, C. Xu, S. Yan, B. Ghanem, Narendra Ahuja, and M.-H. Yang, "Structural sparse tracking," in *CVPR*, 2015.

[5] K. Zhang, L. Zhang, and M.-H. Yang, "Fast compressive tracking," *IEEE TPAMI*, vol. 36, no. 10, pp. 2002–2015, 2014.

[6] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *CVPR*, 2012.

[7] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *BMVC*, 2014.

[8] K. Zhang, L. Zhang, M.-H. Yang, and D. Zhang, "Fast tracking via spatio-temporal context learning," in *ECCV*, 2014.

[9] M. Kristan, J. Matas, A. Leonardis, T. Vojir, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Cehovin, "A novel performance evaluation methodology for single-target trackers," *IEEE TPAMI*, vol. PP, no. 99, pp. 1–1, 2016.

[10] Y. Wu, J. Lim, and M.-H. Yang, "Online object tracking: a benchmark," in *CVPR*, 2013.

[11] J. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE TPAMI*, vol. 37, no. 1, pp. 583–596, 2015.

[12] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang, "Fast visual tracking via dense spatio-temporal context learning," in *ECCV*, 2014.

[13] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, "Learning spatially regularized correlation filters for visual tracking," in *ICCV*, 2015.

[14] F. Michael, V. Toma, N. Georg, P. Roman, B. Amanda, H. Gustav, A. Jorgen, K. Matej, M. Jiri, and L. Ale, "The thermal infrared visual object tracking vot-tir2015 challenge results," in *ICCV Workshop*, 2015.

[15] Z.Q. Sun, H.X. Yao, S.P. Zhang, and X. Sun, "Robust visual tracking via context objects computing," in *ICIP*, 2011.

[16] L. Cerman, J. Matas, and V. Hlavac, "Sputnik tracker: Having a companion improves robustness of the tracker," in *SCIA*, 2009.

[17] H. Grabner, J. Matas, L.V. Gool, and P. Cattin, "Tracking the invisible: Learning where the object might be," in *CVPR*, 2010.

[18] D.S. Bolme, J.R. Beveridge, B.A. Draper, and Y.M. Lui, "Visual object tracking using adaptive correlation filters," in *CVPR*, 2010.

[19] Y. Li and J. Zhu, "A scale adaptive kernel correlation filter tracker with feature integration," in *ECCV*, 2014.

[20] A.S. Montero, J. Lang, and R. Laganiere, "Scalable kernel correlation filter with sparse feature integration," in *ICCV*, 2015.

[21] M. Kristan, J. Matas, A. Leonardis, and T. Vojir, "A novel performance evaluation methodology for single-target trackers," *IEEE TPAMI*, vol. 1, 2015.

[22] J.Y. Lee and W. Yu, "Visual tracking by partition-based histogram backprojection and maximum support criteria," in *ROBIO*, 2011.

[23] T. Vojir and J. Matas, "The enhanced flock of trackers," *Registration and Recognition in Images and Videos*, vol. 532, no. 2, pp. 113–136, 2014.

[24] A. Gonzalez, R. Martin-Nieto, J. Bescos, and J. M. Martinez, "Single object long-term tracker for smart control of a ptzcamera," in *International Conference on Distributed SmartCameras*, 2014.

[25] T. Vojir, J. Noskova, and J. Matas, "Robust scale-adaptive mean-shift for tracking," *Pattern Recognition Letters*, vol. 49, no. 3, pp. 250–258, 2014.