Journal of Engineering Sciences

Vol 11, Issue 1, JAN/2020
ISSN NO:0377-9254

A Neoteric Approach for Visual Tracking using SURF Features

1L Ganesh Kumar, 2Shaik Taj Mahaboob

1PG Student, 2Assistant Professor
1ganeshkumarlakshmana@gmail.com
1Electronics and Communication Engineering,
1JNTUA College of Engineering Pulivendula, Pulivendula, India

***Abstract*— Latterly, a discriminatively acquired correlation filter (DCF) contains enticed ample Heed the monitoring culture for visual objects. DCF's successful outcome may be attributed to the confidence of using a huge aggregate of samples to inform and educate the peak regression model and estimate an object's location. To answer the regression muddle, these samples are all provoked by circularly altering from an examining area. These samples of replicas generate some negative results that exhaust its robustness of trackers based on DCF. In this article, different types of the visual tracking approaches employed previously to the present proposed tracker are proposed which has a single conventional layer. Instead of closely understanding its linear regression model, the Convolutionary Regression with Visual Tracking(CRT) track tries to view the regression setbacks by minimizing the gradient descent (GD) level to one-channel-output convolution level. In CRT method the kernel range of the convolution layer is adjusted to the object range. Antithetical with DCF, all "true" samples obtained from the whole picture can be consolidated. An interpretative limitation of the GD process is that negative observations are more frequent in convolutionary samples and the availability of positive observations is reduced. To order to address this problem, an unusual objective feature is introduced in the CRT tracker to eliminate the simple negative and reinforce the positive. This work can be more enhanced by extracting SURF features to track the object more accurately.**

***Keywords*— DCF-trackers, Gradient descent, Surf features.**

I. INTRODUCTION

Concerning inclusive visual tracking quest, the target here is to identify the state of the identified entity in an image sequence. Visual tracking is extremely superior because the number of positive samples is extremely restrained and negatives are nearly unrestrained when competed to general computer insight drawback.

Discriminative algorithms had performed a notable part in visual tracking recently. These discriminative algorithms can be classified in two types. Firstly it represents Item with produced attributes such as initial RGB colours, HOG [1] and Color Names [2], or VGGNet [3] and ResNet [4] which are deeply learned convolutional features. Secondly the initial image can be used to learn a discriminative classifier. DCF's basic idea is to build a classifier for peak regression.

In CRT [5], these issues are addressed in a unusual method by introducing a framework. The new approach presented in this method is somewhat different from DCF. Instead of searching a logical solution for the regression problem, it tries to obtain a significant resolution through gradient descent (GD).

Kai chen et al., [5] provided a different System for learning large-scale regression models for single-layer visual tracking. This novel method is minimal but performs better than utmost remaining DCF based trackers. They performed experiments on four popular datasets for visual tracking, in which the proposed tracker performed better than utmost remaining DCF based trackers in many aspects.

Discriminative Correlation Filters (DCF) is a learning technique that efficiently utilizes all cyclic shifts of the training samples. It has being played a key role in visual object tracking. Due to circular correlation within formulation, DCF is capable to get complete use of restricted training data. The main disadvantage is that it relies on periodic assumption which leads to production of unsolicited boundary effects, leading to an inaccurate portrayal of image.

The method Speeded Up Robust Features (SURF)[6] is a simple and robust optimization for representing objects and comparing invariant functions of local consistency. Similarly, to several other local descriptor-based strategies, the interest targets of a provided images are identified as basic characteristics of a scale-invariant description. Such a multi-scale visualization is given by transforming the initial object with separate multi-scale distributions (box filters). The real trick is to develop invariant direction descriptors using local gradient measurements (intensity and direction). The primary interest of the SURF method is the fast measurement of the operator utilizing box detectors, It facilitates real-time features such as object tracking and identification.

Multi-scale analysis, such as the SIFT approach, To classify attributes in SURF, together with the first and second level differential operatives, depends on a scale-space model. The uniqueness of the SURF technique is that these functions are facilitated by the box filter strategies. For this function, we will be using the term box-space to differentiate this from the regular Gaussian scale-space. While the space of the Gaussian level is achieved by modifying the initial objects to Gaussian kernels, the object storage is also achieved by modifying the actual image through box filters for different scales.

The article is as follows structured. In Section 2, in Visual Tracking, we give a review of the previous framework. We define the approach proposed for visual tracking in Section 3. In order to analyze accuracy and robustness in Section 4, the input images are analyzed at different VGGNet. In addition, a simple and effective first-

line indexing technique is proposed, based on the contrast between the point of interest and its surroundings. The article is concluded in the section on Conclusion.

II. Related Work

The visual tracking system and different techniques used in visual tracking are discussed in this chapter.

A. *Kernelized Correlation Filters*

Kernelized Correlation Filters (KCF) can handle multiple channel features that increase the accuracy of tracking results. The proximity of two image patches can be calculated by performing correlation between them based on convolutionary theorem. In the DFT correlation filter-based tracker, object patches are passed to the frequency domain and correlation can be measured in this domain. In the scope of the target scale variability, this tracker is unreliable due to the restriction of the bounding box[7]..

B. *Structured Tracking with Kernels*

Standardized Output Detection with Kernels (Struck) is also an inverse slant and frames that total monitoring problem as among the standardized output estimation where the job is to directly predict the variation in object presentation within frames. A novel named Struck which includes learning and monitoring in this process, eliminating need for ad-hoc upgrade techniques. In this standardized output process, SVM system estimation is proposed. The Struck monitor follows closely, but with few major differences, the overall design of a monitoring-by-detection method described in the previous paragraph. First change is that rather than learning a differential classifier, we acquire a classification function, which calculates the object's transition between consecutive objects directly. The difference in that they use a rational method for limiting the number of assistance vectors we maintain at the end of both the initial period. As we will see in the near future, our solution uses a Kernelized SVM that has to specifically manage a set of vectors [8].

C. *Spatially Regularized DCF*

The Spatial regularization function is introduced to penalize filter coefficients within the boundaries of the goal. It allows for larger samples to be trained and detected, as well as increasing discriminative power and more robust tracking. Since the background parameters in the system are reduced by allocating large quantities, a naive enhancement of the standard deviation would also lead to a slight increase in the effective filter size. resulting in a high emphasis on background features, reducing the ability of the learned template to discriminatively[9].

D. *DeepSRDCF*

The DCF-based strategy is based on handcrafted image elucidation features. Instead, DCF-based tracking uses convolutionary layer activations. This network includes five layers of convolution, providing the best tracking results. DeepSRDCF is more reliable than SRDCF and other trackers. It incorporates both deep characteristics of convolution and the framework of SRDCF. Complexity increases due to the extraction of deep features and it limits the performance in real time even on high-end GPUs [10].

E. *Continuous Convolutional Operators*

A new formulation is introduced in the C-COT method to train continuous convolution filters. It's an indirect optimisation method to address the obtaining issue in the discrete spatial domain. It also incorporates deep function maps with multi-resolution, leading to exclusive performance. It has a sub-pixel location advantage and effectively handles all available information[11].

III. Convolutional Regression

We will analyze the regression model used in this paper in this section.

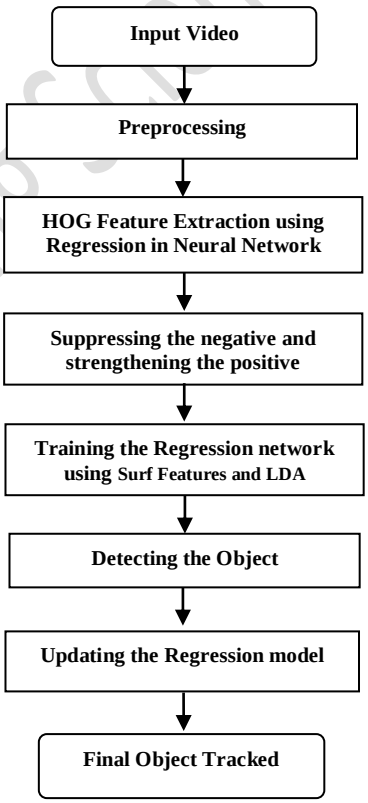

Figure 1 : Hierarchical method of the Proposed Method

The flow diagram of the proposed method can be seen in Figure 1 above. The input video is transformed into frames in the pre-processing stage. In the next step, the regression is used to extract features in the Neural network. Then use Gradient Descent(GD) to suppress the negative and enhance the positive. The regression network is equipped with the SURF function extractor. LDA is also applied for dimensionality reduction. The entity is identified after training the network and the regression model is modified with these findings. Such three phases of learning, identification and upgrading are recognized in combination as a visual monitoring system focused on convolutional regression.

A. Regression Model

This paper considers the regression model in CRT[5] that can be used for visual monitoring as in DCF[7]. Because of a primary image with a defined target, we can obtain several samples of training $X \in \mathbb{R}^{mxn}$, and the corresponding targets for regression $Y \in \mathbb{R}^{m}$. In this case, m is the number of learning samples and n is the sample size. The aim is to learn the coefficients ω for the function of regression $f(x) = \omega^T . x$, by lowering the function below,

$$\arg_{\omega} \min \| X\omega - Y\|^2 + \lambda\|\omega\|^2. \quad (1)$$

Here, $\|.\|$ is the Euclidean norm, and λ is a regularization parameter that regulates overfitting. This problem can be solved in a closed form.

$$\omega = (X^T X + \lambda I)^{-1} X^T Y. \quad (2)$$

Nonetheless, the problem of regression with equation (2) becomes computationally prohibitive when *m* and *n* increase. A workaround is proposed in the DCF method to produce samples by shifting the query patch from one base circularly. It is then possible to simplify the equation (2) for effective calculation. Here we try to solve the regression problem differently.

Instead of generating cyclic samples from a single patch as in DCF, the training and detection samples are collected by sliding a window over the image provided. Using a one-channel convolution sheet, the effects of the sample regression can be estimated. Therefore, to eliminate convolutionary features, the kernel size is set to the object size in feature space instead of the normal size of 3×3 or 5×5 kernel.

B. Suppressing the negative and strengthening the positive

The number of negative samples is nearly infinite in visual tracking, whereas the number of positive samples is very limited. Consequently, positive sample results would be overshadowed by the prevailing negatives. That makes training the regression model difficult to accurately predict the positive.

An improved purpose feature is proposed to address these issues. With low regression error, many negatives can be expected, a truncated loss function to remove these basic negative effects, termed as

$$T(e) = \begin{cases} e, & \text{if } |e| \geq th \\ 0, & otherwise \end{cases} \quad (3)$$

Where *th* is a manually defined threshold. Now this function is applied to the regression errors that are the first term in equation (1), it will exclude the contributions made by simple negatives to change the coefficients ω using GD. Potentially, the positive and hard negative will not be affected.

A weight function is established to boost the contributions of the positive. The weight function is labeled as,

$$W(y) = \exp(a \, . \, y). \quad (4)$$

Here *y* denotes a sample's regression target. A positive sample is usually labelled with a higher y than a negative sample.

The main objective of implementing this weight function is that the right samples should be predicted in visual tracking, i.e. instead of predicting negatives it should predict positives accurately.

Ultimately, it is possible to define the objective function as,

$$\arg_{\omega} \min \| W(Y) \; \Theta \; T(X\omega - Y)\|^2 + \lambda\|\omega\|^2, \quad (5)$$

where Θ indicates the Hadamard product. Remember that the above equation (5) simplifies equation (1) when *a* is set to 0 and *th* is set to 0.

The signal-to-noise(SNR) is used to calculate how well the event is expected. The signal here means the object patch's regression result, and the noise means the mean regression results of a given training patch, so the SNR can be defined as,

$$SNR(M) = \exp(\max(|M) - \text{mean}(M)) \quad (6)$$

Where, for approximating the signal, the peak M is used, and the mean M is for approximating the noise.

IV. Experiments

We conduct comprehensive experiments with 50 sequences on common OTB 50 datasets and 100 sequences on OTB 100. It should be remembered that the dataset OTB 50 is a very difficult subset of OTB 100 and is distinct from the dataset OTB - 2013.

A. Experiment setup

The threshold *th* in equation (3) is set to 0.1 and the regularization λ is fixed to 1.0 in equation (5). Technically, due to its larger size, training patch may take longer to remove the functionality. The size of the search patch can be matched with the training patch to make few improvements. By matching the volume, the extracted features can be reused in the upgrade phase.

B. Feature Selection

This framework can be easily integrated into the handcraft features such as original RGB colors, HOG[2] etc..,. On the other hand, convolutional features learned from deep CNN such as VGGNet[3] and ResNet[4] can outperform this framework. The Convolutionary layers in VGGNet are transferred in such a way that the image patches can extract deep convolutionary features.

The network used for feature selection is the design of ImageNet VGG-m-2048 in which the phase and smaller receptive area in the first convolutionary layer has decreased, Conv2 uses larger phases (two instead of one) to keep the computation time appropriate. As with other architectures, Conv3, Conv4 and Conv5 have a single step. The last hidden layer(Full7) initially has 4096-D dimensional representation of the object, which is high dimensional. Therefore, for this study, lower dimensions 2048, 1024 and 128, 2048 are

considered. Output enhancement can be observed by choosing this VGG-m-2048 architecture. The experiments on various ImageNet-vgg- architectures have been performed. Table 1 displays the experimental results and its performance can be seen in Figure 2.

C. Principal Component Analysis(PCA)

The principal component analysis is used to reduce the dimensionality of the dataset of many variables. Such variables are combined and balanced to create a new set of variables known as orthogonal Principal Components(PC). In a covariance matrix, these PCs are said to be their own vectors and therefore they are said to be orthogonal.

Originally in PCA, the data remains consistent to the extent where their mean is zero. Then the matrix of covariance is determined by calculating own values and own vectors, now the vector of the function is generated by selecting components. Ultimately, a combination of the transposition of the function vector and the transposition of scaled data form the main components.

D. Linear Discriminant Analysis(LDA)

Three steps are to be performed to project the original data matrix to a lower dimensional space. The first step is to determine the separability among various classes, the difference between both the mean and the samples on each class, known as Within class variance, in the second step. In the final step, the lower dimensional space is constructed to maximize the variance of the class also reduce the Inside class variability.Feature Based Evaluation on OTB

Visual tracking's main difficulties are commonly from many aspects. Examining how satisfactory a tracker deals with the various disputes will be noteworthy.

The datasets considered are evaluated based upon the 11 attributes. In terms of accuracy and robustness, the test results can be seen in Table I.

E. Qualitative Evaluation on OTB

The track outcomes of the datasets from OTB -100 and OTB- 50 are shown in the Figure 3. The reference frame along the tracking frame is shown in Figure 3. Each reference frame is that dataset's initial frame and is compared to the remaining frames to detect the object. The objects are solid in the Girl2 series, but the background is mixed up. In DragonBaby and Bird1 sequences, the targets are deformed and rotated, allowing proper visual identification of the images to be troublesome. In this form, the finest resolution uses the negative samples that have been identified that the tracker does not glide into the backdrop. In this algorithm, to renew the discriminative prototype, the extensive negative samples near the target are integrated. Which in DragonBaby and Bird1 can effectively track the targets. Different parameter measurements are measured, and we can observe the Sensitivity and Specificity, Precision and Recall in the Figure 4.

TABLE I.

EVALUATIONS OF THE IMAGENET-VGG-F, IMAGENET-VGG-M, IMAGENET-VGG-S AND IMAGENET-VGG-M-2048 WITH DIFFERENT FEATURES ON OTB-50.

Dataset	ImageNet-vgg-f		ImageNet-vgg-m		ImageNet-vgg-s		Imagenet-vgg-m-2048	
	Accuracy	Robustness	Accuracy	Robustness	Accuracy	Robustness	Accuracy	Robustness
Bird1 (DEF,FM,OV)	86.3464	0.810	86.7508	0.803	83.7112	0.820	83	0.82
Crossing (SV,DEF,FM, OPR,BC)	98.6877	0.989	98.8266	0.988	98.6786	0.987	99	0.993
David3 (OCC,DEF, OPR,BC)	74.6384	2.198	81.5231	2.103	83.3267	2.114	73	2.24
Dog (DEF,SV,OPR)	84.78	3.585	83.4549	3.520	83.4549	3.52	83.54	3.520
DragonBaby (OCC,SV,MB, FM,IPR,OPR,OV)	79.0924	1.838	78.5873	1.838	78.3752	1.837	89.93	1.87
Girl2 (SV,OCC,DEF, MB,OPR)	80.7358	3.386	85.1807	3.301	80.7755	3.323	84.99	3.284

Vol 11, Issue 1, JAN/2020
ISSN NO:0377-9254

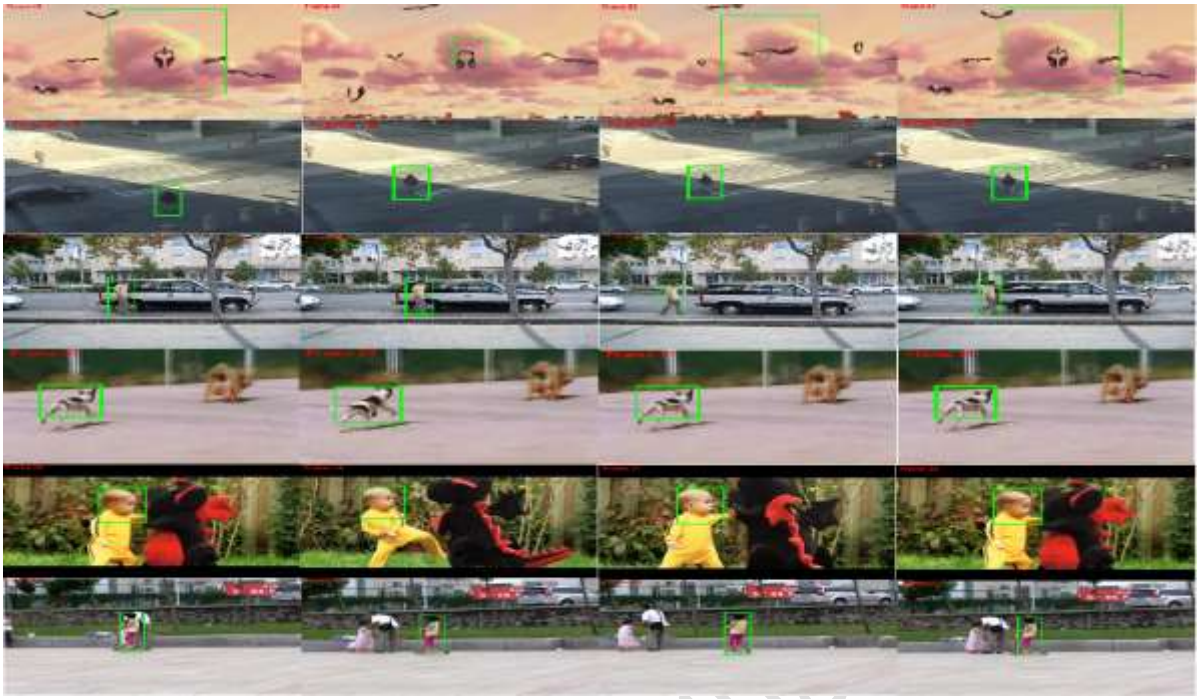

Figure 2 : Output results of VGG-m-2048 architecture performed on Bird1, Crossing, David3, Dog, DragonBaby and Girl2 respectively.

Figure 3 : Track results of the algorithm on the datasets David3, Jogging, Human9 and Tiger1 respectively.

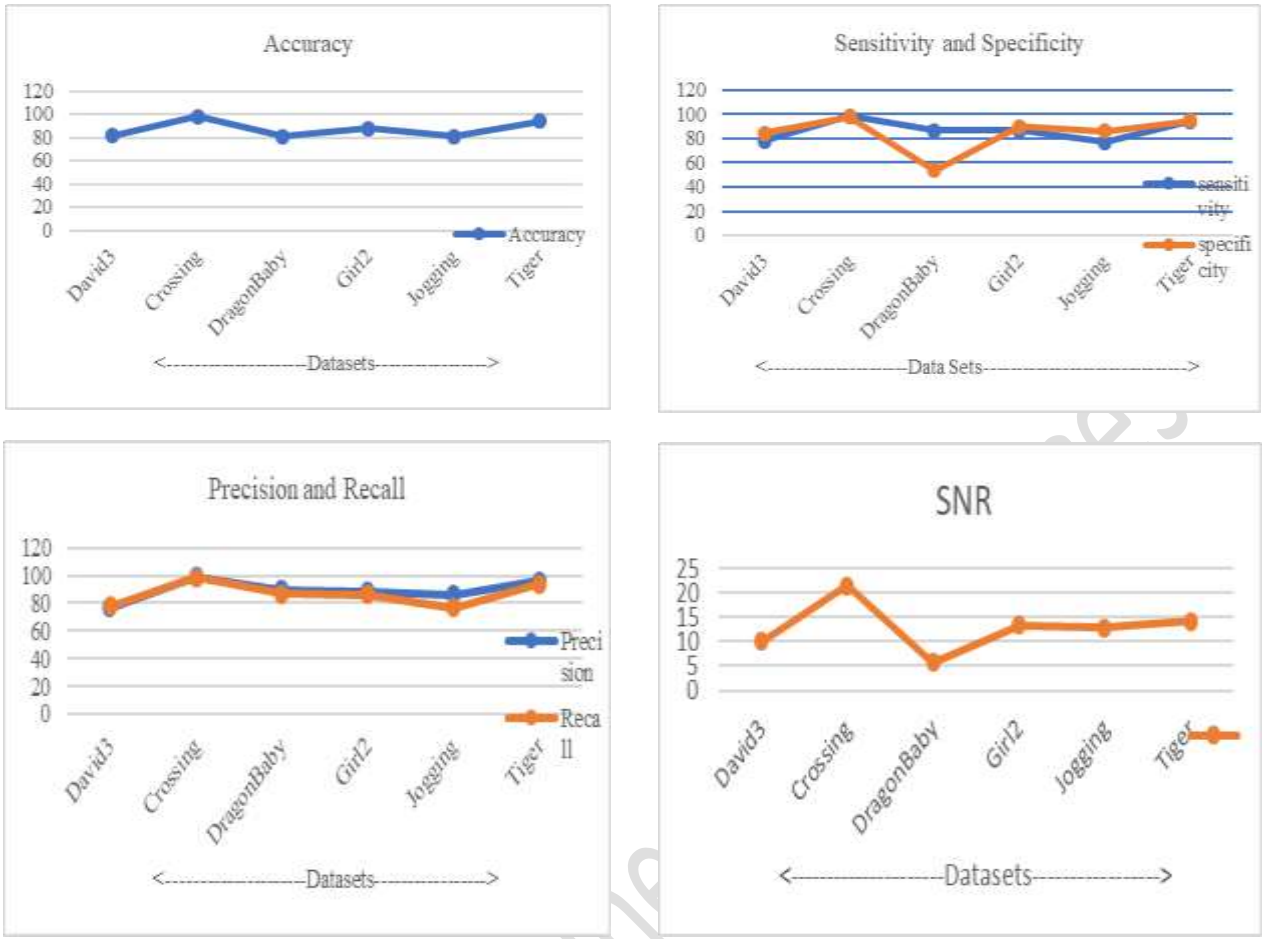

Figure 4 : Graph results of the parameters Accuracy, Sensitivity and Specificity, Precision and Recall and SNR.

CONCLUSION

A new visual monitoring technique using surf features in this article. In the algorithm, a simple ridge regression method for visual monitoring is fitted with a single layer of down-propagating regression defects of convolution using the gradient descent technique. The tracking system can include virtually unlimited "actual" samples compared to the DCF approach. We also suggest an enhanced target feature to remove simple negative effects and boost positive effects to speed up the training process. Our detailed experiments show that the current approach is much more powerful than that of the DCF method references

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645,Sep. 2010.

[2] M. Danelljan, F. S. Khan, M. Felsberg, and J. van de Weijer, “Adaptive color attributes for real-time visual tracking,” in Proc. IEEE Conf.Comput. Vis. Pattern Recognit., Jun. 2014, pp.1090–1097.

[3] K. Simonyan and A. Zisserman, “Very deep convolutional networks forlarge-scale image recognition,” in Proc. Int. Conf. Learn. Represent.,2015, pp. 1–14.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CoRR, vol. abs/1512.03385, pp. 1–12, Dec. 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[5] Kai Chen and Wenbing Tao," Convolutional Regression for Visual Tracking", in Proc. IEEE Trans. on Image Proc.vol.27, NO. 7, 2018, pp. 3611-3620.

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool,"Speeded-Up Robust Features (SURF)",in Computer Vision and Image Understanding,June 2008,pp 346-359.

[7] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, “High speed tracking with kernelized correlation filters,” IEEE Trans.Pattern Anal.Mach. Intell., vol. 37, no. 3, pp. 583–596, Mar. 2015.

[8] S. Hare, A. Saffari, and P. H. S. Torr, “Struck: Structured output tracking with kernels,” in Proc. IEEE Int. Conf. Comput.Vis.,Nov. 2011,pp. 263–270.

[9] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, “Learning spatially regularized correlation filters for visual tracking,” in Proc. IEEE Int.Conf. Comput. Vis., Dec. 2015, pp. 4310–4318.

[10] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg, “Convolutional features for correlation filter based visual tracking,” in Proc.IEEE Int.Conf. Comput. Vis. Workshop, Dec. 2015, pp. 621–629.

[11] M. Danelljan, A. Robinson, F. S. Khan, and M. Felsberg, “Beyond correlation filters: Learning continuous convolution operators for visual tracking,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 472–488.