BIT: BIO-INSPIRED TRACKER

Bolun Cai, Xiangmin Xu, Xiaofen Xing, Chunmei Qing*

School of Electronic and Information Engineering
South China University of Technology, Guangzhou, China
caibolun@gmail.com,{xmxu,xfxing,qchm}@scut.edu.cn

ABSTRACT

Visual tracking is a challenging problem due to various factors such as deformation, rotation and illumination. As is well known, given the superior tracking performance of human vision, bio-inspired model is expected to improve the computer visual tracking. However, the design of bio-inspired tracking framework is challenging, due to the incomplete comprehension and hyper-scale of senior neurons, which will influence the effectiveness and real-time performance of the tracker. According to the ventral stream in visual cortex, a novel bio-inspired tracker (BIT) is proposed, which simulates shallow neurons (S1 and C1) to extract low-level bio-inspired feature for target appearance and imitates senior learning mechanism (S2 and C2) to combine generative and discriminative model for position estimation. In addition, Fast Fourier Transform (FFT) is adopted for real-time learning and detection in this framework. On the recent benchmark[1], extensive experimental results show BIT performs favorably against state-of-the-art methods in terms of accuracy and robustness.

Index Terms— Bio-inspired model, visual tracking

1. INTRODUCTION

Visual object tracking is one of the fundamental problems of computer vision, with wide-ranging applications including video surveillance, human-machine interfaces and robot perception. Although visual tracking has been investigated intensively in the past decade, it is still an enormous challenge in real application because of various factors such as pose, occlusion, scale and illumination. Recent tracking algorithms can be split into two main modules generally: feature extraction and tracking model.

Current tracking features can be categorized into handcrafted and automated. The design of handcrafted features for tracking (e.g. HoG[2], Haar-like[3], color histogram[4]) is difficult, depending on the time-consuming parameter adjustment; automated features learn appearance model from input images, which can be unsupervised as PCA[5] or supervised as sparse coding[6], but require a good underlying model and decrease the real-time performance. The bio-inspired model avoids the parameter adjustment of handcrafted feature and the parameter learning of automated feature. Xing et al.[7] proposed a tracking-by-detection algorithm based on bio-inspired C2 feature, which incorrectly regards senior learning mechanism as a feature extraction, leading to low accuracy and unsatisfying real-time performance. Based on the biology expert knowledge and heuristics, a low-level bio-inspired feature simulating shallow neurons including S1 and C1 units is proposed for visual tracking, which exhibits a well trade-off between invariance and discrimination.

Existing tracking model can be generally categorized as generative or discriminative. For generative models[8, 9, 10], tracking is formulated as searching for the most similar region to the target object within a neighborhood. Discriminative models[11, 12, 13] treat tracking as a classification problem to distinguish the target object from the background. Therefore, an outstanding model should exploit the advantages of both generative and discriminative methods. Li et al.[14] proposed a simplified biologically inspired feature (SBIF) for object representation, but ignores senior learning mechanism in ventral stream. Combining generative and discriminative method, we proposed a bio-inspired tracking framework: the response of S2 units is a generative model via convolution and the C2 classifier simulates neuronal connection as a discriminative model.

In this paper, a bio-inspired tracker based on the ventral stream is proposed, which outperforms state-of-the-art methods on the recent benchmark. Corresponding to feature extraction and tracking model of traditional trackers, a low-level bio-inspired feature is used for simulating shallow neurons (S1 and C1) and a joint model is used for senior learning (S2 and C2). Importantly, our model exploits FFT to speed up the bio-inspired model and dense sampling.

2. BIO-INSPIRED TRACKER

Visual processing in cortex is modeled as a hierarchy of increasingly sophisticated representations. A recent theory[15] of the feed-forward path of object recognition in visual cortex accounts for the ventral stream processing from primary visual cortex (V1) to prefrontal cortex (PFC). In the ventral stream, a HMAX model[16] proposed for object recognition

*Xiangmin Xu is the corresponding author.

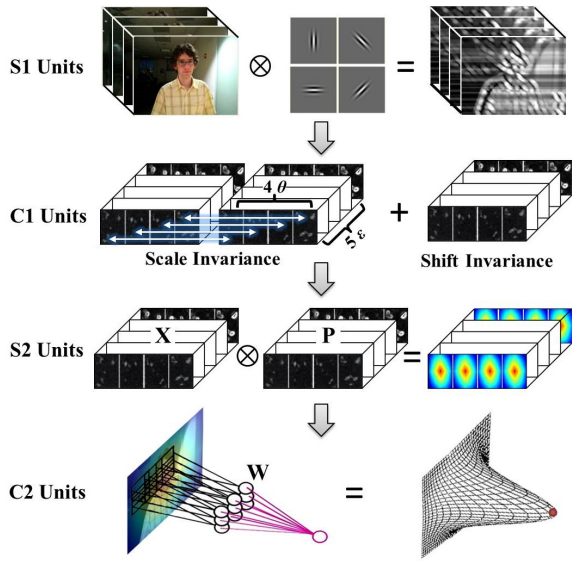

Fig. 1. Bio-inspired Tracker

is particularly designed for visual tracking, which contains alternating layers called simple (S) and complex (C) cell units. According to the primate visual pathway, the bio-inspired framework proposed is shown as Fig.1, which includes S1 units modeling primary visual cortex, C1 units simulating cortical complex cells, S2 and C2 units corresponding to the learning mechanism of senior neurons.

2.1. S1 units: classical simple cells

In the primary visual cortex (V1) [17], simple cell receptive field has the basic characteristics of multi-directional, multi-scale and multi-frequency selection. S1 units can be described by Gabor filters [18], which have been shown to provide an appropriate model of cortical simple cell receptive fields and are described by the following equation (1):

$$G\left(x,y,\theta,s(\delta,\lambda,\gamma)\right)=\exp\left(-\frac{\left(X^2+\gamma^2Y^2\right)}{2\sigma^2}\right)\times\cos\left(\frac{2\pi}{\lambda}X\right),\quad s.t.\quad X=x\cos\theta+y\sin\theta, Y=-x\cos\theta+y\sin\theta \tag{1}$$

where the filter patch coordinate (x, y), the orientation θ, scales s with 3 parameters (bandwidth δ, wavelength λ and aspect-ratio γ). According to [18], we arranged a series of Gabor filters to form a pyramid of scales, spanning a range of sizes from 7×7 to 25×25 pixels in steps of two pixels to model the receipt-field ξ of the simple cells (the parameter values shown in Table 1 for details). The filters come in 4 orientations ($\theta = \pi/4, \pi/2, 3\pi/4, \pi$) and 10 scales s that are arranged in 8 bands ε, thus leading to 40 different S1 receptive field types total. The responding result of classical simple cells can be described as (2).

$$S1\left(x,y,\theta,s\right)=I\left(x,y\right)\otimes G\left(x,y,\theta,s\right), \tag{2}$$

Table 1. Summary of parameters used in BIT framework

Band ε	**Scale** s	**Receipt-field** ξ	δ	λ	γ
1	1	7×7	2.8	3.5	0.23
	2	9×9	3.6	4.6	0.28
2	3	11×11	4.5	5.6	0.32
	4	13×13	5.4	6.8	0.37
3	5	15×15	6.3	7.9	0.41
	6	17×17	7.3	9.1	0.46
4	7	19×19	8.2	10.3	0.51
	8	21×21	9.2	11.5	0.55
5	9	23×23	10.2	12.7	0.60
	10	25×25	11.3	14.1	0.64

where $I\left(x,y\right)$ is the original gray-scale image of tracking sequences.

2.2. C1 units: cortical complex cells

The cortical complex cells (V2) receive the response from simple cells and play the role of primary linear feature integration. C1 units [19] correspond to complex cells, which show the invariance to scale and shift: complex cells tend to have broader spatial frequency (scale invariance) and larger receptive fields (shift invariance). Ilan et al.[20] proposed the spatial integration properties of complex cells can be described by a series of maximum pooling operation.

- *Scale invariance:* There are 5 bands for a total of 10 different scale filter sizes, each of which contains two adjacent S1 filters sizes shown in Table 1. In this stage, the maximum value is recorded over the two scales maps within the same spatial location.
- *Shift invariance:* The C1 unit responses are computed by subsampling these maps using a cell grid Σ of size 5×5. From each grid cell, one single measurement is obtained by taking the maximum of all 25 elements.

  In summary, C1 response can be described as follow:

$$C1\left(x,y,\theta,\varepsilon\right)=\max_{(x,y)\in\Sigma}\left(\max_{s=\{2\varepsilon,2\varepsilon-1\}}S1\left(x,y,\theta,s\right)\right) \tag{3}$$

2.3. S2 units: shape-tuned learning

The tuning properties of neurons in the ventral stream of visual cortex, from V2 to infer-temporal (IT) cortex, play a key role for visual perception in primates and in particular for their object recognition abilities [21]. This training process can be regarded as a generative model, in which S2 units pool over afferent C1 units within its receptive field. Each S2 unit response depends in a radial basis function (RBF) [18] on the Euclidean distance between a new input X and a stored prototype P. For an image patch from the previous C1 layer, the response r of the corresponding to S2 units is given by:

$$r=\exp\left(-\beta\|X-P\|^2\right), \tag{4}$$

where β defines the sharpness of the tuning coefficient. At runtime, S2 response maps are computed across all positions by (4) for each band of C2 units.

2.4. C2 units: task-dependent learning

The task-specific circuits from IT to prefrontal cortex (PFC) require learning for the discrimination between target objects and background clusters. According to bioresearch[21], the routine running in PFC as a classifier is trained on a particular task in a supervised way and receives the activity of a few hundred neurons in IT. Thus, a convolutional neural network (CNN) could correspond to the task-specific circuits found in C2 units with neurons from IT to PFC as

$$C2\left(x,y\right)=\sum_{\theta,\varepsilon}W\left(x,y\right)\otimes S2\left(x,y,\theta,\varepsilon\right), \tag{5}$$

where W is the connection weights of neural network. In addition, a fast estimate method of W will be introduced in next subsection.

2.5. Real-time bio-inspired tracker via FFT

The real-time performance is an important index of object visual tracking method. A lot of tracking approach[10, 22, 4] has been tracking-by-detection, which stems directly from the development of discriminative methods in machine learning. Almost all of the proposed tracking-by-detection methods were based on a sparse sampling strategy. In each frame, a small number of samples are collected in the targets neighborhood by particle filter, because the cost of not doing so would be prohibitive. Therefore, speeding up the dense sampling of S2 and C2 response calculation is a key point of BIT. In this subsection, a real-time BIT based on dense sampling via Fast Fourier transform (FFT) will be introduced.

S2 units: According to (4), we know S2 units respond corresponds to a kernel method based on RBF, which can be rewrite similar to linear function as follow (6), when the RBF is a standard normal function ($\beta=1/2\sigma^2$).

$$\begin{aligned} r&=\exp\left(-1/2\sigma^2\|X-P\|^2\right)\\ &=\exp\left(-1/2\left(X^TX+P^TP-2X^TP\right)\right)\\ &\sim\exp\left(X^TP\right)\sim X^TP \end{aligned} \tag{6}$$

Furthermore, linear kernel is usually preferred in time-critical problems such as tracking, because the weights vector can be computed explicitly. At time t , S2 units dense respond map was calculated by a linear function instead of RBF as follows.

$$S2_{t+1}\left(x,y,\theta,\varepsilon\right)=C1_{t+1}\left(x,y,\theta,\varepsilon\right)\otimes C1_t^P\left(x,y,\theta,\varepsilon\right) \tag{7}$$

We note (7) can be transformed to the frequency domain, in which FFT can be used for fast convolution. that is,

$$\mathcal{F}\left[S2_{t+1}\left(\cdot,\theta,\varepsilon\right)\right]=\mathcal{F}\left[C1_{t+1}\left(\cdot,\theta,\varepsilon\right)\right]\odot\mathcal{F}\left[C1_t^P\left(\cdot,\theta,\varepsilon\right)\right], \tag{8}$$

where $\mathcal{F}[]$ denotes the FFT function and $\odot$ is the element-wise product.

C2 units: As with S2 units, FFT algorithm also can be used for fast convolution and deconvolution in C2 units. Note that the network comprising units with a Gaussian-like tuning function together on their outputs. In order to estimate neuronal connection weights W, the C2 units response map of an object location is modeled as

$$C2\left(x,y\right)=\exp\left(-\frac{1}{2\sigma_s^2}\left(\left(x-x_o\right)^2+\left(y-y_o\right)^2\right)\right), \tag{9}$$

where σ_s is a scale parameter and (x_o,y_o) is the center of tracking target. Therefore, the neuron connection weights W was showed as

$$\mathcal{F}\left[W\left(x,y\right)\right]=\sum_{\theta,\varepsilon}\frac{\mathcal{F}\left[C2\left(x,y\right)\right]}{\mathcal{F}\left[S2\left(x,y,\theta,\varepsilon\right)\right]} \tag{10}$$

The object location $(\hat{x},\hat{y})$ in the $(t+1)$-th frame is determined by maximizing the new C2 response map.

$$\left(\hat{x},\hat{y}\right)=\arg\max_{(x,y)}C2_{t+1}\left(x,y\right), \tag{11}$$

where $C2_{t+1}\left(\cdot\right)=\mathcal{F}^{-1}\left[\sum_{\theta,\varepsilon}\mathcal{F}\left[W_t\left(\cdot\right)\right]\odot\mathcal{F}\left[S2_{t+1}\left(\cdot,\theta,\varepsilon\right)\right]\right]$ and $\mathcal{F}^{-1}[]$ denotes the inverse FFT function.

Updating method: Dependent on the spatial and frequency domains, a classical tracking model updating method is used in this paper. At the t-th frame, the BIT is updated by

$$\begin{cases} C1_{t+1}^P\left(\cdot,\theta,\varepsilon\right)=\rho C1\left(\hat{x},\hat{y},\theta,\varepsilon\right)+\left(1-\rho\right)C1_t^P\left(\cdot,\theta,\varepsilon\right)\\ \mathcal{F}\left[W_{t+1}\left(\cdot\right)\right]=\rho\mathcal{F}\left[W\left(\hat{x},\hat{y}\right)\right]+\left(1-\rho\right)\mathcal{F}\left[W_t\left(\cdot\right)\right] \end{cases}, \tag{12}$$

where ρ is a learning parameter, $C1\left(\hat{x},\hat{y},\theta,\varepsilon\right)$ is the C1 units spatial model computed by (3) and $\mathcal{F}\left[W\left(\hat{x},\hat{y}\right)\right]$ is the frequency model of neural weights computed by (10).

3. EXPERIMENTAL RESULTS

The proposed tracker is implemented in MATLAB 2013A on a PC with Intel Core2 CPU (2.66 GHz) with 2 GB memory, and runs about 10 frames per second (fps) in this platform.

We compared the proposed method with 10 state-of-the-art trackers (Struck[11], SCM[22], TLD[12], CXT[13], VTD[8], VTS[9], CSK[23], ASLA[10], LOT[4], OAB[24]) on the CVPR2013 benchmark [1] that includes 50 sequences. Each sequence is tagged with a number of attributes indicating to the presence of 11 different challenges, including Illumination Variation (IV), Scale Variation (SC), Occlusion (OCC), Deformation (DEF), Motion Blur (MB), Fast Motion (FM), In-Plane Rotation (IPR), Out-of-Plane Rotation (OPR), Out-of-View (OV), Background Clutters (BC), Low Resolution (LR). The best way to evaluate trackers is still a debatable subject. Averaged measures like mean center

Table 2. Tracker precisions ($e = 20$) over 11 subsets (**Red** indicates the best while blue indicates the second best)

	BIT	Struck[11]	SCM[22]	TLD[12]	CXT[13]	VTD[8]	VTS[9]	CSK[23]	ASLA[10]
ALL	**0.693**	0.656	0.648	0.608	0.577	0.576	0.575	0.545	0.532
IV	**0.607**	0.558	0.592	0.537	0.505	0.557	0.572	0.481	0.516
SV	0.652	0.639	**0.672**	0.606	0.550	0.597	0.582	0.503	0.552
OCC	0.631	0.565	**0.639**	0.563	0.494	0.546	0.533	0.500	0.460
DEF	**0.612**	0.521	0.586	0.512	0.422	0.501	0.487	0.476	0.445
MB	0.537	**0.551**	0.339	0.518	0.509	0.375	0.375	0.342	0.278
FM	0.502	**0.604**	0.331	0.551	0.519	0.353	0.351	0.381	0.253
IPR	**0.620**	0.617	0.596	0.584	0.612	0.600	0.578	0.547	0.511
OPR	**0.629**	0.597	0.617	0.596	0.576	0.620	0.603	0.540	0.518
OV	0.388	0.539	0.429	**0.576**	0.510	0.462	0.455	0.379	0.333
BC	**0.689**	0.585	0.578	0.428	0.443	0.571	0.578	0.585	0.496
LR	0.516	**0.545**	0.305	0.349	0.371	0.168	0.187	0.411	0.156

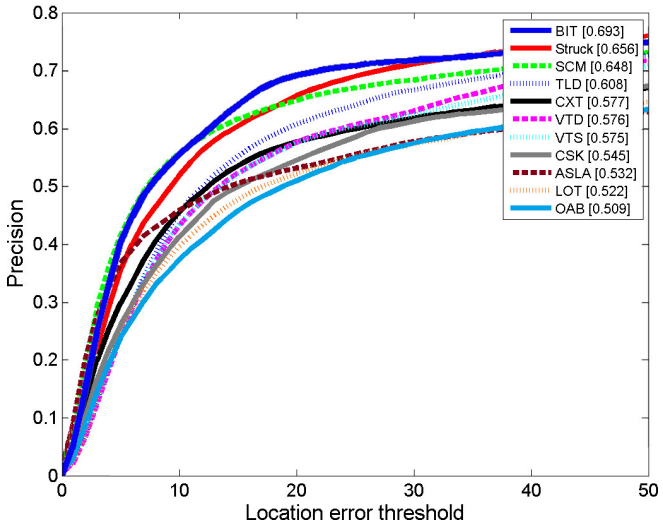

Fig. 2. Precisions plots over all 50 sequences

location error or average bounding box overlap penalize an accurate tracker that fails for short-time more than an inaccurate tracker. According to [1], the precision plot shows the percentage of frames on which the Center Location Error (CLE) of a tracker is within a given threshold e, where CLE is defined as the center distance between tracker output $(\hat{x}, \hat{y})$ and ground truth (x_g, y_g).

Fig.2 shows the precision plots containing the mean error over all the 50 sequences, and a representative precision score ($e = 20$) is used for ranking. In the precision plot, the proposed BIT outperforms Struck[11] by 3.7% in mean CLE at the threshold of 20 pixels. On the other hand, the tracking drift of SCM[22] is less than Struck in the high-precision ($e < 20$), and BIT also acquire the same accuracy as SCM, since SCM and BIT both combine the advantages of generative and discriminative models. The different is SCM employs sparse coding, while BIT adopts bio-inspired model.

Table 2 summarizes the performances between BIT and the top 8 trackers over 11 typical video subsets. Clearly, BIT almost achieved excellent performances in 11 typical challenge subsets, especially on IV, DEF, IPR, OPR and BC. Multi-direction Gabor filters used in S1 units contribute to the robustness of illumination (IV) and rotation (IPR and OPR). C1 units provide the scale and shift competitive mechanism to deal with scale variation (SV) and deformation (DEF). Moreover, the generative model in S2 units and the discriminative model in C2 units rise to the challenges of occlusion (OCC) and background clutters (BC) respectively.

4. CONCLUSION

For the first time, we successfully apply bio-inspired model to real-time visual tracking. Depending on bioresearch, the proposed novel bio-inspired tracker models the ventral stream of primate visual cortex, extracting low-level bio-inspired feature in S1 and C1 units, simulating the learning mechanism of senior neurons in S2 and C2 units. Furthermore, the complicated bio-inspired tracker is still real-time since FFT is used to online learning and detection. Numerous experiments with state-of-the-art algorithms on challenging sequences demonstrated that BIT achieves favorable results in terms of accuracy and robustness. However, C2 units using a single layer convolutional network cannot simulate neurons connection in PFC perfectly, which provides a good starting platform for further research into deep neural network.

5. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (#61171142, #61401163), the Science and technology Planning Project of Guangdong Province of China (#2011A010801005, #2012B061700102).

6. REFERENCES

[1] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in *Computer Vision and Pattern Recognition*. IEEE, 2013, pp. 2411–2418.

[2] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition*. IEEE, 2005, vol. 1, pp. 886–893.

[3] Rainer Lienhart and Jochen Maydt, "An extended set of haar-like features for rapid object detection," in *ICIP*. IEEE, 2002, vol. 1, pp. I–900.

[4] Shaul Oron, Aharon Bar-Hillel, Dan Levi, and Shai Avidan, "Locally orderless tracking," in *Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1940–1947.

[5] 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.

[6] Xue Mei and Haibin Ling, "Robust visual tracking using $l1$ minimization," in *International Conference on Computer Vision*. IEEE, 2009, pp. 1436–1443.

[7] Xiaofen Xing, Suo Qiu, Kailing Guo, and Xiangmin Xu, "Online object tracking algorithm based on biologically-inspired c2 feature," *Journal of South China University of Technology (Natural Science)*, vol. 8, pp. 63–68, 2012.

[8] Junseok Kwon and Kyoung Mu Lee, "Visual tracking decomposition," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1269–1276.

[9] Junseok Kwon and Kyoung Mu Lee, "Tracking by sampling trackers," in *ICCV*. IEEE, 2011, pp. 1195–1202.

[10] Xu Jia, Huchuan Lu, and Ming-Hsuan Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1822–1829.

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

[12] Zdenek Kalal, Jiri Matas, and Krystian Mikolajczyk, "P-n learning: Bootstrapping binary classifiers by structural constraints," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 49–56.

[13] Thang Ba Dinh, Nam Vo, and Gérard Medioni, "Context tracker: Exploring supporters and distracters in unconstrained environments," in *Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1177–1184.

[14] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan, "Robust visual tracking based on simplified biologically inspired features," in *ICIP*. IEEE, 2009, pp. 4113–4116.

[15] Maximilian Riesenhuber and Tomaso Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[16] Thomas Serre and Maximilian Riesenhuber, "Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex," Tech. Rep., DTIC Document, 2004.

[17] John G Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.

[18] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio, "Robust object recognition with cortex-like mechanisms," *Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 411–426, 2007.

[19] David H Hubel and Torsten N Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106, 1962.

[20] Ilan Lampl, David Ferster, Tomaso Poggio, and Maximilian Riesenhuber, "Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex," *Journal of neurophysiology*, vol. 92, no. 5, pp. 2704–2713, 2004.

[21] Thomas Serre, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, Gabriel Kreiman, and Tomaso Poggio, "A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex," Tech. Rep., DTIC Document, 2005.

[22] Wei Zhong, Huchuan Lu, and Ming-Hsuan Yang, "Robust object tracking via sparsity-based collaborative model," in *Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1838–1845.

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

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