Automatic Initialization and Tracking Using Attentional Mechanisms

Vijay Mahadevan Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
vmahadev,nuno@ucsd.edu

Abstract

A biologically inspired approach for automated visual tracking is proposed. In this approach it is hypothesized that target initialization and tracking are a consequence of saliency mechanisms that guide the deployment of visual attention. The recently proposed discriminant center-surround saliency model, is used to derive the tracking framework. In this framework, automatic tracker initialization is achieved using bottom-up saliency with motion features, while the tracking problem is formulated as one of continuous target-background classification, implemented using saliency in two stages. The first, or learning stage, combines a focus of attention mechanism and bottom-up saliency to identify a maximally discriminant set of features for target detection. The second, or detection stage, uses a feature based attention mechanism and a target-tuned top-down discriminant saliency detector, to detect the target. Overall, the tracker iterates between learning discriminant features from the target location in a video frame and detecting the location of the target in the next frame. To implement this tracker, well known properties of the statistics of natural images are exploited leading to computational efficiency. Experimental results comparing the proposed method to the state of the art in tracking are presented, showing improved performance.

1. Introduction

Object tracking is a classical problem in computer vision, and a pre-requisite for many of its important applications, such as surveillance, activity or behavior recognition and video retrieval. Decades of research on this topic have produced a diverse set of approaches and a rich collection of tracking algorithms [37]. Many of the these are based on *appearance modeling*. They learn (and maintain) a model of target appearance, which is used to locate the target as time evolves [17, 8, 4, 19]. The main limitation of these methods is that they rely uniquely on models of object appearance, and do not take the background into account. This limits tracking accuracy when backgrounds are cluttered, or targets have substantial amounts of geometric deformation, such as out-of-plane rotation. To address this limitation, various authors have noted that it is frequently easier to model the differences between target and background than to model the target itself. This has led to the idea of *discriminant tracking*, where the tracking problem is framed as one of continuous object detection, through incremental *target vs. background* classification [7, 2, 13]. Discriminant tracking has two main steps. Given an initial target bounding box, say at time t, the first step consists of *classifier design*: a classifier is trained by selecting visual features that discriminate between target and background, and a decision rule is learned based on these features. In the second step, denoted *target detection*, the classifier is applied to every location of the visual field, so as to determine the most likely location of the target at time $t + 1$. The target bounding box is moved to this location, and the process iterated. This generic formulation has been used to design various trackers [7, 2, 13, 14, 3]. Although these efforts have demonstrated that discriminant tracking can achieve state-of-the-art performance in computer vision [13], this performance is still far from that of the tracking mechanisms implemented by biological vision. In the biological world, object tracking is closely related to the task of fixating objects of interest. The goal is to keep an object on the fovea of the observer, even when either or both are moving. This is achieved with a combination of overt and covert eye movements, and underlies the mechanisms for identification of moving objects [25]. Due to the evolutionary advantage of solving these tasks accurately, it is not surprising that biological vision has evolved extremely efficient tracking mechanisms, in terms of accuracy, robustness, and speed. It has been hypothesized that the effectiveness of these mechanisms, even under the most adverse conditions, involving clutter, low-light etc., is a consequence of the availability of robust saliency mechanisms, that cause pre-attentive pop-out of certain locations of the visual field [25]. These salient locations become the *focus of attention* (FoA) for the post-attentive stages of visual processing, where top-down feedback from higher level cortical layers is used to solve problems such as object recognition or visual search [35].

In this work, we expand on a recently proposed discriminant tracking algorithm based on saliency [22] by postulating that *tracking is simply a manifestation of the continuous computation of saliency over time*. More precisely, we frame discriminant tracking of [22] as a byproduct of the center-surround saliency mechanisms that are prevalent in biological vision [12, 6]. This is done with recourse to a recent computational formulation of visual saliency, denoted *discriminant saliency* [12], which has enabled a number of contributions to both biological and computer vision. We start by showing that discriminant tracking can be implemented with a combination of operations that are well documented in the biological attention literature: *center-surround saliency* [18], a spatial *spotlight of attention* [26], and *feature-based attention* [32]. It is then shown that, under the discriminant saliency formulation, these operations are mapped into statistical operations such as *feature selection* or *target detection*. This enables the derivation of *optimal trackers* that can be implemented with *simple* and *highly efficient* computations, two important requirements for the practical feasibility of any tracker. The saliency formulation is next shown to also establish a *unified framework for automatic tracker initialization, classifier design and target detection*. While the steps of classifier design and target detection are addressed by all discriminant trackers in the literature, previous solutions cannot cope with the initialization. Finally, it is shown that the proposed discriminant tracker outperforms a number of state-of-the-art tracking approaches in the literature.

We start by reviewing the main concepts of *discriminant saliency.* A more extensive discussion can be found in [12, 10, 9, 23].

2. Visual Saliency

The perception of complex scenes by biological vision systems is heavily dependent on attentional mechanisms. These mechanisms allocate the limited perceptual resources available to the scene regions that matter the most, increasing efficiency and robustness to clutter. Attention is itself driven by saliency mechanisms, which assign to each region of the visual field a degree of saliency, or importance. The different regions of the scene are then explored sequentially, according to their saliency. There are two types of saliency mechanisms, commonly denoted *bottom-up* and *top-down*. Bottom-up saliency is completely stimulus driven, i.e. independent of the higher level goals of the perceptual system. It is, for example, responsible for the high saliency of a “danger” sign posted on a wall, which *pops-out* [24] even when we are not looking for danger signs. Top-down saliency mechanisms can be tuned by feedback from high-level cortical areas, according to the tasks to be performed. For example, the eye fixations of a subject trying to identify a person in a photograph will be overwhelmingly located on the faces present in that picture [36]. Two main types of tuning are possible: a *spatial focus of attention* mechanism, also known as the spotlight of attention [26], and *feature-based attention* [32] which manipulates attention by inhibiting or enhancing groups of features. In the following sections, we show that both spatial and feature-based attention play a prominent role in saliency-based tracking. We first review the discriminant formulation of both bottom-up and top-down saliency in greater detail.

2.1. Mathematical formulation of bottom up saliency

Let $\mathcal{V}$ be the visual stimulus and l a location of interest. Two windows are defined around this location: a *center window* $\mathcal{W}_l^1$ containing l, and a surrounding annular window $\mathcal{W}_l^0$ containing *background*. The union of the two windows is denoted the *total* window, $\mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$. Stimuli in the center window are drawn from a *center class*, of label $C(l) = 1$. Stimuli in the surround window are drawn from a *background class*, of label $C(l) = 0$. A set of features $\boldsymbol{Y}$, from a predefined feature space $\mathcal{Y}$, are computed for each of the windows $\mathcal{W}_l^i, i \in \{0, 1\}$. Features $\boldsymbol{Y}$ extracted from the center have probability $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|1)$ and those from the background have probability $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|0)$. The *saliency* of location l, $S(l)$, is quantified by the mutual information between feature responses, $\boldsymbol{Y}$, and class label, C,

$$
\begin{aligned}
S(l) &= I_l(\boldsymbol{Y}; C) && (1)\\
&= \sum_{i=0}^{1} \int p_{\boldsymbol{Y},C(l)}(\mathbf{y}, i) \log \frac{p_{\boldsymbol{Y},C(l)}(\mathbf{y}, i)}{p_{\boldsymbol{Y}}(\mathbf{y}) p_{C(l)}(i)} d\mathbf{y} && (2)\\
&= \sum_{c=0}^{1} p_{C(l)}(i) KL[p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|i) || p_{\boldsymbol{Y}}(\mathbf{y})]. && (3)
\end{aligned}
$$

where $\mathrm{KL}\,(p\,\|q\,) = \int_{\mathcal{X}} p_{\boldsymbol{X}}(x) \log \frac{p_{\boldsymbol{X}}(x)}{q_{\boldsymbol{X}}(x)} dx$ is the Kullback-Leibler (KL) divergence between the probability distributions $p_{\boldsymbol{X}}(x)$ and $q_{\boldsymbol{X}}(x)$.

2.2. Mathematical formulation of top down saliency

For top-down saliency problems, such as object recognition [11, 9], the target class, of label $C = 1$, is the object class to recognize, and the background class, with label $C = 0$, the class of natural images. Features $\boldsymbol{Y}$ have probability $p_{\boldsymbol{Y}|C}(\mathbf{y}|1)$ under the target hypothesis and probability $p_{\boldsymbol{Y}|C}(\mathbf{y}|0)$ under the background hypothesis. Unlike bottom-up saliency, where the absence of any objects can be salient (e.g. a void region is salient within a textured background), recognition requires the detection of the object of interest. This implies that top-down saliency measures must have a bias towards target presence.

This bias is accomplished with a two-step saliency measure. A likelihood ratio test is first used to identify the set of likely target locations $\mathbf{S} = \left\{ l \,\middle|\, \frac{P_{C,\mathbf{Y}}(1,\mathbf{y}(l))}{P_C(1)P_{\mathbf{Y}}(\mathbf{y}(l))} > \frac{P_{C,\mathbf{Y}}(0,\mathbf{y}(l))}{P_C(0)P_{\mathbf{Y}}(\mathbf{y}(l))} \right\}$. These are the locations where the likelihood of the feature responses is larger under the hypothesis of target presence than target absence. As before, the saliency of location l is defined by the amount of information in the visual stimulus for optimal classification into one of the two classes, using the information measure

$$I(C;\mathbf{Y}=\mathbf{y}(l)) = \sum_{i=0}^{1} p_{C|\mathbf{Y}}(\mathbf{y}(l)|i) \log \frac{p_{\mathbf{Y},C}(\mathbf{y}(l),i)}{p_{\mathbf{Y}}(\mathbf{y}(l))p_C(i)}. \quad (4)$$

However, to guarantee that only locations likely to contain the target are declared salient, the saliency computation is restricted to $\mathbf{S}$. This leads to the saliency measure [9, 15]

$$S(l) = \begin{cases} I(C;\mathbf{Y}=\mathbf{y}(l)) & \text{if } l \in \mathbf{S} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Locations where this measure is large have both 1) larger likelihood under the target than background hypothesis, and 2) feature responses that are highly informative for classification.

2.3. Efficient computation of saliency measures

When the features $\mathbf{Y}$ are bandpass in nature (e.g. DCT, Gabor, wavelet), the computation of the saliency measures of (3) and (5) can be simplified by using the statistics of bandpass responses to natural images. This follows from the well known observation that the probability distribution of feature responses of a bandpass feature, to natural images, follows a generalized Gaussian distribution (GGD) [16]

$$p_Y(y;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{ -\left(\frac{|y|}{\alpha}\right)^{\beta} \right\}, \quad (6)$$

where $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\mathrm{d}t$, $t > 0$, is the Gamma function, α a *scale* parameter, and β a *shape* parameter. The β parameter controls the rate of decay of the GGD, from the peak value (e.g. Laplacian when $\beta = 1$ or Gaussian when $\beta = 2$). It has been shown that $\beta \in (0.5, 0.8)$ provides a good fit to large corpora of natural images [29]. We found $\beta = 0.7$ to work best and we adopt this parameter value throughout this work. Given β, the only parameter that remains to be learned is the scale α. This can be done by the method of moments [28].

Further, by exploiting a well known property of bandpass features extracted from natural images: that such features exhibit *consistent* patterns of dependence across an extremely wide range of natural image classes [5, 16] we can approximate (1) by:

$$I(\mathbf{Y};C) \approx \sum_{k=1}^{N} I(Y_k;C). \quad (7)$$

where, the term $I(Y_k;C)$ is the marginal mutual information (MMI) between the k^{th} feature and the class label [34]. It measures how discriminant the k^{th} feature is individually.

It can then be shown that the top-down saliency measure of (5) can be written as [22]:

$$S_k(l) = \begin{cases} \sum_{i=0}^{1} h_i[\xi_k|y_k(l)|^{\beta} - T_k] & \text{if } |y_k(l)| > t_k \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where, y_k is the feature response of the k^{th} feature, $h_i(x) = s\{(-1)^{1-i}x\} \log\left\{\frac{1}{\pi_i}s\{(-1)^{1-i}x\}\right\}$, and

$$\xi_k = \frac{1}{\alpha_{k,0}^{\beta}} - \frac{1}{\alpha_{k,1}^{\beta}}, \quad T_k = \log\frac{\alpha_{k,1}\pi_0}{\alpha_{k,0}\pi_1}, \quad (9)$$

where $\pi_i = P_C(i)$ is the prior for class i, and $\alpha_k, \alpha_{k,i}$ the scale parameters of $p_{Y_k}(y_k), p_{Y_k|C(l)}(y_k|i)$.

3. Discriminant tracking using saliency

The central hypothesis of this work is that discriminant tracking can be implemented with a combination of bottom-up and top-down saliency detection. In this section, we build on this hypothesis to propose a saliency-based discriminant tracker. We first discuss how saliency can be used to perform automatic initialization of targets, and then show how the same framework can be used for classifier design and target detection.

3.1. Automatic tracker initialization

Most tracking algorithms assume a known initial target location l^* and bounding box $\mathcal{W}_{l^*}^1$ [7, 2]. However, these are not available in most tracking applications. While many initialization strategies, such as background subtraction and blob or motion detection, have been proposed [7], they are mostly heuristic. A more principled approach, based on bootstrapping a weak and generic target model for automatic initialization, was proposed in [31]. However, it requires a pre-specified target model, and some degree of supervision to adapt it to different scenes. Saliency-based tracking provides a more natural solution to the initialization problem: to declare as targets the locations of largest bottom-up saliency [22]. This is implemented by evaluating (3) at all locations of the visual field, and finding the most salient (or the set of most salient locations if multiple objects are to be tracked).

$$(l^*) = \operatorname*{argmax}_l \sum_k I_l(Y_k, C) \quad (10)$$

where,

$$I_l(Y_k, C) = \sum_{i=0}^{1} \pi_i \left(\log \left(\frac{\alpha_k^l}{\alpha_{k,i}^l} \right) + \frac{1}{\beta} \left[\left(\frac{\alpha_{k,i}^l}{\alpha_k^l} \right)^\beta - 1 \right] \right). \tag{11}$$

where the parameters $\alpha_{k,i}^l, \alpha_k^l$ are learned from the windows associated with a center-surround operator centered at location l. Overall, the initialization procedure finds the regions whose motion and appearance is most distinct from those of the surrounding background.

The use of (10) has a number of appealing properties. First, it can be seen as an optimal (in the discriminant sense) form of background subtraction. In fact, it is a simplification of a state-of-the-art formulation of background subtraction that performs well even on highly dynamic backgrounds [21]. In this work, we use simple biologically plausible motion energy features in place of the cumbersome dynamic textures of [21]. The proposed simplification sacrifices some ability to model complex dynamics for the sake of biological plausibility and computational tractability. However the use of spatiotemporal features still enables it to account for both target appearance and motion, and is robust to camera motion. This follows from the fact that only motion different from that of the background can be declared salient. For example, an object followed by a panning camera is considered salient.

3.2. Tracking using saliency

Given an initial target location, l^*, obtained using the procedure outlined in the previous section at time t, the first step of discriminant tracking is to design a target/background classifier. The target and background hypotheses are defined by the feature responses in a window centered at l^*, the *target window,* and a surrounding annular *background window*. Hence, like bottom-up saliency, discriminant tracking requires the computation of the discriminant power of each feature in $\boldsymbol{Y}$ with respect to a *center-surround discrimination* problem. The main difference is that, while bottom-up saliency performs the computation at *each* location of the visual field, discriminant tracking only requires it at location l^*. This is equivalent to computing bottom-up saliency after application of a *spatial focus of attention* mechanism tuned to the target location. Given a measure of how discriminant each feature is for target/background discrimination at time t, the next step is to find the target in the next frame, i.e. at time $t+1$. This is formulated as a target detection problem. It requires the selection of the most discriminant features in $\boldsymbol{Y}$, and a decision rule for target detection. Since the discriminant power of each feature is already known, feature selection reduces to suppression of non-discriminant features and enhancement of discriminant ones. This type of manipulation is exactly the function of a *feature-based attention* mechanism. Finally, target detection can be implemented with a top-down saliency measure trained from the feature responses in the target and background windows at time t. The position of the target at time $t+1$ is determined by a search for the location of largest saliency within a neighborhood of the target position at time t. This restriction of the search space reduces the computation needed to identify the new target location, by ignoring regions peripheral to the current focus of attention. It is consistent with the "center bias" observed in the human visual system, where a saccade to a new fixation location is biased to be close to the current center of view [30, 33].

3.3. The core tracking procedure

The discussion of the previous section suggests that discriminant tracking can be implemented with discriminant saliency measures. Starting with the target location l^* at time t, and the associated target ($\mathcal{W}_{l^*}^1$) and background ($\mathcal{W}_{l^*}^0$) windows, the tracker is implemented as follows.

- **Learning:** at time t, estimate the probability distributions $p_{\mathbf{Y}|C(l)}(\mathbf{y}|i), i \in \{0, 1\}$ using the feature responses in $\mathcal{W}_{l^*}^i$, as training sample, and the distribution $p_{\mathbf{Y}}(\mathbf{y})$ from the responses in $\mathcal{W}_{l^*} = \mathcal{W}_{l^*}^0 \cup \mathcal{W}_{l^*}^1$.
- **Feature selection:** Among the N available features, select the subset of $K < N$ that maximizes the saliency measure of (3).
- **Classification:** using these K features compute, at time $t+1$, the top-down saliency of each location l of the visual field, using the saliency measure of (5). Move the target/background windows to the location of largest saliency within a neighborhood of l^*, and iterate the process.

The automatic initialization discussed in Section 3.1 is a special case of discriminant tracking. In the absence of prior information about which features are discriminant for target detection, the tracker simply uses all of them.

4. Experiments and Results

4.1. Automatic Initialization

We performed a set of experiments designed to evaluate automatic tracker initialization using the proposed discriminant saliency tracker (DST). Since none of the other methods in the literature have this capability, no comparison was performed for these sequences. Examples of DST results are shown in Figure 1. The tracker uses the bottom-up discriminant saliency procedure of Section 3.1 to identify the object to track. The background bounding box was assumed to have an edge 4 times larger than the corresponding edge of the target box. The region of maximal saliency and its background were then input to the DST algorithm,

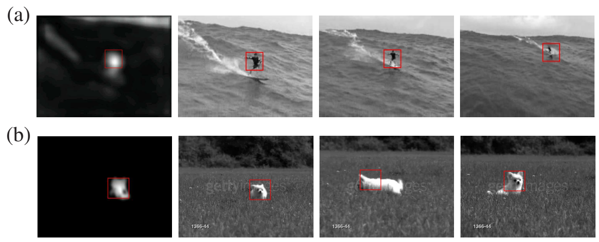

Figure 1. Automatic initialization and tracking. The bottom-up saliency map obtained using biologically plausible motion features is shown on the left column. Target bounding boxes are shown in red. a) "surfer" b) "dog". Target locations in subsequent frames are shown in red.

which tracks the target through the remaining frames. The DST used a two-level Gaussian pyramid, leading to a total of $N = 3 + 64 \times 2 = 131$ features (8×8 DCT features per level plus three spatiotemporal Gabor features). The number of selected salient features, K, is a tunable parameter. Good performance was obtained for any $K \geq 3$, albeit tracking accuracy improved with the number of features, at the expense of increased computation. To guarantee a realistic balance between tracking performance and computation, K was set to 5 in all subsequent experiments. The search neighborhood, $\mathcal{W}^s_{l^*}$, was set to a rectangular region centered at the current target position l^* with size twice that of the object bounding box.

The leftmost column of Figure 1 shows the bottom-up saliency map, and the columns on the right show a few of the subsequent frames (target bounding box shown in red). The tracker initializes the target correctly, and tracks it through substantial variations of scale and pose (note the 3D rotation in "dog").

Table 1 presents the error measures obtained for the sequences evaluated. The error of DST with automatic initialization is compared to that obtained when the tracker is manually initialized with the groundtruth target bounding box. There is no substantive difference. Overall, these results demonstrate the ability of the DST to perform robust target initialization and tracking, in scenes with complex motion. Videos of all sequences are available in [1]

Table 1. Comparison of automatic and manual tracker initialization.

Name	Auto Init	Manual Init
dirtbike	**0.037**	0.038
surfer	0.087	**0.086**
dog	0.115	**0.093**
skiing	**0.079**	0.083

4.2. Comparison to previous trackers

The DST was compared to four trackers in the literature: three discriminant trackers, the MILTracker of [3], the method of Collins et al. [7], and the ensemble tracker of [2], and the incremental visual tracker (IVT) of [27]. The latter represents the state of the art in appearance-based tracking. Software for the MILTracker and IVT was obtained from the authors' webpages. Since no implementations are publicly available for the Collins and ensemble trackers, these algorithms were implemented according to the descriptions in [7, 2].

Table 2. Average tracking error of the five trackers compared. 0 indicates perfect tracking, 1 complete lack of overlap between groundtruth and target bounding box produced by the tracker.

Sequence	IVT	Collins	Ensemble	MIL	DST
motinas	0.55	0.39	0.67	0.52	**0.12**
athlete	0.98	0.70	0.93	0.89	**0.19**
ram	0.68	0.80	0.80	0.51	**0.15**

The performance of all five methods was evaluated against manual groundtruth. The tracking error for a frame at time t was defined as the normalized pixel difference between the groundtruth target bounding box, G^t, and that produced by the tracker, B^t. Performance is evaluated by the average tracking error over a sequence of T frames,

$$\epsilon = \frac{1}{T} \sum_t \frac{\sum_{ij} G^t_{i,j}(1 - B^t_{i,j})}{\sum_{ij} G^t_{i,j}}. \tag{12}$$

where the error $\epsilon = 0$ for perfectly correct tracking, while for complete loss of tracking, $\epsilon = 1$.

The test video sequences were selected from diverse sources (e.g previous works, standard databases, and the web). All sequences include challenging tracking scenarios, such as varying illumination, complete object rotation, or change in perspective. For instance, the "motinas_toni_change_ill" sequence of [20] shows a person turning by 360°, in extremely low light (Figure 2(a)), while the "athlete" sequence includes extreme variations of appearance due to occlusion and strong video compression artifacts (Figure 2 (b)). To increase the difficulty of the tracking task, all sequences were converted to grayscale. To account for this, the Collins tracker was implemented with DCT features, instead of the R,G,B color features proposed in [7]. To compare all five algorithms, they were manually initialized with target bounding box in the first frame. Figure 2 illustrates the tracking results on three of the sequences considered. The qualitative performance of IVT and the ensemble tracker is quite poor, as these methods lose the target in all scenes. Somewhat better performance is achieved by the Collins and MIL trackers. However, these methods lose the target when it undergoes extreme appearance variations, due to illumination changes or rotation. On the other hand, DST tracks the targets successfully in all sequences.

Table 2 presents the errors measured on the sequences of Figure 2. The table shows that DST has the best performance. Videos of all tracking results are available from [1].

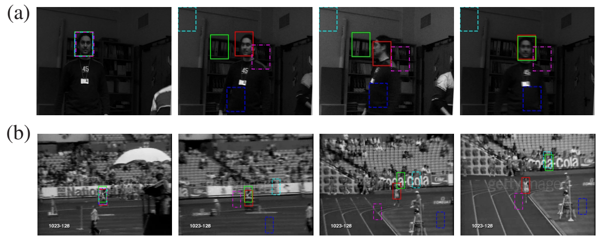

Figure 2. Tracking results on a) "motinas_toni_change_ill" [20] - the person turns around and the illumination changes drastically, b) 'athlete"- a person running inside a stadium. The video is very noisy and the target appearance changes widely, Target locations: DST - thick red box, Collins - thick green box, ensemble - cyan dashed box, IVT - blue dashed box, MIL - magenta dashed box.

5. Conclusion

In this work, we have shown that discriminant tracking follows naturally from the discriminant formulation of visual saliency. This was exploited to construct a simple and computationally efficient framework for tracking, which is consistent with what is known about the attentional mechanisms of biological vision. Experimental comparison with previous trackers shows that the proposed biologically plausible discriminant saliency tracker is significantly more robust. An implementation of this tracker in C, without any optimization, currently runs at ~ 1.5 frames per second (fps), on a standard PC without special hardware. On the same machine, the running times of other discriminant trackers are comparable ($\sim$ 4fps for MIL and $\sim$ 3fps for the Collins tracker).

References

[1] http://www.svcl.ucsd.edu/projects/tracking/results.html.

[2] S. Avidan. Ensemble tracking. *IEEE PAMI*, 29(2):261–271, 2007.

[3] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. *CVPR*, 0:983–990, 2009.

[4] M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *IJCV*, 26(1):63–84, 1998.

[5] R. Buccigrossi and E. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE TIP*, 8:1688–1701, 1999.

[6] J. Cavanaugh, W. Bair, and J. Movshon. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, 88:2530–2546, 2002.

[7] R. Collins, Y. Liu, and M. Leordeanu. On-line selection of discriminative tracking features. *IEEE PAMI*, 27(10):1631 – 1643, October 2005.

[8] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE PAMI*, 25(5):564–577, 2003.

[9] D. Gao, S. Han, and N. Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE PAMI*, 31(6):989, 2009.

[10] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7):1–18, 6 2008.

[11] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *NIPS*, 2005.

[12] D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 21:239–271, Jan 2009.

[13] H. Grabner and H. Bischof. On-line boosting and vision. *CVPR*, 1:260–267, 2006.

[14] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *ECCV*, pages 234–247, 2008.

[15] S. Han and N. Vasconcelos. Biologically Plausible Saliency Mechanisms Improve Feedforward Object Recognition. *Vision Research*, 2010.

[16] J. Huang and D. Mumford. Statistics of Natural Images and Models. In *CVPR*, pages 541–547, 1999.

[17] M. Isard and A. Blake. Condensation conditional density propagation for visual tracking. *IJCV*, 29:5–28, 1998.

[18] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI*, 20(11):1254–1259, 1998.

[19] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE PAMI*, 25(10):1296–1311, 2003.

[20] E. Maggio and A. Cavallaro. Hybrid particle filter and mean shift tracker with adaptive transition model. In *ICASSP*, 2005.

[21] V. Mahadevan and N. Vasconcelos. Background subtraction in highly dynamic scenes. *IEEE CVPR*, 1, 2008.

[22] V. Mahadevan and N. Vasconcelos. Saliency-based discriminant tracking. *CVPR*, 0:1007–1013, 2009.

[23] V. Mahadevan and N. Vasconcelos. Spatiotemporal saliency in dynamic scenes. *IEEE PAMI*, 32:171–177, 2010.

[24] H. C. Nothdurft. Texture segmentation and pop-out from orientation contrast. *Vision Research*, 31(6):1073–1078, 1991.

[25] S. E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, 1999.

[26] M. Posner, C. Snyder, and B. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109(2):160–174, 1980.

[27] D. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3):125–141, May 2008.

[28] K. Sharifi and A. Leon-Garcia. Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video. *IEEE CSVT*, 5(1):52–56, 1995.

[29] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu. On advances in statistical modeling of natural images. *Journal of mathematical imaging and vision*, 18(1):17–33, 2003.

[30] B. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[31] K. Toyama and Y. Wu. Bootstrap initialization of nonparametric texture models for tracking. In *ECCV*, 2000.

[32] S. Treue and J. Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575–579, 1999.

[33] P. Tseng, R. Carmi, I. Cameron, D. Munoz, and L. Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7), 2009.

[34] M. Vasconcelos and N. Vasconcelos. Natural image statistics and low-complexity feature selection. *IEEE PAMI*, 31(2):228–244, 2008.

[35] J. M. Wolfe. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.

[36] A. Yarbus. *Eye movements and vision*. Plenum, New York, 1967.

[37] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):13, 2006.