# Biologically Inspired Object Tracking Using Center-Surround Saliency Mechanisms

Vijay Mahadevan, *Member*, *IEEE*, and Nuno Vasconcelos, *Senior Member*, *IEEE*

**Abstract**—A biologically inspired discriminant object tracker is proposed. It is argued that discriminant tracking is a consequence of top-down tuning of the saliency mechanisms that guide the deployment of visual attention. The principle of discriminant saliency is then used to derive a tracker that implements a combination of center-surround saliency, a spatial spotlight of attention, and feature-based attention. In this framework, the tracking problem is formulated as one of continuous target-background classification, implemented in two stages. The first, or learning stage, combines a focus of attention (FoA) 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. The statistics of natural images are exploited to derive an implementation which is conceptually simple and computationally efficient. The saliency formulation is also shown to establish a unified framework for classifier design, target detection, automatic tracker initialization, and scale adaptation. Experimental results show that the proposed discriminant saliency tracker outperforms a number of state-of-the-art trackers in the literature.

**Index Terms**—Object tracking, discriminant tracking, saliency, attention, motion saliency, automatic target initialization, scale adaptive tracking, discriminant center-surround architecture, video modeling

## 1 Introduction

Object tracking is a classical problem in computer vision and a prerequisite 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 [74]. Many of 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 [9], [19], [39], [42]. The main limitation of these methods is that they uniquely rely 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 versus background* classification [6], [18], [31]. 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 [6], [7], [18], [31], [32].

In the biological world, object tracking is a requirement for fixating objects of interest. The goal is to keep an object on the fovea of the observer, even when either or both are moving [53]. Given the evolutionary advantage of solving this problem, it is not surprising that biological vision has evolved extremely efficient tracking mechanisms in terms of accuracy, robustness, and speed. In the biological vision literature, it has been suggested that tracking is 1) implemented by *attentional mechanisms* [5], [14] and 2) dependent on the distinctiveness of target appearance features [49]. It is also known that a distinct target can be tracked as it changes appearance, even when spatially superimposed on a distractor [10]. Conversely, it has been shown that attentional tracking fails when target features cannot be individuated [13], [70]. With regard to motion, targets can be easily tracked among distractors of identical appearance as long as they are spatiotemporally distinguishable from the latter [38], [55]. While this suggests that both spatial and spatiotemporal target features are used in object tracking [10], it is believed that biological tracking mechanisms do not rely on *motion extrapolation* [44]. In fact, experiments based on the "bouncing-streaming" [58] paradigm have shown that the perceived correspondence of an object in successive time slices depends much more on the similarity of its features (shape, orientation, color, texture, etc.) than on the

- *V. Mahadevan is with Yahoo! Labs, Embassy Golf Links Business Park, Bangalore 560071, India. E-mail: vijay.mahadevan@gmail.com.*
- *N. Vasconcelos is with the Electrical and Computer Eningeering Department, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093. E-mail: nvasconcelos@ucsd.edu.*

*Manuscript received 2 Feb. 2011; revised 23 Dec. 2011; accepted 28 Mar. 2012; published online 19 Apr. 2012.*
*Recommended for acceptance by S. Avidan.*
*For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-02-0076.*
*Digital Object Identifier no. 10.1109/TPAMI.2012.98.*

predictability of the resulting trajectory [23]. In summary, biological tracking requires target-distractor discrimination, in terms of appearance or motion.

Noting that these are also the distinguishing properties of the center-surround saliency mechanisms that guide the deployment of attention [40], and which are prevalent in biological vision [15], [30], we frame discriminant tracking as a *by-product* of saliency detection. This is done with recourse to a recent computational formulation of visual saliency, denoted *discriminant saliency* [30], 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* [41], a spatial *spotlight of attention* [54], and *feature-based attention* [66]. 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 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 classifier design, target detection, automatic tracker initialization, and scale adaptation*. 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 and scale adaptation problems. Finally, it is shown that the proposed discriminant tracker outperforms a number of state-of-the-art tracking approaches in the literature.

The paper is organized as follows: Section 2 reviews the tracking literature. Visual saliency and the discriminant saliency principle are then briefly reviewed in Section 3. Section 4 introduces the saliency-based discriminant tracker, and derives efficient implementations for feature selection, classifier design, target detection, and parameter learning. A number of extensions that improve tracking robustness, enable scale adaptation, and automate tracker initialization are also presented. An extensive experimental evaluation is then presented in Section 5, and some conclusions are finally drawn in Sections 6 and 7.

## 2 Related Work on Object Tracking

Many popular approaches to object tracking are based on *appearance modeling*. They learn and maintain a model of target appearance, which is used to locate the target as time evolves. Conditional density propagation [39] is one of the most popular methods in this class. Targets are represented by some type of visual features, e.g., their contours or deformable templates [75], and the temporal evolution of these features is modeled with a particle filter. Alternatively, target appearance is frequently represented by kernel weighted color histograms, which are combined with the mean shift procedure to identify the most likely position of the target in the next frame [19]. Representations of the target and/or background with probabilistic models, e.g., a mixture of Gaussian (MoG) models, have also been proposed [33], [62]. Equally popular are subspace methods, which maintain a low-dimensional representation of target appearance [9], [36]. Recently, there has been an interest in making these representations adaptive by updating subspaces incrementally, using online principal component analysis [57]. More sophisticated appearance models include a combination of short-term descriptors and long-term stable representations [42], specialized representations tailored to specific entities such as people [56], or multiple image patch representations such as "FragTrack" [3].

Appearance-based trackers have limited accuracy when backgrounds are cluttered or targets have substantial amounts of geometric deformation, such as out-of-plane rotation. *Discriminant trackers* frequently achieve better performance in these scenarios [31] by framing tracking as incremental *target versus background* classification [6], [18]. The superior performance of discriminant trackers over models that rely on motion prediction is consistent with what is known about biological tracking. One of the earliest discriminant trackers, proposed by Collins et al. [18], relies on a feature set composed of histograms of filter responses to the R, G, B channels of the visual stimulus. Discriminant features are selected with a variant of the Fisher discriminant, and the classifier is implemented with a likelihood-based decision rule. Fisher discriminants are also used to classify foreground from background in [45] and [51]. The "ensemble tracking" method of Avidan [6] uses a combination of histograms of oriented gradients [20] and R, G, B pixel values as features. A set ("ensemble") of weak hyperplane classifiers is trained to separate target from background and combined into a decision rule using AdaBoost [24]. Grabner and Bischof [31] have proposed an alternative ensemble tracker, based on online boosting. This maintains a set of weak learners that are updated at every time step. More recently, online boosting has been combined with a semi-supervised update of the weak learners to increase tracker robustness [32]. A multiple instance learning (MIL)-based approach has also been proposed in [7] to minimize the ensemble tracker sensitivity to outliers due to misalignment of the target bounding box.

The robustness of biological tracking mechanisms has inspired computer vision researchers to augment conventional trackers with *focus of attention* (FoA) mechanisms. For instance, Toyama and Hager [64] proposed an incremental FoA procedure to combine multiple trackers, leading to increased robustness.

## 3 Discriminant Saliency

We start by reviewing the main concepts of *discriminant saliency*. A more extensive discussion can be found in [25], [26], [30], and [48].

### 3.1 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* [52] 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 [72]. Two main types of tuning are possible: a *spatial focus of attention* mechanism, also known as the spotlight of attention [54], and *feature-based attention* [66], 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.

### 3.2 Discriminant Saliency

Discriminant saliency is a mathematical formulation for visual saliency. Two classes of visual stimuli are first defined: a *target* class of stimuli of interest and a *background* or null hypothesis of nonsalient stimuli. The visual stimulus is not observed directly, but through projection into a number of features. Saliency is the result of optimal classification (in a decision-theoretic sense) of feature responses into the *target* and *background* hypotheses [27]. More precisely, the saliency of each location of the visual field is equated to the *expected classification accuracy* for the features extracted from it. Locations of smallest probability of error are most salient.

This formulation can be applied to various vision problems by suitable definition of target and null hypotheses. For example, it can be used to implement one-versus-all object detection by defining the target as an object class and the null hypothesis as the set of natural images [27]. This is an instance of top-down saliency, due to the necessity of specifying task-related object classes. Alternatively, target and null hypotheses can be defined as the visual stimuli contained in a pair of *center* and *surround* windows at every location of the visual field [30]. This is a purely stimulus driven definition, which implements bottom-up saliency. Implementations of the discriminant saliency principle have various properties of interest for both biological and computer vision. In the area of biological modeling, they can be mapped into a biologically plausible neural architecture, which has been shown to 1) replicate the computations of the standard neurophysiological model of the visual cortex [30], 2) predict a large body of psychophysics of human saliency [26], and 3) accurately predict human fixations in natural scenes [28]. In computer vision, they have been shown successful for interest point detection [29], background subtraction in highly dynamic scenes [48], and object recognition [25], [34]. We next review the discriminant formulation of both bottom-up and top-down saliency in greater detail.

### 3.3 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, $C$:

$$S(l) = I_l(\boldsymbol{Y}; C) = \sum_{i=0}^{1} \int p_{\boldsymbol{Y},C(l)}(\mathbf{y}, i) \log \frac{p_{\boldsymbol{Y},C(l)}(\mathbf{y}, i)}{p_{\boldsymbol{Y}}(\mathbf{y}) p_{C(l)}(i)} d\mathbf{y} \quad (1)$$

$$= \sum_{c=0}^{1} p_{C(l)}(i) KL[p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|i) \| p_{\boldsymbol{Y}}(\mathbf{y})], \quad (2)$$

where $\mathrm{KL}(p_{\boldsymbol{X}} \| q_{\boldsymbol{X}}) = \int_{\mathcal{X}} p_{\boldsymbol{X}}(x) \log \frac{p_{\boldsymbol{X}}(x)}{q_{\boldsymbol{X}}(x)} dx$ is the Kullback-Leibler (KL) divergence between the probability distributions $p_{\boldsymbol{X}}(x)$ and $q_{\boldsymbol{X}}(x)$. Since the KL divergence measures the disparity between its two arguments and $P_{\boldsymbol{Y}}(\mathbf{y})$ is the average of $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|1)$ and $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|0)$, the mutual information can also be interpreted as a measure of distance between the distribution of feature responses under the two classes. Hence, $S(l)$ measures the extent to which the features $\boldsymbol{Y}$ discriminate between the two classes.

### 3.4 Mathematical Formulation of Top-Down Saliency

For top-down saliency problems such as object recognition [25], [27], 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. Feature $\boldsymbol{Y}$ has 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 toward 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 $\mathbf{S} = \{l | P_{\boldsymbol{Y}(l)|C}(\mathbf{y}|1) > P_{\boldsymbol{Y}(l)|C}(\mathbf{y}|0)\}$. 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:

$$I(C; \boldsymbol{Y}(l) = \mathbf{y}) = \sum_{i=0}^{1} p_{\boldsymbol{Y}(l)|C}(\mathbf{y}|i) \log \frac{p_{\boldsymbol{Y}(l),C}(\mathbf{y}, i)}{p_{\boldsymbol{Y}(l)}(\mathbf{y}) p_C(i)}. \quad (3)$$

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 [25], [34]:

$$S(l) = \begin{cases} I(C; \boldsymbol{Y}(l) = \mathbf{y}), & \text{if } l \in \mathbf{S}, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

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.

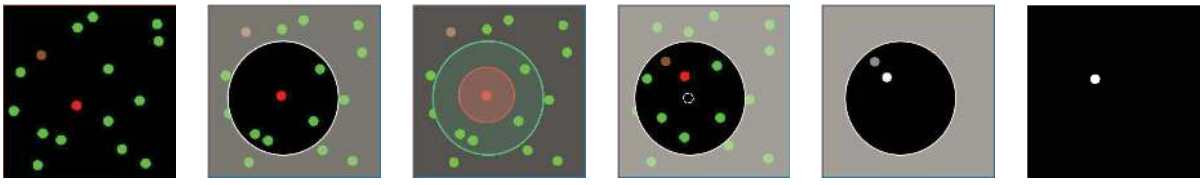


Fig. 1. Illustration of saliency-based tracking. (a) Two disks, one red and one brown are salient among green distractors; (b) defining the red disk as the target, at time $t$, focuses spatial attention on it; (c) computing center surround saliency at this location leads to the selection of the feature "red" as the most salient; (d) the position of the disks at time $t+1$, shown with the focus of attention from time $t$; (e) feature-based attention suppresses all but the red feature channel, which has nonzero response only at the locations of the red and brown disks; (f) the location of the target has the largest saliency inside the focus of attention.

## 4 Discriminant Tracking

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.

### 4.1 The Connection to Saliency

We start by relating discriminant tracking to saliency. Given an initial target location $l^*$ 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 nondiscriminant 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 [63], [67].

The overall process is illustrated in Fig. 1. The display in (a) shows two disks (one red, one brown) moving against a background of green distractors. Assume that the red disk is the target and that the feature set $\mathcal{Y}$ consists of a number of color detectors. At time $t$, the spatial focus of attention mechanism narrows the field of view to the neighborhood of the target, as shown in (b). This makes the target salient. Computation of center-surround saliency as in (c) finds the red color to be the most discriminant feature. Training a top-down saliency measure for target/background classification in this area produces a detector of red disks. For simplicity, we assume this to be a threshold on the red channel of the visual stimulus. Target detection at time $t+1$ starts with the application of feature-based attention, which strengthens the red channel and inhibits all others. This is illustrated in (d) and (e), where we present the display at time $t+1$, and its projection on the selected feature, i.e., its red color channel. Note how the feature-based manipulation of attention eliminates much of the clutter in the scene. In fact, only the red disk elicits a strong response after feature selection. Further application of the top-down saliency detector (red threshold classifier), followed by a search for maximum saliency within a neighborhood of the previous target location, leads to the identification of the red disk, as shown in (f).

### 4.2 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 ($\mathcal{W}^1_{l^*}$) and background ($\mathcal{W}^0_{l^*}$) windows, the tracker is implemented as follows:

1. **Learning.** At time t, estimate the probability distributions $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}|i)$, $i \in \{0,1\}$, using the feature responses in $\mathcal{W}^i_{l^*}$ as training samples and the distribution $p_{\boldsymbol{Y}}(\mathbf{y})$ from the responses in $\mathcal{W}_{l^*} = \mathcal{W}^0_{l^*} \cup \mathcal{W}^1_{l^*}$.
2. **Feature selection.** Among the $N$ available features, select the subset of $K < N$ that maximizes the saliency measure of (2).
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 (4). Find the location of largest saliency within a neighborhood of $l^*$, and set this as the new $l^*$. Set $t = t+1$ and go to Step 1.

Note that the first two stages, learning and feature selection, implement a bottom-up saliency measure, while top-down saliency is used in the third, i.e., classification. The overall procedure has a number of practical limitations. First, the saliency measures of (2) and (4) require the evaluation of the joint probability distribution of the features in $\boldsymbol{Y}$. This is too

complex for most applications of saliency and infeasible for tracking, where there is a premium on computational efficiency. Various simplifications can be achieved by restricting the features to bandpass filters and exploiting the statistical regularities of the responses of such features to natural images. However, a classifier built from bandpass features may not have the robustness necessary to track complex objects subject to nonplanar motion. This type of robustness usually requires more abstract features. Finally, the classifier should operate across multiple scales so as to enable scale adaptation as the distance between objects and camera varies. These issues are addressed in the remainder of this section.

### 4.3 Salient Feature Selection

Feature selection is naturally implemented under discriminant saliency since the saliency measure is itself a measure of discrimination. In fact, extremely efficient implementations are possible when the features belong to the class of bandpass filters. Assuming this to be the case, let the feature space $\mathcal{Y}$ have dimension $N$, and denote $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$. *Salient feature selection* involves the identification of the subset of $K < N$ features that maximizes discrimination between target and background. One possibility to accomplish this is to define $\boldsymbol{Y}_{1,k} = (Y_1, \ldots, Y_k)$, and expand the mutual information of (1) into [68]

$$I(\boldsymbol{Y}; C) = \sum_k I(Y_k; C) + \sum_k (I(Y_k; C|\boldsymbol{Y}_{1,k-1}) - I(Y_k; \boldsymbol{Y}_{1,k-1})), \tag{5}$$

where

$$I(\boldsymbol{Y}; C|\mathbf{Z}) = \sum_i \int P_{\boldsymbol{Y},C,\mathbf{Z}}(\mathbf{y}, i, \mathbf{z}) \log \frac{P_{\boldsymbol{Y},C|\mathbf{Z}}(\mathbf{y}, i|\mathbf{z})}{p_{\boldsymbol{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) p_{C|\mathbf{Z}}(i|\mathbf{z})} d\mathbf{y} d\mathbf{z} \tag{6}$$

is the conditional mutual information between $\boldsymbol{Y}$ and $C$ given the observation of $\mathbf{Z}$. In (5), the term $I(Y_k; C)$ is the marginal mutual information (MMI) between the $k$th feature and the class label. It measures how discriminant the $k$th feature is individually. The terms $I(Y_k; \boldsymbol{Y}_{1,k-1}|C) - I(Y_k; \boldsymbol{Y}_{1,k-1})$ quantify the discriminant information contained in feature dependencies between the $k$th feature and the set of $k-1$ previously selected features [68]. This decomposition allows a substantial simplification of the mutual information by exploiting a well-known property of band-pass features extracted from natural images: that such features exhibit *consistent* patterns of dependence across an extremely wide range of natural image classes [12], [37]. This implies that the dependencies between features carry little information about the class from which the features are extracted, allowing the approximation of (5) by

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

As noted in Section 3.3, the mutual information $I(Y_k; C)$ measures the extent to which feature $Y_k$ discriminates between target and background classes. However, a large mutual information does not imply that the feature is characteristic of the target. In fact, a feature that is totally absent from the target region but prevalent in the background is highly discriminant for target/background classification. In the tracking context, it is usually undesirable to rely on such features since the background can vary drastically as the target, the camera, or both, move. For example, the target can move from an area of the scene where the background is highly textured (e.g., vegetation) to an area where that has virtually no texture (e.g., a white wall). A tracker that relies on features characteristic of the background texture to detect the target can lose the latter as it moves into the textureless regions of the scene. Hence, features that are discriminant but absent from the target region can lead to unstable tracking and should be discarded.

For bandpass features whose responses to natural images have zero mean and probability density functions that decay with the distance to the origin, the detection of features expressed in the target is fairly straightforward. These are the features of larger average response magnitude for target than background. Since the responses have zero mean, they can be identified as the features of larger variance under the target class than under the background class, i.e.,

$$E_{Y_k|C}[y_k^2|1] > E_{Y_k|C}[y_k^2|0]. \tag{8}$$

This condition can be combined with (7) to obtain a very efficient salient feature selection mechanism. Since the mutual information is always nonnegative, the selection of the optimal subset of $K$ ($K < N$) salient features reduces to 1) ordering the $N$ features by decreasing MMI, $I(Y_k; C)$, 2) discarding features that do not satisfy the variance condition of (8), and 3) selecting the first $K$. This is denoted feature selection by maximum marginal diversity in [69].

### 4.4 Efficient Computation of Saliency Measures

In addition to efficient feature selection, the combination of (7) and the statistics of bandpass responses to natural images also simplifies the discriminant saliency measures of (2) and (4). 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) [37]:

$$P_Y(y; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-\frac{|y|^\beta}{\alpha^\beta}\right), \tag{9}$$

where $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} \mathrm{d}t$, $t > 0$, is the Gamma function, $\alpha$ a *scale* parameter, and $\beta$ a *shape* parameter. Note that this holds for both the class-conditional densities $P_{Y|C}(y|i)$, $i \in \{0, 1\}$, and the marginal distribution $P_Y(y)$. Although the latter is a mixture $P_Y(y) = \sum_{i=0}^{1} P_{Y|C}(y|i) P_C(i)$ of GGDs, it is still the density of responses of a bandpass feature to natural images and thus well approximated by a GGD.

The $\beta$ parameter controls the rate of decay of the probability density from its 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 [61]. We found $\beta = 0.7$ to work best and adopt this value in our work. The scale parameter $\alpha$ can be estimated by the method of moments [60]. This exploits the fact that the scale $\alpha_{k,i}$ of the response of feature $Y_k$ to imagery from class $C = i$ is

$$\alpha_{k,i} = \sqrt{\frac{\sigma_{k,i}^2 \Gamma\left(\frac{1}{\beta}\right)}{\Gamma\left(\frac{3}{\beta}\right)}}, \tag{10}$$

with

$$\sigma_{k,i}^2 = E_{Y_k|C}\left[y_k^2|i\right] \approx \frac{1}{n} \sum_{j|y_k^j \in \mathcal{D}_i} \left(y_k^j\right)^2, \tag{11}$$

where $\sigma_{k,i}^2$ is the variance of the responses of $Y_k$ to class $i$, $\mathcal{D}_i = \{y_k^1, \ldots, y_k^n\}$ is a training sample from this class, and we have used the fact that the responses of bandpass filters have zero mean. In summary, given a sample of feature responses from the target and background windows, the estimation of the scale parameters is trivial.

We next note that, under the approximation of (7), the bottom-up saliency measure of (2) reduces to the sum of the marginal mutual informations between features and class label:

$$S(l) = I_l(\boldsymbol{Y}; C) \approx \sum_k I_l(Y_k; C) \tag{12}$$

$$= \sum_k \sum_{i=0}^{1} P_{C(l)}(i) KL[P_{Y_k|C(l)}(y_k|i) \| P_{Y_k}(y_k)]. \tag{13}$$

Combining this with the KL divergence between two GGDs [21],

$$\mathrm{KL}[P_Y(y;\alpha_i,\beta) \| P_Y(y;\alpha,\beta)] = \log\left(\frac{\alpha}{\alpha_i}\right) + \frac{1}{\beta}\left(\left(\frac{\alpha_i}{\alpha}\right)^{\beta} - 1\right), \tag{14}$$

leads to the *simplified bottom-up saliency measure*:

$$S(l) = \sum_k \sum_{i=0}^{1} \pi_i \left(\log\left(\frac{\alpha_k}{\alpha_{k,i}}\right) + \frac{1}{\beta}\left(\left(\frac{\alpha_{k,i}}{\alpha_k}\right)^{\beta} - 1\right)\right), \tag{15}$$

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

With respect to top-down saliency, (4) reduces to

$$S(l) = \sum_k S_k(l), \qquad S_k(l) = \begin{cases} I(C; Y_k(l) = y_k), & \text{if } l \in \mathbf{S}_k, \\ 0, & \text{otherwise,} \end{cases} \tag{16}$$

$$\mathbf{S}_k = \left\{ l \middle| P_{Y_k(l)|C}(y_k|1) > P_{Y_k(l)|C}(y_k|0) \right\}. \tag{17}$$

We next note that [30] when $P_{Y_k|C}(y_k|i)$, $i \in \{0, 1\}$, are GGDs with scale parameters $\alpha_{k,i}$,

$$I(C; Y_k = y_k) = s(g_k(y_k)) \log \frac{s(g_k(y_k))}{\pi_1} + s(-g_k(y_k)) \log \frac{s(-g_k(y_k))}{\pi_0}, \tag{18}$$

with $s(y) = (1 + e^{-y})^{-1}$ a sigmoid function, $\pi_i = P_C(i)$, and

$$g_k(y) = \xi_k |y|^{\beta} - T_k, \quad \xi_k = \frac{1}{\alpha_{k,0}^{\beta}} - \frac{1}{\alpha_{k,1}^{\beta}}, \quad T_k = \log \frac{\alpha_{k,1} \pi_0}{\alpha_{k,0} \pi_1}. \tag{19}$$

From (19) and (10), the variance condition of (8) is equivalent to $\xi_k > 0$, and the sets $\mathbf{S}_k$ can be simplified into

$$\mathbf{S}_k = \{l | |\mathrm{y_k}| > \mathrm{t_k}\} \quad \text{with} \quad t_k = \left(\frac{1}{\xi_k} \log \frac{\alpha_{k,1}}{\alpha_{k,0}}\right)^{\frac{1}{\beta}}. \tag{20}$$

Using this in (16) leads to the *simplified top-down saliency measure*:

$$S_k(l) = \begin{cases} \sum_{i=0}^{1} h_i\left(\xi_k |y_k|^{\beta} - T_k\right), & \text{if } |y_k| > t_k, \\ 0, & \text{otherwise,} \end{cases} \tag{21}$$

with $h_i(x) = s((-1)^{1-i}x) \log(\frac{1}{\pi_i} s((-1)^{1-i}x))$. The form of (21) suggests the interpretation of salient features as matched filters for the detection of visual attributes of the target class. This is due to the constraint $|y_k| > t_k$, which only assigns saliency to the regions where the $k$th feature response has large magnitude. These are regions where the visual stimulus resembles the feature.

In summary, for bandpass features, both salient feature selection and saliency detection are quite simple. Given a sample of responses from feature $Y_k$ in the target and background windows, the parameters $\alpha_k$, $\alpha_{k,i}$, $\xi_k$, $T_k$, and $t_k$ are estimated with (10), (19), and (20). Features $Y_k$ for which $\xi_k \leq 0$ are then discarded. The remaining are ordered by decreasing mutual information $I(Y_k; C)$, using (13) and (14), and the top $K$ selected. Saliency detection is then performed with these features, using (16) and (21). The simplicity of all these operations is crucial for discriminant tracking, where they have to be repeated at each time step.

### 4.5 Spatial Importance Maps

The implementation of a discriminant tracker requires tradeoffs between detector robustness, computational complexity, and adaptivity. Typically, robustness requires decision functions learned from a large training sample. Such functions are difficult to learn and adapt. Adaptation is particularly challenging since both the feature subset added at a given time step and the examples from which it is learned tend to be overwhelmed by those of the previous steps. The more robust a classifier becomes, the more difficult it is to adapt to variations in the statistics of the two classes. However, adaptation is crucial for tracking, where the difficulty is to exactly track objects as they *change appearance* due to variations in lighting, pose, background, etc. The saliency-based discriminant tracker of the previous section is highly adaptive since the learning of salient features is performed at each frame. The price is that, due to limited training data and computation available it can only use a small number of simple features. Hence, as an object detector it is not very robust.

One of its major limitations is that no positional information is stored for the filter responses. As a result, the saliency assessments of (21) do not require spatial consistency of feature responses. For example, it is indifferent if a feature only has large response in the top or bottom half of the target window. Since salient features are usually not expressed in the entire target window, this can lead to noisy saliency maps for target detection. An obvious improvement is to define a feature for each combination of bandpass filter *and* location

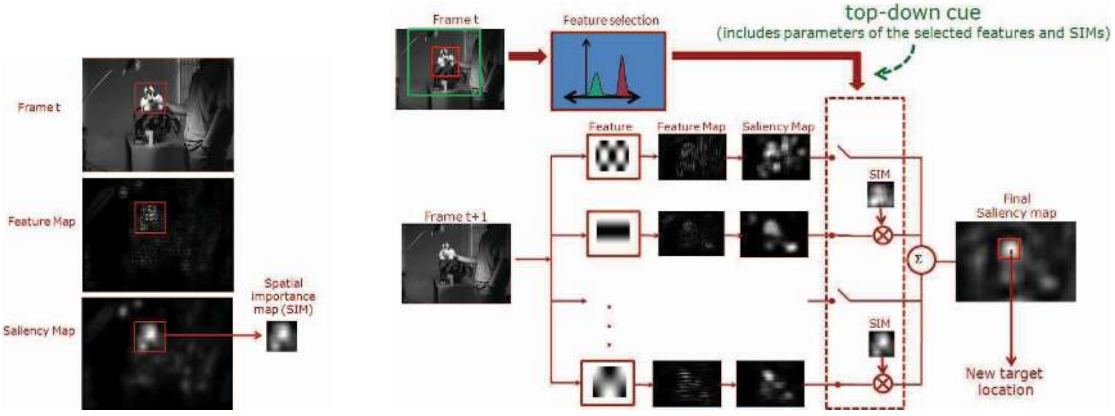


Fig. 2. (a) Spatial importance map (SIM): For each feature, a saliency template of the target is stored at time $t$. (b) Target localization at $t+1$: For each selected feature, a top-down saliency map is computed with (21), and then correlated with the SIM from time $t$ using (23). These saliency maps are combined to produce the overall saliency map, the maximum of which is taken to be the new location of the target.

within the target window, as is popular in face detection [71]. This is, however, infeasible for tracking due to the extensive amounts of computation and training data required. A better alternative is to learn a second layer of features that model *configurations of feature responses*. This is inspired by recent work in HMAX networks [59]. These are biologically inspired object recognition networks, composed of two layers. The first layer can be seen as a (weak) object detector, based on simple bandpass features (Gabor functions) such as those used in this work. The second is an equivalent classifier, but uses more complex features. These are obtained by randomly sampling the responses of the first layer to objects in the target class, and can be interpreted as representative templates of first layer response. In fact, the first layer of the HMAX network can be expanded to perform top-down saliency detection [34], in which case the second-layer filters are *saliency templates*. These summarize the saliency configurations that appear during training, providing a rough characterization of object shape. In this way, the addition of the second HMAX layer increases the robustness of the saliency detector implemented by the first [34].

While the training complexity of a full HMAX network is too large for tracking, the idea of accounting for positional information through the inclusion of saliency templates can still be used. In fact, there is a very natural template to use at time step $t+1$: the map of saliency responses, within the target window $\mathcal{W}^1_{l^*}$, of each salient feature at time $t$. This is denoted *the spatial importance map* and computed as

$$\mathcal{T}_k(l) = \frac{<S_k(l)>_t}{\sum_{l\in\mathcal{W}^1_{l^*}} <S_k(l)>_t}, \quad l \in \mathcal{W}^1_{l^*}, \tag{22}$$

where $<S_k(l)>_t$ is a local average over $4\times 4$ pixels of the $k$th saliency response at time $t$. The proposed normalization guarantees that $\mathcal{T}_k(l)$ sums to 1, giving it the interpretation of a weighting function that emphasizes regions of strong feature response. Since 1) salient features are discriminant for target/background classification and 2) bandpass features respond to image landmarks, such as edges, corners, or texture, these are regions of landmarks that distinguish target from background. In summary, the spatial importance map models the spatial configuration of a set of distinctive target landmarks. This is illustrated in Fig. 2a.

The consistency of the saliency patterns of (21), at times $t$ and $t+1$, can be verified by computing the cross-correlation between the saliency map $S_k$ at time $t+1$ and the spatial importance map $\mathcal{T}_k$ learned at time $t$,

$$R_k(l) = \left\langle S_k|_{\mathcal{W}^1_l}, \mathcal{T}_k \right\rangle, \tag{23}$$

where $S_k|_{\mathcal{W}^1_l}$ is the restriction of $S_k$ to the target window $\mathcal{W}^1_l$, and $<\cdot,\cdot>$ a dot-product. The final saliency measure for the set of $K$ feature responses is

$$S_T(l) = \sum_{k=1}^{K} R_k(l). \tag{24}$$

Its computation is illustrated in Fig. 2b.

The location $l^*_{t+1}$ of largest saliency within a neighborhood $\mathcal{W}^s_{l^*_t}$ of the last known target position $l^*_t$ is selected as the new position of the target at time $t+1$. The feature statistics of target and background windows are updated in an online manner, using

$$\sigma^2_{k,i}(t+1) = \begin{cases} \dfrac{1-\lambda}{n} \displaystyle\sum_{j|y^j_k\in\mathcal{D}_i} \left(y^j_k\right)^2 + \lambda\sigma^2_{k,i}(t), & \text{if } t>0, \\ \dfrac{1}{n} \displaystyle\sum_{j|y^j_k\in\mathcal{D}_i} \left(y^j_k\right)^2, & \text{if } t=0, \end{cases} \tag{25}$$

where $\mathcal{D}_i$ is the sample of examples collected from class $i$ at time $t+1$, and $\lambda$ a decay factor. These statistics are then used for target detection at time $t+2$, and the procedure is iterated.

### 4.6 Scale Adaptive Tracking

Target scale can vary significantly as targets move toward or away from the camera. Trackers that do not adapt to these variations end up relying on a target window that either 1) includes background (when the target shrinks) or 2) excludes foreground (when it grows) and can easily drift. This has motivated a number of scale adaptive extensions of tracking algorithms, ranging from the combination of tracking and scale space representations [11] to specific

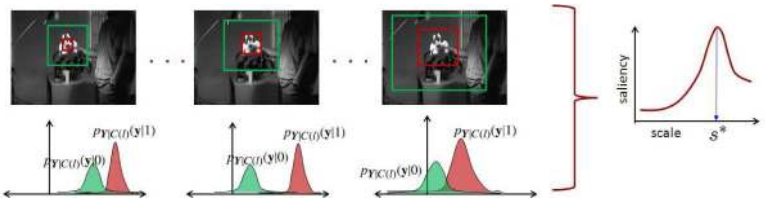


Fig. 3. Saliency-based scale adaptation. The mutual information between the selected salient features and the class label is evaluated over a scale space. The scale at which saliency peaks is chosen as the optimal tracker scale. This is the scale of largest discrimination between target and background.

enhancements applicable only to some trackers, e.g., mean shift [8], [17], [73]. However, scale adaptivity has received little attention in the discriminant tracking literature. Saliency-based tracking offers a natural solution to this problem since scale and saliency are strongly related [43]. In fact, scale adaptation can be achieved as a *byproduct* of discriminant center-surround saliency: The scale of the target is simply that of the center-surround operator that maximizes target/background discrimination. To determine this scale at a given target location, it suffices to search over a discrete scale space $s \in \{s_{min}, s_{max}\}$ of target and background window sizes. For each $s$, the GGD parameters $\alpha^s_{k,i}$, $\alpha^s_k$ are computed from the feature responses in the target and background windows. This can be done efficiently through the use of integral images [71]. For each feature $k$, an integral image of the second moment of feature responses $\mathcal{I}_k(l) = \sum_{j \preceq l} (y^j_k)^2$, where $j \preceq l$ if location $j$ is not below or to the right of location $l$, is first computed. The variance estimate of (25) within a window $\mathcal{D}_i$ of scale $s$ determined by bottom-right, upper-right, bottom-left, and upper-left coordinates $l^s_{br}$, $l^s_{ur}$, $l^s_{bl}$, $l^s_{ul}$ is then

$$\left(\sigma^2_{k,i}\right)^s = \frac{1}{n}\left(\mathcal{I}\left(l^s_{br}\right) - \mathcal{I}\left(l^s_{ur}\right) - \mathcal{I}\left(l^s_{bl}\right) + \mathcal{I}\left(l^s_{ul}\right)\right), \tag{26}$$

where $n$ is the number of pixels in $\mathcal{D}_i$. The GGD parameters are finally estimated with (25) and (10).

Given a set of estimates of the GGD parameters $\alpha^s_{k,i}$, $\alpha^s_k$ at all window sizes $s \in \{s_{min}, s_{max}\}$, the optimal scale is that at which the center-surround saliency measure peaks:

$$\begin{aligned} s^* &= \underset{s:s\in\mathcal{S}_p}{\operatorname{argmax}} \sum_k I_s(Y_k; C), \\ I_s(Y_k; C) &= \sum_{i=0}^{1} \pi_i \left( \log\left(\frac{\alpha^s_k}{\alpha^s_{k,i}}\right) + \frac{1}{\beta}\left(\left(\frac{\alpha^s_{k,i}}{\alpha^s_k}\right)^{\beta} - 1\right)\right), \\ \mathcal{S}_p &= \left\{ s : \frac{\partial(\sum_k I_s(Y_k; C))}{\partial s} = 0, \frac{\partial^2(\sum_k I_s(Y_k; C))}{\partial s^2} < 0 \right\}. \end{aligned} \tag{27}$$

As illustrated in Fig. 3, this is the scale of largest target-background discrimination.

### 4.7 Features

Discriminant tracking can be implemented with any set of bandpass features. In this work, we rely on a combination of discrete cosine transform (DCT) filters to account for spatial information and 3D spatiotemporal Gabor filters to account for motion. DCT features are computed by representing each frame as a Gaussian pyramid and convolving each layer of the pyramid with $8 \times 8$ DCT basis functions. The spatiotemporal features are based on the 3D Gabor filters of [4], [35], which comply with the physiology and psychophysics of the early stages of the visual cortex [4]. Filters tuned to a single spatial frequency of 0.25 cycles/pixel and temporal frequencies of 0 cycles/frames (stationary objects) and $\pm 0.25$ cycles/frames (objects moving to the left or right) were chosen, for a total of three motion-based filters.

It should be noted that while the discriminant tracker does not require explicit modeling of target dynamics (e.g., through Kalman or particle filtering [39]), the inclusion of spatiotemporal features guarantees their *implicit* modeling. For example, if a target is moving to the right at time $t$, the associated spatiotemporal filter is likely to be discriminant at that time. The selection of this filter as a salient feature implies that locations of right-moving objects are more likely to be declared salient at time $t+1$. Hence, the tracker has some ability to *predict* the dynamics of the target. This ability obviously increases with the addition of spatiotemporal filters to the feature set. The limited set used in this work is mostly due to the desire to guarantee low complexity. The implicit modeling of target dynamics is further reinforced by the restriction of the target search to the window $\mathcal{W}^s_{l^*}$. This assumes that targets do not instantaneously jump beyond the region of the focus of attention, i.e., that target motion is smooth.

### 4.8 Automatic Tracker Initialization

Most tracking algorithms assume a known initial target location $l^*$ and bounding box $\mathcal{W}^1_{l^*}$ [6], [18]. 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 [18], they are mostly heuristic. A more principled approach, based on bootstrapping a weak and generic target model for automatic initialization, was proposed in [65]. However, it requires a prespecified 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. This is implemented by evaluating (15) 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). If desired, the search can also be performed over target size, i.e.,

$$\left(l^*, \mathbf{s}^*\right) = \underset{l,\mathbf{s}}{\operatorname{argmax}} \sum_k I_{l,\mathbf{s}}\left(Y_k; C\right), \tag{28}$$

where

$$I_{l,\mathbf{s}}(Y_k; C) = \sum_{i=0}^{1} \pi_i \left( \log\left(\frac{\alpha^{l,\mathbf{s}}_k}{\alpha^{l,\mathbf{s}}_{k,i}}\right) + \frac{1}{\beta}\left(\left(\frac{\alpha^{l,\mathbf{s}}_{k,i}}{\alpha^{l,\mathbf{s}}_k}\right)^{\beta} - 1\right)\right), \tag{29}$$

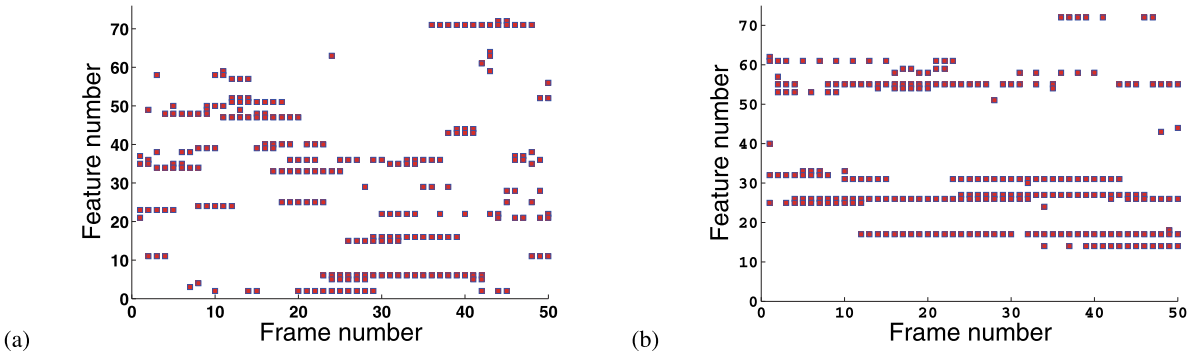


Fig. 4. Features selected in the first 50 frames on (a) "karlsruhe" and (b) "sylvester." The spatial features are numbered from 1 to 64, and correspond to the zig-zag scanning order of the DCT basis functions, while the three spatiotemporal features are numbered from 70 to 72.

and the parameters $\alpha_{k,i}^{l,\mathbf{s}}, \alpha_k^{l,\mathbf{s}}$ are learned, with (10), from feature responses in the windows associated with a center-surround operator of size $\mathbf{s} = (s_x, s_y)$, centered at location $l$. As before, these parameters can be computed efficiently with resort to integral images. Overall, this initialization finds the regions whose motion and appearance is most distinct from those of the background.

The use of (28) has a number of appealing properties. First, it can be seen as a 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 [47]. The proposed simplification sacrifices the ability to model complex dynamics for the sake of computational tractability. Second, while the use of spatiotemporal features enables it to account for both target appearance and motion, it 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. Third, it reduces initialization to 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. This unification of tracker initialization and operation is not possible for most previous trackers.

## 5 Experiments and Results

The performance of the proposed saliency-based tracker was evaluated with an extensive set of experiments. We next report the results of this evaluation.

### 5.1 Comparison to Previous Trackers

The saliency-based tracker was compared to four trackers in the literature: three discriminant trackers, the MILTracker of [7], the method of Collins et al. [18], and the ensemble tracker of [6], and the incremental visual tracker (IVT) of [57]. 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 [6], [18].

The performance of all five methods was evaluated against manual groundtruth. The definition of tracking error followed [22]. At time $t$, the error was defined as the normalized lack of overlap between the groundtruth target bounding box, $G^t$, and that produced by the tracker, $B^t$. The average tracking error, over the $T$ frames in a video sequence,

$$\epsilon = \frac{1}{T}\sum_t \left( 1 - \frac{\sum_{ij} G_{i,j}^t B_{i,j}^t}{\sum_{ij} G_{i,j}^t + \sum_{ij} B_{i,j}^t - \sum_{ij} G_{i,j}^t B_{i,j}^t} \right). \quad (30)$$

was then used as the measure of tracker performance. This ranged between $\epsilon = 0$, for perfectly correct tracking, and $\epsilon = \infty$, for complete loss of tracking.

The test video sequences were selected from diverse sources (e.g previous works, standard databases, and the web). All sequences include challenging tracking problems, such as varying illumination, occlusion, out-of-plane object rotation, or changes in perspective. For instance, the "motinas" sequence of [46] shows a person turning by $360^\circ$ in extremely low light (Fig. 5a), while the "athlete" sequence includes extreme variations of appearance, due to occlusion and strong video compression artifacts (Fig. 5b). The "skater" (Fig. 5d) and "CAVIAR" sequences (from [1]) include severe partial occlusions. To further increase the tracking difficulty, all sequences were converted to grayscale. This required a reimplementation of the Collins tracker with DCT features, instead of the R, G, B color features originally used in [18]. All algorithms were manually initialized with a target bounding box, in the first frame. The background bounding box had sides four times larger than the corresponding sides of the target box.

The saliency-based tracker used a two-level Gaussian pyramid, for a total of $N = 3 + 64 \times 2 = 131$ features ($8 \times 8$ DCT features per level plus three spatiotemporal Gabor features). The number of selected salient features, $K$, is a tunable parameter. To understand its impact on tracking performance, it was varied in the range $[1, 29]$, for two representative sequences. 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. Fig. 4 shows the five features selected in the first 50 frames of two representative sequences. Note that the same, or very similar, features are selected at successive frames, leading to a fairly stable set of selected features over time. The search neighborhood, $\mathcal{W}_{l^*}^s$, was centered at the current target position $l^*$ and had twice the size of the object

TABLE 1
Average Tracking Error of the Five Trackers Compared

| Sequence | IVT | Collins | Ensemble | MIL | Sal | Sal+SIM | Sal+STF | Sal+SIM+STF |
|---|---|---|---|---|---|---|---|---|
| coke11 | 0.97 (82%) | 0.76 (4%) | 0.71 (22%) | 0.68 (0%) | **0.62** (2%) | 0.68 (2%) | 0.63 (2%) | 0.68 (0%) |
| tiger2 | 0.80 (60%) | 0.78 (38%) | 0.88 (72%) | **0.38** (0%) | 0.64 (28%) | 0.77 (20%) | 0.78 (50%) | 0.44 (2%) |
| karls | 0.64 (32%) | 0.47 (11%) | 0.93 (52%) | **0.29** (0%) | 0.52 (37%) | 0.51 (0%) | 0.53 (37%) | 0.31 (0%) |
| dtneu | 0.93 (91%) | 0.27 (0%) | 0.96 (82%) | 0.49 (5%) | 0.21 (0%) | 0.21 (0%) | **0.15** (0%) | 0.26 (0%) |
| plushtoy | **0.11** (0%) | 0.37 (0%) | 0.38 (0%) | 0.17 (0%) | 0.16 (0%) | 0.26 (0%) | 0.21 (0%) | 0.25 (0%) |
| ram | 0.77 (51%) | 0.86 (69%) | 0.87 (64%) | 0.64 (36%) | 0.36 (3%) | 0.77 (59%) | **0.33** (0%) | **0.33** (0%) |
| ballroom | 0.62 (51%) | 0.38 (0%) | 0.70 (26%) | **0.34** (0%) | 0.39 (0%) | 0.46 (0%) | 0.38 (0%) | 0.44 (0%) |
| roadcrossing | 0.51 (0%) | 0.74 (45%) | 0.83 (52%) | 0.46 (0%) | 0.87 (81%) | 0.78 (44%) | 0.77 (56%) | **0.45** (0%) |
| motinas | 0.60 (50%) | 0.47 (21%) | 0.73 (60%) | 0.61 (31%) | 0.95 (88%) | **0.22** (0%) | 0.92 (85%) | 0.24 (0%) |
| athlete | 0.98 (97%) | 0.78 (55%) | 0.94 (90%) | 0.92 (68%) | 0.75 (58%) | 0.41 (0%) | 0.75 (55%) | **0.37** (0%) |
| skater | 0.94 (88%) | 0.49 (0%) | 0.62 (20%) | 0.93 (80%) | 0.47 (0%) | 0.33 (0%) | 0.36 (0%) | **0.30** (0%) |
| CAVIAR | 0.34 (0%) | 0.56 (0%) | 0.96 (67%) | 0.48 (0%) | 0.73 (66%) | 0.33 (0%) | **0.29** (0%) | 0.31 (0%) |
| seq10 | **0.03** (0%) | 0.99 (99%) | 0.94 (29%) | 0.08 (0%) | 0.89 (96%) | 0.14 (0%) | 0.94 (98%) | 0.14 (0%) |
| average | 0.63 | 0.61 | 0.80 | 0.50 | 0.55 | 0.45 | 0.51 | **0.35** |

*0 indicates perfect tracking, 1 complete lack of overlap between groundtruth and target bounding box produced by the tracker. The number in parentheses indicates the percentage of frames for which there was no overlap between groundtruth and target bounding boxes.*

bounding box. Finally, to evaluate the contribution of all tracker components, namely, the spatial DCT features, the spatial importance map (SIM) and the spatiotemporal features (STF), we implemented four variants of the tracker. These are denoted "Sal" (tracker including only the saliency measure of (16) and the spatial DCT features), "Sal+SIM" (saliency plus SIM), "Sal+STF" (saliency plus spatiotemporal) and "Sal+SIM+STF" (saliency plus spatiotemporal features and SIM).

Table 1 presents the errors measured on a set of 13 sequences. For each value of the error, the table also shows in parentheses the fraction of frames for which there was no overlap between the target bounding box obtained by the tracker and the groundtruth bounding box. A number of conclusions can be drawn from these results. First, the baseline "Sal" tracker is already quite competitive with the state of the art. Its average error is only inferior to that of the MIL tracker and by a small amount. While these results could be improved by including more than five DCT features, we found that a better tradeoff between accuracy and complexity is provided by the other extensions. Second, both the addition of STFs and SIMs strengthen tracking performance. The larger improvement of Sal+SIM indicates that it is particularly important to account for the spatial configuration of the target features. Third, the effects of the two tracker extensions are complementary, with the "Sal+SIM+STF" tracker achieving the clear best average performance. In the following discussion, we refer to this version as the discriminant saliency tracker (DST).

Overall, the DST or its variants achieve the top performance on eight sequences. In all these, the DST achieves the best performance over the previous four methods (IVT, Collins, Ensemble, and MIL). Among these, MIL is the best performer, with lowest error rates on three sequences, followed by IVT, with two. When compared to MIL, DST has a very similar tracking error in the sequences where the latter performs the best. On the other hand, in the sequences where it is the top performer, the error of DST can be substantially smaller than that of MIL. With respect to IVT, the DST is also clearly superior. While IVT oscillates between some very small ("plushtoy", "seq10") and mostly large errors, the DST error is small for all sequences. The overall superiority of the DST is captured by the fact that its average error, across sequences, is about 64 percent that of the next best method (MIL). Alternatively, it can be seen from the fact that, while DST never loses track, this happens for all other methods in four of the sequences ("ram," "skater," "motinas," and "athlete").

Fig. 5 illustrates the tracking results on four of the sequences considered. The qualitative performance of IVT and the ensemble tracker is quite poor as these methods lose the target in most scenes. Somewhat better performance is achieved by the Collins and MIL trackers, which only lose the target when it undergoes extreme appearance variations due to partial occlusions, illumination changes, or rotation. On the other hand, DST tracks the targets successfully in all sequences. The results on "seq10," a very long sequence used in [32], show that DST is also able to track over long durations reliably without drifting (Fig. 5c). Overall, it is clear that DST has the best performance. Videos of all tracking results are available from [2].

### 5.2 Scale Adaptive Tracking

To test scale adaptivity, the performance of the DST was evaluated on various sequences of widely varying target size. The comparison was restricted to IVT since no scale adaptive extensions are available for the other methods. The initial position and size of the target were manually specified since IVT has no ability for automatic initialization. Examples of the tracking results are shown in Fig. 6. Note that these sequences are challenging in many ways. Besides wide scale variability, the target can change appearance quite dramatically due to a 360 degree rotation and nonrigid motion (on "gravel" the subject turns, picks up a rock, and throws it in the water), as well as perspective effects (on "dirtbike" the motorcycle approaches the camera from the left and leaves to the right), and the background varies substantially (sky, then sand dunes, then strongly shaded background on "dirtbike"). While IVT loses track in both cases, DST is able to maintain track throughout the sequences, accurately tracking

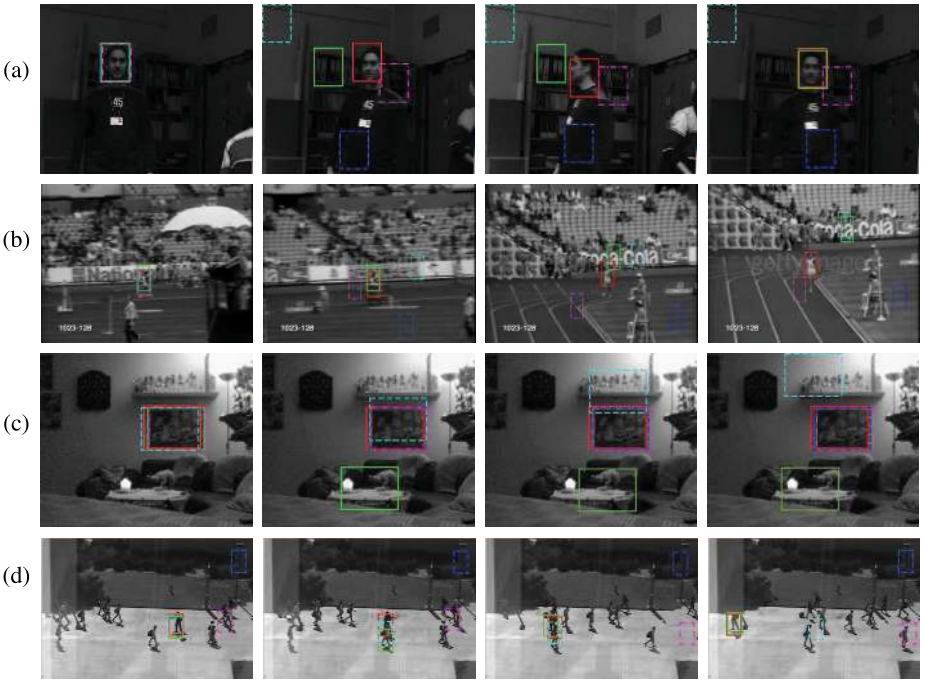


Fig. 5. Tracking results on (a) "motinas_toni_change_ill" [46]—the person turns around and the illumination changes drastically, (b) "athlete"—a person running inside a stadium. The video is very noisy and the target appearance changes widely, (c) "seq10"—an extremely long video sequence used in [32] to test for drifting, (d) "skater" on a pedestrian walkway—the target undergoes partial occlusions on multiple occasions. Target locations: DST—thick red box, Collins—thick green box, ensemble—cyan dashed box, IVT—blue dashed box, MIL—magenta dashed box.

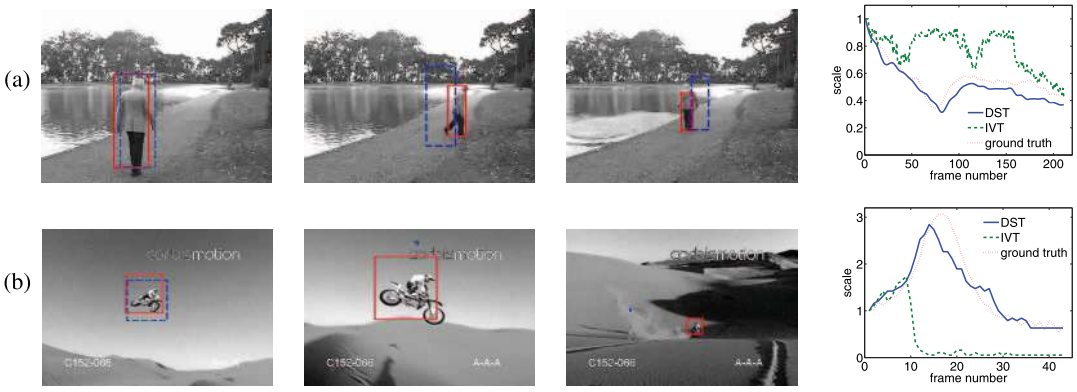


Fig. 6. Scale adaptive tracking on (a) "gravel" and (b) "dirtbike." Target locations: DST—red box, IVT—dashed blue box. Plots of target scale, expressed as the ratio of target size at a frame to size in the initial frame for the respective sequences are shown in the extreme right.

the target position. This robustness is due to the continuous updating of the features used to represent both target and background and the discriminant nature of the tracker. Panels on the extreme right of Fig. 6 present plots of the variation of target scale over time. It is clear that DST is able to handle a wide variability of target scales, closely matching the scale of the groundtruth, while IVT loses track ("gravel") or dwindles into an infinitesimal target box ("dirtbike"). Table 2 summarizes the errors measured on these and two other sequences, confirming the superior performance of DST. Videos of the sequences are again available from [2].

### 5.3 Automatic Initialization

Finally, we performed a set of experiments designed to evaluate automatic tracker initialization using DST. Since none of the other methods have this capability, no comparison was performed for these sequences. Examples of DST results are shown in Fig. 7. The tracker uses the bottom-up discriminant saliency procedure of Section 4.8 to identify the object to track. The region of maximal saliency is then input to the scale adaptive DST algorithm, which tracks the target through the remaining frames. The leftmost column of Fig. 7 shows the bottom-up saliency map, and the columns on the

TABLE 2
Average Tracking Error of IVT and DST When Target Scale Varies Widely

| Name | IVT | DST |
|---|---|---|
| dirtbike | 0.86 (76%) | **0.33** (0%) |
| speedboat | 0.45 (0%) | **0.38** (0%) |
| gravel | 0.76 (8%) | **0.44** (0%) |
| baseball | 0.96 (75%) | **0.44** (0%) |
| average | 0.76 | **0.40** |

*The number in parentheses indicates the fraction of frames in which the groundtruth and target bounding box had no overlap.*

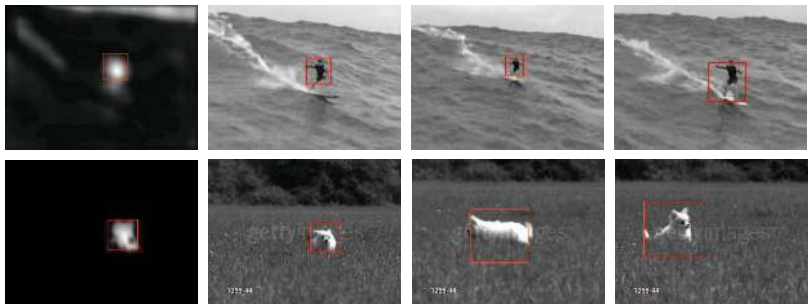


Fig. 7. Automatic initialization and tracking. The bottom-up saliency map used to initialize the tracker is shown in the left column. Target bounding boxes are shown in red. (a) "surfer," (b) "dog." Target locations in subsequent frames are shown in red.

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 3 presents the error measures obtained for these sequences. 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 accurate scale adaptive tracking, in scenes with complex motion. Videos of all sequences are available in [2].

## 6 Discussion and Connections to Other Discriminant Trackers

At an abstract level, the proposed DST is similar to previous discriminant trackers [6], [7], [18]. Like the DST, these are center-surround discriminators, equating target to center and background to surround. In fact, they rely on classifier design and target detection operations that are similar in spirit to those of DST. There are, nevertheless, differences of detail that significantly affect tracking performance.

With respect to feature selection and classifier design, all discriminant trackers analyze the feature set for target/background discrimination. Collins et al. [18] first compute histograms of filter responses on the R, G, B channels of both target and background, and construct a log likelihood ratio between the two class histograms, considering this a new nonlinear feature. Feature discrimination is evaluated by a Fisher discriminant-like *variance ratio* that measures how tightly clustered the log-likelihood ratios are for the two classes. This is equivalent to transforming the features into a nonlinear space and learning a linear classifier in that space. It is optimal, in the minimum probability of error sense, only when the classes are *Gaussian* and have equal covariance, after the feature transformation. Overall, the tracker suffers from the fact that this discrimination measure is somewhat heuristic. The distribution of log-likelihood ratios is hard to characterize [16], the assumption of unimodality (Gaussianity) does not hold in general (i.e., for all features), and is especially troubling when there is background clutter. There is even less evidence in support of the assumption of equal class variance. These observations could account for the limited effectiveness of the tracker.

The ensemble tracker [6] relies on a set ("ensemble") of weak hyperplane classifiers to separate target from background. Each weak learner implements a threshold on a linear combination of the original features. A simpler approach is used by the MIL tracker [7], where each weak learner is a decision stump, i.e., a threshold on one of the original features. Both trackers rely on the classification error rate as measure of discrimination for feature selection. While this is a close approximation to the mutual information used by the DST [69], the feature selection procedure is quite different: Both the ensemble and MIL trackers rely on boosting (AdaBoost and MILBoost, respectively). Boosting is sensitive to outliers. This is a limitation in tracking since the target and background classes of a tracking problem are rarely exclusive. On the contrary, a certain amount of background is usually covered by the target window and vice versa. The sensitivity of boosting to outliers is well known in machine learning, where a number of extensions have been proposed to address the problem [50]. This is indeed the difference between the ensemble and MIL trackers. The latter implements boosting under the MIL formalism exactly to decrease outlier sensitivity. By minimizing this problem, MIL achieves significantly better results.

TABLE 3
Comparison of Automatic and Manual Tracker Initialization

| Name | Auto Init | Manual Init |
|---|---|---|
| dirtbike | **0.33** (0%) | **0.33** (0%) |
| surfer | 0.33 (0%) | **0.32** (0%) |
| dog | 0.38 (0%) | **0.37** (0%) |
| skiing | **0.27** (0%) | 0.28 (0%) |

*The number in parentheses indicates the fraction of frames in which the groundtruth and target bounding box had no overlap.*

## 7 Conclusion

In this work, we have shown that discriminant tracking follows naturally from the discriminant formulation of visual saliency. In particular, tracking can be implemented with a combination of bottom-up center-surround discriminant saliency and spatial attention for learning, feature-based attention for feature selection, and top-down saliency for target detection. 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, and provides a unified solution to the problems of classifier design, target detection, automatic tracker initialization, and scale adaptation. Experimental results show the improved performance of the

proposed discriminant saliency tracker over existing approaches. An implementation of this tracker in C, without any optimization, currently runs at $\sim 1.5$ frames per second (fps), on a standard PC without special hardware. On the same machine, the running times of other discriminant trackers are comparable ($\sim 4$ fps for MIL and $\sim 3$ fps for the Collins tracker).

Among its shortcomings, DST does not explicitly retain target features that appear in the previous frames. Therefore, it cannot handle prolonged partial or complete occlusions. Also, as the approach depends on finding features that can discriminate the target from the background, DST is not suitable when there are objects very similar to the target in the background or for tracking large targets with inadequate backgrounds. Finally, DST has been designed for tracking single targets. To track multiple targets, DST has to be augmented with additional modules such as an identity management scheme.

## REFERENCES


[1] http://homepages.inf.ed.ac.uk/rbf/caviar, 2011.
[2] http://www.svcl.ucsd.edu/projects/tracking/results.html, 2012.
[3] A. Adam, E. Rivlin, and I. Shimshoni, "Robust Fragments-Based Tracking Using the Integral Histogram," *IEEE Conf. Computer Vision and Pattern Recognition,* pp. 798-805, 2006.
[4] E.H. Adelson and J.R. Bergen, "Spatiotemporal Energy Models for the Perception of Motion," *J. Optical Soc. of Am. A,* vol. 2, no. 2, pp. 284-299, 1985.
[5] R. Allen, P. Mcgeorge, D. Pearson, and A.B. Milne, "Attention and Expertise in Multiple Target Tracking," *Applied Cognitive Psychology,* vol. 18, no. 3, pp. 337-347, 2004.
[6] S. Avidan, "Ensemble Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 2, pp. 261-271, Feb. 2007.
[7] B. Babenko, M.-H. Yang, and S. Belongie, "Visual Tracking with Online Multiple Instance Learning," *IEEE Conf. Computer Vision and Pattern Recognition,* pp. 983-990, 2009.
[8] S. Birchfield and S. Rangarajan, "Spatiograms versus Histograms for Region-Based Tracking," *IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1158-1163, 2005.
[9] M. Black and A. Jepson, "Eigentracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation," *Int'l J. Computer Vision,* vol. 26, no. 1, pp. 63-84, 1998.
[10] E. Blaser, Z. Pylyshyn, and A.O. Holcombe, "Tracking an Object through Feature-Space," *Nature,* vol. 408, pp. 196-199, 2000.
[11] L. Bretzner and T. Lindeberg, "Feature Tracking with Automatic Selection of Spatial Scales," *Computer Vision and Image Understanding,* vol. 71, no. 3, pp. 385-392, 1998.
[12] R. Buccigrossi and E. Simoncelli, "Image Compression via Joint Statistical Characterization in the Wavelet Domain," *IEEE Trans. Image Processing,* vol. 8, no. 12, pp. 1688-1701, Dec. 1999.
[13] P. Cavanagh, "Attention-Based Motion Perception," *Science,* vol. 257, no. 5076, pp. 1563-1565, 1992.
[14] P. Cavanagh and G.A. Alvarez, "Tracking Multiple Targets with Multifocal Attention," *Trends in Cognitive Sciences,* vol. 9, no. 7, pp. 349-354, 2005.
[15] J. Cavanaugh, W. Bair, and J. Movshon, "Nature and Interaction of Signals from the Receptive Field Center and Surround in Macaque V1 Neurons," *J. Neurophysiology,* vol. 88, pp. 2530-2546, 2002.
[16] H. Chernoff, "On the Distribution of the Likelihood Ratio," *The Annals of Math. Statistics,* vol. 25, no. 3, pp. 573-578, 1954.
[17] R. Collins, "Mean-Shift Blob Tracking through Scale Space," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, 2003.
[18] R. Collins, Y. Liu, and M. Leordeanu, "Online Selection of Discriminative Tracking Features," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 10, pp. 1631-1643, Oct. 2005.
[19] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 5, pp. 564-577, May 2003.
[20] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2005.
[21] M.N. Do and M. Vetterli, "Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance," *IEEE Trans. Image Processing,* vol. 11, no. 2, pp. 146-158, Feb. 2002.
[22] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge (VOC '07) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.
[23] J. Feldman and P.D. Tremoulet, "Individuation of Visual Objects over Time," *Cognition,* vol. 99, no. 2, pp. 131-165, 2006.
[24] Y. Freund and R.E. Schapire, "1997, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *European Conf. Computational Learning Theory,* pp. 23-37, 1995.
[25] D. Gao, S. Han, and N. Vasconcelos, "Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition," *IEEE Trans. Pattern Anaysis and Machine Intelligence,* vol. 31, no. 6, p. 989, June 2009.
[26] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the Plausibility of the Discriminant Center-Surround Hypothesis for Visual Saliency," *J. Vision,* vol. 8, no. 7, pp. 1-18, 2008.
[27] D. Gao and N. Vasconcelos, "Discriminant Saliency for Visual Recognition from Cluttered Scenes," *Proc. Advances in Neural Information Processing Systems,* 2005.
[28] D. Gao and N. Vasconcelos, "Bottom-Up Saliency Is a Discriminant Process," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.
[29] D. Gao and N. Vasconcelos, "Discriminant Interest Points Are Stable," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 2007.
[30] D. Gao and N. Vasconcelos, "Decision-Theoretic Saliency: Computational Principle, Biological Plausibility, and Implications for Neurophysiology and Psychophysics," *Neural Computation,* vol. 21, pp. 239-271, Jan. 2009.
[31] H. Grabner and H. Bischof, "On-Line Boosting and Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 260-267, 2006.
[32] H. Grabner, C. Leistner, and H. Bischof, "Semi-Supervised On-Line Boosting for Robust Tracking," *Proc. European Conf. Computer Vision,* pp. 234-247, 2008.
[33] B. Han and L. Davis, "On-Line Density-Based Appearance Modeling for Object Tracking," *Proc. 10th IEEE Int'l Conf. Computer Vision,* pp. 1492-1499, 2005.
[34] S. Han and N. Vasconcelos, "Biologically Plausible Saliency Mechanisms Improve Feedforward Object Recognition," *Vision Research,* vol. 50, pp. 2295-2307, 2010.
[35] D. Heeger, "Optical Flow from Spatiotemporal Filters," *Int'l J. Computer Vision,* vol. 1, no. 4, pp. 279-302, 1988.
[36] J. Ho, K. Lee, M. Yang, and D. Kriegman, "Visual Tracking Using Learned Linear Subspaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, 2004.
[37] J. Huang and D. Mumford, "Statistics of Natural Images and Models," *Proc. IEEE Conf. Computer Vision and Pattern Recogition,* pp. 541-547, 1999.
[38] J. Intriligator and P. Cavanagh, "The Spatial Resolution of Visual Attention," *Cognitive Psychology,* vol. 43, pp. 171-216, 1997.
[39] M. Isard and A. Blake, "Condensation Conditional Density Propagation for Visual Tracking," *Int'l J. Computer Vision,* vol. 29, pp. 5-28, 1998.
[40] L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vision Research,* vol. 40, pp. 1489-1506, 2000.
[41] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[42] A.D. Jepson, D.J. Fleet, and T.F. El-Maraghi, "Robust Online Appearance Models for Visual Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 10, pp. 1296-1311, Oct. 2003.
[43] T. Kadir and M. Brady, "Scale, Saliency and Image Description," *Int'l J. Computer Vision,* vol. 45, pp. 83-105, Nov. 2001.
[44] B. Keane and Z. Pylyshyn, "Is Motion Extrapolation Employed in Multiple Object Tracking? Tracking as a Low-Level, Non-Predictive Function," *Cognitive Psychology,* vol. 52, no. 4, pp. 346-368, 2006.
[45] R. Lin, D. Ross, J. Lim, and M. Yang, "Adaptive Discriminative Generative Model and Its Applications," *Proc. Advances in Neural Information Processing Systems,* pp. 801-808, 2004.

[46] E. Maggio and A. Cavallaro, "Hybrid Particle Filter and Mean Shift Tracker with Adaptive Transition Model," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* 2005.
[47] V. Mahadevan and N. Vasconcelos, "Background Subtraction in Highly Dynamic Scenes," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, 2008.
[48] V. Mahadevan and N. Vasconcelos, "Spatiotemporal Saliency in Dynamic Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 1, pp. 171-177, Jan. 2010.
[49] T. Makovski and Y. Jiang, "Feature Binding in Attentive Tracking of Distinct Objects," *Visual Cognition,* vol. 17, no. 1, pp. 180-194, 2009.
[50] H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos, "On the Design of Robust Classifiers for Computer Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 779-786, 2010.
[51] H. Nguyen and A. Smeulders, "Robust Tracking Using Foreground-Background Texture Discrimination," *Int'l J. Computer Vision,* vol. 69, no. 3, pp. 277-293, 2006.
[52] H.C. Nothdurft, "Texture Segmentation and Pop-Out from Orientation Contrast," *Vision Research,* vol. 31, no. 6, pp. 1073-1078, 1991.
[53] S.E. Palmer, *Vision Science: Photons to Phenomenology.* MIT Press, 1999.
[54] M. Posner, C. Snyder, and B. Davidson, "Attention and the Detection of Signals," *J. Experimental Psychology: General,* vol. 109, no. 2, pp. 160-174, 1980.
[55] Z.W. Pylyshyn and R.W. Storm, "Tracking Multiple Independent Targets: Evidence for a Parallel Tracking Mechanism," *Spatial Vision,* vol. 3, no. 3, pp. 179-197, 1988.
[56] D. Ramanan, D. Forsyth, and A. Zisserman, "Tracking People by Learning Their Appearance," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 1, pp. 65-81, Jan. 2007.
[57] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental Learning for Robust Visual Tracking," *Int'l J. Computer Vision,* vol. 77, nos. 1-3, pp. 125-141, May 2008.
[58] A.B. Sekuler and R. Sekuler, "Collisions between Moving Visual Targets: What Controls Alternative Ways of Seeing an Ambiguous Display?" *Perception,* vol. 28, no. 4, pp. 415-432, 1999.
[59] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 3, pp. 411-426, Mar. 2007.
[60] K. Sharifi and A. Leon-Garcia, "Estimation of Shape Parameter for Generalized Gaussian Distributions in Subband Decompositions of Video," *IEEE Trans. Circuits and Systems for Video Technology,* vol. 5, no. 1, pp. 52-56, Feb. 1995.
[61] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On Advances in Statistical Modeling of Natural Images," *J. Math. Imaging and Vision,* vol. 18, no. 1, pp. 17-33, 2003.
[62] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 246-252, 1999.
[63] B. Tatler, "The Central Fixation Bias in Scene Viewing: Selecting an Optimal Viewing Position Independently of Motor Biases and Image Feature Distributions," *J. Vision,* vol. 7, no. 14, pp. 1-17, 2007.
[64] K. Toyama and G.D. Hager, "Incremental Focus of Attention for Robust Visual Tracking," *Int'l J. Computer Vision,* pp. 189-195, 1996.
[65] K. Toyama and Y. Wu, "Bootstrap Initialization of Nonparametric Texture Models for Tracking," *Proc. European Conf. Computer Vision,* 2000.
[66] S. Treue and J. Trujillo, "Feature-Based Attention Influences Motion Processing Gain in Macaque Visual Cortex," *Nature,* vol. 399, no. 6736, pp. 575-579, 1999.
[67] P. Tseng, R. Carmi, I. Cameron, D. Munoz, and L. Itti, "Quantifying Center Bias of Observers in Free Viewing of Dynamic Natural Scenes," *J. Vision,* vol. 9, no. 7, article 4, 2009.
[68] M. Vasconcelos and N. Vasconcelos, "Natural Image Statistics and Low-Complexity Feature Selection," *IEEE Trans. Pattern Analysis and Machine Intelligent,* vol. 31, no. 2, pp. 228-244, Feb. 2008.
[69] N. Vasconcelos, "Feature Selection by Maximum Marginal Diversity," *Proc. Advances in Neural Information Processing Systems,* 2002.
[70] F.A.J. Verstraten, P. Cavanagh, and A.T. Labianca, "Limits of Attentive Tracking Reveal Temporal Properties of Attention," *Vision Research,* vol. 40, no. 26, pp. 3651-3664, 2000.
[71] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 511-518, 2001.
[72] A. Yarbus, *Eye Movements and Vision.* Plenum, 1967.
[73] A. Yilmaz, "Object Tracking by Asymmetric Kernel Mean Shift with Automatic Scale and Orientation Selection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 1-6, 2007.
[74] A. Yilmaz, O. Javed, and M. Shah, "Object Tracking: A Survey," *ACM Computing Surveys,* vol. 38, no. 4, article 13, 2006.
[75] Y. Zhong, A. Jain, and M. Dubuisson-Jolly, "Object Tracking Using Deformable Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 5, pp. 544-549, May 2000.

**Vijay Mahadevan** received the BTech degree from the Indian Institute of Technology, Madras, the MS degree from Rensselaer Polytechnic Institute, Troy, New York, and the PhD degree from the University of California, San Diego, in 2002, 2003, and 2011, respectively, all in electrical engineering. Currently, he is with Yahoo! Labs, Bangalore. From 2004 to 2006, he was with the Multimedia DSP group at Qualcomm Inc, San Diego, California. His research interests include computer vision, machine learning, signal and image processing, and biomedical image analysis. He is a member of the IEEE.

**Nuno Vasconcelos** received the licenciatura degree in electrical engineering and computer science from the Universidade do Porto, Portugal, and the MS and PhD degrees from the Massachusetts Institute of Technology in 1988, 1993, and 2000, respectively. From 2000 to 2002, he was a member of the research staff at the Compaq Cambridge Research Laboratory, which became the HP Cambridge Research Laboratory in 2002. In 2003, he joined the Electrical and Computer Engineering Department at the University of California, San Diego, where he heads the Statistical Visual Computing Laboratory. His work spans various areas, including computer vision, machine learning, signal processing and compression, and multimedia systems. He is the recipient of a US National Science Foundation CAREER award, a Hellman Fellowship, and has authored more than 75 peer-reviewed publications. He is a senior member of the IEEE.