Saliency-based Discriminant Tracking

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

Abstract

We propose a biologically inspired framework for visual tracking based on discriminant center surround saliency. At each frame, discrimination of the target from the background is posed as a binary classification problem. From a pool of feature descriptors for the target and background, a subset that is most informative for classification between the two is selected using the principle of maximum marginal diversity. Using these features, the location of the target in the next frame is identified using top-down saliency, completing one iteration of the tracking algorithm. We also show that a simple extension of the framework to include motion features in a bottom-up saliency mode can robustly identify salient moving objects and automatically initialize the tracker. The connections of the proposed method to existing works on discriminant tracking are discussed. 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 pre-requisite for important applications of computer vision, such as surveillance [13], activity or behavior recognition [27]. Many years of research on the tracking problem have produced a diverse set of approaches and a rich collection of tracking algorithms [33]. A popular subset among these are the so-called *appearance based* methods, which learn and maintain a model of target appearance and use it to locate the target as time evolves. For instance, targets can be represented by their contours, and the temporal evolution of these contours modeled with particle filters [20]. Alternatively, target appearance can be represented by kernel weighted histograms, which are popular in the context of mean shift algorithms [8]. More sophisticated appearance models include a combination of long term stable representations and short term descriptors [21], or low-dimensional subspace representations that are updated incrementally [25]. All of these methods 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 proposed the formulation of "discriminant tracking" - object tracking as continuous object detection, by posing the problem as one of incremental "target vs. background" classification [7, 3, 18]. Given a target bounding box at video frame t, a classifier is trained to distinguish target features from those of the background. This classifier is then used to determine the location of the target in frame $t + 1$. The bounding box is moved to this location, the classifier updated, and the process iterated.

In the biological world, object tracking is tightly related to attentional tasks, such as the guidance of eye movements. Due to the evolutionary advantages of solving these tasks accurately, it is not surprising that biological vision systems have developed extremely efficient tracking mechanisms, in terms of both accuracy and speed. The effectiveness of these mechanisms, even under the most adverse conditions (e.g. highly cluttered scenes, low-light, etc), is a consequence of the availability of robust saliency mechanisms, that cause pre-attentive pop-out of salient locations in the visual field [24]. 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 tracking or visual search [32] with modest amounts of computation. The robustness of the biological solutions has motivated computer vision researchers to augment conventional tracking algorithms with FoA mechanisms. For instance, Toyama and Hager [28] proposed an incremental FoA procedure to combine multiple trackers, leading to increased robustness. Nevertheless, there has been little work aimed at deriving a principled understanding of what computational mechanisms could be used by biological vision to solve the tracking problem, how these mechanisms relate to the state-of-the-art algorithms from computer vision, and how these connections could be exploited to achieve increased com-

puter vision performance.

In this work, we present a contribution along these three dimensions. We consider tracking in the context of center-surround saliency mechanisms that are prevalent in biological vision [17, 5]. In particular, we consider a recently proposed computational principle for visual saliency, denoted by *discriminant saliency* [17]. This principle has been shown to have a number of attractive properties for both the biological and computer vision communities. In the area of biological vision, is has been shown 1) to lead to computational models of saliency that replicate an extensive collection of psychophysics from both saliency and visual search [14], and 2) to have a 1-1 to one mapping to the standard neurophysiological model of the area V1 of the brain [17]. For computer vision, it has been shown to produce algorithms that achieve state-of-the-art performance in the problems of interest point detection [16], object recognition [15], and background subtraction [23].

In this work, we show that discriminant tracking can be posed as a particular instance of this generic principle. In particular, we show that it provides a *unified* and *principled* framework for the solution of the three problems posed by the design of a discriminant tracker: target initialization, feature selection and target detection. This unifies tracking with prior work on background subtraction, enabling highly robust automatic target initialization. By exploiting connections between discriminant saliency and the statistics of natural images, it also enables highly computationally efficient tracking algorithms without compromise of discrimination optimality. This is shown not to be the case for previous computer vision solutions to discriminant tracking, which the proposed discriminant tracking algorithm is shown to outperform experimentally.

2. Tracking Using Discriminant Saliency

Discriminant saliency [17] poses the saliency problem as one of optimal decision-making between two classes of visual stimuli: a class of *stimuli of interest*, and a *background* or null hypothesis, consisting of stimuli that are not salient. This is implemented by establishing a binary classification problem which opposes the stimuli of interest to the null hypothesis. The saliency of each location in the visual field is then equated to the discriminant power (expected classification accuracy) of a set of visual features, extracted from that location, for the differentiation between the two classes.The locations that can be classified, with lowest expected probability of error, as containing stimuli of interest are denoted as salient.

The discriminant saliency principle is generic and can be applied to various vision problems, by suitable definition of class of interest and null-hypothesis. For example, it can be used to implement one-vs-all object detection, by defining the class of interest to be an object class, and the null hypothesis as a collection of other object classes [15]. In the biological vision literature, this is commonly referred as *top-down saliency*, due to the requirement of feedback from high-level cortical areas for the specification of object classes. On the other hand, the principle can be equally applied to the solution of *bottom-up* saliency, which is pre-attentive and purely stimulus driven. This is implemented by defining the classification problem as one of discrimination between the visual stimulus contained in a pair of *center* (class of interest) and *surround* (null hypothesis) windows, at every location of the visual field [17]. For computer vision, this type of saliency is of interest for the solution of problems such as background subtraction, where the goal is to identify *any* object that does not belong to the background. It has been shown that discriminant saliency can be mapped into a biologically plausible neural architecture, which replicates both the computations of the standard neurophysiological model of area V1 of the brain and a large body of psychophysics of human saliency [17].

Assuming that the initial location of a target object is known, the tracking problem reduces to two of the three questions listed above, namely, feature selection and target detection. Since this assumption underlies all current implementations of discriminant tracking [7, 3, 18], we start by discussing how top-down discriminant saliency can be used to solve these two problems, in the remainder of this section. Later, in Section 3, we show how bottom-up discriminant saliency can be used for automatic tracker initialization.

2.1. Discriminant Saliency

Let $\mathcal{V}$ be a d dimensional visual stimulus ($d = 3$ for grayscale, $d = 4$ for color video) and let l indicate the initial position of the target. Two windows are defined around this location: a *target window* $\mathcal{W}_l^1$ containing the target, and a surrounding annular window $\mathcal{W}_l^0$ containing *background*. A classification problem opposing the two classes, target class with label $C(l) = 1$ and background class with label $C(l) = 0$, is posed at location l. A set of features $\mathbf{Y}$ from a predefined feature space $\mathcal{Y}$ (e.g. raw pixel values, Gabor, DCT, wavelet, or SIFT features), are computed for each of the windows $\mathcal{W}_l^i, i \in \{0, 1\}$. Features extracted from the target window are assumed to be drawn with probability density $p_{\mathbf{Y}|C(l)}(\mathbf{y}|1)$ and those from the background window with probability density $p_{\mathbf{Y}|C(l)}(\mathbf{y}|0)$.

The *saliency* of location l, $S(l)$, is defined as the extent to which the features $\mathbf{Y}$ can discriminate between the two classes. This is quantified by the mutual information between feature responses, $\mathbf{Y}$, and class label, C,

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

and can be shown to approximate the expected probability of correct classification of the optimal target/background classifier. More precisely, the mutual information of (1) is an approximation to one minus the Bayes error rate

$$L^* = 1 - E_{\mathbf{y}}[\max_i P_{C|\mathbf{Y}}(i|\mathbf{y})], \tag{2}$$

of the classification problem, where $E_{\mathbf{y}}$ denotes expectation with respect to $P_{\mathbf{Y}}(\mathbf{y})$ [31]. The mutual information can also be written as

$$S(l) = \sum_{c=0}^{1} p_{C(l)}(i) KL[p_{\mathbf{Y}|C(l)}(\mathbf{y}|i)||p_{\mathbf{Y}}(\mathbf{y})] \tag{3}$$

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. Learning Salient Features

The connection between discriminant saliency and the Bayes error rate for target/background classification, leads to a very natural criteria for salient feature selection: the features that enable optimal discrimination between target and background are those of largest mutual information with the class label. These *salient features* can be seen as either the most informative features for the target/background classification, or as the feature set of (approximately) lowest Bayes error rate for this classification.

Discriminant salient feature selection can also be performed efficiently. Let the feature space $\mathcal{Y}$ have dimension N and denote by $\mathbf{Y} = (Y_1, \ldots, Y_N)$ the random process from which all vectors of feature responses are drawn. Defining $\mathbf{Y}_{1,k} = (Y_1, \ldots, Y_k)$, the mutual information of (1) can be expanded ([30]) into

$$I(\mathbf{Y}; C) = \sum_k I(Y_k; C) + \sum_k [I(Y_k; \mathbf{Y}_{1,k-1}|C) - I(Y_k; \mathbf{Y}_{1,k-1})] \tag{4}$$

where

$$I(\mathbf{Y}; C|\mathbf{Z}) = \sum_i \int P_{\mathbf{Y},C,\mathbf{Z}}(\mathbf{y}, i, \mathbf{z}) \log \frac{P_{\mathbf{Y},C|\mathbf{Z}}(\mathbf{y}, i|\mathbf{z})}{p_{\mathbf{Y}|\mathbf{Z}}(\mathbf{y}|\mathbf{z}) p_{C|\mathbf{Z}}(i|\mathbf{z})} d\mathbf{y} d\mathbf{z} \tag{5}$$

is the conditional mutual information between $\mathbf{Y}$ and C given the observation of $\mathbf{Z}$. In (4), the term $I(Y_k; C)$ represents the discriminant power of the k^{th} feature individually, and is denoted its *marginal diversity* (MD). The terms $I(Y_k; \mathbf{Y}_{1,k-1}|C) - I(Y_k; \mathbf{Y}_{1,k-1})$ quantify the discriminant information contained in feature dependencies between the k^{th} feature and the set of $k-1$ previously selected features [30]. This decomposition allows as 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 [4, 19]. This implies that the dependencies between features carry little information about the class from which the features are extracted, allowing the approximation of (4) by

$$I(\mathbf{Y}; C) \approx \sum_{k=1}^{N} I(Y_k; C) = \sum_k \sum_i P_C(i) KL\left[P_{Y_k|C}(y|i)||P_{Y_k}(y)\right] \tag{6}$$

Note that this approximation does not require the assumption of feature independence, it simply follows from the constancy of feature dependences across natural image classes. The approximation is studied in detail in [30].

Since the mutual information is always non-negative, it follows that the selection of the optimal subset of K ($K < N$) salient features has very little complexity [31]. It consists of 1) ordering the N features by decreasing $I(Y_k, C)$, and 2) selecting the first K. This procedure is denoted as feature selection by maximum marginal diversity (MMD) in [31]. The terms in the right hand side of (6) only require marginal density estimates. In this work, we adopt a feature set composed of 8×8 DCT features at multiple scales. The fact that the DCT features belong to the set of bandpass features (as would Gabor coefficients, wavelet features, or image derivatives) makes these marginal density estimates extremely simple to compute.

2.3. Efficient Computation of the MD

The probability distribution of feature responses of a bandpass feature, to natural images, is well known to follow a generalized Gaussian distribution (GGD) [19]

$$P_Y(y; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|y|}{\alpha}\right)^{\beta}\right\}, \tag{7}$$

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 parameter β controls the rate of decay from the peak value, and defines a sub-family of the GGD (e.g. Laplacian when $\beta = 1$ or Gaussian when $\beta = 2$). The GGD parameters can be estimated from a sample of feature responses by the method of moments [26], using

$$\sigma^2 = \frac{\alpha^2 \Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})} \quad \text{and} \quad \kappa = \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})}, \tag{8}$$

where σ^2 and κ are, respectively, the variance and kurtosis of Y

$$\sigma^2 = E_Y[(Y - E_Y[Y])^2], \quad \text{and} \quad \kappa = \frac{E_Y[(Y - E_Y[Y])^4]}{\sigma^4}.$$

Furthermore, when the class-conditional densities $P_{Y|C}(y|i)$ and the marginal $P_Y(y)$ are GGDs $P_Y(y; \alpha_i, \beta_i)$

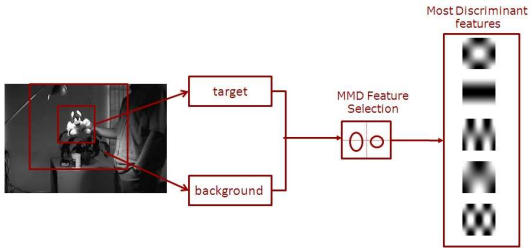

Figure 1. MMD feature selection for target/background discrimination. Feature responses are computed at the target location, from the center (target) and surround (background) windows. Features are ordered by their MD, and the most discriminant are selected.

and $P_Y(y;\alpha,\beta)$ respectively, the KL divergences of (6) have closed form [10]

$$KL[P_Y(y;\alpha_i,\beta_i)||P_Y(y;\alpha,\beta)] \quad (9)$$
$$= \log\left(\frac{\beta_i\alpha\Gamma(1/\beta)}{\beta\alpha_i\Gamma(1/\beta_i)}\right) + \left(\frac{\alpha_i}{\alpha}\right)^{\beta}\frac{\Gamma((\beta+1)/\beta_i)}{\Gamma(1/\beta_i)} - \frac{1}{\beta_i}.$$

These properties enable an extremely efficient computation of the marginal diversity of (6). The maximum MD (MMD) feature selection procedure is illustrated in Figure 1.

2.4. Target tracking by saliency detection

Once the salient features that best discriminate target from background at time step t have been computed, the goal is to identify the target location at time $t+1$. This reduces to detecting the locations of feature response y that can be most confindently be assigned to the target class in video frame $t+1$, given the discriminant features selected at time t. Under the information theoretic definition of discriminantion, classification confidence is measured by

$$I(C;Y=y) = \sum_{i=0}^{1} p_{C|Y}(y|i)\log\frac{p_{Y,C}(y,i)}{p_Y(y)p_C(i)},$$

Given the response Y_k of the k^{th} feature to the frame at time $t+1$, this results in the saliency measure

$$S_k(y) = \begin{cases} I(C;Y_k=y_k) & \text{if } y_k \in \mathbf{S}_k \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

with

$$\mathbf{S}_k = \left\{ y \left| \frac{P_{C,Y_k}(1,y_k)}{P_C(1)P_{Y_k}(y_k)} > \frac{P_{C,Y_k}(0,y_k)}{P_C(0)P_{Y_k}(y_k)} \right. \right\}. \quad (11)$$

$\mathbf{S}_k$ contains the set of points that are classified as belonging to the target class (C = 1) by the likelihood ratio test $P_{Y_k|C}(y_k|1)/P_{Y_k|C}(y_k|0) > 1$ and $I(C;Y_k=y_k)$ encodes the *confidence* of the classification, according to the k^{th} feature. Points such that the likelihood under the target hypothesis is much larger than that under the background hypothesis are very informative for target detection, and have large saliency.

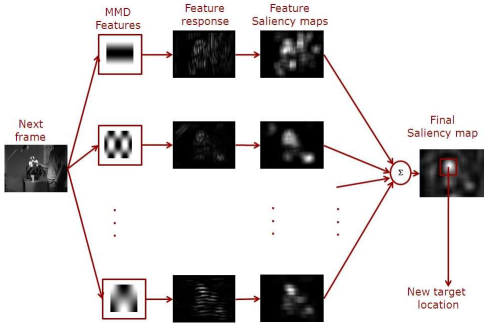

Figure 2. Target localization in the next frame. For each selected feature, a top-down saliency map is computed with (14). These saliency maps are combined to produce the overall saliency map, the maximum of which is taken to be the new location of the target.

For GGD features this saliency measure can be computed very efficiently, using the fact that [17]

$$I(C;Y=y) = s[g(y)]\log\frac{s[g(y)]}{\pi_1} + s[-g(y)]\log\frac{s[-g(y)]}{\pi_0}, \quad (12)$$

where $s(y) = (1+e^{-y})^{-1}$ is a sigmoid function, $\pi_i = P_C(i)$ is the prior for class i, and

$$g(y) = \left(\frac{|y|}{\alpha_0}\right)^{\beta_0} - \left(\frac{|y|}{\alpha_1}\right)^{\beta_1} + T, \quad (13)$$

with $T = \log\frac{\alpha_0\beta_1\pi_1\Gamma(1/\beta_0)}{\alpha_1\beta_0\pi_0\Gamma(1/\beta_1)}$. The total confidence measure for the set of K feature responses $\mathbf{y}$ is

$$S_T(\mathbf{y}) = \sum_{k=1}^{K} S_k(y_k). \quad (14)$$

The computation of this measure is illustrated in Figure 2.

The salient features selected at time t can be seen as matched filters for the detection of the salient visual attributes of the target, according to the appearance of the latter at that time. This follows from the fact that $\mathbf{S}_k$ is a set of the form

$$\mathbf{S} = \left\{ y \left| P_{Y_k|C}(y|1) > P_{Y_k|C}(y|0) \right. \right\}, \quad (15)$$

and, for GGD features, this reduces to $\mathbf{S}_k = \{y_k \,||y_k| > t_k\}$, where t_k is a threshold that depends on the parameters of the two GGDs. Hence, only regions of large magnitude feature response are considered salient. This implies that the features are matched to the visual stimuli considered salient and pertain to the target class. The location of largest saliency at time $t+1$ is selected as the new position of the target. Feature selection is then repeated from target and background windows centered at this location, to learn the appearance model at time $t+1$. The resulting features are then used for saliency detection at time $t+2$ and the procedure is iterated. The entire tracking algorithm is summarized in Algorithm 1.

Algorithm 1 Tracking Using Discriminant Saliency

Input: Current target location l, $t = 0$, initial frame $\mathcal{I}_0$ containing M pixel locations
while Next frame exists **do**
  Set $t = t + 1$.
  Feature Selection: Given a set of DCT features $Y_k, k \in \{1, \ldots, N\}$,target location l from the null hypothesis, and a target number of features K.
  Obtain target patch $\mathcal{W}_l^1$ and surround $\mathcal{W}_l^0$ from $\mathcal{I}_{t-1}$.
  for $k = \{1, \ldots, N\}$ **do**
    Estimate GGD parameters of $P_{Y_k|C}(y|i)$, from responses of Y_k to $\mathcal{W}_l^i, i \in \{0, 1\}$, using (8).
    Estimate GGD parameters of $P_{Y_k}(y)$, from responses of Y_k to $\mathcal{W}_l^1 \cup \mathcal{W}_l^0$, using (8).
    Compute $I(Y_k, C)$, using (6) and (9).
  end for
  Output: return the K features of largest $I(Y_k, C)$.
  Target detection in new frame: Given the frame $\mathcal{I}_t$, a set of K discriminant features Y_k for the target class, and the GGD parameters of $P_{Y_k|C}(y|i), i \in \{0, 1\}$, and $P_{Y_k}(y)$.
  for $k = \{1, \ldots, K\}$ **do**
    for $m = \{1, \ldots, M\}$ **do**
      Compute the response y_m of Y_k at location of pixel l_m of $\mathcal{I}_t$, and $P_{Y_k|C}(y_m|i), i \in \{0, 1\}$, using (7).
      Compute $S_k(y_m)$, using (10)
    end for
  end for
  for $m = \{1, \ldots, M\}$ **do**
    Compute total confidence measure $S_T(\mathbf{y}_m)$ at l_m with (14).
  end for
  Output: Set l =argmax$_{l_m}$ $S_T(\mathbf{y}_m)$.
end while

3. Automatic tracker initialization

Most tracking algorithms assume that the initial target position l and a bounding box are manually provided [7,3]. This is frequently not practical in real applications, where manual supervision is expensive or unavailable. While many ad-hoc initialization strategies, such as background subtraction, and blob or motion detection, have been proposed [7] most of these have limited scope. For example, they tend to fail when the background is itself dynamic, as is the case of many natural scenes [23]. A more principled approach, based on bootstrapping a weak and generic target model for automatic initialization, has been proposed by Toyama and Ying [29]. It, however, requires a target model to begin with, and some degree of supervision to adapt to different scenes.

Under the discriminant saliency principle, there is no fundamental difference between tracker initialization and the tracking operation itself. The only difference is that, while the latter is a top-down saliency procedure, the former is a problem of bottom-up saliency. In fact, it has been shown that bottom-up discriminant saliency with suitable models for spatiotemporal stimulus statistics is a state of the art solution for the problem of (unsupervised) background subtraction [23].

To compute saliency in the unsupervised or bottom-up

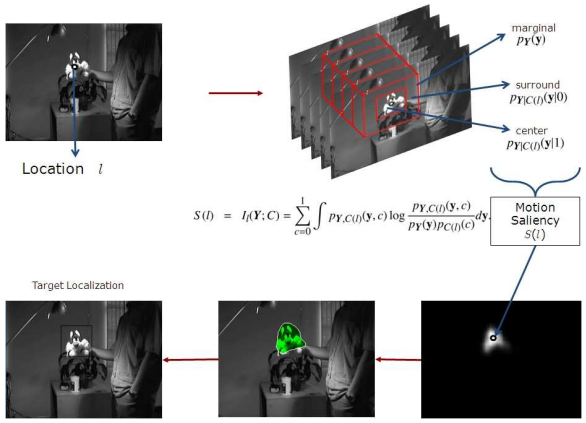

Figure 3. Illustration of automatic target identification for initializing the tracker. By using spatiotemporal features (e.g.optical flow or dynamic textures) to represent center and surround windows, (3) is used compute the saliency of every location l. The saliency measure $S(l)$ is highest for regions containing salient moving objects. By finding the location of highest motion saliency, the initial position and scale of the target can be estimated.

mode, a classification problem is posed at every location l of the visual field, between a *center* window $\mathcal{W}_l^1$ around l, and a *surround* annular window $\mathcal{W}_l^0$. The union of the two windows is denoted the *total* window, $\mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$. Through the inclusion of spatiotemporal features in the feature space, this classification problem can identify locations which are most different from their surround, *in terms of both spatial and temporal stimulus statistics*. The regions of highest saliency can then be associated with a potential target.

The requisite spatiotemporal features can be selected based on the nature of the scene. For scenes shot with static cameras, optical flow features can be used [14]. In this case, the features y used in the saliency formulation (3) are the magnitude and direction of optical flow vectors. For more complex backgrounds, following [23], we use the dynamic texture (DT) model of [11] as the probability model $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|c)$ for the spatiotemporal stimuli $\mathbf{y}(\tau)$ in (3). DT parameters are learned from center, surround, and total windows, to obtain the densities $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|1)$, $p_{\boldsymbol{Y}|C(l)}(\mathbf{y}(\tau)|0)$, and $p_{\boldsymbol{Y}}(\mathbf{y}(\tau))$, respectively. $S(l)$ is finally computed with (3). This procedure is illustrated in Figure 3. Further details are available in [23], where the procedure is shown to have great robustness to complex background dynamics, and camera motion. In our experience, the computation of spatiotemporal saliency, from the initial frames of a video sequence, is a robust automatic procedure to identify the *moving targets* of typical interest for surveillance and monitoring applications. The locations of these targets are then used to initialize the discriminant tracker described in Section 2.

4. Experiments and Results

To validate the proposed algorithm, we performed two types of experiments - the first set comparing the discriminant saliency tracker (DST) with other tracking approaches when the target location is known and 2) automatic initialization and tracking on video clips without any prior knowledge.

Comparison to Existing Trackers: We compared the performance of DST with three other trackers : two discriminant trackers, (the method of Collins et al. [7], and the ensemble tracker [3]), and the incremental visual tracker (IVT) [25], a representative of the state of the art in appearance based tracking.

The test clips for tracking were selected from diverse sources (e.g previous works, standard database, and from the web). All clips include challenging situations such as varying illumination, complete object rotation and change in perspective. For instance, the "motinas_toni_change_ill" of [22] shows a person turning around 360° in extremely low light (Figure 4(a)), while the "gravel" clip has perspective distortion induced by the person moving away from the camera (Figure 5). Since the test clips are grayscale, we implemented a version of the Collins tracker that uses DCT features instead of the R,G,B color features proposed in the original publication [7]. All four algorithms were initialized with target location and bounding box in the first frame. The background bounding box was assumed to have an edge 3 times larger than the corresponding edge of the target box. Each training image from target or background was decomposed using a two-level Gaussian pyramid and 8×8 DCT features computed at each location (for a total of $N = 64 \times 2 = 128$ features). The number of MMD features selected for each frame was set to $K = 5$. To enforce temporal coherence, the discriminant features were learned using the target appearance of the current frame and 2 past frames, and tracking was performed using the method of Algorithm 1.

The results of tracking on three of the clips tested are shown in Figures 4 and 5. For these clips, the qualitative performance of IVT and the Collins tracker is extremely poor and they fail to track the target in all three scenes. The ensemble tracker fails to track the object when it undergoes extreme appearance variation due to illumination changes or target rotation (e.g."motinas_toni_change_ill" in Figure 4(a), "karlsruhe" in Figure 4(b)), while DST tracks the targets successfully in all the clips. For each clip, a quantitative estimate of tracking error was also obtained using groundtruth data. Tracking error was defined as the average pixel difference, between the groundtruth bounding box and the bounding box obtained by the tracker. The results for the three clips are tabulated in Table 1. DST clearly outperforms all other trackers. The videos (and larger pictures) of all results are available from the attached supplementary [2].

Figure 4. Results of tracking on a)"motinas_toni_change_ill" [22] - the person is turning around and the illumination changes drastically b) "karlsruhe" [1] - the car makes a U-turn. The target locations obtained by the four methods on four frames are shown : DST - thick red (pale gray when not in color) box, Collins - thick black box, ensemble - white dashed box, IVT black dashed box.

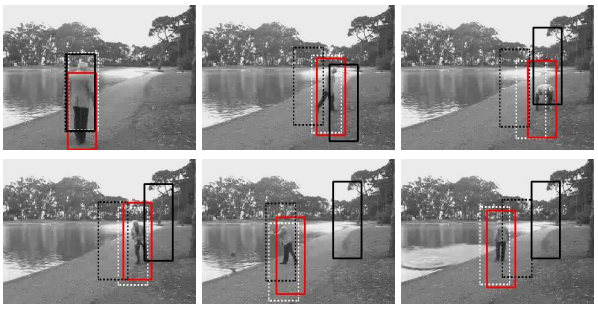

Figure 5. Results of tracking on "gravel". The target locations obtained by the four methods on six frames are shown : DST - thick red (pale gray when not in color) box, Collins - thick black box, ensemble - white dashed box, IVT black dashed box.

Results for Automatic Initialization and Tracking: The result of tracking using automatic initialization for a static camera scene is shown in Figure 6(a). From the initial frames of the clip, a motion saliency map is generated using optical flow features, and the regions of maximal motion saliency are identified as potential targets. These are input to the DST algorithm, which then tracks the targets through the remaining frames.

For scenes with extremely dynamic backgrounds, a dynamic texture based motion saliency algorithm is used. Figure 6(b) shows the motion saliency map obtained using this procedure for a surfing scene, and a few of the subsequently tracked frames. These results demonstrate the ability of the discriminant saliency framework to perform robust target initialization even for scenes with extremely dynamic backgrounds. The video results are available in the attached supplement [2].

5. Connections to other discriminant trackers

At an abstract level, the proposed discriminant saliency tracker is similar to the previously proposed discriminant trackers [7, 3]. In this section we provide an analysis of these two trackers, to highlight the connections, and show

Clip Name	IVT	Collins	Ensemble	DST
motinas_toni_change_ill	× 0.70	× 0.71	× 0.52	✓0.15
karlsruhe	× 0.29	× 0.34	× 0.56	✓0.04
gravel	× 0.62	× 0.71	✓0.10	✓0.03

Table 1. Performance comparison of four tracking algorithms on three clips. In addition to the average tracking error for each method, a '×' (loses track) or ' ✓' (maintains track) is shown to indicate tracking continuity as observed visually.

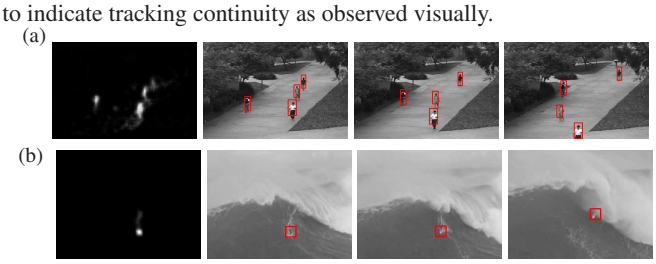

Figure 6. Results of automatic initialization and tracking on a) "pedestrians". The motion saliency map obtained using the bottom-up formulation with optical flow features is shown on the extreme left. b) "wave", the motion saliency map computed using dynamic textures is shown on the extreme left. Positions of the targets for three frames are shown in red boxes.

that each stage - center-surround training, feature extraction, goodness of discrimination to select features and finally, goodness of fit to identify target locations, is equivalent to the corresponding stage of the proposed tracker.

Discriminant trackers in the literature are effectively center-surround operations, defining target as the center and background as the surround. This architecture is a direct analogue of the discriminant saliency framework. Further, selecting the best features and detecting the target in the next frame are also closely related to their counterparts in the discriminant saliency framework as discussed below.

5.1. Feature Selection and Discriminability

All discriminant trackers include a feature selection procedure. Features can be considered as a transformation of the observation space to (a presumably lower dimensional) feature space $\mathcal{Y}$,where discriminating the target from the background is easier than in the original space. The transformation can be linear or non-linear. For instance, Collins et al. [7] use linear combinations of R, G, B pixel values as the features. "Ensemble tracking" [3] uses a set of non-linear features composed of histograms of oriented gradients [9] along with the R,G,B pixel values. In the proposed approach, we use DCT features.

In all three methods, the set of features is analyzed in terms of its discriminability for the classification task - separating the target from the background. Collins et al. [7] first compute histograms of filter responses applied to the R,G,B color channels of both target and background, and construct a log likelihood ratio between the two class histograms, considering this as a new non-linear feature. The feature discriminability is 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 non-linear space and learning a linear classifier that minimizes the classification error in the feature space, under the assumption that the new feature has a *Gaussian* distribution.

In [3], a set ("ensemble") of weak hyperplane classifiers are trained to separate target from background in the feature space. However, each classifier is obtained after weighting the points by a diagonal matrix of weights. This corresponds to a linear transformation and each re-weighting is equivalent to creating a new feature. The discriminability in this case is directly equal to the error rate of classification.

Hence, while the features themselves might be different, all approaches to discriminant tracking use *discriminability based on the minimum probability of error criterion to select the best features*, albeit under different assumptions.

5.2. Target Detection as a goodness of fit

The next step involves using the selected features to perform target detection in a new frame. In [7], the confidence measure used to classify points in the next frame is simply the log-likelihood ratio between the probability of target and the probability of background as learned from the current frame. This acts as a matched filter and finds regions that best correspond to the probabilistic description of the target, while corresponding least to that of the background. This definition of the confidence measure is similar to saliency measure of (14).

In ensemble tracking, the selected features are a set of weak classifiers and the confidence measure for locations in the new frame is simply a weighted combination of the (normalized) classification margin at that location. The margin represents the level of belief in the classification result, and is directly analogous to saliency of (6). [1]

In summary, both ensemble tracking and the Collins tracker, are fundamentally similar to the proposed discriminant tracker. However, the formulation of discriminant tracking as a center-surround saliency problem has several merits over other discriminant trackers. This is discussed below.

5.3. Merits of the Discriminant Saliency Tracker Over Other Discriminant Methods

The Collins tracker uses a heuristic discriminability measure similar to a Fisher's discriminant. While this measure has been empirically shown to work for color features, it lacks a generic principled justification. Furthermore, the distribution of log-likelihood ratios is hard to characterize [6]. The assumption of unimodality also does not hold in general (i.e. for all features), and is especially troubling

[1] As the weak classifier used is a hyperplane classifier, the probability of correct classification, and hence the mutual information, are related to the margin using the error function (erf).

when there is background clutter. This partially accounts for the results above where, for the DCT features used, the Collins tracker performed as poorly as the IVT. In addition, the use of histogram based features is computationally inefficient, and the procedure cannot be extended to include spatiotemporal features, such as dynamic textures for motion assisted tracking.

In ensemble tracking, the selected weak classifiers are combined using AdaBoost. This could be a disadvantage, in the tracking context, for two reasons : a) boosting is computationally expensive, and b) it tends to overfit the limited training data available. In result, the tracker does not perform well when there are large variations of appearance, such as the rotating objects of Figures 4 and 5. On the other hand, MMD-based feature selection is computationally efficient and, as seen from the results above, seems to achieve a better trade-off between classification accuracy and generalization. A similar outcome has been reported for image classification, where a mutual information based feature selection procedure been shown to outperform boosting based methods [12].

6. Conclusion

In this work, we have shown that discriminant tracking can be framed as a saliency problem, and solved using biologically inspired computational principles. The resulting framework provides a principled unifying methodology to perform all three tasks involved in tracking: initialization, feature selection and target detection. Experimental results show that tracking using the DST is robust and accurate, outperforming previous state-of-the-art trackers. Being unsupervised and invariant to egomotion, DST could be used for applications such as automated surveillance and monitoring from moving cameras.

References

[1] http://i21www.ira.uka.de/image_sequences.

[2] See attached supplementary material.

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

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

[5] 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.

[6] H. Chernoff. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25(3):573–578, 1954.

[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. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(5):564–577, 2003.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, June 2005.

[10] M. N. Do and M. Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE Transactions on Image Processing*, 11(2):146–158, 2002.

[11] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, 2003.

[12] F. Fleuret and I. Guyon. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

[13] L. M. Fuentes and S. A. Velastin. From tracking to advanced surveillance. In *ICIP (3)*, pages 121–124, 2003.

[14] 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.

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

[16] D. Gao and N. Vasconcelos. Discriminant interest points are stable. In *CVPR*, June 2007.

[17] D. Gao and N. Vasconcelos. Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 2008.

[18] H. Grabner and H. Bischof. On-line boosting and vision. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 1:260–267, 2006.

[19] J. Huang and D. Mumford. Statistics of Natural Images and Models. In *Computer Vision and Pattern Recognition*, pages 541–547, 1999.

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

[21] 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.

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

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

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

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

[26] 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.

[27] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE PAMI*, 22(8):747–757, 2000.

[28] K. Toyama and G. D. Hager. Incremental focus of attention for robust visual tracking. In *IJCV*, pages 189–195, 1996.

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

[30] M. Vasconcelos and N. Vasconcelos. Natural image statistics and low complexity feature selection. *IEEE PAMI*, In press.

[31] N. Vasconcelos. Feature selection by maximum marginal diversity. In *NIPS*, 2002.

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

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