Discriminative Spatial Attention for Robust Tracking

Jialue Fan[1], Ying Wu[1], and Shengyang Dai[2],[*]

[1] Northwestern University
2145 Sheridan Road, Evanston, IL 60208
[2] Sony US Research Center
1730 N. 1st St, San Jose, CA 95112

Abstract. A major reason leading to tracking failure is the spatial distractions that exhibit similar visual appearances as the target, because they also generate good matches to the target and thus distract the tracker. It is in general very difficult to handle this situation. In a selective attention tracking paradigm, this paper advocates a new approach of discriminative spatial attention that identifies some special regions on the target, called *attentional regions* (ARs). The ARs show strong discriminative power in their discriminative domains where they do not observe similar things. This paper presents an efficient two-stage method that divides the discriminative domain into a local and a semi-local one. In the local domain, the visual appearance of an attentional region is locally linearized and its discriminative power is closely related to the property of the associated linear manifold, so that a gradient-based search is designed to locate the set of local ARs. Based on that, the set of semi-local ARs are identified through an efficient branch-and-bound procedure that guarantees the optimality. Extensive experiments show that such discriminative spatial attention leads to superior performances in many challenging target tracking tasks.

1 Introduction

Our computer vision research on target tracking always aims to develop methods that can work as good as the human. Large research efforts have been devoted to region-based tracking and have produced many outstanding methods, e.g., the mean-shift tracker [17], the kernel-based tracker [3], and the ensemble tracker [19], etc. The major research has been largely focused on effective image region matching to handle large variations in images, and efficient search to locate the target. However, many real applications in video analysis always demand trackers that are more robust and can perform for a longer duration.

Among many reasons that lead to tracking failure, one of the most difficult cases is due to the distractions in the environment that present similar visual appearances as the target and thus exhibiting good matching to the target. These

---

[*] This work was conducted while the third author was a graduate student at Northwestern University.

distractions can be from the background clutter or from similar objects in the scene. As the distractions produce false positives in target detection, they lead to wrong association to the tracker, and thus fail the tracker. Because they do give good matches to the target, it is difficult to detect such a distraction failure promptly based on their matching scores.

It is known that our human dynamic visual perception is selective [18], which allows the processing in our visual system to be concentrated on relevant and important visual information. The selection occurs in all stages in visual processing, and it can be based on both innate principles as well as learned heuristics. It is the visual selection that makes our visual system efficient and adaptive in following moving targets. Among many possible kinds of visual selections, spatial attention focuses the computation on some selected local image regions on the target, called *Attentional Regions* or ARs. Tracking the target is fulfilled by the tracking of these ARs. This mechanism appears to be a key in handling clutters, distractions and occlusions in target tracking.

To introduce spatial attention to the design of tracking algorithms, in addition to the matching and searching of ARs, the selection of ARs is a critical issue for persistent tracking. We often observe an interesting phenomenon in various region-based tracking methods that the initialization of the target region may largely influence the tracking performance. A slightly different initialization of the target region sometimes ends up with a much better or worse result. Unfortunately, this phenomenon has not received much attention in the literature, although it conveys a strong message that the selection of ARs cannot be arbitrary. This paper is concerned on finding ARs on the target so as to achieve more robust and persistent tracking.

More specifically, an AR is a local image region that has the largest discriminative power among others in its spatial domain. This spatial selection task is not trivial. For a given target, the number of its candidate attentional regions (i.e., any sub image region on the target) are enormous. Although we can examine all ARs in a brute-force way, we cannot afford its $O(n^2)$ complexity in practice because n (i.e., the number of candidates) is huge, and thus a more efficient method is desirable.

This paper presents a novel and efficient solution to the spatial selection of discriminative attentional regions. In the feature space, the feature of an AR has a large *margin* to its nearest neighbors, and we can use this margin in the feature space to represent the discriminative power of an AR. The larger the margin, the more distinctive an AR is in its spatial domain. An AR needs to be distinctive in both its small spatial neighborhood (i.e., local) and a larger domain (i.e., semi-local) that is determined by the possible motion of this attentional region. In the local domain, the local neighbors of an attentional region approximately span a local linear manifold, so that we recast the discriminative power to be a condition number measure of this local linear manifold, and design an efficient gradient-based search for all local ARs. In the semi-local domain, as the approximation does not hold, we design an effective branch-and-bound search that largely reduces the complexity while achieving the optimality. Our extensive

experiments show that the selected discriminative attentional regions are more resilient to distractions and lead to robust tracking.

The novelty of this work includes the following four aspects. (1) Because most existing tracking methods focus on matching but spatial distractions also exhibit good matches, these methods are challenged. This paper explicitly handles the distractions by discovering attentional regions that are resilient to distractions. (2) The proposed approach to locating ARs considers both local and semi-local distractions. This new approach leads to an efficient solution that integrates a gradient-based search and a branch-and-bound search. (3) Based on the spatial selection, this paper presents a new robust tracking algorithm that uses multiple ARs and is adaptive to the appearance changes of the target and the dynamic scene.

2   Related Work

In this section, we briefly review recent approaches related to our work. Region-based tracking has been studied in [17,3,5,7,8,13]. In [7], the spatial configuration of the regions is done by optimizing the parameters of a set of regions for a given class of objects. However, this optimization needs to be done off-line. In [8], a method for a well known local maximally stable extremal region (MSER) has been proposed. As the backward tracking is integrated, it restricts its application to off-line tracking.

There is a vast literature on salient region selection [10,15,4,11,12,6,1,2]. In these works, spatial selection expects the regions to be located at corner-like points. They emphasize the repeatability of the regions in matching. The repeatability of the regions is related to the local discrimination introduced in this paper. But this paper goes one step further. Beside the local discrimination, this paper also studies the semi-local case.

It is worth mentioning that the proposed AR selection mechanism is different from the feature selection paradigms [9]. Feature selection aims to choose global features that best discriminate the object from the background. The target is treated as a whole in those approaches. While in the proposed method, the target is represented by a set of spatial attentional regions. Such a difference in modeling leads to the difference in the selection. In feature selection methods, discriminative features are selected to separate the target and the background, but the AR selection chooses local distinctive image sub-regions (rather than the features). Since the spatial distracters exhibit similar visual appearances as the target, choosing whatever features always results in similar feature vectors. Therefore, feature selection methods are limited in handling this case. On the contrary, the proposed spatial selection method pinpoints to the actual spatial distinctions, and thus is well able to cope with such spatial distracters.

The most closely related work to the proposed method may be [5]. In [5], a general framework of spatial selective attention was advocated for tracking. The early selection process extracts a pool of ARs that are defined as the salient image regions which have good localization properties, and the late selection process dynamically identifies a subset of discriminative attentional regions through

a discriminative learning on the historical data on the fly. However, this work is a large leap from [5], not only because this work presents a much more in-depth study of spatial selection, but also it makes the general selective attentional tracking framework more practical and more effective in practice. The main differences include: (1) The tracking method in [5] is a very specific implementation, and many components in this framework need further investigation and improvement. Moreover, it selects the ARs that are only local discriminative, and it is quite limited in handling the semi-local distraction which is much more common and more challenging in practice. On the contrary, the proposed method selects the ARs that are both local and semi-local discriminative. (2) We explicitly define discriminative margin, which is a new concept, and consider the local discriminative and semi-local discriminative in a unified way. On the contrary, the late selection in [5] is not as principled as the proposed approach.

3   Attentional Region (AR)

3.1   Spatial Discrimination

An attentional region (or AR) is a local image region which has the largest discriminative power among others in its spatial domain. At the first step, we need to define a general discriminative measure.

Given a region $R(\mathbf{x})$ located at position $\mathbf{x}$ in an image, we denote the set of its neighboring regions by $\{R(\mathbf{y}), \mathbf{y} \in \mathcal{N}(\mathbf{x})\}$, where $\mathcal{N}(\mathbf{x})$ is the spatial neighborhood of $\mathbf{x}$, and we call it the *discriminative domain*. The visual features of $R(\mathbf{x})$ is represented by the feature vector $\mathbf{f}(\mathbf{x})$. Denote by $D(\cdot, \cdot)$ the metric to measure the difference of two feature vectors. Then we define the general discriminative score $\rho(\mathbf{x})$ of the AR $R(\mathbf{x})$ by:

$$\rho(\mathbf{x}) \triangleq \min_{\mathbf{y} \in \mathcal{N}(\mathbf{x})} D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y})). \tag{1}$$

It is clear that the larger the $\rho(\mathbf{x})$ is, the more discriminative the AR $R(\mathbf{x})$ is from its neighbors. If $\rho(\mathbf{x}) = 0$, i.e., there is a perfect match in the neighborhood, then this AR has no discriminative power.

However, in practice, we recognize the fact that the most similar one is very likely to be located in a very close vicinity $\mathcal{L}(\mathbf{x})$, i.e., $\min_{\mathbf{y} \in \mathcal{N}(\mathbf{x})} D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y}))$ is very likely equal to $\min_{\mathbf{y} \in \mathcal{L}(\mathbf{x})} D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y}))$. Then this discriminative score can only reflect the local discrimination. To characterize the semi-local discrimination, we should exclude $\mathcal{L}(\mathbf{x})$ when we define the discriminative score. Let $\mathcal{S}(\mathbf{x}) = \mathcal{N}(\mathbf{x}) \backslash \mathcal{L}(\mathbf{x})$. So in practice, we define the discriminative scores $\rho_S(\mathbf{x})$ and $\rho_L(\mathbf{x})$ for semi-local and local domains, respectively:

$$\rho_S(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{S}(\mathbf{x})} D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y})), \tag{2}$$

$$\rho_L(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{L}(\mathbf{x})} D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y})). \tag{3}$$
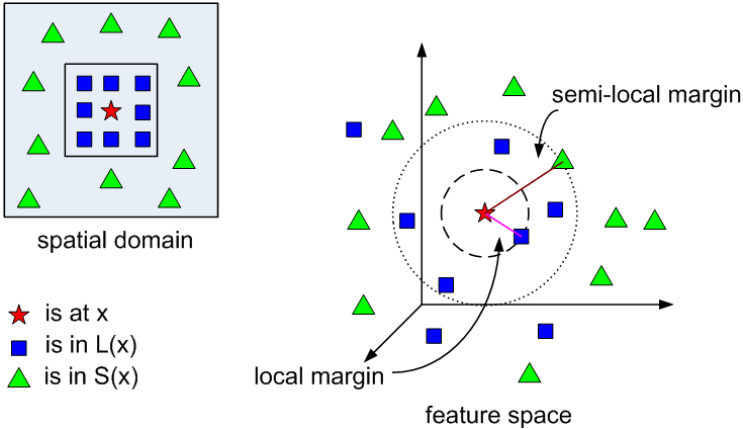
Fig. 1. The discriminative margins for a certain AR

Figure 1 illustrates this concept. In the spatial domain, the red star represents $\mathbf{x}$, the blue squares represent some $\mathbf{y} \in \mathcal{L}(\mathbf{x})$, and the green triangles represent some $\mathbf{y} \in \mathcal{S}(\mathbf{x})$. We also show them in the feature space where the distance between two points is determined by the distance measure $D(\cdot, \cdot)$. The hypersphere O_L is centered at $\mathbf{x}$ with a radius $\rho_L(\mathbf{x})$. Therefore, all the blue squares are out of the hypersphere, and there is at least one blue square on the boundary of the hypersphere. It is clear that the discriminative score $\rho_L(\mathbf{x})$ reflects the *margin* between the target and the set of its local neighbors in the feature space. The larger the $\rho_L(\mathbf{x})$ is, the more local discriminative the AR $R(\mathbf{x})$ is. Similarly, the hypersphere O_S is centered at $\mathbf{x}$ with the radius $\rho_S(\mathbf{x})$. The discriminative score $\rho_S(\mathbf{x})$ reflects the *margin* between the target and the set of its nearest semi-local neighbors in the feature space.

3.2   Attentional Region

An AR needs to be distinctive in both its local spatial neighborhood (i.e., the local domain) and a larger domain (i.e., the semi-local domain).

We denote the set of *local* ARs by $\mathcal{X}_L = \{\mathbf{x} : \rho_L(\mathbf{x}) > \epsilon_L\}$ where $\epsilon_L > 0$ is a threshold for the local domain. Similarly, denote the set of *semi-local* ARs by $\mathcal{X}_S = \{\mathbf{x} : \rho_S(\mathbf{x}) > \epsilon_S\}$. By definition, an AR needs to be discriminative at both local and semi-local domains. Therefore, the set of ARs $\mathcal{X} = \mathcal{X}_L \cap \mathcal{X}_S$.

The intuitive explanation of the difference between AR and a common region is shown in Fig. 2. In Fig. 2, three representative patches are chosen, and the matching scores between the selected patches and their neighbors are visualized.

As shown in Fig. 2, the matching error surfaces of the AR and the common regions behave quite differently: The region at the chin has a poor local discriminative power since its neighbors along the boundary looks quite similar. The region at the eye has a poor semi-local discriminative power, because there is a similar eye corner in the valid semi-local domain and it acts as the distractor.
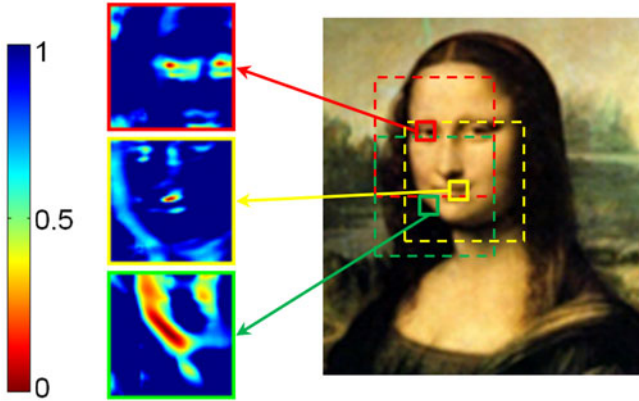
Fig. 2. Three regions and their matching error surfaces with their corresponding neighbor regions

Fig. 3. ARs are related to their associated discriminative domain. The leftmost is the local ARs. When $\mathcal{N}(\mathbf{x})$ becomes larger, there exists a less number of ARs. As shown in the rightmost, only three ARs survive in the largest range we specified. The positions are at the mouth and the joint part between the leg and the body of the zebra.

The region at the mouth has both strong semi-local and local discriminative power as good matches are only focused in a very small neighborhood. Traditional methods [10,4,11,12] may examine those local ARs but are unable to identify the semi-local ones, because they only consider the local properties.

Whether a region is discriminative or not is related to the range of the associated discriminative domain $\mathcal{N}(\mathbf{x})$. A region is an AR in a spatial domain if and only if there are no distractors (i.e., good matches) in this domain. When the domain becomes larger, some distractors may be present, and thus reduce the discriminative power of this region in the larger domain. If the discriminative power becomes below the threshold, this region is no longer an AR. Thus, when we keep enlarging the discriminative domain, we have fewer and fewer ARs. Figure 3 shows one example to illustrate this situation.

4   Spatial Selection of ARs

For a given target, denote the set of its candidate regions (i.e., any sub image region on the target) by Λ. The spatial selection task, i.e., finding the ARs in Λ, is not trivial. Comparing all regions with all of their neighbors in a brute-force

way is computationally infeasible, because the number of the candidate regions is huge, and thus a more efficient method is needed.

We propose a two-step method to find ARs: (1) we first obtain all local ARs $\mathcal{X}_L$ based on an efficient gradient-based search. (2) Then we select a subset of $\mathcal{X}_L$, whose element has strong semi-local discriminative power to be ARs through an efficient branch-and-bound search that guarantees the optimality.

4.1 Gradient-Based Search for Local ARs

For a region located at $\mathbf{x}$, assume $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^d$. The visual features of its local spatial neighbors constitute a linear manifold (up to two dimensional) at $\mathbf{f}(\mathbf{x})$ in the feature space. Assume $\Delta\mathbf{x} = [\Delta u, \Delta v]^T$, we have

$$\mathbf{f}(\mathbf{x} + \Delta\mathbf{x}) \approx \mathbf{f}(\mathbf{x}) + \mathbf{\Phi}\Delta\mathbf{x}, \tag{4}$$

where $\mathbf{\Phi} \triangleq [\frac{\partial \mathbf{f}}{\partial u} \frac{\partial \mathbf{f}}{\partial v}]$ is a $d \times 2$ matrix.

Using L2 metric for matching, the local discriminative margin $\rho_L(\mathbf{x})$ becomes:

$$\rho_L(\mathbf{x})^2 = \min_{\mathbf{x}+\Delta\mathbf{x}\in\mathcal{L}(\mathbf{x})} \|\mathbf{f}(\mathbf{x} + \Delta\mathbf{x}) - \mathbf{f}(\mathbf{x})\|^2 \approx \min_{\mathbf{x}+\Delta\mathbf{x}\in\mathcal{L}(\mathbf{x})} (\Delta\mathbf{x})^T \mathbf{A} \Delta\mathbf{x}, \tag{5}$$

where $\mathbf{A} \triangleq \mathbf{\Phi}^T\mathbf{\Phi}$ is a 2×2 matrix which characterizes this local linear manifold.

Case 1: $rank(\mathbf{A}) = 1$. It is clear that $\rho_L(\mathbf{x}) = 0$.

Case 2: $rank(\mathbf{A}) = 2$. The minimum is obtained at the inner boundary of $\mathcal{L}(\mathbf{x})$ due to the discretization of $\mathbf{x}$. Assume the inner boundary of $\mathcal{L}(\mathbf{x})$ to be $\|\Delta\mathbf{x}\| = 1$. Then we have

$$\rho_L(\mathbf{x})^2 = \min_{\|\Delta\mathbf{x}\|=1} (\Delta\mathbf{x})^T \mathbf{A} \Delta\mathbf{x}. \tag{6}$$

We perform SVD on $\mathbf{\Phi}$ and obtain two singular values σ_1 and σ_2. Without loss of generality, we assume $\sigma_1 \geq \sigma_2$. As $\mathbf{A} = \mathbf{\Phi}^T\mathbf{\Phi}$, σ_1^2 and σ_2^2 are the eigenvalues of $\mathbf{A}$.

We can easily see that $\rho_L(\mathbf{x})^2 = \sigma_2^2$. Therefore, maximizing the margin $\rho_L(\mathbf{x})$ is equivalent to maximizing σ_2. It is clear that when $det(\mathbf{A})$ becomes larger, $\rho_L(\mathbf{x})$ will become larger, and then the problem becomes meaningless. But considering the fact that $det(\mathbf{A})$ is bounded, i.e., $det(\mathbf{A}) \leq \chi^2$, and the fact that $det(\mathbf{A}) = (\sigma_1\sigma_2)^2$, we have

$$\sigma_2^2 = \chi\frac{\sigma_2^2}{\chi} \leq \chi\frac{\sigma_2^2}{\sigma_1\sigma_2} = \chi\frac{1}{\sigma_1/\sigma_2}. \tag{7}$$

It is clear that maximizing σ_2 amounts to minimizing the condition number σ_1/σ_2 of $\mathbf{\Phi}$.

The above analysis reveals the relation between the discriminative power of a region and the singularity property of its local linear manifold.

In practice, only obtaining the criterion for local AR placement is insufficient, since it is not attractive to exhaustively evaluate this criterion all over the image.

In [4], a gradient descent algorithm has been proposed to efficiently find good placement where the condition number of $\mathbf{\Phi^T\Phi}$ is locally minimized. We follow that algorithm in this paper. First we randomly initialize a set of AR candidates. Following the gradient of the condition number these ARs converge to their corresponding local minima. The set of local minima is $\mathcal{X}_L$.

The matrix $\mathbf{\Phi}$ depends on the choices of the feature space and the matching metric. In this paper we use the contextual flow [16] as the feature vector, because it is robust to small changes on local appearance that invalidate the constancy in brightness.[1]

4.2   Branch-and-Bound Selection of ARs from $\mathcal{X}_L$

Based on the set of local ARs $\mathcal{X}_L$ obtained in Sect. 4.1, we obtain ARs with a strong semi-local discriminative power from $\mathcal{X}_L$. We solve a more general and flexible problem as follows:

Given the set $\mathcal{X}_L = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ (i.e., $|\mathcal{X}_L| = N$), we want to choose the ARs $\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_M \in \mathcal{X}_L$ with the M largest discriminative score $\rho_S(\cdot)$.

Since the linear approximation is invalid in the semi-local discriminative domain $\mathcal{S}(\mathbf{x})$, differential approaches are not appropriate. A brute-force exhaustive method is: $\forall \mathbf{x} \in \mathcal{X}_L$, we calculate $\rho_S(\mathbf{x})$, and then select the most discriminative ones. The complexity is $O(|\mathcal{S}(\mathbf{x})| \times N)$, and is still intensive in practice.

Here we use a branch-and-bound search which largely reduces the complexity while maintaining the same optimal result as by the exhaustive search.

Let $\mathcal{S}(\mathbf{x}) = \{\mathbf{x} + \Delta\mathbf{l}_1, \cdots, \mathbf{x} + \Delta\mathbf{l}_n\}$, where $n = |\mathcal{S}(\mathbf{x})|$, and $\Delta\mathbf{l}_i$ is the relative position between the target AR and its ith neighbor in the semi-local discriminative domain. Denote $\rho_i(\mathbf{x}) = \min\limits_{\mathbf{y} \in \{\mathbf{x}+\Delta\mathbf{l}_1, \cdots, \mathbf{x}+\Delta\mathbf{l}_i\}} D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{y}))$. Then we have $\rho_i(\mathbf{x}) = \min\{\rho_{i-1}(\mathbf{x}), D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}+\Delta\mathbf{l}_i))\}$, thus $\rho_1(\mathbf{x}) \geq \cdots \geq \rho_n(\mathbf{x}) = \rho_S(\mathbf{x})$. In the beginning, we initialize an empty priority queue P to store the candidates. For each $\mathbf{x}_i \in \mathcal{X}_L$, we calculate $\hat{\rho}(\mathbf{x}_i) = \rho_1(\mathbf{x}_i)$ as the upper bound of $\rho_S(\mathbf{x}_i)$. Then we sort $\{\hat{\rho}(\mathbf{x}_i)\}$ in the descending order and push them sequentially into P so that the top state has the largest $\hat{\rho}(\cdot)$. For each $\mathbf{x}$, we associate a variable $\gamma(\mathbf{x})$ to count the number of elements in $\mathcal{S}(\mathbf{x})$ which has been searched around $\mathbf{x}$.

At every iteration, we retrieve the top state $(\mathbf{x}, \hat{\rho}(\mathbf{x}))$ from P, where $\hat{\rho}(\mathbf{x})$ is the current upper bound of $\rho_S(\mathbf{x})$, and $\hat{\rho}(\mathbf{x}) = \rho_{\gamma(\mathbf{x})}(\mathbf{x})$. If $\gamma(\mathbf{x}) = n$, meaning that we have already sought all the neighbors in $\mathcal{S}(\mathbf{x})$, we output $\mathbf{x}$ into the set of ARs and remove $\mathbf{x}$ from P.

Otherwise $\gamma(\mathbf{x}) < n$, we increase $\gamma(\mathbf{x})$ by 1, calculate $\mathbf{f}(\mathbf{x} + \Delta\mathbf{l}_{\gamma(\mathbf{x})})$, and update the upper bound

$$\hat{\rho}(\mathbf{x}) := \min\{\hat{\rho}(\mathbf{x}), D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x} + \Delta\mathbf{l}_{\gamma(\mathbf{x})}))\}. \tag{8}$$

Then we insert $(\mathbf{x}, \hat{\rho}(\mathbf{x}))$ into P maintaining the property that P is sorted with the descending order of $\rho(\cdot)$ (replace the old $\mathbf{x}$). Then we retrieve the top state

---

[1] In [16], $\mathbf{\Phi}$ is the contextual gradient and can be computed directly in a closed form.

Table 1. The branch-and-bound algorithm for selecting ARs

Input $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N, \Delta l_1, \cdots, \Delta l_n$
Output $\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_M$

1. FOR $i = 1$ TO N DO
     calculate $\mathbf{f}(\mathbf{x}_i), \mathbf{f}(\mathbf{x}_i + \Delta l_1)$,
     set $\hat{\rho}(\mathbf{x}_i) = D(\mathbf{f}(\mathbf{x}_i), \mathbf{f}(\mathbf{x}_i + \Delta l_1))$,
     $\gamma(\mathbf{x}_i) = 1$
   Initialize P as empty priority queue. $c = 0$.
2. Sort $\{\hat{\rho}(\mathbf{x}_i)\}$ in descending order.
   Let $\hat{\rho}(\bar{\mathbf{x}}_1) \geq \cdots \geq \hat{\rho}(\bar{\mathbf{x}}_N)$,
   FOR $i = N$ TO 1 DO
       push $(\bar{\mathbf{x}}_i, \hat{\rho}(\bar{\mathbf{x}}_i))$ into P
3. Retrieve top state $(\mathbf{x}, \hat{\rho}(\mathbf{x}))$ from P.
4. If $\gamma(\mathbf{x}) = n$
       $c = c + 1$, $\hat{\mathbf{x}}_c = \mathbf{x}$, goto 3.
   Else goto 5.
5. If $c = M$, Return. Else goto 6.
6. $\gamma(\mathbf{x}) = \gamma(\mathbf{x}) + 1$
   Calculate $\mathbf{f}(\mathbf{x} + \Delta l_{\gamma(\mathbf{x})})$
   Set $\hat{\rho}(\mathbf{x}) = \min\{\hat{\rho}(\mathbf{x}), D(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x} + \Delta l_{\gamma(\mathbf{x})}))\}$
   Insert $(\mathbf{x}, \hat{\rho}(\mathbf{x}))$ into P so that P is still sorted
   w.r.t. $\hat{\rho}(\cdot)$. Goto 3.

---

again iteratively until a number of M ARs are found. The algorithm is summarized in Table 1.

The top state $\mathbf{x}$ of P has the largest upper bound of $\rho_S(\mathbf{x})$, because for the remaining $\mathbf{x}_i$s in P, $\rho_S(\mathbf{x}_i)$ is bounded by $\hat{\rho}(\mathbf{x})$. As each time we only consider the most promising $\mathbf{x}$ of P, this significantly reduce the complexity. The complexity is $O(\sum_{i=1}^{N} \gamma(\mathbf{x}_i))$, and this method guarantees the optimality.

In practice, the complexity versus the exhaustive search is measured by the ratio $r = \frac{1}{nN} \sum_{i=1}^{N} \gamma(\mathbf{x}_i)$. The value of r is 0.18 on average for our testing sequences, e.g., for sequence `zebra`, $r = 0.18$. For sequence `dolphin`, $r = 0.16$. This means that our method significantly reduces the complexity in searching for ARs. Extra operations in our method (*i.e.*, insertion and sorting) have little computational complexity, as those operations take much less time than computing D.

5   Discriminative Attentional Visual Tracking

As the ARs are not similar to the other regions in their discriminative domain, the tracking performance of ARs is very robust. We propose a new attentional tracking method by using AR, and it has three important steps: At the first step, we extract ARs from images. Secondly, the contextual flow tracking algorithm [16] is applied to track each ARs independently. Finally, the beliefs of all the ARs are fused to determine the target location.
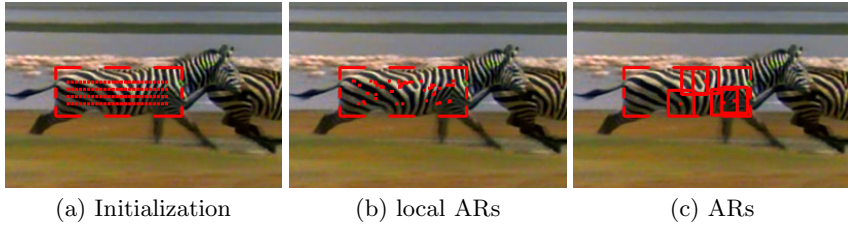
(a) Initialization   (b) local ARs   (c) ARs

Fig. 4. AR selection

5.1 AR Selection/Tracking

At the first frame, the target is initialized by the user. We evenly initialize N_{max} tentative ARs inside the target (Fig. 4(a)). The local ARs are shown in Fig. 4(b). Figure 4(c) shows top five ARs. For each AR, we record the geometrical relation between the ARs and the target (the relative position and the scale).

For each AR, the tracking is done based on the contextual flow method [16].

### 5.2 Attentional Fusion and Target Estimation

After obtaining the motion of each AR, we apply a Hough-voting scheme [14] to estimate the target location based on the matching scores of ARs and the recorded geometry. The estimated AR location casts a probabilistic vote about the target centroid position with respect to the AR center. The better the matching performance of a certain AR, the higher the probabilistic score. After the votes from all ARs are aggregated into a Hough image, the target location can be estimated as the peak in this image. This scheme is appropriate to handle occlusion. If some ARs are occluded, their matching scores will be very low, thus the probabilistic votes from those ARs are very low, and contribute less to the object location prediction than the ARs which are not occluded.

The scale of the target is estimated by a voting-like approach based on the scale estimation for each AR. To obtain a robust estimation, we only count the ARs which have high matching scores.

5.3 Model Adaptation

As the appearance changes, due to view differences, illumination variations and shape deformation, can ruin the observation, the model adaptation mechanism is necessary. We adapt the model by updating the ARs when necessary. The matching score of each AR measures the variation of its appearance. If the matching performance is good enough, we call the AR *active*. Otherwise, for a certain AR, if the matching score has been low for a long period of time (e.g., consecutive 10 frames), we call it *inactive* since it probably undergoes appearance changes or short term occlusion.

At the current frame, after target estimation, we check the matching score for each AR to see if it remains active. When there are m inactive ARs at the current frame, we remove them and select m new ARs from the target.
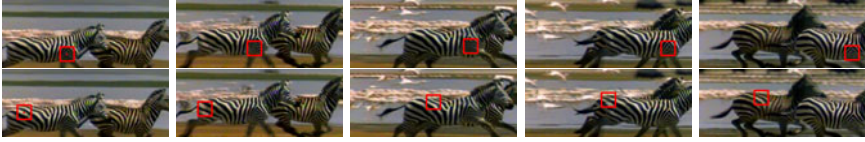
Fig. 5. Comparison of different placement of one AR. (Top) the AR from our method (bottom) the local AR

6   Experiments

For tracking initialization, we evenly initialize $N_{max} = 100$ tentative ARs inside the target. The size of the ARs is 25×25. For a certain AR, the size of its discriminative domain $\mathcal{N}(\mathbf{x})$ is determined by its possible motion and the maximum search range for tracking. The larger the possible motion, the larger $\mathcal{N}(\mathbf{x})$ we use.

Without code optimization, our C++ implementation comfortably runs at around 15 fps on average on Pentium 3G for 320×240 images.

We compare our method with an attentional visual tracker (AVT) [5] that reported excellent tracking performance. For fair comparison, we use the contextual flow as the feature vector, and use the Hough voting scheme in the fusion process for both methods. In addition, we have included the late selection procedure in AVT for comparison.

6.1   Using the Most Discriminative AR

ARs are resilient to distractors, because by definition an AR is not confused by its neighboring regions in its discriminative range. In this experiment, we compare the tracking performance by selecting different ARs and demonstrate the effectiveness of our method. The AR with the largest discriminative power is shown at the top row of Fig. 5. We choose some local ARs for comparison (one example is shown at the bottom row of Fig. 5). It is observed that at the top row, the texture of the best AR is quite different from its neighborhood. While at the bottom row, the texture of the AR contains stripes which is not quite discriminative in its semi-local discriminative domain. Therefore, the tracking performance shows that the local AR is unstable during tracking (keep drifting) while the AR at the top row succeeds and is very stable.

6.2   Handling Local Appearance Changes

Tracking targets undergoing local deformation is difficult in practice. However, if the local deformation only occurs in some parts of the target, the ARs on other parts can still make the tracking robust. These stable ARs contribute more in the fusion process as they have strong matching, so the tracking performance is still good. The comparison result is shown in Fig. 6 and 7. In Fig. 6, although the target appearance changes at some parts, the bottom-right AR is persistently

Fig. 6. A comparison of DAVT and AVT [dancing]. (Top) AVT (bottom) the proposed method
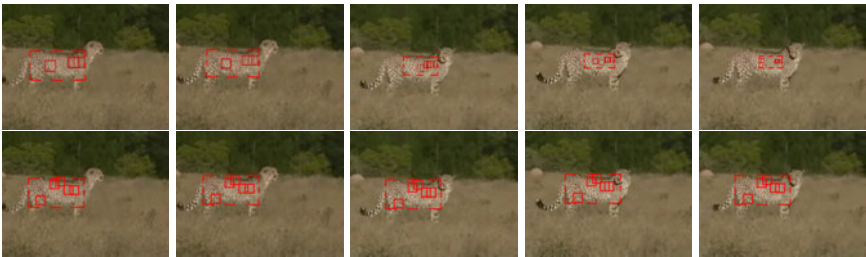

Fig. 7. A Comparison of DAVT and AVT [cheetah]. (Top) AVT (bottom) the proposed method

robust and thus dominates the fusion and gives good tracking results. The white ARs indicate those that have relative bad matching. Although these white ARs sometimes do not have strong matching, in most cases they are robust, since they are located at the boundary of the face and there are no distractors nearby. In Fig. 7, the textures of the cheetah are very similar. The ARs found by the proposed method are near the back and thigh of the cheetah. These regions look different from the body of the cheetah, so they hardly drift to some other regions inside the body. However, for AVT, it only selects some local ARs. We observe that there are some distractors in the semi-local domain of these local ARs and AVT fails as shown in Fig. 7.

We manually labeled the ground truth of our testing sequences to evaluate the tracking performance. Figure 8 shows the comparison of tracking error between DAVT and AVT in tracking error over time on the `bicycle` sequence (we use a different initialization as in [5]). At the 330th Frame, AVT is distracted and fails, but our method keeps the track persistently.

6.3   Handling Scale, Rotation and Occlusion

The scale estimation can be handled since the selected ARs are stable and rarely distracted. As in our matching method, the contextual descriptor is rotation invariant if we only use color contexts, the ARs give accurate matching despite of the motion. Then we estimate the rotation by measuring the relative position
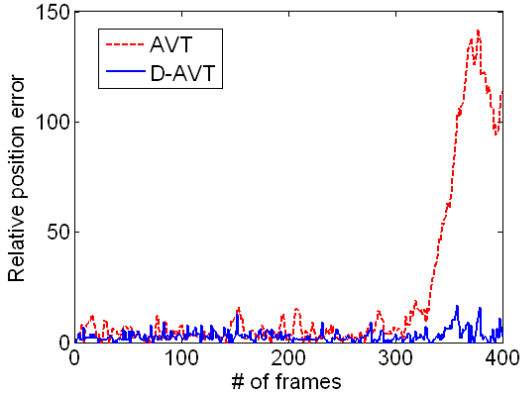
Fig. 8. Comparison: tracking errors between DAVT and AVT
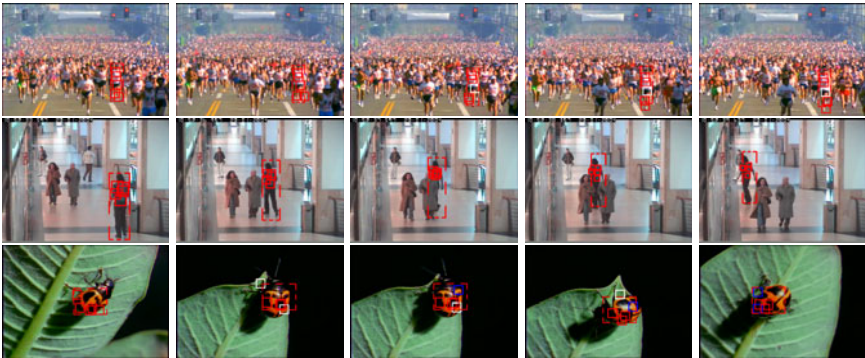
Fig. 9. Three examples of the proposed method

between the ARs. The occlusion can be handled by the fusion process. The model adaptation is also illustrated. Three examples are shown in Fig. 9. On the bottom row, the blue ARs indicate those that have been updated.

7   Conclusion

Spatial distraction is a major culprit for tracking failure, because distractors also exhibit good matching. This paper presents a novel approach of discriminative spatial attention to overcome this challenge, by selecting a set of discriminative attentional regions on the target. The discrimination power of an attentional region is defined by the margin of its feature from that of those in its discriminative domain. By integrating local and semi-local discrimination, this paper proposes an efficient method in finding ARs. Extensive tests demonstrate that the proposed discriminative spatial attention scheme significantly improves the

robustness in tracking. The further analysis [20] reveals that the existing works on local saliency detection share the common purpose of achieving good localization properties. Therefore, our AR selection scheme is very flexible so that those methods can be alternatively adopted for finding local ARs.

References

1. Peters, R., Itti, L.: Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention. In: CVPR (2007)
2. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on PAMI, 1254–1259 (1998)
3. Hager, G., Dewan, M., Stewart, C.: Multiple kernel tracking with ssd. In: CVPR (2004)
4. Fan, Z., Yang, M., Wu, Y., Hua, G., Yu, T.: Efficient optimal kernel placement for reliable visual tracking. In: CVPR (2006)
5. Yang, M., Yuan, J., Wu, Y.: Spatial selection for attentional visual tracking. In: CVPR (2007)
6. Gao, D., Vasconcelos, N.: Discriminant interest points are stable. In: CVPR (2007)
7. Parameswaran, V., Ramesh, V., Zoghlami, I.: Tunable kernels for tracking. In: CVPR (2006)
8. Donoser, M., Bischof, H.: Efficient maximally stable extremal region (mser) tracking. In: CVPR (2006)
9. Collins, R., Liu, Y.: On-line selection of discriminative tracking features. In: ICCV (2003)
10. Shi, J., Tomasi, C.: Good features to track. In: CVPR (1994)
11. Kadir, T., Brandy, M.: Saliency, scale and image description. IJCV (2001)
12. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. IJCV (2005)
13. Kwon, J., Lee, K.: Tracking of a non-rigid object via patch-based dynamic appearance modeling and adaptive basin hopping monte carlo sampling. In: CVPR (2009)
14. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: CVPR (2009)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
16. Wu, Y., Fan, J.: Contextual flow. In: CVPR (2009)
17. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: CVPR (2000)
18. Palmer, S.: Vision science: photons to phenomenology. The MIT Press, Cambridge (1999)
19. Avidan, S.: Ensemble tracking. In: CVPR (2005)
20. Fan, J., Wu, Y., Dai, S.: Discriminative spatial attention for robust tracking. Northwestern University Technical Report (2010)