Unifying Spatial and Attribute Selection for Distracter-resilient Tracking

Nan Jiang
Northwestern University
Evanston, IL, USA.
nan.jiang@northwestern.edu

Huazhong University of Sci. & Tech.
Wuhan, Hubei, P. R. China.

Ying Wu
Northwestern University
Evanston, IL, USA.
yingwu@northwestern.edu

Abstract

Visual distracters are detrimental and generally very difficult to handle in target tracking, because they generate false positive candidates for target matching. The resilience of region-based matching to the distracters depends not only on the matching metric, but also on the characteristics of the target region to be matched. The two tasks, i.e., learning the best metric and selecting the distracter-resilient target regions, actually correspond to the attribute selection and spatial selection processes in the human visual perception. This paper presents an initial attempt to unify the modeling of these two tasks for an effective solution, based on the introduction of a new quantity called Soft Visual Margin. *As a function of both matching metric and spatial location, it measures the discrimination between the target and its spatial distracters, and characterizes the reliability of matching. Different from other formulations of margin, this new quantity is analytical and is insensitive to noisy data. This paper presents a novel method to jointly determine the best spatial location and the optimal metric. Based on that, a solid distracter-resilient region tracker is designed, and its effectiveness is validated and demonstrated through extensive experiments.*

1. Introduction

The key in visual target tracking is to perform accurate identification of the target in the incoming image frames. This can be done by matching candidate regions with the given target region. Tracking fails when there are no good matches or the good ones turn out to be false positive matches. In the former case, the target may disappear or be occluded; while in the latter case, the tracker is confused by the non-target candidates that appear very similar to the target, namely the *visual distracters* or *distracters* for short. The distracters can be induced by cluttered backgrounds, surrounding crowds, or camouflage. Distracters are detrimental and generally difficult to handle. An example is shown in Fig. 1, where other horses persistently generate false positives to the target highlighted in the yellow box, and easily fail the tracker. This paper is concerns on handling such scenarios with distracters.

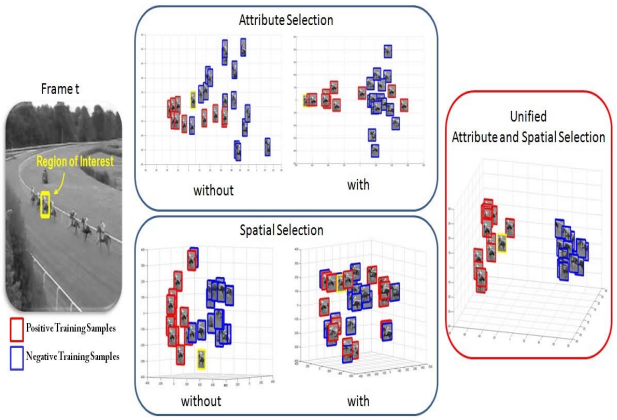

Figure 1. This paper handles visual distracters. This paper shows that the unification of attribute selection (via metric learning) and spatial selection (via optimal region placement) is able to robustly separate the true target and the distracters.

In fact, the handling of visual distracters is closely related to three critical factors in region matching: *visual features*, *matching metric* and *tracking region placement*. Visual features represent the matching subjects, and have been extensively studied in the past several decades. Various methods have been used in tracking, e.g., local invariance [17], shape [14], region appearances [8], local statistics [6], or even some level of recognition [2]. This paper does not focus on specific features, but assumes abstract features without losing generality. It is common in practice that features are matched based on a pre-defined criterion or metric, e.g., the Euclidean metric or Bhattacharyya metric. Recent studies [15, 16] have shown that learning an adaptive metric can be beneficial to discriminate the true target from the distracters. The third factor is the placement of the tracking region [10]. The *tracking region* (or attentional region [9]) is the actual region used for matching, while the delineated target region is the rough user input for appointing the target. Simply using such rough delineated region for matching leads to an interesting phenomenon: under a

given metric, certain regions may produce reliable matching, while some may easily be distracted. This implies that the placement of the tracking region needs to be optimized. This paper attempts to find a principled solution to unify optimal metric and optimal region placement, i.e., to determine the matching metric as well as jointly identify the appropriate tracking regions so as to achieve reliable and accurate region tracking.

This problem is actually directly related to the attentional selection in the human visual perception. Psychological evidences show that the human visual perception is selective [19], so that the human visual system can easily construct accurate correspondences. There are two major attentional selection mechanisms. One is the *spatial selection* that identifies special image regions (namely attentional regions) for visual processing; and the other one is the *attribute selection* that chooses or forms appropriate visual features. In our problem, optimal region placement plays the role of *spatial selection*, and learning matching metric is actually *attribute selection*. The study in this paper attempts to give a plausible explanation of these mechanisms and to provide a unified solution to these two tasks.

The basic idea of our study is based on the introduction of a new quantity, *Soft Visual Margin*, that unifies both feature discrimination and spatial separability between the target and its spatial distracters. We treat a given image region and its close spatial vicinity as the target, but those not in the vicinity as the distracters. The proposed *Soft Visual Margin* gives a soft version of the margin between the two classes in an analytical form. It quantifies the mixup situation of the target and its distracters. The larger the margin, the stronger the discrimination between the target and its distracters. As it is a function of matching metric as well as the spatial location, it is clear that different image regions and different metrics produce different margins. This paper presents a novel formulation, where the joint determination of the optimal metric and the best region placement of the attentional region is achieved via *Soft Visual Margin* maximization. In this way, the identified attentional region with the learned matching matric gives the best class discrimination, and thus matching this attentional region shall give reliable and accurate results. Based on this, a solid region-based distracter-resilient tracking method is designed.

This paper gives a new computational model to naturally unify spatial selection and attribute selection. Besides that, to the best of our knowledge, it is an original attempt to jointly determine the matching metric and select the attentional regions for tracking. Moreover, it provides an efficient solution to an effective region tracker.

2. Background

There have been two parallel attempts to enlarge the class discrimination between the target and the background.

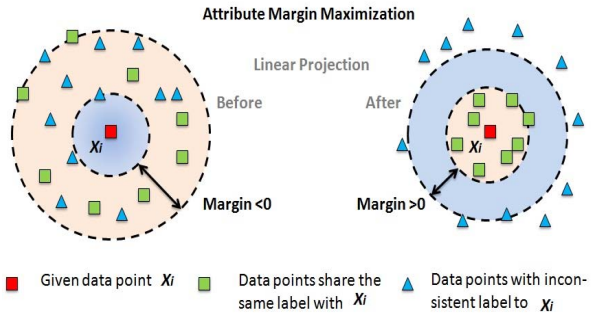

Figure 2. An illustration of attribute margin maximization via metric learning.

One is to determine some discriminative image regions in the spatial domain to represent the target of interest [9, 10, 13]. For example, an image region can be associated with a linear system for motion estimation. Thus the stability of this linear system can be used to characterize its reliability [10], and a gradient-based local search is designed to locate such reliable regions efficiently. In addition, an entropy-based measure, called *intrackability* [13], is proposed to measure the uncertainty of the motion estimation, and thus characterizing the matching reliability. But in general, these methods are computationally demanding. As matching failure is often associated with the distracters that exhibit very similar visual appearance to the target, the concept of discriminative attentional region (or D-AR) is introduced [9]. A D-AR has a large margin to its visual distracters, and can be found through an efficient branch-bound search [9]. All these methods assume a pre-determined metric in the feature space.

The other attempt is to determine a more discriminative feature space in which the target can be best separated from the background, and this is generally done through only adaptive learning. Different methods have different objectives for discriminative learning. For example, the method in [5] uses the Fisher discriminant and performs exhaustive search over all possible feature combinations. When the feature space is large, the applicability of this method is limited due to its exhaustive search scheme. Recently, interesting attempts of integrating metric learning in tracking have been made [22, 15, 16]. These methods aim to find the best projection of the original feature space to achieve certain objectives. For example, one objective is to collapse all the data in its class to a single data point [11, 22]. Another is to maximize the performance of nearest neighbor classification [12, 15, 16]. These objectives are not specifically designed for visual tracking to distinguish the target from its distracters, and they simply use the delineated image region as the target.

This paper describes an original attempt to unify these two approaches by introducing the concept of *Soft Visual Margin* as a single objective for both spatial selection

and attribute selection. Unlike the discriminative margin introduced in [9] that is not analytical and is sensitive to noise, this *Soft Visual Margin* has much nicer properties: it is insensitive to noise and is even differentiable. Based on the maximization of this *Soft Visual Margin*, the unification can be achieved by jointly optimizing the matching metric and identifying the attentional regions, which can not be done in the existing margin-based metric learning methods[23, 20, 12, 21].

3. Soft Visual Margin Maximization

3.1. Hard Margin

Denote by $\mathbf{x} \in \mathbb{R}^N$ the visual feature at a spatial location of interest $\mathbf{c}$, where N is the number of pre-determined primitive attributes or features. For this location $\mathbf{c}$, as the set of locations at its spatial vicinity are still regarded as the target itself, we treat them as positive samples and denote the set by $\mathcal{C}^+$. Meanwhile, we also have a set of regions at other locations that are regarded as mismatches or distracters, even if they may exhibit similar visual appearance as the target, and call them negative samples and the set $\mathcal{C}^-$.

In the feature space $\mathbb{R}^N$, we denote $\mathbf{x}_{ij} \triangleq \mathbf{x}_i - \mathbf{x}_j$. Instead of taking for granted to use the Euclidean distance for matching, we use the Mahalanobis distance:

$$D_{ij}(\mathbf{x}_i, \mathbf{x}_j|\mathbf{A}) = \|\mathbf{A}x_i - \mathbf{A}x_j\|^2 = \mathbf{x}_{ij}^T\mathbf{A}^T\mathbf{A}\mathbf{x}_{ij}, \quad (1)$$

where $\mathbf{A}$ is a linear transform to project the original feature space to a new one, and it characterizes a distance metric. It is difficult to specify this metric, but we will see in later sections how it can be learned. As $\mathbf{A}x_{ij}$ is a linear projection of x_{ij}, learning this metric acts like another layer of feature extraction.

Based on the positive and negative sets $\mathcal{C}^+$ and $\mathcal{C}^-$, we can characterize their separability or class discrimination by using the margin between the two sets in the feature space. For the location of interest and its feature $\mathbf{x}$, the margin is considered as the difference between the distance from the furthest positive sample and the distance from the closest negative sample, i.e.,

$$\gamma(\mathbf{x}|\mathbf{A}) \triangleq \min_{j\in\mathcal{C}^-} D_j(\mathbf{x}_j, \mathbf{x}|\mathbf{A}) - \max_{j\in\mathcal{C}^+} D_j(\mathbf{x}_j, \mathbf{x}|\mathbf{A})$$

Conceptually this objective function is fine, but it has two issues. First, it is sensitive to noise. Second, it is not analytical and very difficult to manipulate. This motivates our introduction of *Soft Visual Margin* in the next section.

3.2. Soft Visual Margin

For the negative set $\mathcal{C}^-$, we can use the following *soft min* to approximate the probability that $\mathbf{x}$ treats $\mathbf{x}_j \in \mathcal{C}^-$ as its nearest neighbor:

$$p_j = \begin{cases} \frac{\exp(-\alpha\|\mathbf{A}x-\mathbf{A}x_j\|^2)}{\sum_k \exp(-\alpha\|\mathbf{A}x-\mathbf{A}x_k\|^2)} & \text{if } j \in \mathcal{C}^- \\ 0 & \text{if } j \in \mathcal{C}^+ \end{cases}$$

where $k \in \mathcal{C}^-$, and $\alpha > 0$. Therefore, the soft version of the distance from the nearest negative sample can be written as $\sum_{j\in\mathcal{C}} p_j D_j$, where $\mathcal{C} = \mathcal{C}^+ \bigcup \mathcal{C}^-$.

In the same fashion, we can use *soft max* to approximate the probability that $\mathbf{x}$ treats $\mathbf{x}_j \in \mathcal{C}^+$ as its furthest neighbor:

$$q_j = \begin{cases} \frac{\exp(\beta\|\mathbf{A}x-\mathbf{A}x_j\|^2)}{\sum_k \exp(\beta\|\mathbf{A}x-\mathbf{A}x_k\|^2)} & \text{if } j \in \mathcal{C}^+ \\ 0 & \text{if } j \in \mathcal{C}^- \end{cases}$$

where $k \in \mathcal{C}^+$, and $\beta > 0$, and the soft version of the distance from the furthest positive sample is $\sum_{j\in\mathcal{C}} q_j D_j$.

Based on that, we define *Soft Visual Margin* for a given spatial location (or pixel location) $\mathbf{c}$ with visual feature $\mathbf{x}$ as:

$$\epsilon(\mathbf{x}, \mathbf{A}) \triangleq \sum_{j\in\mathcal{C}(x)} (p_j - q_j)D_j. \quad (2)$$

In contrast to the hard margin, the above proposed *Soft Visual Margin* is analytical and differentiable. In addition, because it integrates all samples, it is less sensitive to noise than the hard margin.

It is clear that the *Soft Visual Margin* at a particular pixel location measures the discrimination power of its visual feature to separate the target from its nearby distracters. And it is also clear that this discrimination power also depends on the metric for the feature distances. In view of this, we can unify the following two tasks:

- **Attribute selection**: given the location of the target, we determine the best metric that maximizes the *Soft Visual Margin*, i.e., identifying new feature projection matrix $\mathbf{A}$ so as to maximize the visual discrimination;
- **Spatial selection**: under a given metric, the location of the target may not present strong discrimination. Consequently, the matching of the target is likely to be distracted. By maximizing the *Soft Visual Margin*, we determine a location nearby that exhibits a local maximum discrimination. It can be regarded as an attentional region. This selection allows us to find this best attentional region for matching.

The following sections describe the details of maximizing *Soft Visual Margin* for attribute selection, spatial selection and its unification for visual tracking.

3.3. Attribute Selection and Metric Learning

As mentioned before, the primitive visual features (e.g., the color distribution or the textures) at a given spatial location are the attributes of the target at this location. An interesting question is how we can determine the best metric of these features to achieve the best discrimination between the target and its nearby visual distracters. This is actually an attribute selection task. This basic idea is shown in Fig. 2. The primitive visual features need to be known in advance, e.g., stacking all the pixel intensities to form a vector or 32-bin color histogram and 32 Gabor filter responses. To learn the best metric, our objective is:

$$\max_{\mathbf{A}} \epsilon(\mathbf{A}|\mathbf{c}) = \sum_{j \in \mathcal{C}(\mathbf{c})} (p_j - q_j) D_j, \tag{3}$$

where $\mathcal{C}(\mathbf{c})$ is the set of positive and negative samples collected around the given target location $\mathbf{c}$, as described before. This optimization problem can be solved by a gradient-based method.

For the soft max form:

$$d \triangleq \sum_j q_j D_j = \frac{\sum_j D_j e^{\beta D_j}}{\sum_k e^{\beta D_k}},$$

we obtain the following derivative:

$$\frac{\partial d}{\partial D_j} = q_j \Big[1 + \beta (D_j - d) \Big].$$

Similar form of the derivative can be found for the soft min operation. Denote by $\mathbf{x}_{0j} \triangleq \mathbf{x} - \mathbf{x}_j$. Then after some manipulations, we obtain the following gradient:

$$\begin{aligned} \frac{\partial \epsilon}{\partial \mathbf{A}} &= 2\mathbf{A} \sum_j \Bigg\{ \Big[-\alpha p_j D_j x_{0j} x_{0j}^T \\ &+ p_j x_{0j} x_{0j}^T + \alpha p_j D_j \sum_k p_k x_{0k} x_{0k}^T \Big] \\ &- \Big[\beta q_j D_j x_{0j} x_{0j}^T + q_j x_{0j} x_{0j}^T \\ &- \beta q_j D_j \sum_l q_l x_{0l} x_{0l}^T \Big] \Bigg\} \end{aligned} \tag{4}$$

3.4. Spatial Selection and Optimal Placement

Finding the best metric for a given target location may not be enough, because under the same metric, different spatial locations exhibit different discrimination power. The locations that are more resilient to noise and distracters are more discriminative. In other words, these locations observe larger margins. Then a problem of interest is: given an initial spatial location, can we find a location nearby that exhibits a local maximum of the *Soft Visual Margin*?

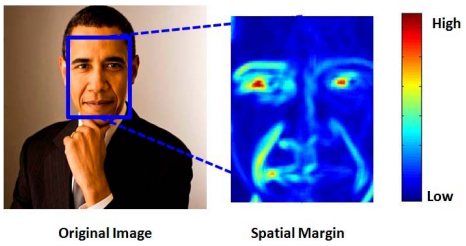

Figure 3. An example of the discrimination power of spatial selection.

Of course, this can be done by performing an exhaustive search. For every spatial location, we compute its *Soft Visual Margin*. Fig. 3 shows an example of the *Soft Visual Margin* map for every pixel location of a face image region, where the eye corns and mouth corners give large margins. However, this process is quite computationally demanding. Is there a way to accelerate?

As the *Soft Visual Margin* is differentiable, we have obtained an efficient gradient-based search solution to spatial selection. Without losing generality, we use the following special case to explain our solution. Instead of using advanced visual features, here we use the image template by stacking all the pixels to form the feature vector. At spatial location $\mathbf{c}$, we use $I_j(c)$ to substitute $\mathbf{x}_j$. Sure, other features can be used as well.

Given a metric, as $\mathbf{A}$ is obtained from the attribute selection, it is clear that p_j and q_j are functions of $\mathbf{c}$. We rewrite:

$$D_j(\mathbf{c}) = \|\mathbf{A}I(c) - \mathbf{A}I_j(c)\|^2. \tag{5}$$

Then we have:

$$p_j(\mathbf{c}) = \begin{cases} \frac{\exp(-\alpha \|\mathbf{A}I(c) - \mathbf{A}I_j(c)\|^2)}{\sum_k \exp(-\alpha \|\mathbf{A}I(c) - \mathbf{A}I_k(c)\|^2)} & \text{if } j \in \mathcal{C}^- \\ 0 & \text{if } j \in \mathcal{C}^+ \end{cases}$$

$$q_j(\mathbf{c}) = \begin{cases} \frac{\exp(\beta \|\mathbf{A}I(c) - \mathbf{A}I_j(c)\|^2)}{\sum_l \exp(\beta \|\mathbf{A}I(c) - \mathbf{A}I_l(c)\|^2)} & \text{if } j \in \mathcal{C}^+ \\ 0 & \text{if } j \in \mathcal{C}^- \end{cases}$$

The objective function can be written as:

$$\max_{\mathbf{c}} \epsilon(\mathbf{c}|\mathbf{A}) \triangleq \sum_j \Big(p_j(\mathbf{c}) - q_j(\mathbf{c}) \Big) D_j(\mathbf{c}) \tag{6}$$

It's easy to derive that:

$$\frac{\partial D_j}{\partial \mathbf{c}} = 2\Big(I'(\mathbf{c}) - I_j'(\mathbf{c}) \Big) \mathbf{A}^T \mathbf{A} \Big(I(\mathbf{c}) - I_j(\mathbf{c}) \Big), \tag{7}$$

$$\frac{\partial p_j}{\partial \mathbf{c}} = -\alpha p_j \Big(\frac{\partial D_j}{\partial \mathbf{c}} - \sum_k \frac{\partial D_k}{\partial \mathbf{c}} \Big), \tag{8}$$

$$\frac{\partial q_j}{\partial \mathbf{c}} = \beta q_{ij}\Big(\frac{\partial D_j}{\partial \mathbf{c}} - \sum_l \frac{\partial D_l}{\partial \mathbf{c}}\Big). \quad (9)$$

Therefore, we obtain:

$$\begin{aligned}\frac{\partial \epsilon}{\partial \mathbf{c}} &= 2\sum_j \Bigg[(p_j - q_j)\frac{\partial D_j}{\partial \mathbf{c}} - \alpha p_j \frac{\partial D_j}{\partial \mathbf{c}} \\ &+ \alpha p_j D_j \sum_k p_k \frac{\partial D_j}{\partial \mathbf{c}} - \beta q_j \frac{\partial D_j}{\partial \mathbf{c}} \\ &+ \beta q_j D_j \sum_l q_l \frac{\partial D_j}{\partial \mathbf{c}}\Bigg] \quad (10)\end{aligned}$$

The localization of a local maximum can be performing by searching along this gradient.

3.5. Unification for Visual Tracking

To design a robust visual tracking method, we can unify the above attribute selection and spatial selection processes through maximizing the *Soft Visual Margin*. This is a supervised process and we need both positive and negative data. Based on the above descriptions on *Soft Visual Margin*, attribute selection and spatial selection, we have an overall objective for attentional selection at frame t:

$$(\mathbf{A}^*, \mathbf{c}^*)_t = \arg\max_{\mathbf{A},\mathbf{c}} \epsilon(\mathbf{A}, \mathbf{c}) \quad (11)$$

This can be done through a two-step iteration process based on the coordinate descent scheme:

- A-Step: Attribute selection that learns the best metric for the fixed location obtained in S-Step, i.e., $\mathbf{A}^* = \arg\max_{\mathbf{A}} \epsilon(\mathbf{A}|\mathbf{c})$;
- S-Step: Spatial selection that determines the location of the most stable region (i.e., attentional region) which exhibits the largest margin based on the new metric obtained in A-Step, i.e., $\mathbf{c}^* = \arg\max_{\mathbf{c}} \epsilon(\mathbf{c}|\mathbf{A})$;

Once both attribute and spatial selections are obtained for the target at frame t, we can perform matching at the frame $t+1$. The visual tracking is based on matching the new attentional region by the learned new metric. The search can be performed via either local exhaustive search or through gradient-descent search.

In this paper, we use the image template and stack all the pixels to form its feature vector. The target and the candidate regions are all resized to 20×20, leading to a $\mathbb{R}^{400}$ feature space. Exhaustive local search is employed to track the target in the new frames. Once the target is localized at time t, we perform attribute selection and spatial selection iteratively to determine the best metric and the optimal attentional region, and use them to process the next frame.

Dataset	Name	Points	Dimensions	Classes
1	Fertility	100	9	2
2	Diabetes	768	8	2
3	ILPD	578	8	2
4	Vertebral Column	310	6	3
5	SPECTF Heart	187	44	2
6	Balance Scale	625	4	3

Table 1. Benchmark Datasets

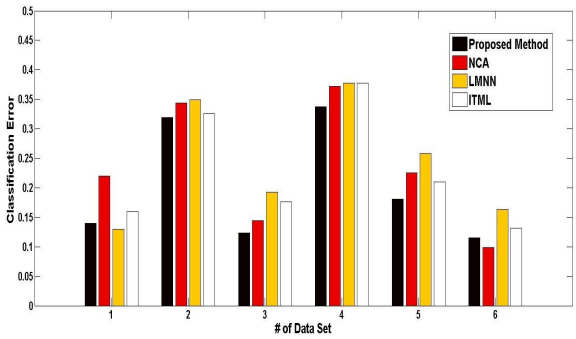

Figure 4. Classification error rate on benchmark datasets.

4. Experiment

4.1. Attribute selection via margin maximization

To demonstrate the effectiveness of the proposed attribute selection method alone, we compare the proposed method with several widely used state-of-the-art metric learning methods, including NCA[12], LMNN[23] and ITML[7]. The comparison is performed on six benchmark datasets from the UCI repository[1] that are commonly used for evaluation in the machine learning community. The properties of the datasets are shown in Table 1.

Fig. 4 shows the classification error rate averaged over 10 runs on the testing data by cross validation. All the methods try to learn a Mahalanobis distance but under different objectives. NCA [12] uses soft classification for metric learning. LMNN [23] pursues the optimal metric by maximizing the margin between inconsistent classes in a non-analytical way. And ITML solves the metric learning based on entropy. Our method incorporates the ideas of soft classification and margin maximization. As shown in Fig. 4, on 4 datasets, our method outperforms all these metric learning methods. On the rest two datasets, it gives comparable performance to the best among these baseline methods.

4.2. Spatial selection via margin maximization

To demonstrate the effectiveness of spatial selection along with attribute selection, we test our method on real videos shown in Fig. 5. The red bounding box in Fig. 5 represents the original target region. To locate the attention-

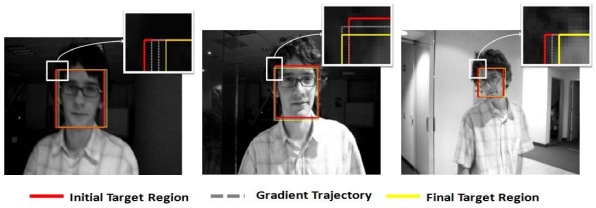

Figure 5. Sample frames to demonstrate the spatial selection process.

Figure 6. Comparison results among basic template tracker(*Green*), spatial selection based tracker(*Yellow*), attribute selection based tracker(*Red*) and the proposed method(*Cyan*).

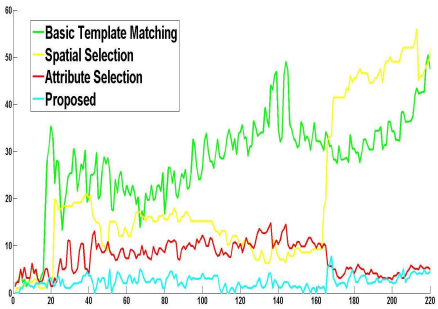

Figure 7. Quantitative comparison to baseline methods.

Figure 8. Comparison results between DAT(*Red*) and the proposed method(*Blue*).

al region nearby that has the largest discrimination power, the proposed method performs an efficient gradient-based search (in Eq. 10). The updating process is illustrated by the gray dash lines. After iterating between the location optimization and the metric optimization, the final location of the attentional region is depicted by the yellow box in Fig. 5. In our experiments, we observe that this process generally converges within 6 iterations. Target matching can be performed on this attentional region for robust tracking.

4.3. Soft Visual Margin for Visual Tracking

The proposed method employs *Soft Visual Margin* for visual tracking. To handle the scaling, we enlarge and shrink the target template by the factor of 1.05 and 0.95, respectively. We choose the best match as the optimal estimation

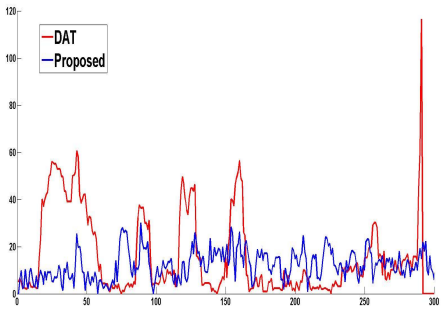

Figure 9. Quantitative comparison to spatial selection methods.

for target location and scale. The proposed method takes 3.8206 seconds per frame on average to complete both attribute and spatial selection for each frame, as well as the target localization by Matlab and C++ implementation on a desktop PC with Intel Core i5 2.5GHz CPU and 6G memory. The majority of the computation attributes to the metric learning. As it is not necessary to update the metric on every frame. When there is no drastic changes between two consecutive frames, the needs of metric is not urgent, and using the same metric will not sacrifice the matching accuracy much.

Therefore, our implementation takes a tradeoff where the spatial selection is performed for every frame, while attribute selection is only performed after large changes or a certain number of frames. Specifically, if the value of the objective function in Eq.3 is no less than the 0.8 times of the previous value, we only perform the spatial selection. This leads to a huge speedup. Under the same setting, it only takes 0.7732 second per frame on average. It will run comfortably in real-time by a pure C++ implementation.

4.3.1 Comparison to baseline methods

To demonstrate the necessity of unifying the attribute and spatial selection, we compare the following 4 tracking strategies, including: (1) basic template matching based visual tracking, without attribute or spatial selection; (2) spatial selection based visual tracking method, with no attribute selection; (3) attribute selection based tracking, no spatial selection; (4) unified attribute and spatial selection based tracking. The objective and subjective results are shown in Fig. 6 and Fig. 7 respectively. It's obvious that: (1) the basic template matching based tracking method easily fails due to the distractions from the false positives; (2) the spatial selection does improve the matching in several cases; (3) the attribute selection method can tell the foreground apart from the background, but it may not be stable due to online learning; (4) by unifying attribute and spatial selection, the proposed method is robust to distracters and is stable and consistent.

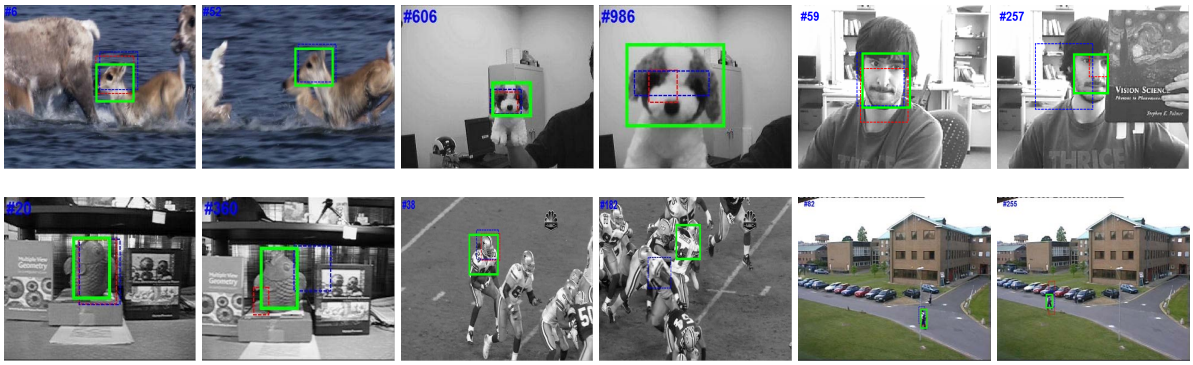

Figure 10. Comparison among TUDAMM(*Red*), SRML tracker(*Blue*) and the proposed method(*Green*).

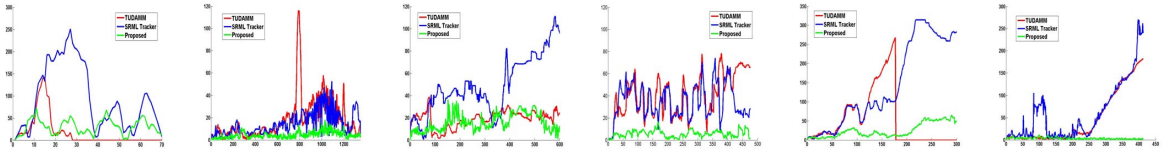

Figure 11. Quantitative comparison among TUDAMM(*Red*), SRML tracker(*Blue*) and the proposed method(*Green*).

4.3.2 Comparison to attribute selection methods

Metric learning methods have recently been integrated into visual tracking. Some representative ones include TUDAMM [22] and SRML Tracker [16]. TUDAMM tries to collapse all the training samples in the same class to one single point. In most of the cases, such an objective is difficult to achieve. SRML tracker is mainly developed from [15], targeting on a robust metric learning. These methods can improve the matching but they are not especially designed for visual tracking. When the localization of the target is not accurate (e.g., shifted by several pixels), this will introduce noisy or even wrong samples for metric adjustment, and in turn it will largely jeopardize the tracking performance.

Fortunately, this can largely be alleviated by the spatial selection in our proposed method, because the spatial selection always identifies the most stable and discriminative nearby attentional region. The collection of training samples for metric learning is no longer centered on the tracking result, but instead on the attentional region. This will in turn improve metric learning. The tracking results of the proposed methods comparing to TUDAMM and SRML trackers are shown in Fig. 10. It is evident that the proposed tracker is much more robust and reliable than the TUDAMM and SRML trackers. The quantitative comparisons in Fig. 11 are consistent with the subjective results, and clearly show the superiority of the proposed method.

4.3.3 Comparison to spatial based method

Besides comparing to the attribute selection based tracking methods, we also compare the proposed method to DAT [9], as shown in Fig. 8 and Fig. 9. The boy in blue t-shirt has a very similar appearance to the children nearby. We observe that DAT is able to tell the target apart from the background to some extend, but its discrimination power is not as good as that in the proposed method in general.

4.3.4 Comparison to general tracking methods

We also compare the proposed method to some state-of-the-art tracking algorithms, including LOT [18], MIL [3], L1APG [4] and CT [25]. Most test sequences and the results of these methods are obtained from the benchmark reported in [24]. As shown in Fig. 12, due to the unification of attribute and spatial selection, the proposed method has more discrimination power to distinguish the target from the background, as demonstrated in the *Ballet* sequence. When there are distracters with similar appearance, these baseline algorithms can hardly give accurate results. In contrast, the proposed method tracks the target correctly and persistently. This is evident in most other test sequences. The quantitative comparisons among LOT, MIL, L1APG, CT and the proposed method are given in Fig.13. It clearly demonstrates the superior performance of the proposed method.

5. Conclusion

This paper presents a new approach attempting to integrate attribute selection and spatial selection for accurate and reliable region matching in robust tracking. It gives a plausible computational unification of these two important processes of attentional selection, based on the introduction of a new discriminative measure of *Soft Visual Margin*. Maximizing it over matching metric and region placement leads to the joint optimization for spatial selection and met-

Figure 12. Comparison among LOT(*Red*), MIL(*Blue*), L1APG(*Green*), CT(*yellow*) and the proposed method(*Cyan*).

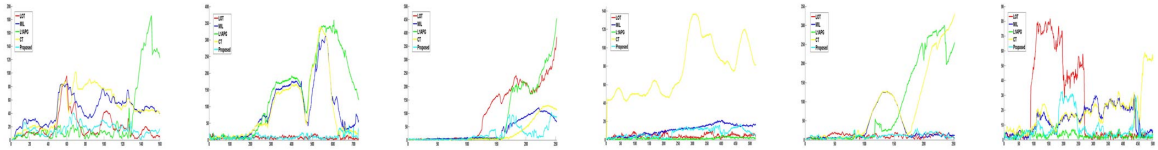

Figure 13. Quantitative comparison among LOT(*Red*), MIL(*Blue*), L1APG(*Green*), CT(*yellow*) and the proposed method(*Cyan*).

ric learning. This paper gives an efficient gradient-based solution to this problem. Theoretical analysis and extensive experiments demonstrate the effectiveness of this new approach for visual tracking.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (grant No. 61201377), and was supported in part by National Science Foundation grant IIS-0916607, IIS-1217302, and DARPA Award FA 8650-11-1-7149.

References

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007. 5

[2] S. Avidan. Support vector tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(8):1064–1072, 2004. 1

[3] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 983–990, 2009. 7

[4] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1830–1837, 2012. 7

[5] R. T. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *IEEE Conference on Computer Vision and Pattern Recognition*, 27(10):1631 –1643, oct. 2005. 2

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003. 1

[7] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *International Conference on Machine Learning*, pages 209–216, 2007. 5

[8] M. Donoser and H. Bischof. Efficient maximally stable extrmal region(MSER) tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 1

[9] J. Fan, Y. Wu, and S. Dai. Discriminative spatial attention for robust tracking. In *European Conference on Computer Vision*, 2010. 1, 2, 3, 7

[10] Z. Fan, M. Yang, Y. Wu, G. Hua, and T. Yu. Efficient optimal kernel placement for reliable visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 658 – 665, 17-22 2006. 1, 2

[11] A. Globerson and S. Roweis. Metric learning by collapsing classes. *Advances in Neural Information Processing Systems*, 18:451–458, 2006. 2

[12] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighborhood component analysis. In *Advances in Neural Information Processing Systems*, 2005. 2, 3, 5

[13] H. Gong and S. Zhu. Intrackability: Characterizing video staistics and pursing video representations. *International Journal of Computer Vision*, 2012. 2

[14] M. Isard and A. Blake. Condensation:conditional density propagation forvisual tracking. *International Journal of Computer Vision*, 29(1):5–28, Aug. 1998. 1

[15] N. Jiang, W. Liu, and Y. Wu. Adaptive and discriminative metric differential tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 1, 2, 7

[16] N. Jiang, W. Liu, and Y. Wu. Order determination and sparsity-regularized metric learning adaptive visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1956–1963, 2012. 1, 2, 7

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 1

[18] S. Oron, A. Bar-Hillel, D. Levi, and S. Avidan. Locally orderless tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1947, 2012. 7

[19] H. E. Pashler. *The Psychology of Attention*. MIT Press, Cambridge, MA, 1998. 2

[20] L. Torresani and K. C. Lee. Large margin component analysis. In *Advances in Neural Information Processing Systems*, pages 1385–1392. MIT Press, Cambridge, MA, 2007. 3

[21] J. Wang, A. Woznica, and A. Kalousis. Parametric local metric learning for nearest neighbor classification. *Advances in Neural Information Processing Systems*, abs/1209.3056, 2012. 3

[22] X. Wang, G. Hua, and T. X. Han. Discriminative tracking by metric learning. In *European Conference on Computer Vision*, 2010. 2, 7

[23] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, Vancouver, BC, 2006. 3, 5

[24] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 7

[25] K. Zhang, L. Zhang, and M.-H. Yang. Real-time compressive tracking. In *European conference on Computer Vision*, pages 864–877, Berlin, Heidelberg, 2012. Springer-Verlag. 7