ABSTRACT
Psychophysical findings have shown that human vision system has an ability to improve target search by enhancing the representation of image components that are related to the searched target, which is the so-called feature-based visual attention. In this paper, motivated by these psychophysical findings, we propose a robust visual tracking algorithm by simulating such feature-based visual attention. Specially, we consider the general sparse basis functions extracted on a large set of natural image patches as features. We define that a feature is related to the target when succeeding activations of that feature cannot increase system’s entropy. The target is finally represented by the probability distribution of those related features. The target search is performed by minimizing the Matusita distance measure between the distributions of the target model and candidate using Newton-style iterations. The experimental results verify that the proposed method is more robust and effective than widely used mean shift based methods.

Index Terms—Visual tracking, visual attention, entropy gain, Newton-style iterations

1. INTRODUCTION
Visual tracking is a critical step in many computer vision applications, such as automated surveillance, video indexing, human computer interfaces and vehicle navigation. In the past few decades, numerous tracking algorithms have been proposed. The three main frameworks into which most algorithms fall are 1) moving detection-and-track; 2) target representation-and-search; 3) filter-and-data association. Although these elegant tracking algorithms have been proposed, visual tracking still faces too many challenges such as partial occlusion, illumination changes, pose changes, and camouflage environment, which cause the substantial performance degradations of traditional tracking algorithms, and furthermore restrict their usage in many practical applications.

However, visual tracking is a very basic functionality of human visual system (HVS). Revealing and exploiting the perception mechanism of HVS in visual tracking should be an effective way to improve the robustness of a tracking system. There are some significant findings that give us inspiration. Firstly, many psychophysical studies [1, 2] have demonstrated that feature-based visual attention—that is, the ability to improve target search by enhancing the representation of image components that are related to the target—should be particularly useful when searching for the target. Based on this finding it is possible for us to improve the robustness of a tracking system by selecting those features that are related to the target and ignoring those that are not from the view of visual attention. Secondly, the receptive fields of simple cell in visual cortex can be characterized as being spatially localized, oriented and bandpass, and it forms a sparse, distributed representation of natural images [3]. Standard methods [4, 5] can obtain such sparse representation which has similar properties with the responses of the receptive fields.

Motivated by aforementioned findings, in this paper, we propose a novel and effective algorithm that simulates the feature-based visual attention for visual tracking. The algorithm first collect a large number of training samples (patches) from natural images in advance. Using standard method, we then trained a set of general sparse basis functions that can yield a sparse representation of any natural image patch. We refer to such general sparse basis functions as features in this work. After dividing the target region into multiple patches and representing each patch as a linear combination of those trained general sparse basis functions, we compute the probability distribution of feature activities and select features that are related to the tracked target by judging whether each feature’s entropy gain is less than 0. The target is finally represented by the probability distribution of those selected features. The target search is performed by minimizing the Matusita distance measure between the distributions of the target model and candidate using Newton-style iterations.

There are also some methods that are similar to the proposed method. In [6], in order to be robust to partial occlusion, Adam et al. evenly divide the target region into multiple patches and represent each patch by a gray histogram. Each patch is used to vote on the candidate positions and the final tracking result is obtained by combining the vote maps of all patches. In [7], Yang et al. present a visual tracking algorithm that selects attentional spatial regions to represent the
target. Our proposed method is quiet different from these two methods. In [6, 7], they divide the target into multiple regions and use color distributions on selected regions to represent the target. In our work, the spatial division is just for extracting essential features underlining in patches. The target is finally represented by the probability distribution of those selected features.

The paper is organized as follows. Section 2 introduces the proposed method. Section 3 includes experimental comparisons between the proposed tracking algorithm with two mean shift based algorithms, and Section 4 concludes the paper.

2. PROPOSED METHOD

2.1. General sparse basis function extraction

In order to simulate the feature-based visual attention in visual tracking, we should choose an effective feature space. Unlike traditional methods which usually use color space, in this work, we use general sparse basis functions as features to represent the tracked target. The motivations of using general sparse basis functions as features are twofold: 1) The sparse representation is statistically independent and efficient 2) The sparse representation has similar properties with the responses of the receptive fields in visual cortex.

We first collect 132000 8 × 8 RGB image patches from natural images as training set. Using the method [4], we learn $M = 8 \times 8 \times 3$ basis functions $\mathbf{A} = (a_1, a_2, \ldots, a_M)$ and corresponding filters matrix $\mathbf{W}$. The sparse basis functions resemble short edges, and the filters are similar to receptive fields of simple cells [5]. For any RGB image patch $I$, its response $s_j$ to $j$th filter $w_j$ can be computed as

$$s_j = w_jI$$

(1)

2.2. Target representation and search

Suppose two images, with one designed as the “model image” that includes the tracked target, while the other is the “target image” in which we need to localize the target. In the model image, the target model is centered at location $x$ and divided into $N \times 8 \times 8$ RGB image patches $I_1, I_2, \ldots, I_N$ with step $l$. The divided patches can partially overlap when $l < 8$ or non-overlap when $l \geq 8$. The corresponding center point locations of $N$ patches are $x_1, x_2, \ldots, x_N$. Using the trained filters matrix $\mathbf{W}$, we can obtain a responses matrix

$$S_x = (s_{ij})_{N \times M}$$

(2)

where $s_{ij} = w_jI_i$ is the response of $i$th patch to $j$th filter. The square of the response, energy, reflects the intensity of this response. We define the normalized energy matrix

$$E_x = (\xi_{ij})_{N \times M}$$

(3)

where

$$\xi_{ij} = \frac{s_{ij}^2}{\sum_{j=1}^{M} s_{ij}^2}$$

(4)

The normalization operation makes all response energies of one patch to $M$ filters sum to 1.

With these definitions, a kernel-weighted feature distribution $q = (q_1, q_2, \ldots, q_M)^T$ of the target model can be defined as

$$q_j = \frac{1}{C} \sum_{i=1}^{N} K(\frac{x_i - x}{h}) E_x(i, j)$$

(5)

where $K$ is the kernel function, $h$ is the bandwidth, and $C$ is the normalization constant. The kernel function assigns higher weights to patches that are nearer to the target center. We choose kernel $K(x)$ with Epanechnikov profile $k(x)$. The constant $C$ is derived by imposing the condition $\sum_{j=1}^{M} q_j = 1$.

$$C = \sum_{i=1}^{N} K(\frac{x_i - x}{h})$$

(6)

Define a column vector $\mathbf{K} = (K_1)^{N \times 1}$ where $K_i = \frac{1}{h}K(\frac{x_i - x}{h})$. Then the feature distribution $q$ can also be denoted by matrix form

$$\mathbf{q} = \mathbf{E}^T \mathbf{K}$$

(7)

We define that a feature is related to the target when succeeding activations of that feature cannot increase system’s entropy. The intuition behind this definition is straightforward: From the view of signal coding, a feature is unexpected feature and needed to be coded when succeeding activations of that feature can increase system’s entropy [8]. Such features correspond to uncertain appearance changes and are not suitable to robust tracking. In contrast, the related features by our definition are relatively unchanged features (predictable features from the view of signal coding) and are best for robust tracking. In this work, we introduce the entropy gain of each feature to judge whether the feature is related to the target or not. The joint entropy of distribution $q$ is

$$H(q) = -\sum_{j=1}^{M} q_j \log q_j.$$ 

We compute the entropy gain of $j$th feature:

$$g_j = \frac{\partial H(q)}{\partial q_j} = \begin{cases} -H(q) - q_j(1 + \log q_j) - \log q_j & \text{if } q_j \neq 0 \\ 0 & \text{else} \end{cases}$$

(8)

We define the related feature indexes as set $F = \{j | g_j < 0\}$ and non-related feature indexes as set $\overline{F} = \{j | g_j \geq 0\}$. The size of set $F$ is $M'$. An example of illustrating that related features are more suitable to track than non-related features is shown in Fig. 1. We reconstruct the original image (Fig. 1(a)) with related features and non-related features respectively. The reconstructed images are shown in Fig. 1(b).
Fig. 1. The reconstructions of the original image with related features and non-related features. (a) The original image. (b) The reconstructed image with 58 related features. (c) The reconstructed image with 134 non-related features.

and Fig. 1(c). Even if with a small part of features (in our experiment, 58 related features of total 192 features), the reconstructed image reserves the most information of the original image. However, the reconstructed image with most features (134 non-related features) loses the most information of the original image. The illustration shows that representing the target with related features is more effective and suitable for tracking.

With the defined related features, correspondingly, we obtain a new filters matrix $\tilde{W}$ consisting of filters in $W$ but with indexes in set $F$. With new filters matrix $\tilde{W}$, we can re-compute the new normalized energy matrix $\tilde{E}_x$ using similar procedure as in Eq. 3. Note that the size of matrix $\tilde{E}_x$ is $N \times M'$. Then we can obtain the probability distribution of related features $\tilde{q} = (\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_{M'})^T$

$$\tilde{q} = \tilde{E}_x^T K$$  \hspace{1cm} (9)

Similarly, we can also compute the feature distribution $\tilde{p}(x + \Delta x)$ of the target candidate centered at location $x + \Delta x$ in target image

$$\tilde{p}(x + \Delta x) = \tilde{E}_{x+\Delta x}^T K$$  \hspace{1cm} (10)

where $\tilde{E}_{x+\Delta x}$ is the normalized energy matrix computed at location $x + \Delta x$ in the target image with the new filters matrix.

Once we obtain the distributions of the target model $\tilde{q}$ in the model image and the target candidate $\tilde{p}(x + \Delta x)$ in the target image, we can perform the target search by checking all target candidates and finding a candidate that best matches the target model as the tracked result. In this work, the target search is performed by minimizing the Matusita distance measure between the distributions of the target model and candidate using Newton-style iterations [9] which is more efficient than widely used mean shift method [10].

### 3. EXPERIMENTAL RESULTS

In order to verify the robustness of our tracking method, we present some tracking results on three test sequences which correspond to three kinds of challenges: partial occlusion, camouflage environments and pose changes. The proposed method is compared with the original mean shift tracker [10] and the new mean shift tracker [11]. We also introduce the quantitative criterion in term of relative distance [7] to evaluate algorithm’s performance. A perfect tracking expects the relative distance to be around 0.

The first experiment uses the walking sequence [7]. The walking person is subjected to irregular severe occlusion when passing behind the bush at 156th frame and the guidepost at 618th frame. We initialize the model with a region of size $48 \times 168$. The bandwidth is $h = 87$ and the division step is $l = 8$. As show in Fig. 2, the original mean shift tracker successfully tracks the person before 156th frame where there is no occlusion. However, it loses the target at 156th frame due to bush occlusion and then recovers the target at 394th frame. At 600th frame, it loses the target again due to the severe occlusion by the guidepost. The new mean shift tracker is more robust to partial occlusion than the original mean shift tracker and successfully tracks the person before 600th frame. However, it loses the target when the guidepost occludes the person at 600th frame. In contrast, our tracker reliably and successfully tracks the person in entire sequence which is also verified by the quantitative results in Fig. 5(a).

The camouflage environment is very challenging for visual tracking since there are similar objects around the target. We verify that our tracker is robust to camouflage background by tracking a zebra with another zebra nearby in zebra sequence. The target model is initialized with a region of size $180 \times 46$. The bandwidth is $h = 92$ and the division step is $l = 8$. As shown in Fig. 3, the original mean shift tracker and new mean shift tracker lose the right target and track the other zebra when the right target passes the other zebra at 104th frame. Our tracker accurately tracks the right target. The reason for this is that two mean shift based methods both use color RGB as features to represent the target. When nearby background presents similar color, the trackers cannot accurately localize the target. However, our tracker represents the target with features that are related to the target. Such representation can effectively discriminate the target with background even if in camouflage environment. The quantitative comparison is shown in Fig. 5(b) from which we can see that our tracker gets the lowest relative distances for entire sequence than comparison methods.
Fig. 3. Tracking results on zebra sequence for frame #8, 80, 104, 134 and 148. The first row are the results obtained by out tracker. The second and third rows are the results obtained by the original mean shift tracker and new mean shift tracker respectively.

Fig. 4. Tracking results on face sequence for frame #1, 16, 50, 96 and 208. The first row are the results obtained by out tracker. The second and third rows are the results obtained by the original mean shift tracker and new mean shift tracker respectively.

In the third experiment, the sequence face involves strong pose changes which is very challenging for tracking. We initialize the target model of the frontal face with a region of size $76 \times 84$. The bandwidth is $h = 56$ and the division step is $l = 4$. As shown in Fig. 4, the original mean shift tracker tracks the background region when the head turn left or right. Our tracker achieves better tracking results than the original mean shift tracker and comparable tracking results with the new mean shift tracker. The quantitative comparisons are shown in Fig. 5(c) which also verifies that our tracker is superior to the original mean shift tracker.

We also measured the processing speed of the proposed algorithm. On walk sequence which consists of 671 frames of size $320 \times 240$, a frame rate of 27 fps was achieved on a standard PC with a 3.0 GHz processor and 1 GB memory. This makes the algorithm well-suited to real-time applications.

4. CONCLUSION

In this paper, we propose a novel and robust visual tracking algorithm by simulating the feature-based visual attention mechanism of human visual system. There are three innovations in this work. First, unlike previous methods that usually use color features, we extract general sparse basis functions as features. The obtained features are statistic independent and have similar properties with the responses of the receptive mechanisms of human visual system. There are three innovations in this work. First, unlike previous methods that usually use color features, we extract general sparse basis functions as features. The obtained features are statistic independent and have similar properties with the responses of the receptive fields of simple cell. Second, we represent the target just with those features that are related to the target and ignore those that are not. Such features selection mechanism provides the ability to be robust to partial occlusion and effectively discriminate the target from the background. Finally, by optimizing the object function using Newton-style iterations, the proposed tracking algorithm is computationally efficient. The comparison results between the proposed algorithm and two mean shift based tracking algorithms on three challenging sequences verify the effectiveness of the proposed algorithm.

5. REFERENCES


