Discriminative feature selection for visual tracking

Junkai Ma¹,²,³, Haibo Luo¹,³, Wei Zhou⁴, Yingchao Song¹,²,³, Bin Hui¹,²,³ and Zheng Chang¹,³

1 Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China
2 University of Chinese Academy of Sciences, Beijing, China
3 Key Laboratory of Opto-Electronic Information Processing CAS, Shenyang, China
4 AVIC Jiangxi HONGDU Aviation Industry Group LTD, Nanchang, China

E-mail: junkai.ma@hotmail.com

Abstract. Visual tracking is an important role in computer vision tasks. The robustness of tracking algorithm is a challenge. Especially in complex scenarios such as clutter background, illumination variation and appearance changes etc. As an important component in tracking algorithm, the appropriateness of feature is closed related to the tracking precision. In this paper, an online discriminative feature selection is proposed to provide the tracker the most discriminative feature. Firstly, a feature pool which contains different information of the image such as gradient, gray value and edge is built. And when every frame is processed during tracking, all of these features will be extracted. Secondly, these features are ranked depend on their discrimination between target and background and the highest scored feature is chosen to represent the candidate image patch. Then, after obtaining the tracking result, the target model will be update to adapt the appearance variation. The experiment show that our method is robust when compared with other state-of-the-art algorithms.

1. Introduction

Visual tracking plays an important role in various computer vision tasks such as video surveillance, action recognition, traffic moniton etc. Although a variety of tracking algorithms are proposed during recent decades, the robustness of the tracking is still a challenge. This is mainly caused by the illumination variation, clutter background, pose changes, scale variation et al in the tracking processing. To solve these problems, it is necessary to design robust and efficient tracking algorithms.

The existing tracking algorithms can be categorized into two groups in term of their appearance model. One is called generative model [1, 2], which construct an appearance model based on the target appearance information, and track the target by finding the most similar candidate in the every frame of the video. Mei et al. proposed a tracking method with sparse model and represented the target by a template dictionary [2]. Ross et al. present a method that learns an incrementally a low-dimensional subspace as the representation of the target, this method can copy with the appearance changes during tracking [1]. The other is called discriminative model [3, 4], which consider the tracking as classification problem to distinguish the target from the background. Henriques et al. exploit the property of circulant matrices and proposed an algorithm called kernelized correlation filter (KCF) [3]. KCF learned a regressive function in frequency domain and incorporate multi-channel feature what makes the algorithm efficient and robust. Hare et al. employed a structured support vector machine as the classifier
to distinguish the target and background [4]. This structure output classifier can output the accurate position of the target rather than the positive and negative label of the image patch, so it attained a preferable tracking result.

Most algorithms proposed recently are focus on the observation model, which determines whether a candidate is the target or the background. As an important component in tracking algorithm, the feature used to represent the target is also closely related to the performance of the tracking algorithm. The proper feature can enhance the performance of the tracker. On the contrary, improper feature will degrade the performance of the tracker. Numerous works studied the feature selection in tracking [5, 6, 7, 8]. In this paper we employed an on-line feature selection mechanism to choose the most discriminative feature during the tracking processing to make the tracker more accurate.

2. Proposed method

2.1. Algorithm overview

An overview of the propose method is shown in Fig. 1. As the common implement in tracking processing, the target is selected in the 1-st frame and the tracking result is given by algorithm iteratively frame by frame. When the \( t \)-th frame is coming, the search region is assigned around \( l_{t-1} \), the target location in the previous frame. Firstly, some candidate is sampled in the search region and the most discriminative features, which is provided by the previous frame, are extracted form each candidate. Then, the confidence score of each candidate is calculate by the classifier according to the target appearance model. The image patch which has the highest confidence score is considered as the target. Then some positive and negative samples are extracted. At last by using the negative and positive examples, the most discriminative feature is chosen according to the feature selection mechanism to represent the target in the next frame and the target appearance model can also be update depended on these samples.

The MIL tracker[9] is the baseline of the proposed method. Following by the MIL, a strong
classifier is build to discriminate target from the background. The strong classifier consist of several weak classifier

$$H(x) = \sum_{i=1}^{K} h_i(x),$$

where $x$ is the sample. Each weak classifier is related to a feature instance in the feature pool. The detail of the proposed method is shown in the Algorithm 1.

**Algorithm 1** Proposed Tracking Algorithm

1. Point the location target manually in the 1-st frame $l_0$.
2. Sample the image patch sets $S^+ = \{x | \|l_x - l_0\| < \alpha\}$, $S^- = \{x | \beta < \|l_x - l_0\| < \gamma\}$ as the positive and negative sample set respectively.
3. Establish the $n$ weaker classifier $h_k, i = 1, 2, ... n$ for each $f_i$ in the feature pool $\mathcal{F}$.
4. Top $k$ most discriminative features are be selected based on fisher criterion.
5. for $i = 2$ to end of the sequence do
6. Sample some candidate image patches in the search region near the $l_i$.
7. Extract the top $k$ features of every candidate.
8. Use the weak classifiers which are related to the $k$ feature(selected by the last step) to constitute the strange classifier $H^i = \sum_{i=1}^{k} h_k$
9. Sample the image patch sets $S^+ = \{x | \|l_x - l_i\| < \alpha\}$, $S^- = \{x | \beta < \|l_x - l_i\| < \gamma\}$ as the positive and negative sample set respectively.
10. Create the feature pool on the sample set. Rank the features in the feature pool and select the top $k$ most discriminative feature for next tracking.
11. Update the all of the $n$ weak classifiers.
12: end for

2.2. Feature pool

Different kind of feature has variant performance to represent the image according to different scenario. So more the diverse the feature is, more robust the feature can represent the target. In our algorithm, the feature pool consist of the gradient information(HoG) [10], local contrast information (Haar) [11], which shows good performance in studies [4, 12] in recent years.

In order to make the HoG and the Haar features fit to the tracking algorithm, we use the modified HoG and Haar feature proposed by the studies [12, 13, 8]. In traditional HoG, the histogram of gradient is summed up in a small rectangle, which is called cell, in the image patch. The feature of the holistic image patch is composed by the histogram in a set of overlapping cell. Fig. 2(a) show the detail. In our method we replace this regular cell selection with a randomly selection method, which is shown in the Fig. 2(b). By this method, the histogram of each random cell is considered as feature instance of the image patch. For Haar feature, the base implement is shown in Fig. 3(a), each rectangle indicates a type of Haar feature, whose value is the weighted sum of the pixel value in the rectangle. The pixels in the white rectangle have the weight of 1 and those in the black rectangle have the weight of -1. For a image cell, $p$ types of Haar feature can be obtained. As the same in the HoG feature, the Haar feature of whole region is represented by a set of overlapping cells. In this case the types of feature is fixed. We also using the randomly selected cells to replace them. As shown in Fig. 3(b), a Haar feature instance is composed of several sub-rectangles, which are randomly selected in the target image
patch. The weight of each sub-rectangle is also a random value between -1 and 1. Every feature element is the weighted sum of the pixel value in those sub-rectangle.

Figure 2. (a) The locations of cells in the original HoG are distributed in a regular pattern. (b) The locations of cells in our method are distributed in a random pattern.

Figure 3. (a) An example of the original Haar feature. Five typical types of the Haar feature are listed in the left. Applied these type of feature in every small region in the target bounding box and obtain a feature vector. (b) The type of Haar feature, the location and the weight of every small patch, is designed in a random way. A group of rectangles have same color belong to a type of Haar feature.

Calculating these features is a time consuming processing in tracking algorithm. The computing time is increasing along with the growing of the feature regions number, while a large number of the feature region is necessary to guarantee the diverse of the feature. So, a fast feature extraction algorithm is needed. In [14] a general framework, which is called integral channel, is proposed to fast calculate the region-based feature. By using the integral channel, both Haar feature and HoG feature can be computed efficiently.

Given an input image $I$, a channel of $I$ is the output of a channel generation function $\Omega$, which is a linear or non-linear transformation of the input. So the channel of $I$ is defined as $C = \Omega(I)$. The widely used channels are R-G-B channels, Gabor filter channel, and DoG (Difference of Gaussian) filter channel etc. An integral image of these channels can be used to speed up the computation of the region based feature. The HoG feature is a weighted histogram where bin index is determined by the gradient angle and weight by gradient magnitude, so the channels are given by $Q_\theta(x, y) = G(x, y) \cdot 1(\Theta(x, y) = \theta)$, where $G(x, y)$ and $\Theta(x, y)$ are gradient magnitude and quantized gradient angle, respectively. $1(\cdot)$ is the indicate function. Based on this way, integral channels can be used to simplify the computation of the feature. The histogram of a given region can be obtain by plus or minus the value of four corners in every integral channel. The Haar feature can be obtained in the same way.
2.3. Discriminative feature ranking

By giving the feature pool of the samples set. The discrimination of each feature is measured by fisher criterion in a supervised way. We labeled the target as the positive and background as negative. The i-th feature extracted from target and background is defined as $f_{i,+}$ and $f_{i,-}$ respectively. The fisher discrimination score is given by:

$$p_i = \frac{|\mu_{i,+} - \mu_{i,-}|^2}{\sigma_{i,+}^2 + \sigma_{i,-}^2}$$  \hspace{1cm} (2)

where $\mu_i$ and $\sigma_i$ are the mean and standard deviation of i-th feature. For the total number of $n$ features, the fisher discrimination score is calculate by Eq. 2. Then those features are ranked based on their fisher discrimination score. In the end, the top $k$ most discriminative feature is selection to represent the target.

3. Experiment

We experiment our method on several sequences in the OTB-50 dataset[15]. There are some details of our experiment. We set the total number of the feature to $n = 100$ and using the top $k = 20$ most discriminative feature to represent the target. We test our algorithm on five typical sequences, david3, subway, suv, coke and trellis in the dataset.

In order to shown performance of the proposed method, we compare the result with other 4 state-of-art tracker ASLA[16], L1APG[17], IVT[1] and MIL[9]. The tracking results on the 5 test sequences are shown in Fig. 4 to Fig. 8. From the result, we can find that our method is more robust than the compared methods. In the trellis sequence, when the illumination changes our tracking can catch the target correctly. In the coke and david3 sequence the target undergoes deformation and in-plane rotation, our method can still tracking the target. In the suv and subway sequences, the target have some slightly occusion, the result shown that our method can handle this kind of challenge.

![Figure 4. Tracking results comparison on Trellis sequence](image)

The ground truth of the target location is given in the OTB-50 dataset. For evaluate the propose method more accurate, we plot the center location error (CLE) of each tracker on every sequence in Fig. 9. The CLE is widely used to evaluate the performance of tracker. If the ground truth location $l_g$ and the tracking result $l_t$ are given, the CLE is defined as:

$$CLE = d(c_g, c_t), \hspace{1cm} (3)$$

in which the $d(\cdot, \cdot)$ is the Euclidean distance and $c_g, c_t$ are the center of the ground truth and tracking result, respectively.
From the Fig. 9, we can conclude that our method has the best result among the compared trackers. Even in the end of each sequence, other trackers undergo heavily drift while our method has low center location error. Table 1 shows the average center location error of all the 5 trackers in every sequence.

4. Conclusion
In this paper, we propose a robust tracking algorithm which can select the discriminative features dynamically during the tracking. A rich feature pool which contains different information of the image is built. A features ranked method is used to choose the most discriminative ones between target and background. The experiment shows that our method outperforms some state-of-the-art algorithms in some challenge tracking scenarios.
Figure 8. Tracking results comparison on subway sequence

Figure 9. The plot of center distance error on the 5 sequences

Table 1. Average center location error of 5 trackers on 5 sequences

<table>
<thead>
<tr>
<th></th>
<th>ASLA</th>
<th>IVT</th>
<th>MIL</th>
<th>L1APG</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trellis</td>
<td>7.6</td>
<td>119</td>
<td>71</td>
<td>62</td>
<td>7.7</td>
</tr>
<tr>
<td>coke</td>
<td>60</td>
<td>82</td>
<td>46</td>
<td>50</td>
<td>18</td>
</tr>
<tr>
<td>david3</td>
<td>87</td>
<td>51</td>
<td>29</td>
<td>90</td>
<td>4.5</td>
</tr>
<tr>
<td>subway</td>
<td>138</td>
<td>130</td>
<td>7.5</td>
<td>147</td>
<td>2.9</td>
</tr>
<tr>
<td>suv</td>
<td>74</td>
<td>57</td>
<td>82</td>
<td>93</td>
<td>3</td>
</tr>
</tbody>
</table>

References


