Robust object tracking based on sparse representation

Shengping Zhang, Hongxun Yao, Xin Sun and Shaohui Liu

{spzhang,yhx,xsun,shaohl}@vilab.hit.edu.cn

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ABSTRACT

In this paper, we propose a novel and robust object tracking algorithm based on sparse representation. Object tracking is formulated as an object recognition problem rather than a traditional search problem. All target candidates are considered as training samples and the target template is represented as a linear combination of all training samples. The combination coefficients are obtained by solving for the minimum $l_1$-norm solution. The final tracking result is the target candidate associated with the non-zero coefficient. Experimental results on two challenging test sequences show that the proposed method is more effective than the widely used mean shift tracker.

Keywords: Object tracking, sparse representation, $l_1$-norm minimization,

1. INTRODUCTION

Object tracking in the computer vision community usually refers to the efforts of consistently finding the motion state of the tracked target in consecutive video frames. In other words, given a target template in advance, the goal of object tracking is to obtain the position and size of the target in current frame by checking all target candidates and find the one which is the most similar with the target template. It is used in many practical applications, such as automated surveillance, video analysis, human computer interfaces and vehicle navigation and so on.

In the previous literature, numerous tracking algorithms have been proposed and they usually fall into two major groups: deterministic methods and stochastic methods. In deterministic methods, the target object is located in current frame by maximizing the similarity between the target model and a target candidate. Kernel-based tracking algorithms have obtained increasing interests due to they are simple and effective. The feature histogram based target representations are regularized by spatial masking with an isotropic kernel. The masking induces spatially-smooth similarity functions suitable for gradient-based optimization, hence, the target localization problem can be formulated using the basin of attraction of the local maxima. Employing a metric derived from the Bhattacharyya coefficient as similarity measure, the mean shift procedure is used to perform the optimization.

In stochastic methods, the widely used particle filter is a Bayesian sequential importance sampling technique for estimating the posterior distribution of state variables characterizing a dynamic system. It provides a convenient framework for estimating and propagating the posterior probability density function of state variables regardless of the underlying distribution. It consists of essentially two steps: prediction and update. Especially, it maintains multiple hypotheses in the state space and assigns each hypothesis a weight reflecting the probability of observing the target on the hypothetical position. Then the posterior probability density function of the target state is estimated with these hypotheses and corresponding weights.

For object tracking, the target appearance model is very important for robust tracking. In two kinds of tracking methods, the target appearance is usually modeled as color histogram. However, color information is not enough to discriminate the target from background, especially in complex background. An effective appearance model that combines texture and color information is proposed for face tracking. A mixture model of three components with an online EM algorithm is proposed to model the appearance variation during tracking. This algorithm selects most discriminative feature to model the appearance and achieve impressive performance.

In fact, object tracking is a search problem which search the target template in all target candidates. The final result is one of the target candidates such that it has the most similar appearance with the target template.
From the view of pattern recognition, object tracking can also be considered as a recognition problem which recognizes the target template from the library that consists of all target candidates. Recently, object recognition algorithms based on sparse representation have achieved great successes. Specially, face recognition via sparse representation\(^{11}\) has obtained significant performance improvement because of its two excellent properties: 1) The choice of features is no longer critical, what is critical, however, is whether the number of features is sufficiently large and whether the sparse representation is correctly computed; 2) This framework can handle errors due to occlusion and corruption uniformly by exploiting the fact that these errors are often sparse with respect to the standard (pixel) basis. These two properties are also very important for object tracking. Some previous tracking algorithms\(^{12}\) try to learn features that can discriminate tracked target from background. On the other hand, partial occlusion is very challenging for object tracking because the appearance information of the target will lose in the presence of occlusion. There are also some previous methods trying to handle occlusion.\(^{13,14}\)

In this paper, motivated by the success of sparse representation in object recognition,\(^{11}\) we proposed a novel and robust object tracking algorithm. Be differ from previous algorithms, we consider object tracking as a pattern recognition problem. The target template is used as the test sample, and all target candidates are used as training samples. Recognizing the test sample from training samples is based on sparse representation. Specially, the test sample is represented by a linear combination of all training samples. The combination coefficients are obtained by choosing the minimum $l_1$-norm solution. The final tracking result is the target candidate associated with the non-zero coefficient.

The proposed algorithm is motivated by successful applications of sparse representation in computer vision. There are also some relevant works including background subtraction,\(^{15}\) media recovery,\(^{16}\) texture segmentation,\(^{17}\) and lighting estimation,\(^{18}\) etc.

The paper is organized as follows. Section 2 introduces the proposed method including problem formulation and proposed solution based on sparse representation, respectively. Section 3 includes experimental comparisons between the proposed tracker and traditional mean shift tracker, and Section 4 concludes the paper.

### 2. PROPOSED METHOD

This section describes the proposed algorithm for object tracking based on sparse representation. Since the proposed method considers object tracking as an object recognition problem which is significantly different from traditional tracking algorithm, we first formulate the problem, and then give the detailed solution to this problem.

#### 2.1 Problem Formulation

In the context of object tracking, given a target template in advance, the goal of object tracking is to find the state of the target in consecutive video frames. The state usually consists of position and size of the target. In this paper, we assume the size of the target keeps unchanged. Therefore the state is just the position of the target. In the previous literature, a main kind of tracking algorithms are deterministic methods which search the state of the target in the current frame from all the candidate positions. A candidate position is in a local neighbor of the position of the target in the last time. In this paper, be different from traditional methods, we consider the target tracking in the context of object recognition and will formulate it below.

Let $y$ denote the target template of size $M \times N$ and $x_{t-1}$ the target state in the time instant $t - 1$. Given a search radius $R$, in the current time $t$ we can obtain $L$ target candidate states $x_{1}^{t}, x_{2}^{t}, \ldots, x_{L}^{t}$ in the local neighbor centered on $x_{t-1}$. An example of training samples of tennis sequence is show in Fig. 1. Let $y_{i}^{t} \in \mathbb{R}^{d}(d = M \times N)$ denote the feature vector of the candidate region centered on $x_{i}^{t}$ with the same size with the target template. The problem of tracking target template $y$ in the time instant $t$ can be considered as an object recognition problem that recognizes a given object $y$ from all training samples $\{y_{1}^{t}, y_{2}^{t}, \ldots, y_{L}^{t}\}$. From the view of pattern recognition, all training samples can be divided into two classes: the first class consists of only the right target that has the most similar appearance with the target template and the second class consists of all the rest of training samples. The goal of object tracking is to find the label of the right target in entire training set.
2.2 Proposed Solution Based on Sparse Representation

We have formulated object tracking as an object recognition problem, then we will give a novel solution to this problem. Object recognition methods based on sparse representation have gained increasing interests due to their significant successes in the field of face recognition. In this paper, we proposed a novel approach to object tracking based on sparse representation.

When given a search radius $R$, we can obtain a large number of target candidates. If we assume that the target template and all target candidates lie on a special low-dimensional feature space, often called a manifold, then the target template can be represented as a linear combination of all target candidates:

$$y = \alpha_1 y_1^t + \alpha_2 y_2^t + \cdots + \alpha_L y_L^t. \quad (1)$$

All training samples are arranged as columns of a matrix $A = [y_{i-1}^t, \ldots, y_{L-1}^t] \in \mathbb{R}^{d \times L}$. Then, the linear representation of $y$ can also be rewritten in matrix form as

$$y = A\alpha, \quad (2)$$

where $\alpha = [0, \ldots, 0, \alpha_i, 0, \ldots, 0]^T \in \mathbb{R}^L$ is the coefficient vector whose entries are zero except one having the most similar appearance with the target template.

As the entries of the vector $\alpha$ encode the corresponding relation between the target template $y$ and target candidates $\{y_1^t, y_2^t, \ldots, y_L^t\}$, it is tempting to attempt to obtain it by solving the linear system of equation $y = A\alpha$. Because the linear system $y = A\alpha$ is usually underdetermined, its solution is not unique. Recent
development in the emerging theory of sparse representation and compressed sensing shown that this problem can be solved by sparse solution via $l^1$-minimization:  
\[
\hat{\alpha} = \arg \min \|\alpha\|_1 \quad \text{subject to} \quad y = A\alpha.
\] (3)

This problem can be solved in polynomial time by standard linear programming methods. Once the vector $\hat{\alpha}$ is obtained, we can find the index of the non-zero coefficient and the target candidate associated with the index is used as the tracking result. For example, if the index is $i$, then the tracking result at the current time instant $t$ is obtained as
\[
x_t = x^i_t
\] (4)

The whole algorithm is show in Alg. 1.

**Algorithm 1**: Robust object tracking based on sparse representation

**Input**: The target state in the time instant $t-1$: $x_{t-1}$, and a given a search radius: $R$

1. Obtain $L$ target candidate states $x^1_t, x^2_t, \ldots, x^L_t$ in the local neighbor centered on $x_{t-1}$.
2. Obtain $L$ training samples $\{y^1_t, y^2_t, \ldots, y^L_t\}$
3. Compute linear representation coefficients $\hat{\alpha}$ based on Eq. 3
4. Find the index $i$ of the non-zero entry from vector $\hat{\alpha}$

**Output**: The tracking result in the current time $t$ is:
\[
x_t = x^i_t
\]

### 3. EXPERIMENTAL RESULTS

In order to verify the effectiveness of the proposed method, we conduct experiments on two challenging video sequences which involve fast moving and pose changes. The proposed method is compared with the state-of-the-art algorithm: mean shift tracker. The comparison is done based on qualitative evaluation by looking at tracked results provided by the algorithms, and quantitative evaluation in term of relative distance. The relative distance is defined as the normalized distance

\[
\text{Relative Distance} = \sqrt{\left(\frac{x - x_0}{s_x}\right)^2 + \left(\frac{y - y_0}{s_y}\right)^2}
\]

where $(x, y)$ is the location of the tracked target, $(x_0, y_0)$ and $(s_x, s_y)$ are the location and size of the ground truth. A perfect tracking expects the relative distance to be around 0. In all experiments below, the feature vector $y^i_t$ of $i$th candidate region is the gray image of size of $10 \times 10$ which is obtained by downsampling the candidate region into $10 \times 10$. The search radius $R$ for all experiments is set to be 30.

The first experiment is conducted on the *tennis* sequence in which the tennis moves very fast. This tennis sequence is very challenging for mean shift tracker because the offset between consecutive two frames is very big. Due to the mean shift vector is less then the offset, mean shift tracker will lose the target. As shown in Fig.2, mean shift tracker failed to tracking the tennis at 17th, 32th and 47th. However, our tracker successfully tracked the target in the entire sequence. It should be noted that the mean shift tracker successfully tracked the target at 28th frame because the tennis just fell into the search range of mean shift procedure. The quantitative comparison on this *tennis* sequence is also shown in Fig.3, which also verifies that our tracker significantly outperforms the mean shift tracker.
Figure 2. Qualitative comparison on tennis sequence. The first row are the tracking results by mean shift tracker at 11th, 17th, 28th, 32th and 47th frames. The second row are the corresponding results obtained by the proposed method.

Figure 3. Quantitative comparison on tennis sequence.
Figure 4. Qualitative comparison on person sequence. The first row are the tracking results by mean shift tracker at 35th, 53th, 57th, 89th and 94th frames. The second row are the corresponding results obtained by the proposed method.

Figure 5. Quantitative comparison on person sequence. The chart shows the relative distance over time for the two tracking methods.
The second experiment is conducted on the person sequence which involves pose and illumination changes. As shown in Fig. 4, the mean shift tracker cannot obtain accurate tracking results in the entire sequence. However, our tracker successfully tracked the target in the entire sequence. The main reason for this is that when there are pose and illumination changes, object appearance cannot supply enough information for the mean shift tracker. For the proposed method, due to the use of sparse representation, our tracker is robust to such appearance changes, which is also verified in face recognition. The quantitative comparison on this person sequence is also shown in Fig. 5, which also verifies that our tracker is significantly superior to the mean shift tracker.

We should point out that in our matlab implementations, our tracker is slower than the mean shift tracker due to some matrix computation in our method. However, using high-performance computers, our tracker can also achieve real-time processing speed. On the other hand, the main contribution of this paper lies in its novelty that it is the first time to use sparse representation in the field of object tracking.

4. CONCLUSION
In this paper, we propose a novel and effective object tracking algorithm which is significantly different from traditional tracking algorithms. It considers object tracking as a object recognition problem which goal is to recognize the target template from all target candidates. It exploits sparse representation, a popular technique in the field of object recognition in recent, to give the solution to this problem. To be our best knowledge, it is the first time for us to exploit sparse representation in object tracking. Experimental results on two challenging test sequences verify that the proposed method is significantly superior to widely used mean shift tracker.

5. ACKNOWLEDGMENTS
This work is supported by the National Natural Science Foundation of China (60775024), National Basic Research Program of China (2009CB320906).

REFERENCES


