A FAST OBJECT TRACKING APPROACH BASED ON SPARSE REPRESENTATION

Zhenjun Han, Jianbin Jiao+, Qixiang Ye

Graduate University of Chinese Academy of Sciences, Beijing, 100049, China
+Corresponding Author: Fax: +86-10-88256278, Email:jiaojb@gucas.ac.cn

ABSTRACT

This paper proposes a new approach based on object sparse representation (OSR) for object tracking. The OSR method implemented by L1-norm minimization is robust to the partial occlusion and deterioration in object images. Firstly, we dynamically construct a set of samples in a predicted searching window in a new video frame, on which the sparse representation of the tracked object can be calculated by the OSR method. This procedure can automatically select the subset of the samples as a basis which most compactly expresses the object with small residuals and rejects all other possible but less compact representations. In terms of this sparse and compact representation, the instantaneous tracking result is achieved in the new video frame. Extensive comparative experiments demonstrate the effectiveness of the proposed approach especially in occlusion context.

Index Terms— Object tracking, sparse representation, L1-norm minimization

1. INTRODUCTION

Visual object tracking is to automatically find the same object in adjacent frames from a video sequence after the object's location is initialized. It plays an important role in many video based applications, such as human computer interaction systems, intelligent surveillance and robotics [1] etc.

The previous research on object tracking has fallen into three different categories: motion models, searching methods and object representation.

Motion models improve the tracking stabilization by predicting the object's location in a new frame based on its historical motion characteristics. Early works used a Kalman filter [2] to provide solutions that are optimal for a linear and Gaussian model. The particle filter, also known as the sequential Monte Carlo method, has been applied to tracking problems under the name Condensation [3].

Searching methods use various matching strategies to find object's location in a new video frame. Robust and fast similarity measure such as mean-shift algorithm [4] has been applied to find the optimal solution for tracking.

Object representation is another key part of tracking. Color [5], contour [6] and feature point [7] features are employed to represent the object. In addition, there have been enormous efforts on finding the "optimal" features for tracking by discriminating the object with its background. In [8], an appearance-adaptive model is incorporated in filter frameworks to realize adaptive visual tracking.

Based on modern investigation and study in human vision system (HVS), sparse representation of an object has been brought out [9], where Wright et al. showed that using parsimony as the principle for choosing a limited subset of models from a model set is more effective for object representation. Considering the basic issue of a tracking problem is to locate a specific object in a searching window in a new frame, it is reasonable to make a hypothesis that the tracked object can be represented as a linear superposition of the samples just inside the searching window. In addition, the searching window is always much larger than the tracked object region, leading to a sparse coefficient vector of the linear superposition, since coefficients corresponding to samples obtained from background (named negative samples) tend to be zeros. Inspired by the basic issue of tracking and the work of [9], we cast the object tracking as finding a sparse representation of the object based on a sample set.

In this paper, our method is different from [10] mainly in the construction of the sample set and the location scheme of the object in current frame. We dynamically construct the set based on the instantaneous tracking result, which can ensure the adaption to the appearances variations of the object and its background in current tracking frame and lead to more stable tracking results. In addition, we cast the samples with their coefficients as an estimator of the object's distribution in the searching window. Comparisons in details between the above two tracking approaches will be given in section 3.

The rest of the paper is organized as follows. Details of our approach are described in section 2; Experiments are given in section 3, and conclusions in section 4.

2. OBJECT SPARSE REPRESENTATION FOR TRACKING

The key of object tracking is to find the object location (x,y) and its scale s in a searching window of a new video frame. The flow chart of our tracking algorithm is shown in Fig.1. We will discuss details of the approach in section 2.1, 2.2,

and 2.3.

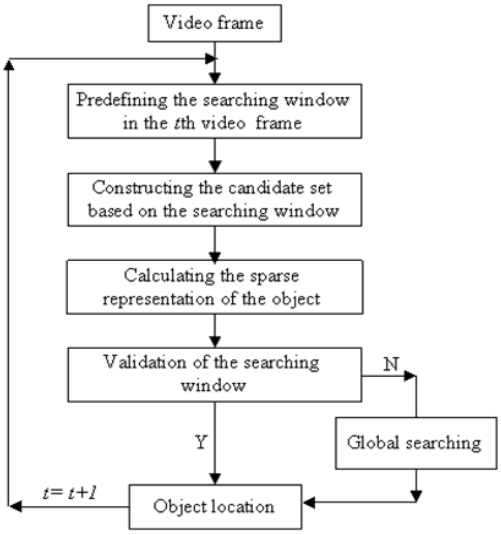


Fig. 1. Flow chart of the proposed tracking algorithm.

2.1. Sample set construction

A sample set is dynamically constructed in a new tracking frame based on a searching window (the black rectangle in Fig.2a), which is a rectangle of size WxH surrounding the previous tracking result. This window can be determined with a Kalman filter method with constant velocity motion model [2]. Each sample in the set is defined as a sub-window of the searching window (the red rectangle in Fig.2b). A sample rectangle in the window is specified by $r=(x,y,s)$ with $0<x<W$, $0<y<H$, $s>0$. This sample set is almost infinitely large. For practical reasons, it is reduced as follows:

1. The (x,y) varies with the step of n pixels in horizontal and vertical orientations;

2. The s is uniformly sampled from 0.8 to 1.2 times of the tracked object's size, when (x,y) is fixed.

These restrictions lead to reasonable number of samples in the set. Supposing that we totally obtain K samples for constructing the set $\{S_t^i, i=1...K\}$ for the tth video frame, where most of the samples are associated with the background, and a few of them are parts of or the whole object (shown in Fig.2c). When the set is fixed, a composite visual feature set (HOGC [8]) A_t^i, which is a 120 dimension vector including both color and gradient histograms, is extracted to descript its corresponding sample S_t^i. HOGC can capture both the color and contour characteristics. Then we can obtain a feature set $A_t=\{A_t^i, i=1...K\}$ for all the samples at frame t.

2.2. Object sparse representation (OSR)

Since the key issue of a tracking problem is to locate a specific object in a searching window, therefore, when we obtain

Fig. 2. (a) The searching window of the object. (b) A sample. (c) Examples in the sample set.

the sample set at frame t, the tracked object can be formally represented as

$$A_t\psi_t \approx F \tag{1}$$

where F is the feature vector of the tracked object in the composite feature space, $\psi_t=\{\psi_t^i, i=1...K\}$ is a coefficient vector associated with A_t and F, and ψ_t^i is the coefficient of the ith sample in the set at frame t. Although above model can also be complex ones, we shall first assume that a linear system is considered in our paper from both efficiency requirement of a real application and ease of representation.

In a real tracking condition, the searching window is always much larger than the tracked object region, therefore there are finite samples which contain the characteristic of the tracked object and play roles in the linear superposition, supposing there are r nonzero coefficients in $\psi_t=\{\psi_t^i, i=1...K\}$ we can reasonably infer that $r \ll K$. In this case, we say that the object has an r-sparse representation on the sample set. As the set contains some redundant samples, representing objects in parsimony way becomes an important step for a good tracking system. Minimizing L0-norm is the principle to obtain a sparse representation, which is, however, a *NP*-hard problem. Recent development in the theory of compact sensing shows that the solution of L1-norm minimization subject to a linear system of the samples can be used to find sparse enough representation of the object. In terms of the set A_t and F, a sparse representation is computed as follows:

$$\min ||\psi_t||_1, \quad subject\ \ to \quad A_t\psi_t = F \tag{2}$$

where $\|\cdot\|_1$ represents the L1-norm.

Since real images are noisy, it may not be possible to exactly express the object directly with a sparse representation of the samples. To model the noise in the video frame, we empirically consider a noise term as ε=0.1, therefore the Eq. (2) is modified as follows:

$$\min ||\psi_t||_1, \quad subject\ \ to \quad ||A_t\psi_t - F||_2 \le \varepsilon \tag{3}$$

where $\|\cdot\|_2$ represents the L2-norm. This model can be solved in polynomial time by a linear programming or quadratic programming method [11].

By solving Eq. (3), the vector of r-sparse coefficients can be obtained. An example of the coefficient vector is given in Fig. 3, where we use 100 samples to calculate the sparse

representation of a tracked object. It can be seen from Fig. 3c that there are about 10 coefficients of the vector are nonzero, showing the high sparsity.

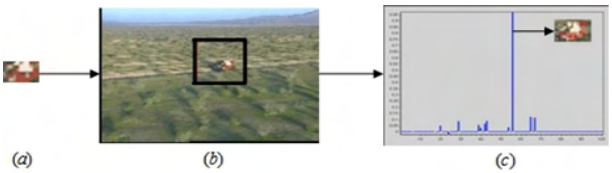


Fig. 3. (a) The tracked object. (b) The searching window in the video frame. (c) The sparse coefficient vector.

2.3. Validation of the searching window and Object tracking based on OSR

Validation of the searching window or the object is another critical problem in the visual object tracking. In the dominant approaches, the residuals between the object with the instantaneous tracking result are used for validation. In [9], they show the coefficients in the vector ψ_t=$\{\psi_t^i, i$=1...$K\}$ are better statistics for validation than the residuals. In the proposed tracking approach we effectively determine the validation of the searching window based on the coefficient vector. Supposing ψ_t is the sparse solution of Eq. (3) at frame t, we define the sparsity ratio (SR) of ψ_t as follows:

$$SR(\psi_t) = \frac{\sum_i \delta(\psi_t^i)}{K}, \tag{4}$$

$$\delta(\psi_t^i) = \begin{cases} 1, & if|\psi_t^i| > 0.1 \\ 0, & otherwise \end{cases} \tag{5}$$

We choose a threshold τ=0.2 and accept the searching window at frame t as valid if $SR(\psi_t) < \tau$ and reject as invalid otherwise shown in Fig. 4d.

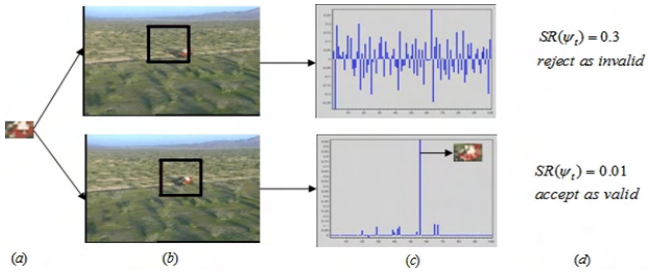


Fig. 4. (a) The tracked object. (b) The searching window in the tth video frame. (c) The sparse solution. (d) The validation of the searching window.

When we accept the searching window of the current frame as valid, the object sparse representation (OSR) calculated by L1-norm minimization can automatically select samples which most compactly express the object. Therefore, samples with bigger coefficients always appear more representative to the tracked object which is shown in Fig.3c. To track the object, we propose an approach using the samples with their coefficients as an estimator of the distribution of the object in the searching window. Therefore, the instantaneous object position and scale at frame t can be calculated as follows:

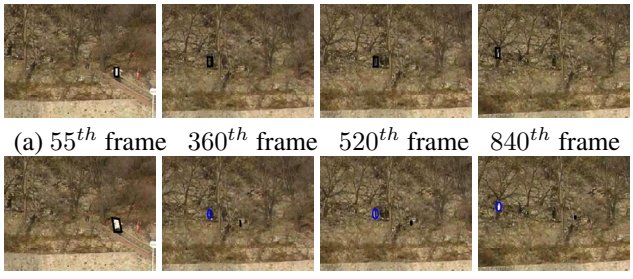
$$O_t(x,y,s) = \sum_i S_t^i(x,y,s) * \psi_t^i \tag{6}$$

where $O_t(x,y,s)$ is the instantaneous tracking result at frame t with location (x,y) and scale s, $S_t^i(x,y,s)$ is the ith sample at frame t.

3. EXPERIMENTS

In this section, experiments with comparisons are carried out to validate the proposed approach. The experimental videos are from VIVID, CAVIAR and SDL data set [12]. The objects include moving humans and vehicles. Two challenging tracking examples shown in Figs. 5 and 6 are used to compare the proposed tracking approach with the method in [10].

The first video in Fig. 5 from the SDL data set includes serious object occlusions (360^{th} and 840^{th} frames) and appearance variations (520^{th} frame). Our approach can track the object robustly; while the method in [10] loses the object and the tracking window quickly degenerates to a small point (360^{th}, 520^{th} and 840^{th} frames). The tracking results of this video show that the proposed approach can effectively deal with partial occlusions and appearance variations.



(a) 55^{th} frame 360^{th} frame 520^{th} frame 840^{th} frame

(b) 55^{th} frame 360^{th} frame 520^{th} frame 840^{th} frame

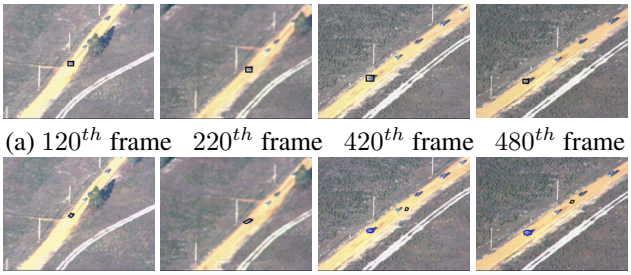
Fig. 5. Tracking results with partial occlusions and appearance variations of the tracked object. The tracking result is marked with black rectangle and the ground truth is with blue ellipse, once there are tracking errors. (a) Results of our proposed method, and (b) Results of the method in [10].

In the second video from the VIVID data set shown in Fig. 6, the car being tracked always goes straight along a road. During the tracking process, the object images have heavy deterioration (220^{th} and 480^{th} frames), which results in the degeneration of its representation in a feature space. The track-

ing results of this video show that the proposed approach can effectively deal with deterioration in object images.

Tracking efficiency is another criterion for a real-time tracking application. In the experiments, our proposed tracking approach can work about 20 frames per second averagely, while about 0.3 frames per second in [10] on a computer with Core(TM)2 Duo CPU (2.53GHz) and 3GB memory.

To quantitatively evaluate the proposed approach, we use the average displacement error rate (DER) [8] of 10 video clips from the above 3 data sets to reflect the performance of each method. We compare our method with other three representative ones, including SIFT based tracking method [7], Kalman Filter based tracking method [2] and Particle Filter based method [3]. The results of four methods are shown in Fig.7. It can be seen from the figures that the average DER of our method (about 0.05 to 0.12) is smaller than that of the other three methods in almost the whole tracking process, showing better performance of our proposed approach.



(a) 120^{th} frame 220^{th} frame 420^{th} frame 480^{th} frame

(b) 120^{th} frame 220^{th} frame 420^{th} frame 480^{th} frame

Fig. 6. Tracking results with deterioration in object images. The tracking result is marked with black rectangle and the ground truth is with blue ellipse, once there are tracking errors. (a) Results of our proposed method, and (b) Results of the method in [10].

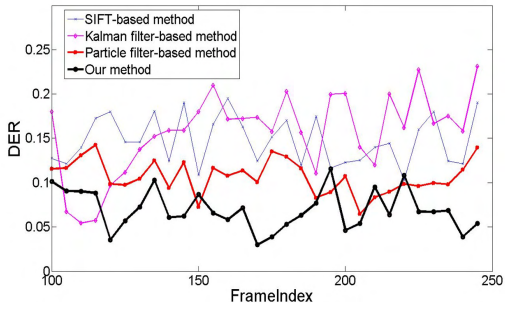


Fig. 7. Average DER of four tracking methods.

4. CONCLUSIONS

In this paper, we have proposed a novel object tracking approach based on OSR. The tracking results with comparisons are provided, which indicates that the proposed tracking approach achieves state-of-the-art, especially when there are partial occlusions, distortions and appearances variations of both objects and their backgrounds.

The new concepts and techniques introduced include the tracking sample set and object sparse representation. A knowing issue in the proposed method is the whole object occlusion problem, which will be considered in the future works.

5. AKNOWLEDGEMENT

This work is supported in part by National Basic Research Program of China (973 Program) with Nos. 2011CB706900, 2010CB731800, and National Science Foundation of China with Nos. 61039003, 60872143.

6. REFERENCES

[1] N. Papanikolopoulos, P. Khosla, and T. Kanade, "Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision," *IEEE Transactions on Robotics and Automation*, 1993.

[2] E. Cuevas, D. Zaldivar, and R. Rojas, "Kalman filter for vision tracking," *Technical Report B, Fachbereich Mathematikund Informatik*, 2005.

[3] M. Isard and A. Blake, "Condensation-conditional density propagation for visual tracking," *IJCV*, 1998.

[4] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," *IEEE Conference on CVPR*, pp. 142–149, 2000.

[5] R. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *PAMI*, 2005.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, 2005.

[7] Y. J. Li, J. F. Yang, R. B. Wu, and F. X. Gong, "Efficient object tracking based on local invariant features," *IEEE Conference on SCIT*, pp. 697–700, 2006.

[8] Z.j. Han, Q.x. Ye, and J.b. Jiao, "Online feature evaluation for object tracking using kalman filter," *IEEE Conference on ICPR*, 2008.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on PAMI*, 2008.

[10] X. Mei and H. b. Ling, "Robust visual tracking using l1 minimization," *IEEE Conference on ICCV*, 2009.

[11] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, 2001.

[12] http://coe.gucas.ac.cn/SDL-HomePage/index.asp.