Robust Visual Object Tracking via Sparse Representation and Reconstruction

Zhenjun Han, Qixiang Ye, and Jianbin Jiao

University of Chinese Academy of Sciences
{hanzhj,qxye,jiaojb}@ucas.ac.cn

Abstract. Visual object tracking plays an essential role in vision based applications. Most of the previous research has limitations due to the non-discriminated features used or the focus on simple template matching without the consideration of appearance variations. To address these challenges, this paper proposes a new approach for robust visual object tracking via sparse representation and reconstruction, where two main contributions are devoted in terms of object representation and location respectively. And the sparse representation and reconstruction (SR^2) are integrated into a Kalman filter framework to form a robust object tracker named as SR^2KF tracker. The extensive experiments show that the proposed tracker is able to tolerate the appearance variations, background clutter and image deterioration, and outperforms the existing work.

Keywords: Visual Tracking, Sparse Representation, Sparse Reconstruction.

1 Introduction

Object tracking is important to many computer vision applications ranging from visual surveillance [1], human computer interaction [2] and automatic robotics [3]. Although the state of the art has advanced significantly in the past decade [4]-[9], some challenging problems still remain open.

Tracking in complex environment requires two crucial components: discriminative object representation and effective spatial location. The object representation models the characteristics of a target, which include the shape, pose or even the reflectance properties. Appearance based representation [4] gains popularity recently because 2D images can be used straightforwardly. However, appearance variations are often caused by a number of factors in the natural environment, including shape deformation, background variation, lighting changes as well as occlusion [5]-[6]. It is imperative for a robust appearance representation to model such variations. Another goal of object tracking is to locate the target spatially. Considering the noise from the backgrounds and the object itself, it is useful for employing the probabilistic estimation framework, e.g., Kalman filter [7] or particle filter [8]. However, the template matching (the core of the estimation framework) is widely known for the drifting problem due to error accumulation over the time [10].

R. Wilson et al. (Eds.): CAIP 2013, Part II, LNCS 8048, pp. 282–289, 2013.

In this paper, we treat the representation as the classification between two-class linear regression models [5]-[6] and apply sparse signal optimization to address this problem. The rationale behind this scenario is that we compactly represent the object with discriminative features. Secondly, to facilitate the accurate location, the target object is reconstructed as a superposition of the searching candidates instead of the straightforward template matching. The reconstruction can handle the errors due to object occlusion and image corruption uniformly since the noise or errors are often sparse w.r.t. to the standard (pixel) basis. The proposed sparse representation and reconstruction (SR^2) are finally integrated into a Kalman filter to form a robust visual tracker named as SR^2KF tracker. Figure 1 shows the framework of our approach. The most related work is [9], where a robust visual tracking method is proposed by modeling the tracking as a sparse approximation problem (we refer it as SAP tracker in this paper). In SAP tracker, although the challenges of object occlusion are addressed seamlessly, the object appearance variation is not fully considered.

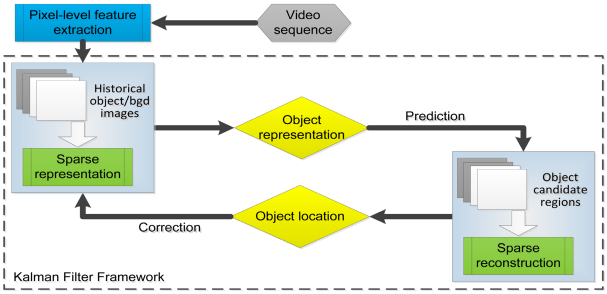


Fig. 1. The flowchart of the proposed tracking approach

The rest of the paper is organized as follows. In Section 2, we introduce the sparse representation. Section 3 presents the reconstruction strategy for object location. Section 4 describes the SR^2KF tracker. The experimental results are reported and analyzed in Section 5. Finally, we conclude the paper in Section 6.

2 Sparse Representation for Object Appearance

As the core of visual tracking, we treat the extraction of discriminative representation as a two-class (object versus background) classification problem. The existing work [11] has contended that the sparse representation selects the subset which represents the input signal most compactly and is naturally discriminative. We exploit the discriminative nature of sparsity for object appearance.

2.1 Online Construction and Update of the Training Samples

The sparse representation is obtained in an optimization framework based on the aggregation of the historical tracking results called as training samples in this

paper. Let M be the number of time steps (frames) of the historical tracking. The training sample set is defined as $T = \{(\ell_j^+, \ell_j^-)\}, j = 1, ..., M$, where ℓ_j^+ and ℓ_j^- are the object (positive sample) and the corresponding background region (negative sample) obtained at the jth step, respectively. We adopt the histograms of oriented gradient and color (HOGC) [12] as the pixel-level feature for the object and background regions [12].

An update scheme is performed to keep the latest appearance information of the object and background in the training set. At each tracking step, a pair of positive and negative samples (ℓ_j^+, ℓ_j^-) are randomly selected and re-placed with the latest tracking result and the corresponding background. For only two samples in the set updated each time, the tracking stability is thus ensured and the feature drifting is avoided even when the noisy samples are updated.

2.2 Sparse Solution via an Improved *l*1-Norm Minimization

The existing work [13] has demonstrated that the sparse problem can be effectively resolved by $l1$-norm minimization. In the context of feature representation, we formulate our problem as following

$$\min \|S\|_1, \quad s.t. \begin{cases} S^T \ell_j^+ \geq \alpha \\ S^T \ell_j^- \leq -\alpha \end{cases}, \tag{1}$$

where $||\cdot||_1$ represents the $l1$-norm and $S \in R^N$ is the vector of sparse coefficients which measure the discriminative ability of each component in HOGC features. The constraints in Eq. (1) ensure that the training samples can be correctly classified and the shortest distance between two classes is 2α.

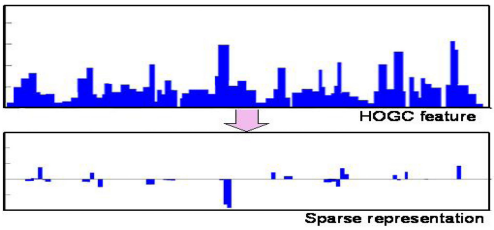


Fig. 2. Sparse representation of the object appearance via $l1$-norm minimization

Given the sparse solution S and the object O in the current step (frame), the representation is calculated as

$$F = S \otimes \text{HOGC}(O), \tag{2}$$

where HOGC(O) represents the histogram features of object O and $\otimes$ denotes the dot-product operation. Figure 2 shows an example of the sparse representation given the HOGC features. The obtained sparse representation is finally used as a prior to locate the object in the successive frame.

3 Sparse Reconstruction for Object Location

The existing work [14] shows that the sparse reconstruction is robust to noise and insensitive to partial occlusion. We thus exploit it to identify the most informative candidates for accurate object location.

3.1 Instantaneous Sample Set for Reconstruction

In our approach, with the pre-identification in the instantaneous frame, all the possible candidates will be firstly selected to form a sample set, which is used for reconstruction based location. The instantaneous sample set is dynamically built up in a search window (the black rectangle region in Figure 3). A sample in the search window (the red rectangle region in Figure 3) is specified by a triplet $r = (x, y, s)$ [12], where $x \in (0, W)$ and $y \in (0, H)$ are the image coordinates and $s > 0$ is the scale of the candidate. Let K be the number of selected candidates, the instantaneous sample set is defined as $A = \{a_k\}, k = 1, ..., K$.

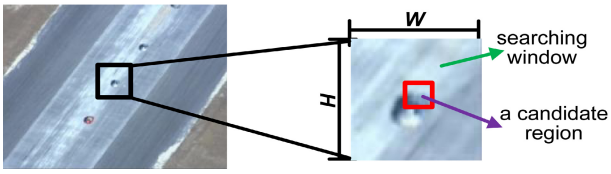


Fig. 3. The searching window for instantaneous sample set construction

3.2 Sparse Reconstruction for Object Location

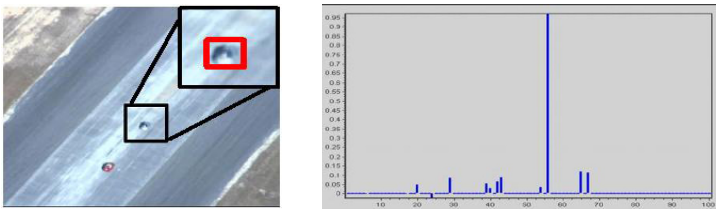
The aim of the sparse reconstruction is to approximate the potential object location by a linear superposition of the instantaneous samples given the prior object representation as follows

$$A\psi = F, \tag{3}$$

where F is the given sparse representation, and $\psi = \{\varphi_k\}$, $k = 1, ..., K$ is the reconstruction vector in which φ_k is the reconstruction coefficient for the kth sample in the instantaneous set A.

Intuitively, the search window is larger than the target region with some instantaneous samples from non-object regions. Only a few samples are suitable for linear superposition. Even in the case of partial occlusion, the number of effective object samples are limited. Consequently, ψ subjects to sparse property. We can then formulate the sparse reconstruction as

$$\begin{aligned} &\min ||\psi||_1, \\ &s.t. A\psi = F. \end{aligned} \tag{4}$$



(a) (b)

Fig. 4. Sparse reconstruction for object location. (a) The object location calculated by reconstruction strategy. (b) The coefficient vector of sparse reconstruction.

Figure 4(b) shows an example of the solution ψ. Among 100 samples, about 10 samples are enough for the approximation with the corresponding coefficients.

Essentially, the entries in the reconstruction vector are the confidences showing how the corresponding candidates are the object. Therefore, the object location with its scale $O(x, y, s)$ is finally identified as

$$O(x, y, s) = \sum_{k=1}^{K} a_k(x, y, s) \cdot \varphi_k. \tag{5}$$

Figure 4(a) shows the result of object location calculated by Eq. (5).

4 SR^2 Kalman Filter

We integrate the sparse representation and reconstruction (SR^2) into the Kalman filter (KF) and develop a robust object tracker named as SR^{2}KF tracker. The Kalman filter addresses the general problem of estimating the state X of a discrete time process that is governed by a linear stochastic difference equation as

$$X_{t+1} = MX_t + u_t, \tag{6}$$

with a measurement Z that is

$$Z_t = HX_t + v_t. \tag{7}$$

The random variables u_t and v_t represent the state and measurement noise, respectively. They are assumed to be independent and have normal distribution. In SR^{2}KF tracker, the first order Markov model is employed, *i.e.*

$$X_t = \begin{pmatrix} F_t \\ \Delta F_t \\ P_t \\ \Delta P_t \end{pmatrix}, Z_t = \begin{pmatrix} F_t \\ P_t \end{pmatrix}, \tag{8}$$

and

$$M = \begin{pmatrix} I_{N\times N} & I_{N\times N} & 0 & 0 \\ 0 & I_{N\times N} & 0 & 0 \\ 0 & 0 & I_{L\times L} & I_{L\times L} \\ 0 & 0 & 0 & I_{L\times L} \end{pmatrix}, H = \begin{pmatrix} I_{N\times N} & 0 & 0 & 0 \\ 0 & 0 & I_{L\times L} & 0 \end{pmatrix}. \tag{9}$$

where F_t is the sparse representation at time (frame) t and $\Delta F_t = F_t - F_{t-1}$, $P_t = (x_t, y_t)$ denotes the center coordinates of the object bounding box and $\Delta P_t = P_t - P_{t-1}$, $I_{N\times N}$ and $I_{L\times L}$ are the identity matrixes where N is the dimension of the sparse representation and L is 2.

5 Experimental Results

To demonstrate the effectiveness of the proposed approach, we performed comprehensive experiments on several public benchmarks. To avoid the influence from detection, the target objects were initialized manually.

Ten image sequences were selected from the public sets VIVID [15], CAVIAR [16] and SDL [17]. The selected test data consists of a variety of challenging cases in the complex environments, *e.g.*, occlusion among objects, object scale variation and rotation, lighting changes, and clutter background. To quantitatively evaluate the performance, we adopt the displacement error rate (DER) [12] between the tracking result and the ground-truth.

5.1 Evaluation on the Appearances Variation

The average and variance of DER were used to validate the effectiveness of the sparse representation against other two representative feature representation methods: variance ratio feature shift [5] and peak difference feature shift [6].

In Figure 5(a), the sparse representation is with the average DER ranging from 0.04 to 0.1. The variance of DER for our approach is 2.62×10^{-4} which is much lower than the ones of variance ratio feature shift (value at 3.80×10^{-3}) and peak difference feature shift (value at 2.41×10^{-3}). This demonstrates that the proposed object representation is more robust to the appearance variation.

5.2 Evaluation on Partial Occlusion

Partial occlusion is another key issue for inaccurate location and unstable tracking. The Kalman filter [7] and particle filter [8] are two well-acknowledged tracking algorithms which can handle the partial occlusion to some extent. We thus conducted the evaluation on all test data and compared the proposed reconstruction strategy with the Kalman filter and particle filter based tracking algorithms.

In Figure 5(b), the performance of the reconstruction strategy is apparently better than the Kalman filter and particle filter measured by average DER. In addition, the stability measured by variance of DER also outperforms the ones of Kalman filter (5.81×10^{-3}) and particle filter (1.30×10^{-3}).

5.3 Evaluation of SR^{2}KF Tracker

We finally evaluated the performance of SR^{2}KF tracker and compared with the SAP tracker [9] against different complex situations including image deterioration, appearances variations of object and background, and partial occlusion.

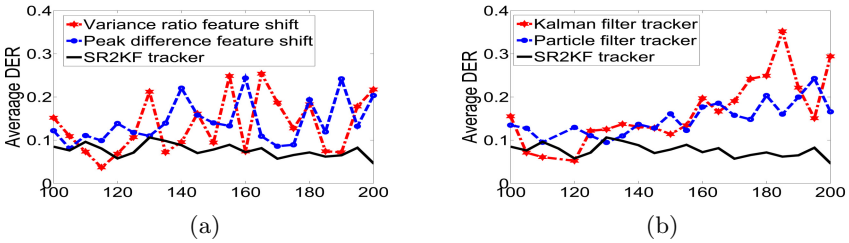


Fig. 5. Experimental results. (a) For appearance variation; and (b) partial occlusion.

The first test sequence (as shown in Fig. 6(a)) is from VIVID data set. During the tracking process, the object subjects to severe image quality deterioration (at 220^{th} frame), which results in the degeneration of its appearance representation in pixel-level space. The sequence shown in Fig. 6(b) from SDL set is quite challenging due to the clutter backgrounds, the similarity of other objects to the tracking target in appearance, and severe occlusion. In these tracking conditions, the SAP tracker failed in tracking ultimately but SR^{2}KF tracker worked successfully. The experimental results demonstrated that our proposed SR^{2}KF tracker is effective for visual object tracking and outperforms [9].

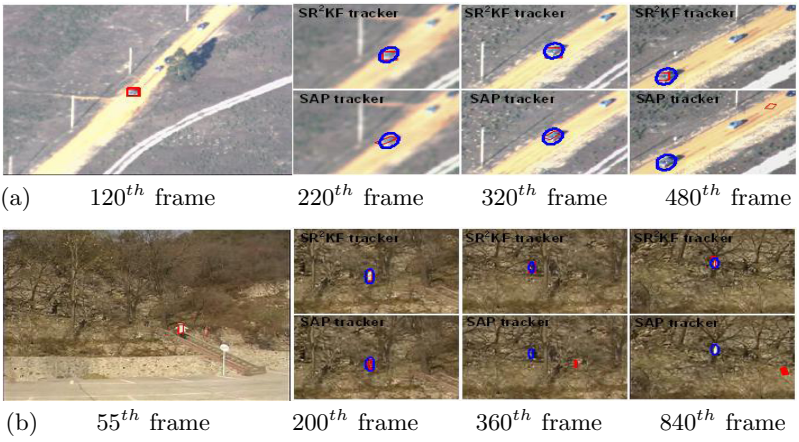


Fig. 6. Tracking results of SR^{2}KF tracker and the SAP tracker in [9]. The tracking result is marked with red rectangle and the ground-truth is with blue ellipse.

6 Conclusions and Future Work

Object representation and location are two primary issues in visual tracking. In this paper, we implemented a visual object tracking approach named as

SR^2KF tracker by integrating the sparse representation and reconstruction into a Kalman filter framework. The sparse representation is able to encode the variant appearance patterns and the reconstruction scenario is robust to object partial occlusion for location. The extensive evaluation demonstrates our approach is powerful for visual tracking problems compared with the existing work.

The future work includes 1) more pixel-level appearance and motion features will be integrated in the current approach to improve the performance, and 2) the case of complete occlusion will be addressed.

Acknowledgements. 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, 61271433 and 61202323.

References

1. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR (1999)
2. Bradski, G.: Real time face and object tracking as a component of a perceptual user interface. In: IEEE Workshop on Applications of Computer Vision (1998)
3. Papanikolopoulos, N., Khosla, P., Kanade, T.: Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision. IEEE Trans. Robotics and Automation, 14–35 (1993)
4. Shi, J., Tomasi, C.: Good features to track. In: CVPR (1994)
5. Collins, R., Liu, Y.: On-line selection of discriminative tracking features. In: ICCV (2003)
6. Collins, R., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Trans. PAMI, 1631–1643 (2005)
7. Cuevas, E., Zaldivar, D., Rojas, R.: Kalman filter for vision tracking. Technical Report (2005)
8. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. In: IJCV, pp. 5–28 (1998)
9. Mei, X., Ling, H.: Robust Visual Tracking and Vehicle Classification via Sparse Representation. IEEE Trans. PAMI, 2259–2272 (2011)
10. Matthews, I., Ishikawa, T., Baker, S.: The template update problem. IEEE Trans. PAMI, 810–815 (2004)
11. Xu, R., Zhang, B., Ye, Q., Jiao, J.: Cascaded l1-norm minimization learning (CLML) classifier for human detection. In: CVPR (2010)
12. Han, Z., Jiao, J., Zhang, B., Ye, Q., Liu, J.: Visual object tracking via sample-based Adaptive Sparse Representation (AdaSR). Pattern Recognition, 2170–2183 (2011)
13. Donoho, D.: For most large underdetermined systems of linear equations the minimal l1-norm near solution approximates the sparsest solution. Comm. on Pure and Applied Math., 797–829 (2004)
14. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. PAMI, 210–227 (2008)
15. VIVID Dataset, `http://vividevaluation.ri.cmu.edu/datasets/datasets.html`
16. CAVIAR Test Case Scenarios, `http://homepages.inf.ed.ac.uk/rbf/CAVIAR`
17. SDL Data set, `http://www.ucassdl.cn/resource.asp`