Robust Visual Tracking with Improved Subspace Representation Model

Jing Cheng*1, Sucheng Kang2

1Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi, China, 214122
2Yancheng Teachers University, Yancheng, China, 224051
*Corresponding author, e-mail: chengjing907@163.com

Abstract

In this paper, we propose a robust visual tracking with an improved subspace representation model. Different from traditional subspace representation model, we use sparse representation, but not the collaborative representation to reconstruct the observation samples, which can avoid the redundant object features in subspace effectively. Moreover, to reject the outliers in the process of tracking, we also propose the combination of sparse box templates and Laplacian residual. To solve the minimization problem of object representation efficiently, a fast numerical algorithm that accelerated proximal gradient (APG) approach is proposed for the Lagrangian function. Finally, experimental results on several challenging video sequences show better performance than LSST and many state-of-the-art trackers.

Keywords: visual tracking, subspace, sparse representation, accelerated proximal gradient

1. Introduction

The visual tracking or the object tracking of video sequence in computer vision and related areas is a crucial task [1-2]. It has many practical applications such as video surveillance, vehicle navigation and human-computer interaction. Although much progress has been made in the past decades, designing a practical visual tracking system is still a challenging problem due to numerous challenges in real world. For example, pose variation, shape deformation, varying illumination, camera motion, and occlusions may increase the difficulty for visual tracking algorithms. This article focuses on the extensive research of single target tracking problem.

Recently, sparse representation and compressed sensing techniques have been proposed and successfully applied in visual tracking. Sparse representation is first introduced into visual tracking by Mei and Ling [3]. From then, it is used to model various trackers. Similar to sparsity-based approach for face recognition developed in [4], these tracking methods express a target by a sparse linear combination of the templates. In [5], an accelerated proximal gradient approach (APG) was proposed to improve the L1 tracker and get much time saved. In [6], Lu et al. present L1 tracker based on multitask reverse representation formulation that search multiple subsets for the whole candidate set to reconstruct templates with minimum error. Benefitting from the stable recovery capability of sparse signal using the L1-norm minimization, these trackers have demonstrated good robustness in various tracking environments.

The success of collaborative representation-based face recognition raise the researchers’ attention, then the collaborative representation of PCA subspace has been further employed in visual tracking [7-8]. But In [8], the IVT does not take the outliers into consideration and this method algorithm is less effective to the occlusions and corruptions. Then in [8], Wang et al. assume the outliers as a linear combination of trivial templates. There are also some shortages in LSST [9], for example: The redundant feature of the subspace lead to bad result of the collaborative representation. Moreover, LSST models residual with Gaussian distribution, in fact, it cannot be used to model the bias effectively, which has been shown in [9].

Motivated by the above work, in this paper, we propose an improved sparse representation tracking algorithm with PCA basis vectors. There are some differences between our work and LSST [9]: Firstly, we use sparse representation to construct the observation samples; secondly, we use the combination of sparse box templates and Laplacian residual to reject the outliers; thirdly, we use the L1-norm minimization approach to solve the minimization problem of problem of object representation efficiently; finally, we use an accelerated proximal gradient approach to solve the Lagrangian function.
samples for the rejecting of redundant object information in subspace. Secondly, inspired by the success work in face recognition [7], we measure the outliers as the combination of sparse box templates and Laplacian residual, but not the combination of trivial templates and Gaussian residual, which can be more effective to reject the main outliers. Finally, an effective numerical method based on APG is introduced to solve the unconstrained L1 regularized problem.

2. Research Method
2.1. Object Representation

In LSST [9], a target region $y \in \mathbb{R}^{dxl}$ can be represented as follows:

$$y = Dc + n + e$$

(1)

Where the columns of $D \in \mathbb{R}^{dxm}$ are orthogonal basis vectors of the subspace, $c \in \mathbb{R}^{mxl}$ is the target coefficient vector, $n$ is an error term modeled by Laplacian, and $e$ is the residual modeled by Gaussian. To reject the redundant features in subspace, and reject the outliers in subspace more effectively, we apply L1-norm to regulate the target coefficient $c$. Moreover, considering the occlusion pixels are always spatial continuous, we use sparse box templates, but not the trivial templates for the mainly error. Figure 1 shows the object representation.

Observation is defined as:

$$c^* = \arg\min_{c} \{ \| y - Dc - n \|_1 + \mu \| n \|_1 + \phi \| c \|_1 \}$$

(2)

Let $e = y - Dc - n$, we can rewrite the Equation (2) as:

$$c^* = \arg\min_{c} \{ \| e \|_1 + \mu \| n \|_1 + \phi \| c \|_1 \} \quad \text{s.t.} \quad y = Dc + n + e$$

(3)

Equation (3) is a constrained convex optimization problem which can be efficiently solved through ALM operation. The corresponding ALM function is:

$$L(e,n,c) = \| e \|_1 + \mu \| n \|_1 + \phi \| c \|_1 + \gamma, y - Dc - n - e > + \frac{\tau}{2} \| y - Dc - n - e \|_2^2$$

(4)

Then we simplify Equation (4) as follows:

$$L(e,n,c) = \| e \|_1 + \mu \| n \|_1 + \phi \| c \|_1 + \frac{\tau}{2} \| y - Dc - n - e + \frac{\gamma}{\tau} \|_2^2$$

(5)

Where $\langle \cdot, \cdot \rangle$ denotes the inner product operator, $\gamma$ is a vector of Lagrange multiplier, $\tau$ is a constant that determines the penalty for large representation error. Across iteratively minimizing the augmented Lagrangian function, the ALM algorithm iteratively estimates the Lagrange multiplier and the optimal solution, as follows:
We apply the APG approach to solve this minimization problem. The numerical algorithm for solving Equation (6) is summarized in Algorithm 1.

\[
\begin{align*}
\gamma_{k+1} &= \gamma_k + \tau_k (y - Dc - n - e) \\
\tau_{k+1} &= \rho \tau_k \\
\end{align*}
\]

\[
\begin{align*}
(c_{k+1}, c_{k+1}, n_{k+1}) &= \arg \min_{c, e, n} L_{c_{k+1}} (c, e, n, \tau_k) \\
\end{align*}
\]

\[
\begin{align*}
\gamma_{k+1} &= \gamma_k + \tau_k (y - Dc - n - e) \\
\tau_{k+1} &= \rho \tau_k \\
\end{align*}
\]

We apply the APG approach to solve this minimization problem. The numerical algorithm for solving Equation (6) is summarized in Algorithm 1.

In algorithm 1, we need to solve:

\[
\begin{align*}
c_{k+1} &= \arg \min_c \|c\|_2 + \frac{L}{2} \|c - Z_{k+1}^{c} + \nabla_c F(Z_{k+1}^{c}, Z_{k+1}^{n}, Z_{k+1}^{e})\|_2^2 \\
e_{k+1} &= \arg \min_e \|e\|_2 + \frac{L}{2} \|e - Z_{k+1}^{e} + \nabla_e F(Z_{k+1}^{c}, Z_{k+1}^{n}, Z_{k+1}^{e})\|_2^2 \\
n_{k+1} &= \arg \min_n \|n\|_2 + \frac{L}{2} \|n - Z_{k+1}^{n} + \nabla_n F(Z_{k+1}^{c}, Z_{k+1}^{n}, Z_{k+1}^{e})\|_2^2 \\
\end{align*}
\]

Algorithm 1. Accelerated proximal gradient methods (APG) for solving (2)

Set initial guesses: \( c_0 = c_{-1} = 0, n_0 = n_{-1}, e_0 = e_{-1}, t_0 = t_{-1} = 1 \)

Input: The PCA subspace \( D \), the candidate sample \( y \)

For \( k = 1, 2, \ldots \), iterate until both the \( c, e \) and \( n \) are convergent to optimal state:

1:\( Z_{k+1}^{c} = c_k + \frac{t_{k+1} - 1}{t_k} (c_k - c_{k-1}) \)

2:\( Z_{k+1}^{n} = n_k + \frac{t_{k+1} - 1}{t_k} (n_k - n_{k-1}) \)

3:\( Z_{k+1}^{e} = e_k + \frac{t_{k+1} - 1}{t_k} (e_k - e_{k-1}) \)

4:\( c_{k+1} = \arg \min_c \|c\|_2 + \frac{L}{2} \|c - Z_{k+1}^{c} + \nabla_c F(Z_{k+1}^{c}, Z_{k+1}^{n}, Z_{k+1}^{e})\|_2^2 \)

5:\( n_{k+1} = \arg \min_n \|n\|_2 + \frac{L}{2} \|n - Z_{k+1}^{n} + \nabla_n F(Z_{k+1}^{c}, Z_{k+1}^{n}, Z_{k+1}^{e})\|_2^2 \)

6:\( e_{k+1} = \arg \min_e \|e\|_2 + \frac{L}{2} \|e - Z_{k+1}^{e} + \nabla_e F(Z_{k+1}^{c}, Z_{k+1}^{n}, Z_{k+1}^{e})\|_2^2 \)

7:\( \gamma_{k+1} = \gamma_k + \tau_k (y - Dc - n - e) \)

8:\( \tau_{k+1} = \rho \tau_k \)

9:\( t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \)

End

Output: The optimal \( c^*, e^* \) and \( n^* \)
Where \( \nabla_c F(c,n,e) = -\tau D^\top (y - Dc - n - e + \frac{\tau}{\gamma}) \), \( \nabla_n F(c,n,e) = -\tau (y - Dc - n - e + \frac{\tau}{\gamma}) \), and \( L \) is a Lipschitz constant.

According to the soft thresholding operator, \( S_\theta(x) \) is defined as:

\[
S_\theta(x) = \text{sign}(x) \max(|x| - \theta, 0)
\]

It is easy to show that the solution of Equation (9), Equation (10) and Equation (11) can be obtained by:

\[
c_{k+1} = S_{\psi/L}(x) \left( Z_{c,k+1}^c - \frac{\nabla_c F(Z_{c,k+1}^c, Z_{n,k+1}^n, Z_{e,k+1}^e)}{L} \right)
\]

\[
e_{k+1} = S_{\psi/L}(x) \left( Z_{e,k+1}^e - \frac{\nabla_e F(Z_{c,k+1}^c, Z_{n,k+1}^n, Z_{e,k+1}^e)}{L} \right)
\]

\[
n_{k+1} = S_{\rho/L}(x) \left( Z_{n,k+1}^n - \frac{\nabla_n F(Z_{c,k+1}^c, Z_{n,k+1}^n, Z_{e,k+1}^e)}{L} \right)
\]

2.2. Bayesian Interference Framework

In this paper, we employ a particle filter based on Bayesian interference framework to track the target object. The technique is mainly to estimate the posterior distribution of random variables related to Markov chain. Let \( y_t = \{y_1, y_2, \ldots, y_t\} \) denotes the observation of the target from first frame to the frame \( t \). At the frame \( t \), we can estimate the target state variable \( x_t \) by using the maximum a posterior estimation

\[
\hat{x}_t = \arg \max_{x_t} p(x_t \mid y_t)
\]

Where \( x_t^i \) indicates the \( i \)-th sample of the state \( x_t \). Based on the Bayes theorem, the posterior distribution \( p(x_t \mid y_t) \) can be estimated recursively by:

\[
p(x_t \mid y_t) \propto p(y_t \mid x_t) \int p(x_t \mid x_{t-1}) p(x_{t-1} \mid y_{t-1}) x_{t-1}
\]

**Motion model:** \( p(x_t \mid x_{t-1}) \) is the motion model that describes the state transition between two continuous frames, \( p(y_t \mid x_t) \) is the observation model which estimates the likelihood of candidates. The state \( x_t \) is included with six different parameters, let \( x_t = \{l_x, l_y, \theta, s, \alpha, \phi\} \) and these six parameters respectively represent \( x \), \( y \) translations, rotation angle, scale, aspect ratio, and skew respectively. The affine motion model can be modeled by Gaussian distribution, it is formulated by the random walk.

\[
p(x_t \mid x_{t-1}) = N(x_t; x_{t-1}, \sigma)
\]

Where \( \sigma \) is diagonal covariance matrix whose elements are the variances of the affine parameters.
**Observation model:** In fact, there are still some Gaussian noise remaining. The residual $e$ can be considered as the error for the reconstruction. In order to get the better result of the target object, the observation likelihood is set to be as:

$$p(y | x) = \exp(-\xi E(c^*, e^*, n^*))$$

(19)

where $E(c^*, e^*, n^*) = \frac{1}{2} \| y - Dc^* - e^* - n^* \|^2 + \alpha \| c^* \|^2 + \beta \| n^* \|^2$, $c^*$, $e^*$ and $n^*$ is the optimal solution of Equation (6), $\xi$ is a constant controls the shape of the Gaussian kernel, $\alpha$ and $\beta$ is the penalty term.

The entire tracking procedure is summarized in Figure 2. At the outset, the state of the target is manually initialized. Then, the candidate samples can be obtained from the motion model. Once the state parameters of residual and sparse coefficient are obtained from Algorithm 1, we can evaluate the likelihood of each candidate state. Finally, the samples are cumulated to update the subspace for handling the change of target object. The whole tracking procedure will keep running until the target state of last frame are obtained.

![Figure 2. Our Tracking Algorithm Framework](image)

**2.3. Model Update**

After finish the tracking of the current frame, we have to incrementally update the subspace in order to learn the appearance of the target object. The subspace $D$ is consisted of PCA basis vectors. When we obtain the best candidate state of current frame, we get the corresponding observation vector $y$ and the residual term $e$ which used to identify outliers. Then we can update the subspace by:

$$y_i = \begin{cases} y_i, & |c_i| = 0 \\ \mu_i, & |e_i| \neq 0 \end{cases}$$

(20)

Where $y_i$, $c_i$, and $\mu_i$ are the i-th element of $y$, $c$ and $\mu$ respectively, $\mu$ is the mean vector computed by [8].

**3. Results and Analysis**

The proposed tracker is implemented in MATLAB and runs at 6 frames per second on a 3.5GHz CPU with 8GB memory. We empirically set $L = 5$, $\rho = 1.5$, $\alpha = 0.025$, $\beta = 0.025$. The location of the target in the first frame is manually labeled. Each observation is normalized to 32 × 32 pixels and the size of box template is 4 × 4 pixels, and 16 PCA basis vectors are used for
the subspace in all the experiments. 600 particles are adopted and our tracker is incrementally updated when 5 samples are accumulated.

To prove the effectiveness of the proposed algorithm, we use nine challenging image sequences contain different challenging factors (e.g., illumination change, severe occlusion, background clutter, etc.) and compare our method with five competitive methods: Incremental visual tracking (IVT) [8], Adaptive Structural Local Appearance (ASLA) [10], Sparsity based Collaborative Model (SCM) [11], Discriminative Sparse Similarity Tracking (DSST) [6], and Least Soft-threshold Squares Tracking (LSST) [9]. For a fair evaluation, we run these codes with the same bounding box in the first frame.

3.1. Quantitative Evaluation

For the purpose of assessing the tracker, we evaluate the above-mentioned algorithms using two criteria: the center error and the overlap rate. Table 1 reports the average center location errors in pixels, where a smaller average error means a more accurate result. On the contrary, a bigger overlap rate means a better results. Given the tracking result (bounding box) \( T \) and the corresponding ground truth bounding box \( G \), the overlap score is defined as:

\[
\text{score} = \frac{\text{area}(R_T \cap R_G)}{\text{area}(R_T \cup R_G)}
\]

Table 2 reports the average overlap rates, where larger average scores mean more accurate results.

| Sequences     | ASLSA | DSST | IVT | SCM | LSST | Ours |
|---------------|-------|------|-----|-----|------|------|
| Occlusion1    | 9.6   | 11.4 | 12.5| 3.2 | 5.3  | 4.3  |
| Football      | 8.9   | 8.4  | 17.3| 15.2| 7.6  | 3.7  |
| DavidIndoorNew| 32.4  | 11.9 | 35.9| 17.7| 3.1  | 2.9  |
| Singer1       | 3.7   | 12.8 | 11.9| 3.3 | 3.5  | 1.8  |
| Car11         | 1.8   | 1.7  | 1.9 | 1.8 | 1.6  | 1.5  |
| Deer          | 8.0   | 8.8  | 135.2| 10.1| 10.0 | 6.2  |
| Jumping       | 5.2   | 6.8  | 6.4 | 3.9 | 4.8  | 3.5  |
| Boy           | 2.8   | 3.2  | 177.2| 51.8| 176.5| 2.7  |
| Average       | 8.1   | 7.4  | 45.4| 12.0| 23.7 | 3.1  |

| Sequences     | ASLSA | DSST | IVT | SCM | LSST | Ours |
|---------------|-------|------|-----|-----|------|------|
| Occlusion1    | 0.94  | 0.82 | 0.80| 0.93| 0.84 | 0.92 |
| Football      | 0.67  | 0.70 | 0.58| 0.61| 0.69 | 0.80 |
| DavidIndoorNew| 0.42  | 0.60 | 0.43| 0.50| 0.75 | 0.76 |
| Singer1       | 0.62  | 0.70 | 0.46| 0.84| 0.81 | 0.88 |
| Car11         | 0.93  | 0.78 | 0.84| 0.80| 0.84 | 0.84 |
| Deer          | 0.63  | 0.63 | 0.24| 0.60| 0.59 | 0.69 |
| Jumping       | 0.67  | 0.61 | 0.62| 0.73| 0.65 | 0.74 |
| Boy           | 0.79  | 0.78 | 0.19| 0.53| 0.31 | 0.79 |
| Average       | 0.71  | 0.72 | 0.52| 0.71| 0.70 | 0.81 |

3.2. Qualitative Evaluation

Severe Occlusion: We test three sequences (Occlusion1, and Football) with heavy or long-time partial occlusion, scale change and rotation. Figure 3(a) demonstrate that the proposed method performs well in terms of occlusion. The SCM methods achieve better performance in some cases as both of them include part-based representations with overlapping patches (Occlusion1). The IVT method is sensitive to partial occlusion (Football) since the OLS distance is not effective to handle outliers. Overall, the LSST and our tracker can perform better than other trackers. As the redundant information in subspace and the ambiguous L2-regularized coefficient to square template, we adopt the Laplacian noise to measure the residual and our tracker is more robust.
**Illumination Change:** Figure 3(b) shows the tracking results in the sequences (DavidIndoorNew, Singer1) with significant illumination variation, scale change and pose change. We can see that the IVT methods are less effective in these cases (e.g., DavidIndoorNew #439, #666 and Singer1 #130). Due to the use of incremental PCA algorithm, the proposed tracker achieves good performance in dealing with the appearance change caused by light change. For the same reason, the LSST methods also perform well.

![Figure 3(a)](image1)

![Figure 3(b)](image2)

![Figure 3(c)](image3)

![Figure 3(d)](image4)

Background Clutter: Figure 3(c) demonstrates the tracking results in Car11 and Deer sequences with background clutter. These videos also pose other challenging factors including illumination variation (Car11), fast motion (Deer). As the proposed L1-norm encourages good matching results when outliers occur, our tracker performs better than other methods in these videos (e.g., Deer #61 and Car11 #373).

Fast Motion: Figure 3(d) presents the tracking results on Boy and Jumping sequences with abrupt motion. Furthermore, the appearance change caused by motion blur poses great challenges for capturing the tracked targets accurately and updating the observation models properly. As the proposed tracker can effectively handle the outliers in the process of tracking, the tracking results are presented in Figure 2(d) (e.g., boy#406 and jumping#216).
4. Conclusion
This paper presents a robust visual tracking method based on the improved subspace representation model. Different from the LSST supposing of Gaussian-Laplacian noise to the coding residual, we use the Laplacian distribution to model the coding residual and add extra box templates for the tolerating of the outliers. Moreover, the used ALM method based on APG approach numerical algorithm is applied to solve the minimization problem of object representation effectively. Extensive experimental results validate the proposed method can reduce redundant information of the subspace and achieve more favorable performance than several competitive methods. In the future, we plan to integrate multiple visual cues (e.g., color and depth) into our object representation for better tracking.

References
[1] Cao Jie, et al. Object Tracking Based on Multiple Features Adaptive Fusion. Telkomnika Indonesian Journal of Electrical Engineering, 2014.
[2] Zhang Hai Tao. A Tracking Algorithm of Moving Target in Sports Video. Telkomnika Indonesian Journal of Electrical Engineering, 2014.
[3] X Mei, H Ling. Robust visual tracking using $L_1$ minimization. In Proc. ICCV. 2009: 1-10.
[4] J Wright, AY Yang, A Ganesh, SS Sastry, Y Ma. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009; 31(2): 210-227.
[5] C Bao, Y Wu, H Ling, H Ji. Real time robust $L_1$ tracker using accelerated proximal gradient approach. In Proc. CVPR. 2012: 1830-1837.
[6] B Zhuang, H Lu, Z Xiao, D Wang. Visual tracking via discriminative sparse similarity map. IEEE Trans. Image Process. 2014; 23(4): 1872-1881.
[7] Zhang L, Yang M, Feng X, Ma Y, Zhang D. Collaborative Representation based Classification for Face Recognition. Computer Science. 2012.
[8] D Ross, J Lim, R Lin, M Yang. Incremental Learning for Robust Visual Tracking. International Journal of Computer Vision. 2008; 77(1-3): 125-141.
[9] Wang D, H Lu, M Yang. Robust Visual Tracking via Least Soft-threshold Squares. IEEE Transactions on Circuits & Systems for Video Technology. 2015: 1.
[10] Lu Huchuan, X Jia, MH Yang. Visual tracking via adaptive structural local sparse appearance model. IEEE Conference on Computer Vision & Pattern Recognition. 2012: 1822-1829.
[11] Yang Ming Hsuan, H Lu, W Zhong. Robust object tracking via sparsity-based collaborative model. IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society. 2012: 1838-1845.