Advances in Engineering Research, volume 86

2nd International Conference on Electrical, Automation and Mechanical Engineering (EAME 2017)

Target Tracking Algorithm Based on Sparse Representation of Cooperative Template

Qisen Lin*, Ming Li and Xiaoxu Li
School of Computer and Communication, Lanzhou University of Technology, Gansu Lanzhou 730050, P.R. China
*Corresponding author

Abstract—**Aiming at the problem that the matching error between the target template and the candidate template. In this paper, a new target tracking algorithm based on sparse representation of cooperative template is proposed. When constructing the target template, the global and local templates were used to set up the collaborative template to describe the target; in order to further deal with the influence of occlusion factors, the sparse coefficient were weighted pretreatment; Making full use of the local information of the target, and combining the global information, the measurement function was designed to calculate the similarity between the candidate target and the target template. Experimental results show that the proposed method is robust to rotation, partial occlusion and scale change.**

Keywords-sparse representation; cooperative template; Target tracking; target template

I. INTRODUCTION

Visual target tracking occupies an important position in the field of computer vision, in the security monitoring [1], the human-computer interaction [2], car navigation [3], and many other fields has been widely used. In practical applications, because of the effects of occlusion, illumination changes, rotation and translation, complex background and other factors, the tracking realism and accuracy are not high enough to meet the needs of practical applications.

In 2009, Mei [4] et al. first applied the sparse representation to target tracking. In Ref. [5], the author was combined with L1 and principal component analysis to reduce the noise interference. In Ref. [4,5], we use the global features of the target to establish the target template, the global appearance template can effectively deal with the global appearance changes due to illumination changes or rotation, but it can not deal with the problem of partial occlusion. In Ref. [6], it was used to detect the occlusion of tracking sequences by using local sparse representation, and occlusion treatment of the target was achieved. In Ref. [7] adopted a local appearance model based on sparse representation and online dictionary learning to track the target, however, when the tracking was not taken into account the importance of local features, it was necessary to calculate the observed likelihood value for each local patch, which reduced the real-time tracking.

This paper proposes an improved algorithm for sparse representation of target tracking. Firstly, the global template based on color histogram and the local template based on sparse representation are combined to construct the new cooperative template to describe the target. And then, in the collaborative template, the weight of the global template is reduced, so as to highlight the role of the local template in the processing of occlusion; finally, the target template is updated online to adapt to the appearance of the target.

II. BASIC THEORY OF GLOBAL TEMPLATE

In the video information, the color information is typically characterized by having a rotation and scale invariance. Therefore, we build a global template based on the color histogram. In the first frame, the image of the target area is transformed from RGB space to HSV space. Then, the H component is quantized to N_0 level, and the number N_0 is the color histogram, which is to establish the H component histogram. Set x_0 as the center of the target template, the color distribution of the object at y_c can be expressed asMaintaining the Integrity of the Specifications

$$q_u = C_h \sum_{i=1}^{n_0} k\left(\left\|\frac{y_c - x_i}{h}\right\|^2\right)\delta[b(x_i) - u] \tag{1}$$

where u is the color vector, n_0 is the number of pixels in the target template, x_i is the x_0 as the center of the coordinates of the i pixel, $k(\cdot)$ is the kernel function, h is the core function of the bandwidth, $b(x_i)$ is the x_i pixel point in the color space of the index value, $\delta(\cdot)$ is the Dirac function, C_h is a normalized constant which makes $\sum_{u=1} q_u = 1$. C_h is given by

$$C_h = \frac{1}{\sum_{i=1}^{n_0} k\left(\left\|\frac{x_i}{h}\right\|^2\right)} \tag{2}$$

In the k-th frame image, the candidate target can be expressed as the formula (1), denoted as $p_u(y)$.

III. The Proposed Method

A. Local Template Based on Sparse Representation

A local model based on sparse representation is constructed by referring to the method in [8]. Firstly, defining a set of target templates, this template set includes n target templates, that is $T=[T_1,T_2,...,T_n]$, overlapping blocks are performed on each target template $T_i\ (i=1,2,...,n)$, the target templates are divided into some local image patches, and each local patch is normalized, then these local image patches are used to form a template dictionary for sparse coding of the local image patches in the target candidate region, so as to get the final sparse dictionary, i.e.,

$$D=[d_1,d_2,...,d_{(\mathrm{n}\times\mathrm{N})}]\in R^{d\times(n\times N)} \tag{3}$$

where d is the dimension of the image patch vector, n is the number of the target templates, N is the number of local image patches sampled in the target area. Each column in D is obtained by the Vectorization of local image patches and the normalization of L_2 norm. In the process of tracking, the candidate targets are overlapping block and vector in the same way, i.e.,

$$Y=[y_{1,}y_2,...,y_N]\in \mathrm{R}^{\mathrm{d}\times\mathrm{N}} \tag{4}$$

Each local patch of a candidate target is been represented by a L_1 regularized minimization,

$$\min_{a_i}\|y_i-Da_i\|_2^2+\lambda\|a_i\|_1,\ \mathrm{s.t.}\ a_i\geq 0 \tag{5}$$

where y_i denotes the i-th vector of local image patch, $a_i\in R^{(n\times N)\times 1}$ is the sparse coefficient of i-th local patch. λ is a scalar, used to adjust the reconstruction error $\|y-Da_i\|_2^2$ and sparse coefficient vector of sparsity, $a_i\geq 0$ represents all the elements of a_i are non negative. Vector $A=[a_1,a_2,...,a_N]\in R^{(n\times N)\times N}$ represents the sparse coefficients of all local patches in the candidate target image. Accordingly, the coefficient A can be represented as sparse as follows:

$$A=\left[\underbrace{a_1^{(1)}\ ...\ a_N^{(1)}}_{A^T[1]}\ \Big|\ \underbrace{a_1^{2}\ ...\ a_N^{2}}_{A^T[2]}\ \Big|\ ...\ \Big|\ \underbrace{a_1^{n}\ ...\ a_N^{n}}_{A^T[n]}\right]^T \tag{6}$$

where $A[i]\in R^{(n\times N)\times 1}$ represents the i-th template in A, and $a_i^{(k)T}\in R^{N\times 1}$ represents the k-th local patch corresponding to the i-th template sparse coefficient.

B. Construction of Collaborative Template

In the framework of particle filter, a new collaborative template is constructed by fusing the global template and local template. In the global template based on the color histogram, we use the equation (7) as the similarity measure function, i.e.,

$$\rho_i=\exp(-\frac{d(q_u,p_u(y))}{2\sigma_0^2}) \tag{7}$$

where σ_0 is the standard deviation, ρ_i indicates the similarity in the current frame between the i-th candidate target and the target template, $d(\cdot)$ represents the Bhattacharyya distance, which can be expressed as

$$d(q_u,p_u(y))=\sqrt{1-\sum_{u=1}^{N_0}\sqrt{p_u(y)q_u}} \tag{8}$$

In local template based on sparse representation, we use an indication function o to indicate whether the local patch is occluded, i.e.,

$$o_i=\begin{cases}1 & \varepsilon_i<\varepsilon_0\\ 0 & otherwise\end{cases} \tag{9}$$

where $\varepsilon_i=\|y_i-Da_i\|_2^2$ represents the reconstruction error of local patch y_i; ε_0 is a predefined threshold which determines the local patch whether is occluded. After the weighted coefficient vector:

$$\varphi_i=\frac{1}{C}\sum_{k=1}^{n}o\odot a_i^{(k)},\quad i=1,2,...,\ N \tag{10}$$

where $\odot$ represents the element-wise multiplication; φ_i is the weighted sparse coefficient of the i-th local patch; C is the normalized coefficient. We define the following function to calculate the similarity between the candidate target and the target template, α_i is a sparse similarity, i.e.,

$$\alpha_i=\frac{\varphi_i}{\sum_{i=1}^{N}\varphi_i} \tag{11}$$

When the target is occluded, the local template plays a more important role for determining the tracking results in the collaborative template, so we will be the global template similarity measure function ρ_i linear normalization to abtain $\rho_i^* \in [0,1]$, and ρ_i^* square to reduce the weight value. ρ_i linear normalization, i.e.,

$$\rho_i^* = \frac{\rho_i - \rho_i^{\min}}{\rho_i^{\max} - \rho_i^{\min}} \tag{12}$$

where $\rho_i^{\min}$ and $\rho_i^{\max}$ respectively indicate the minimum and the maximum of the similarity measure function ρ_i.

A new likelihood function expression for collaborative template is obtained by the multiplicative mechanism, Where C_t represents the likelihood function value of the t-th candidate target, the expression is:

$$C_t = (\rho_t^*)^2 \cdot \alpha_t \tag{13}$$

C. Template Update

This paper adopts the method of dynamic update template to ensure the applicability of the appearance model. The algorithm updates a template every 5 frames. Update the template according to the following:

$$q_n = \eta q_0 + (1-\eta) p_l \text{ , } \rho_i < \rho_0 \tag{14}$$

where η is the update rate, the experience value of 0.2~0.4. q_n is a new template, q_o is the template of the first frame, p_l is the latest frame to get the template. We set a threshold ρ_0, the experimental results show that the ρ_0 is generally 0.4. When the similarity function ρ_i is less than a predetermined threshold ρ_0, then the target is in the occlusion state, the template is updated at this time, and vice versa is not updated.

The algorithms of this paper are summarized as shown in Table I:

TABLE I. TRACKING ALGORITHM IN THIS PAPER

Input: Video image sequence $\{F_t\}(t = 1,2,\ldots,n)$.

Output: Target position $\{X_t\}$ $(t = 1,2,\ldots,n)$ for each frame tracking.

1. Template initialization: Set up the initial m frames of image sequences as training images.
2. Construction of global template and local template.
3. Collaborative template representation: We use the multiplicative mechanism to obtain a new likelihood function expression C_t (equation 13).
4. Particle generation: Get the candidate region $X_t = \sum_{i=1}^{n} w_t^i X_{t-1}$ corresponding to each candidate target.
5. Determine the target motion state: According to the motion model to obtain the current frame tracking target affine coordinates, the target to be tracked is scaled.
6. Update the target template: Update the template by solving the (equations 14).
7. Calculate the observation likelihood: By (equations 13) to calculate the candidate target observation likelihood.

IV. Experimental Results and Analysis

In order to verify the effectiveness of the proposed algorithm, a number of standard video data sets were tested, and we did tracking experiments on 2 challenging video sequences. The selected tracking videos are shown in Table II.

TABLE II. TRACKING VIDEO IN THIS PAPER

Sequences	#Frames	Challenging factors
Faceocc2	819	Partial occlusion, in-plane rotation, out-plane rotation
Singer1	321	Illumination change, scale change, pose variation

The experimental results are compared with three state-of-the-art tracking methods including the IVT [9], L1 [4], MIL [10], IVT is blue, MIL is green, L1 is yellow, our tracker is red. The proposed algorithm is implemented in Matlab (R2014a) and runs on a PC with Intel Core i5-4210U CPU (2.40 GHz) with 4 GB memory. The tracking results of the experiments are expressed in rectangular boxes. The tracking algorithm uses in the observation image of 32×32 the size of the window. The overlapped 16×16 local image patches are extracted in the case of moving target area with 8 as the step size. The parameter λ in the formula (5) is fixed to 0.01, the parameter ε_0 in the formula (9) is fixed to 0.1. The number of particles is set to $N_1 = 500$. The experimental results are analyzed and illustrated from two aspects of qualitative and quantitative analysis.

A. Qualitative Analysis

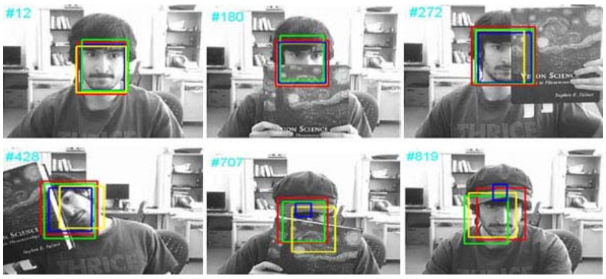

FIGURE I. TRACKING RESULTS OF THE FACEOCC2 SEQUENCE

Faceocc2 video sequence is the face repeatedly occluded and the occurrence of in-plane rotation scene. The tracking results are shown in Figure I , the image frames of the tracking results are #12, #180, #272, #428, #707, #819. The method proposed in this paper shows a better tracking performance. In front of 272 frames, four algorithms are able to stabilize the tracking of the target; IVT, MIL and L1 methods are offset by a small amplitude in the #428 frame; in the #707 frame, the scale of the tracking results of IVT algorithm is severely reduced, and the target is lost in the following tracking process, but MIL, L1 algorithm is relatively better performance.

In the Singer1 video sequence, the stage of the female singer in the performance of the process has undergone dramatic change in light and scale change in the scene. The tracking results are shown in Figure II, the image frames of the tracking results are #1, #96, #125, #175, #244,#321. The IVT algorithm has a drift in the #125 frame, and gradually deviates from the target in the following tracking, but the algorithm is not completely lost; MIL algorithm does not take into account the scale change, starting from the #96 frame, the tracking rectangle is significantly larger than the real target. Compared with the four methods, the method proposed in this paper and the L1 method are both offset smaller, and the tracking is more accurate.

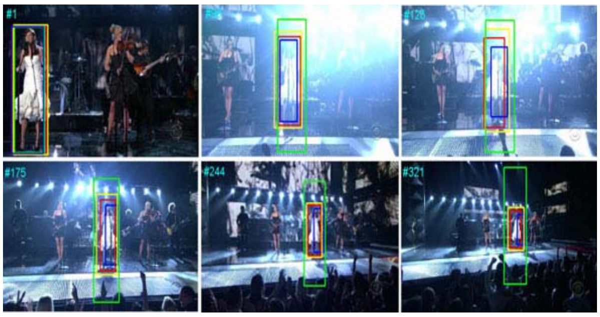

FIGURE II. TRACKING RESULTS OF THE SINGER1 SEQUENCE

B. Quantitative Analysis

In this paper, we take the center position error to quantitatively evaluate the algorithm. The results of the analysis are shown in Table III and Figure III-Figure IV. Figure III-Figure IV is a line graph constructed by the center position deviation and frames.

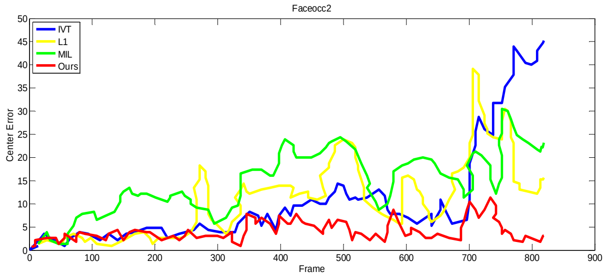

FIGURE III. CENTER ERROR PLOT OF THE FACEOCC2 SEQUENCE

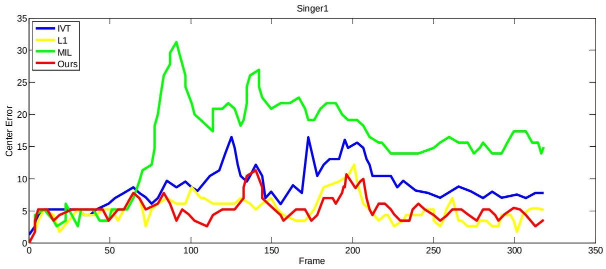

FIGURE IV. CENTER ERROR PLOT OF THE SINGER1 SEQUENCE

Table III shows the 4 methods of tracking test results in average respectively center error of 2 groups of test video, traditional bold underlined represents the best tracking results.

TABLE III. ANALYSIS OF THE MEAN CENTER ERROR OF THE ALGORITHM

	IVT	**L1**	**MIL**	**Proposed**
Faceocc2 Singer1	13.77 8.86	14.60 **<u>5.07</u>**	15.40 16.25	**<u>5.33</u>** 5.30

As can be seen from Table III, proposed method can accurately track the target of Faceocc2 . In the Singer1 video sequence, the results obtained by proposed method and L1 tracker are most similar; in the Faceocc2 sequence, proposed method obtains the minimum mean center error. Compared with the other three tracking devices, the tracker in this paper has the best tracking effect in Faceocc2 sequences. Considering the overall performance, the performance of proposed method is the best in the tracking process.

V. CONCLUSIONS

In this paper, we propose a new algorithm based on collaborative template, and improves the performance of the template and the accuracy of tracking; the local image is weighted by sparse coefficients, which can better deal with the influence of occlusion on tracking; the dynamic template updating strategy is adopted to deal with the change of the appearance of the target. Experimental results show that the proposed algorithm has better tracking performance than IVT, MIL and L1 algorithm in the case of target rotation, scale changes and partial occlusion. However, when this algorithm is used for template update, the adaptive ability of the algorithm is decreased due to the fixed threshold.

ACKNOWLEDGMENT

This research is supported by the National Natural Sciece Foundation of China (No. 61563030).

REFERENCES

[1] B. Tian, Y. Li, B. Li, and D. Wen, "Rear-View Vehicle Detection and Tracking by Combining Multiple Parts for Complex Urban Surveillance", IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 2, **(2014)**, pp. 597-606.

[2] H. Hasan, and S. Abdul-Kareem, "Static hand gesture recognition using neural networks", Artificial Intelligence Review, vol. 41, no. 2, (**2014**), pp. 147-181.

[3] M. Dubska, A. Herout, R. Juranek R, and J. Sochor, "Fully Automatic Roadside Camera Calibration for Traffic Surveillance", IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 3, **(2015)**, pp. 1162-1171.

[4] X. Mei， and H. B. Ling, "Robust visual tracking using L1 minimization", Proceedings of the IEEE International Conference on Computer Vision (ICCV), **(2009)**, pp. 1436-1443.

[5] D. Wang, H. C. Lu, and M. H. Yang, "Online object tracking with sparse prototypes", IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, vol. 22, no. 1, **(2013)**, pp. 314-325.

[6] H. N. Zhao, X. Wang, and M. Liu, "Robust Object Tracking with Occlusion Handling based on Local Sparse Representation", International Journal of Signal Processing Image Processing and Pattern Recognition, vol. 7, no. 3, **(2014)**, pp. 407-420.

[7] T. X. Bai, Y. F. Li, and X. L. Zhou, "Learning local appearances with sparse representation for robust and fast visual tracking", IEEE Transactions on Cybernetics, vol. 45, no. 4, **(2015)**, pp. 663-675.

[8] X. Jia, "Visual tracking via adaptive structural local sparse appearance model", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, **(2012)**, pp. 1822-1829.

[9] D. A. Ross, J. Lim, R. S. Lin, and M. H. Yang, "Incremental learning for robust visual tracking", International Journal of Computer Vision (IJCV), vol. 77, no. 1 **(2008)**, pp. 125-141.

[10] B. Babenko, M. H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning", IEEE Conference on Computer Vision and Pattern Recognition, **(2009)**, pp. 983-990.