International Journal of Multimedia and Ubiquitous Engineering
Vol.12, No.1 (2017), pp.141-152
http://dx.doi.org/10.14257/ijmue.2017.12.1.12

Design and Optimal an Object Tracking Method based on Hybrid Templates: Experimental Analysis of Video Sequences

Dawei Yang and Fan Zhou*

Shenyang Ligong University, Shenyang, 110159, CHINA
Corresponding author: Fan ZHOU, yahoo_zf@163.com

Abstract

To improve the robustness of object tracking method, the study on tracking method based on sparse representation is done in the paper, and a new object tracking method based on hybrid templates is proposed. The sparse representation of global template to candidate target generates reconstruct error, and the sparse representation of local structural sparse dictionary to candidate target generates similarity function. The optimal discriminate result of the logistic decision function which combine two models regard as tracking result, the experimental results and analysis demonstrate the performance of the proposed method.

Keywords*: Object tracking, Sparse representation, Logistic decision function, Hybrid templates*

1. Introduction

According the difference in extracting model [1], object tracking methods can be divided into two classes, that is, the tracking methods based on global model and the tracking method based on local model. The methods based on global model [2-5] look the tracked target as a whole, extract the whole feature and model the representation. This kind of methods are fit for the situation that the information of the target is intact, while unsuitable when there is partial occlusion. The methods based on local models [6-8] segment the target as many fragments according to a certain spatial arrangement, extract feature vector in each picture fragment and model the appearance with these vectors. The methods can deal with the appearance change such as partial occlusion. In practical applications, the combination of these two kinds of extracting feature is used to model the target appearance.

2. Global Template and Feature Selection

The common way of selecting positive and negative samples is handmade in the first frame. In our method, a rectangle is drawn around the tracked target, the size of the rectangle is same as the tracked target region, then a bigger rectangle is drawn around the first one, that is, there is an annular rectangle. The negative samples will be produced randomly within the annular rectangle using particle filtering framework [9]. Given the number of particles is N, there will be N negative templates, and the information in each template include most background and a small part of target. In the paper, the handmade selection of the target region is as positive sample, and there is one positive sample and N negative samples. In consideration of the efficiency, the selected image regions are stretched into the same size ($n \times n$), then these image fragments are transformed into vectors, which contain target information if it is positive sample, while contain partial target information and most background information if it is negative sample.

In each frame, there will be N candidate regions selected around the prior tracking result using particle filtering. To obtain better tracking results, six affine transformation

ISSN: 1975-0080 IJMUE

parameters are used to describe the target moving. That is, the state variable at time t is denoted as $z_t = [x_t, y_t, \alpha_t, \beta_t, \phi_t, \gamma_t]$, these parameters correspond to six affine transformation parameters of rectangle respectively, *i.e.* the offset values in x and y direction, scaling, aspect ratio, rotation angle and angle of inclination from time t-1 to t. These affine transformation parameters are always supposed to be mutual independence.

It is redundancy using gray feature to separate target from background, so the feature selection can use formula (1)

$$\min_s \left\| A^T S - p \right\|_2^2 + \lambda \left\| S \right\|_1 \tag{1}$$

Where, $A \in R_K \times N_n$ is composed of N_n negative templates. K is the number of dimension before feature selection, vector $p \in R^{N_n} \times 1$ represent the property of each template in training sample set. The result of formula (1) is sparse coefficient vector s, in which the nonzero element represent the classification feature vector extracted from original K dimensional feature space. The feature selecting method ensure that the suited classifying feature vector can be selected from dynamic environment adaptively.

The mapping matrix S project the original feature space to feature space after classifying selection, and this can be done by removing all zero elements from diagonal matrix S', the elements in which is determined by formula (2)

$$S'_{ii} = \begin{cases} 0 & s_i = 0 \\ 1 & others \end{cases} \tag{2}$$

Where, the diagonal element S'_{ii} equal zero only when s_i equal zero. The training template set and candidate sample set are projected to discriminative feature space by particle filtering, and the space after projection is composed of training template set $A' = SA$ and $x' = Sx$.

Given a target candidate, its sparse coefficient represented by training sample set can be acquired by formula (3)

$$\min_\alpha \left\| x - A\alpha \right\|_2^2 + \lambda \left\| \alpha \right\|_1 \tag{3}$$

There is a reconstruction error when a candidate sample is represented by sparse dictionary, which can be gain by formula (4)

$$\delta_b = \left\| x - A\alpha \right\|_2^2 \tag{4}$$

3. Local Model and Occlusion Estimation

The sake of simplicity, intensity feature is select to present the local information of image in the paper. The target region is fragmentized locally on the basis of certain spatial arrangement. A number of local image fragments are gained using a sliding window, which overlap the target region. Each fragment is stretched into a form of vector and normalized as $yi \in R^G \times 1$, where, G is the size of the image fragment. In line with sparse representation theory [10-11], the sparse code of each image fragment can be obtained by formula (5)

$$\min_{\beta_i} \left\| y_i - D\beta_i \right\|_2^2 + \lambda \left\| \beta_i \right\|_1 \tag{5}$$

The corresponding histogram of each local fragment is formed by concatenating these sparse codes together, denoting by $\rho = [\beta_1, \beta_2, \ldots, \beta_M]^T$, which represents a histogram of a candidate target region.

To deal with occlusion, it can be done by modifying the constructed histogram to exclude the occluded image fragment. The fragment with maximum reconstruct error is regarded as occlusion, and the responding sparse coefficient is set zero. So the weighted histogram can be generated by formula (6)

$$\phi = \rho \otimes o \tag{6}$$

Where, operator $\otimes$ represents convolution, each element in o represents the corresponding occlusion fragment, which may be gained by formula (7)

$$o_i = \begin{cases} 1 & \varepsilon_i < \varepsilon_0 \\ 0 & others \end{cases} \tag{7}$$

Where, $\varepsilon_i = \left\| y_i - D\beta_i \right\|_2^2$, it is the reconstructed error of y_i, and ε_0 is the predefined threshold value to decide whether the occlusion happened in local fragment. Thus, there is a histogram ϕ_j based on sparse coefficient corresponding to each candidate sample. In consideration of the local spatial information and occlusion, the method is more effective and robust.

A histogram function is adopted to represent the similarity of histogram between candidate target and initial template. The function is same with the literature [12], that is

$$H_C = \sum_{j=1}^{J \times M} \min(\phi_C^j, \varphi^j) \tag{8}$$

Where, ϕ_C^j and φ^j represents the histogram of the candidate region numbered c and template, respectively. The fragments in template can be obtained in labeled region in the initial frame, the histogram corresponding to each image sequence compute only once, while the template of negative need to be updated every other five frames.

4. Tracking Method Based on Logistic Decision Function

In the preceding part of the paper, it is introduced that the reconstructing error based on negative sample to candidate target is regarded as the discriminating criterion. The smaller is the reconstructing error, the more possible is the candidate region sparsely represent by negative sample set. In other words, there are more information about target and less information about background in the candidate region. On the contrary, if the reconstructing error is bigger in the candidate region which reconstructed sparsely by negative sample, which shows that there are more information about background and less information about target.

The construction of local structural model is based on similarity histogram modeling, in which the target region labeled in initial frame is regarded as the positive sample. Firstly, the positive sample is fragmentized based on overlapping window. Secondly, the fragment histogram of positive sample template and candidate region produced by particle filtering are computed respectively. Finally, the result set of target is selected by similarity function. To ensure the robustness of tracking, the global template and local templates are combined together, then the judgement and selection is done by logistic decision function, which adopted in the paper is showed in formula (9)

$$M = \log(1 + e^{wz}) \tag{9}$$

Where, $z = H_c / \delta_b$, H_c represents the similarity function based on local structural model. The bigger is the similarity between the histogram of candidate region and the histogram of positive sample generated in initial frame, the bigger of the possibility of the region is the tracking result. While δ_b represents the reconstructing error of the candidate region which is represented by negative sample template set sparsely. The smaller is the

error, the more possible is that the candidate is the tracking result set. W represents the normalized weighted coefficients. Therefore, combining the global template and local model together using the aforementioned function, we select the maximum value of the logistic decision function and regarding it as the tracking result, which will make the tracking result more accurate and robust.

5. Experimental Results and Analysis

To verify the effectiveness of the method, we do some experiments with the computer with Inter(R) Core(TM) i5-2004 CPU @3.10 GHz 4.00GB, and realize our method in Matlab 2010. Aiming at the difficult problem in tracking, we do the experiments using the representative video sequence, such as deer, board, car, caviar, faceocc2, girl and singer. There are one or more interference factors in the sequence. The particle number generated by particle filtering is set to 100, the number of global templates is set to 200, and the size of the global template is 32×32 pixels. The size of the image fragment in the local model is 6×6 pixels, the number of local model is 196. The initial target region is labeled manually.

5.1. Experimental Sequence Selection and Analysis

Aiming at different tracking scene and different problem, we select different sequence to do experiment. Target moving quickly. In the sequence deer, the target move quickly, and there are some similar disturb factor in the background. It is difficult to track the target in the sequence, the tracked target will lost or track a wrong target. Some tracking result of our method shows in Figure 1.

Figure 1. Tracking Result of Deer Sequences

Occlusion problem. In the sequence faceocc2, there are partial occlusion in the face with a book, and the book moves with the face's movement. In the sequence caviar, in the process of tracking the target, some similar background target appear and occlude the target heavily. Some tracking result of our method shows in Figure 2 and Figure 3.

Figure 2. Tracking Result of Faceocc2 Sequences

Figure 3. Tracking Result of Caviar Sequences

Illumination and scale change. The typical sequence of illumination change and scale is singer1. In the earlier stage of the sequence, the illumination is a little of dark, and it is simple to track the target, many tracking methods can deal with it well. While in the middle stage of the sequence, there are severe change in illumination and scale, it is hard for many tracking methods to track the target accurately, the phenomenon of drift and lose will happen from this stage, such as from frame 120. Some of our tracking results are showed in Figure 4.

Figure 4. Tracking Results of Singer1 Sequences

Rotation. In the sequence of girl, the target *i.e*. the head of the girl, rotate along with the moving from left to right, and from front to back. At the same time, there is scale change with the target's movement. The rotation is harder to track accurately than simply drift motion. Our method can track the target accurately even though rotation happened in the sequence. Some tracking results is showed in Figure 5.

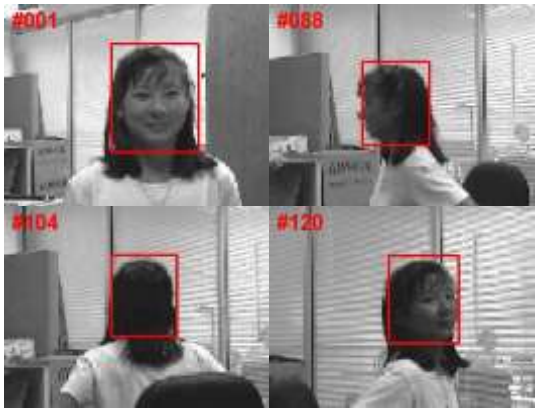


Figure 5. Tracking Results of Girl Sequences

Complicated background. The tracking in complicated background is always a difficult problem in object tracking domain. There are so many objects in complicated background, and the features of target and other objects are similar in a certain extent. The updating of appearance model and template have much impact in tracking results. Once a similar object of background is trained, the tracking will fail. In the paper, two sequences are selected, *i.e*. board and car. In the board sequence, the complicity of the background is the main factor, while the main factor include complicity of background, illumination change and scale in car sequence. Part of the tracking results is showed in Figure 6(board) and Figure 7(car).

Figure 6. Tracking Results of Board Sequences

Figure 7. Tracking Results of Car Sequences

5.2. Contrast Experiment and Qualitative Analysis

The method in this paper combines the global template and local template, so we select two similar ones in many typical methods to do comparison. In 2012 CVPR conference, Xu proposed a tracking method (ASLS) [7], which adopted local template to model the appearance of target. Wang proposed a method (PLS) [13] modeling the appearance based on global template. The comparison among these methods can account for the availability of our method.

These methods are implemented in the same compiled environment as front section. The comparisons are done with five different sequences, that is, david3, faceocc2, singer1, shaking and deer. In the following result images, red rectangle represents the proposed method in the paper, green rectangle represents LPS method, and blue rectangle represents ASLS method.

Figure 8. Contrast Results of Deer Sequence

In the contrast results (Figure 8), the proposed method and ASLS can both track the target accurately, while PLS lose the target from the fifth frame. The contrast results show that the tracking method based on global template perform better than the method based on local structural model in the sequence including target moving quickly.

With regard to the occlusion, three methods show similar performance when the occlusion is slight. While the occlusion is hard, such as the pedestrian is occluded heavily by the roadside trees in sequence david3, ASLS lose the target and can't continue to track when the pedestrian pass through the trees. PLS will lose target randomly, Whilst, the proposed method can track the target in the sequence accurately and shows better performance than others.

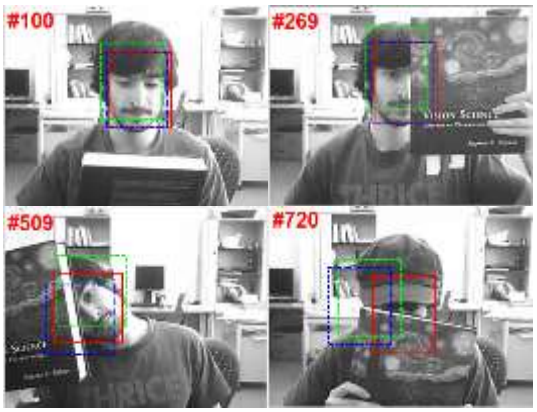


Figure 9. Contrast Results of Faceocc2 Sequence

Figure 10. Contrast Results of David3 Sequence

The proposed method perform better than others in the sequence of complicated background and illumination change. The drift happens in the tracking results of ASLS and PLS in sequence singer1 (Figure 11). In the sequence shaking (Figure 12), PLS begin to lose the target from the fiftieth frame. ASLS displays random deviation and track the target unfaithfully.

Figure 11. Contrast Results of Singer1 Sequence

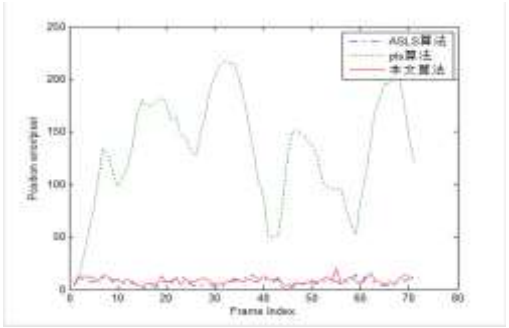
Figure 12. Contrast Results of Shaking Sequence

5.3 Contrast Experiment and Quantitative Analysis

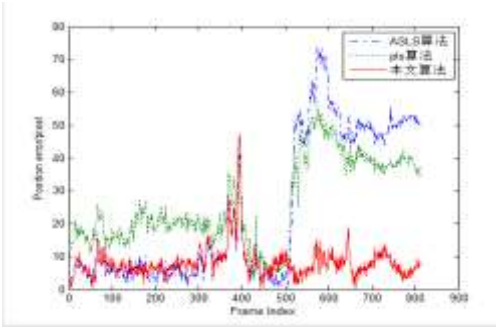
Comparing the tracking results with the benchmark information of each sequence, the result is used to compare the tracking accuracy. There are many computing method, the method in the paper is formula (10)

$$\varepsilon = \sqrt{(x - x_0)^2 + (y - y_0)^2} \tag{10}$$

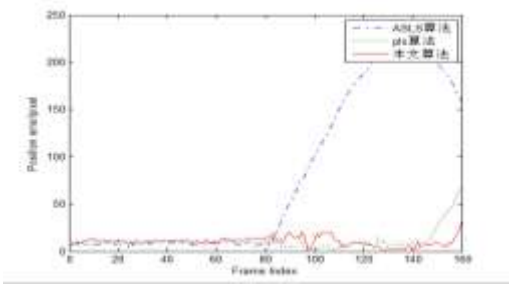
Where, (x, y) represents the center coordinates of the tracking result, (x_0, y_0) represents the center coordinates of the benchmark. The result of the formula is called center error. The diagram of center error of the comparison among the three methods show in Figure 13. The red line, green line and blue line represent the center error of the proposed method, PLS and ASLS respectively. The center error value of these method in the sequence is showed in Table 1.



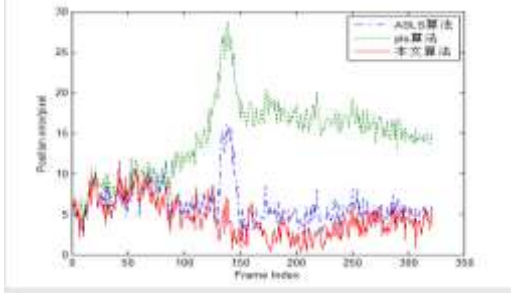
(a) Center Error of Deer



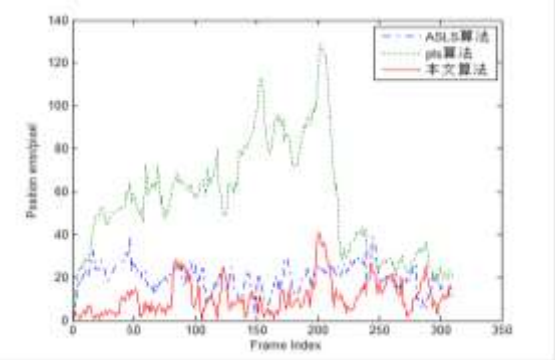
(b) Center Error of Faceocc2



(c) Center Error of Avid3



(d) Center Error of Singer1



(e) Center Error of Shaking

Figure 13. Analysis Diagram of Center Error

From the analysis above, the performance of ours is better than ASLA and PLS. PLS is good at dealing with occlusion, while ours and ASLS perform well to deal with other disturb factors.

Table 1. Center Error Value Unit: Pixel

	s inger1	fa ceocc2	d avid3	s haking	d eer
ours	4.48	8.17	1 0.55	1 0.89	8.35
PLS	14.31	26.19	1 1.13	5 5.96	133.35
ASLS	6.15	23.70	7 9.95	1 9.55	6.91

6. Conclusion

To improve the robustness of target tracking technology, the study on tracking method based on global template and method based on local template is done in the paper. A new tracking method based on logistic decision function is proposed. The global template and local template is combined together to represent the target in the method. The experimental results and analysis demonstrate the performance of the proposed method.

Acknowledgments

The paper is funded by the project of "Natural Science Foundation Guidance Plan of Liaoning Province" (NO:2016010993-301), "the doctoral scientific research foundation of Liaoning province" (NO:LG201609), "the doctoral scientific research foundation of Shenyang Ligong University" (NO:BS-2015-03).

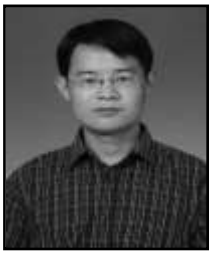
References

[1] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem", PAMI, vol. 26, **(2004)**, pp. 810-815.
[2] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking", PAMI, vol. 25, no. 5, **(2003)**, pp. 564-575.
[3] J. Kwon and K. M. Lee, "Visual tracking decomposition", In CVPR, **(2010)**, pp. 1269-1276.
[4] X. Mei and H. Ling, "Robust visual tracking using L1- minimization", In ICCV, **(2009)**, pp. 1436-1443.
[5] X. Mei, H. Ling, Y. Wu, E. Blasch and L. Bai, "Minimum error-bounded efficient L1 tracker with occlusion detection", In CVPR, **(2011)**, pp. 1257-1264.
[6] B. Liu, J. Huang, L. Yang, and C. A. Kulikowski, "Robust tracking using local sparse appearance model and k-selection", In CVPR, **(2011)**, pp. 2968-2981.

[7] X. Jia, H. Lu and M. H. Yang, "Visual Tracking via Adaptive Structural Local Sparse Appearance Model", In CVPR, **(2012)**, pp. 1822-1829.
[8] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. A. Kulikowski, "Robust and fast collaborative tracking with two stage sparse optimization", In ECCV, **(2010)**, pp. 624-637.
[9] A. Doucet, N. D. Freitas and N. Gordon, "Sequential Monte Carlo methods in practice", New York: Springer, **(2001)**.
[10] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry and Y. Ma, "Robust face recognition via sparse representation", PAMI, vol. 31, no. 2, **(2009)**, pp. 210-227.
[11] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification", In CVPR, **(2009)**, pp. 1794-1801.
[12] W. Zhong, H. Lu and M. Yang, "Robust Object Tracking via Sparse Collaborative Appearance Model", Ranaon on Mag Rong: a Blaon of H Gnal Rong Oy, vol. 23, **(2014)**, pp. 2356-2368.
[13] Q. Wang and F. Chen, "Object Tracking via Partial Least Squares Analysis", Image Processing, IEEE Transactions on, vol. 21, no. 10, **(2012)**, pp. 4454- 4465.

Authors

Dawei Yang, was born in Liaoning Province, China, in 1976, is a lecturer of School of Information Science & Engineering, Shenyang Ligong University. He has published more than 10 papers on international journals, national journals, and conferences in recent years. His research interests include: image processing and pattern recognition.



Fan Zhou, was born in Shanxi Province, China, in 1976, is a associate professor of School of Information Science & Engineering, Shenyang Ligong University. He has published more than 15 papers on international journals, national journals, and conferences in recent years. His research interests include: Wireless communication/image signal processing and electronic countermeasures.