Robust Object Tracking via Multi-block and Sparsity-Based Representation

Wu Wang, Jinxu Tao, Weiquan Ye, Yongjun Jiang, and Zhongfu Ye

***Abstract*—In this paper, we propose a robust object tracking algorithm based on classifier, multi-block and local sparse coding with multiple discriminative dictionary. The first part of the propose method is train a classifier with the dictionary encodes the information of both target information and background information. The second part exploits the block information of the object target. The different blocks in a sample should contribute differently to the visual tracking, the model effectively exploit the similarity and distinctiveness of different blocks. Each block is coded on its own discriminative dictionary to allow flexible coding block and the parameter after sparse coding can be used for the weights allocation simultaneously. Furthermore, the update scheme considers both the latest observations and the original template, thereby enabling to alleviate drift problem. Extensive experiments on challenging sequences show that the robust tracking achieved by our algorithm.**

***Index Terms*—Visual tracking, sparse representation, multi-block.**

I. Introduction

Visual object tracking is a significant computer vision task which can be applied to many domains, such as visual surveillance, human computer interaction, and video compression. Despite extensive research on this topic, it still suffers from difficulties in handling complex object appearance changes caused by factors such as illumination variation, partial occlusion, shape deformation, and camera motion. Therefore, effective modeling of the 2D appearance of tracked objects is a key issue for the success of a visual tracker [1]. A variety of tracking algorithms have been proposed to overcome these difficulties.

These methods can be formulated in two different ways: generative model and discriminative model. Generative methods represent objects with models that find the most similar candidate to the target appearance. The work [2]-[4] belong to the generative model. Recently, some generative model exploit the development of sparse representation in tracking [5]-[7] due to its robustness to occlusion and image pose change. It was proposed to update the target appearance model incrementally for adapting to dynamic environmental changes and object appearance variations. Discriminative methods formulate the tracking as a binary classification problem [8]-[11] to find a best decision boundary that can best separate the target from background. For using the background information, these method can be strong robustness to avoid the target judgment as background. Furthermore, the classifier can be online updated during the tracking procedure to handle the pose change.

This paper is structures as fellows. The next section reviews some related work on sparse representation for visual tracking. Section III introduces the method that learn object appearance model using local sparse coding and discriminative dictionary with a trained classifier. Learning object appearance model using multi-block sparse coding and the discriminative dictionary, then robust weights allocation of different blocks are described in the Section IV. The detail of the proposed tracking algorithm is shown in Section V. Experimental results and some discussions are shown in Section VI. We conclude the paper in the last section.

Manuscript received February 27, 2015; revised April 8, 2015.

Wu Wang, Jinxu Tao, Yongjun Jiang, and Zhongfu Ye are with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, National Engineering Laboratory for Speech and Language Information Processing, China, Hefei, Anhui 230027, China (e-mail: wangwu6@mail.ustc.edu.cn, jyj365@mail.ustc.edu.cn, tjingx@ustc.edu.cn, yezf@ustc.edu.cn).

Weiquan Ye is with China Tobacco Anhui Industrial Co., Ltd, Hefei, Anhui 230088, P. R. China (e-mail: yewq@ustc.edu.cn).

II. Related Work

Recently, sparse representation has been applied to vision problem [12], including image enhancement [13], face recognition [14], and visual tracking [5]. L1 tracker [5] apply sparse representation to visual tracking and deal with occlusions via trivial templates. The sample with minimal reconstruction errors is regarded as tracking result. Further, [15] consider the important of the background information. The dictionary is composed of target templates and background templates which overcomes the drawbacks of L1 tracker that provide more discrimination power than the dictionary used in [5]. Another advantage of [15] is it selects discriminative features through the classification information which can decrease the dimension of the representation. In [16], they using a linear support vector machine (SVM) to train a discriminative model to separate the target object from the background.

Occlusion is one of the most challenge problems in object challenge for some of the information loss. Yang *et al.* [17] proposed a visual tracking approach based on "bag of features" algorithm. It's robust in handling occlusion, scaling and rotation. Adam *et al.* [11] presents the "frag-track" algorithm to handle the occlusions problem. The object is represented by randomly multiple image fragments or patches. It combines fragments based representation and voting map with the integral histogram tool which is robust to partial occlusions. Nevertheless, the template is not updated and sensitive to large appearance variations.

DOI: 10.7763/IJKE.2015.V1.13

III. Learning Object Appearance Model with Discriminative Classifier

The first part of the model is using the local target information to construct foreground dictionary and background dictionary. Like [16], we use overlapped sliding windows on the target object region and background region to obtain foreground patches and background patches. The foreground dictionary $D_p \in R^{G\times J}$ is generated from k-means cluster center via the patches belong to the foreground patches. Similarly, we can obtain $D_n \in R^{G\times L}$ from background patches, where G denotes the size of the patch. J and L is the number of cluster center of the foreground dictionary and background dictionary, respectively. Then we can construct the overall dictionary $D = [D_p, D_n] \in R^{G\times(J+L)}$. Firstly, extract R patches from image region and each patch is converted to a vector $x_i \in R^{G\times 1}$. The sparse coefficient vector α_i is calculated by

$$\min_{\alpha_i} \frac{1}{2}\left\|x_i - D\alpha_i\right\|_2^2 + \lambda_1 \left\|\alpha_i\right\|_1 \quad (1)$$

where $\left\|\bullet\right\|_1$ and $\left\|\bullet\right\|_2$ denote the l_1 and l_2 norms respectively, $\left\|x_i - D\alpha_i\right\|_2^2$ is the reconstruction error and λ_1 is a scalar constant. It is easy to prove that the distribution of the sparse codes is different between the positive sample and negative sample. The M_{pos} positive samples is extracted around the target location within a radius of a few pixels using the sliding window method. Similarly, the M_{neg} negative samples is extracted further away from the object location. Then, compute the sparse code of each image patch to form the training data $\{\alpha_i, y_i\}_i^M$, where the $\alpha_i \in R^{(J+L)R}, y_i \in \{+1, -1\}$ and $M = M_{pos} + M_{neg}$. With the training data, our linear classifier can defined as:

$$\min_w \frac{1}{M}\sum_{i=1}^{M} L(y_i, w, \alpha_i) + \frac{\lambda_2}{2}\left\|w\right\|_2^2 \quad (2)$$

where w is the classifier parameter, λ_2 controls the strength of the regularization term. $L(\bullet)$ is the loss function which is defined as:

$$L(y, w, \alpha) = \log(1 + e^{-yw^T\alpha'}), \quad (3)$$

where $\alpha' = [\alpha^T, 1]$ is the augmented vector. The corresponding classification score of any candidate is

$$H = \frac{1}{1 + e^{-w^T\alpha'}} \quad (4)$$

With the initialized of the classifier, the score can be regard as the similarity measure for tracking. A candidate with largest score indicates that it is can be considered as the tracking result.

IV. Multi-block Sparse Coding and Weight Allocation

Motivated by the success of frag-track to deal with occlusion, we present the second part of the model based on multi-block sparse coding.

A. Multi-block Sparse Coding

Each block dictionary set is composed of N_p positive template and N_n negative template. To construct each foreground dictionary, we extract the image patches using the dividing target region method. Firstly, the select image region are normalized to the same size (36×36 in our experiments). Then each image region is divided into K blocks (4 in our experiments). Each block is stacked to form the corresponding template vector. Then aggregate the template vector to form the target basis set $\mathrm{D}_p^k = [\mathrm{t}_1^k, \ldots, t_{N_p}^k] \in R^{d\times N_p}$, where t_i^k is the i th vectorized foreground region of the k th block and d is the dimension of the vector. Similarly, the negative template is extracted further away from the object location and

also divided into K blocks. The background basis set is $\mathrm{D}_\mathrm{n}^k = [\mathrm{t}_1^k, \ldots, t_{N_n}^k] \in R^{d\times N_n}$. Each block can obtain an overall dictionary $\mathrm{D}^k = [\mathrm{D}_p^k, \mathrm{D}_n^k] \in R^{d\times(N_p+N_n)}$. In current frame, we draw N candidate particle around the tracked result in the previous frame with a particle filter through affine transformation[15]. Each candidate particle divided into K blocks and the block sample is $x^k = [X_1^k, \ldots, X_N^k] \in R^{d\times N}$, similarly. The affine parameters can be modeled with six scalar Gaussian distributions. Then the coefficients $\alpha^k = [\alpha_1^k, \ldots, \alpha_N^k] \in R^{(N_p+N_n)\times N}$ is computed by

$$\min_{\alpha^k} \left\|\mathrm{x}^k - \mathrm{D}^k\alpha^k\right\|_2^2 + \lambda_3 \left\|\alpha^k\right\|_1 \quad (5)$$

where λ_3 is the sparse parameter. (2) is actually the Lasso regression problem that can be solved efficiently by Least angle regression (LARS). Then we can obtain K sparse code for *K* blocks.

B. Occlusion Analysis

In order to deal with occlusion, we construct multiple block to represent the sample. When the patch is not occluded, the sparsity of the different blocks show in Fig. 1. The left image show the original image and the sample of four blocks. The right show the four sparse vector of one sample through the solve of (2). By checking the coefficients in α^k, we found that the sparse coefficients mainly concentrated in the first 200 which is the number of the positive template (i.e. N_p), the coefficients corresponding to the background is almost all of

zeros. This indicates that target can be well representation by foreground dictionary D_p^k. When the patch is occluded like the Fig. 2. The sparse coding of the complete patch are dispersed in the foreground and background region. However, the coefficients of different blocks are discriminative. It's easy to find that the block which is not occluded can be can be well represented by the foreground dictionary. In contrast, the sparse coefficients of occluded blocks are scattered in the region of all the foreground and background.

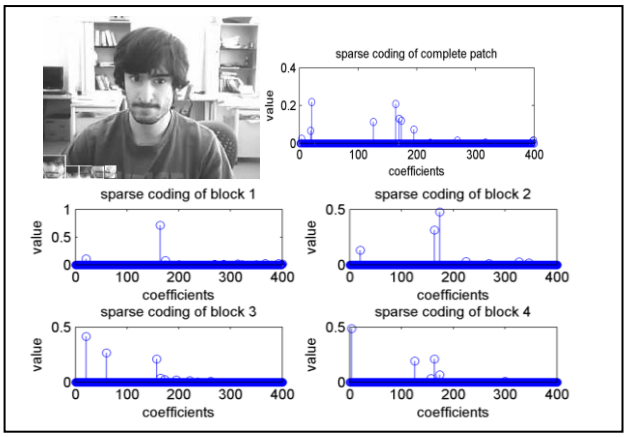

Fig. 1. The patch is not occluded. Top left: The original image and target candidatesthe sample with four blocks. Top right: The sparse coding of the complete patch. Bottom: The sparse coding of the four block.

In order to handle the occlusion problem, we exploit the discriminative information mentioned above. The block with large sparse coefficients in background region can be considered as occlusion. With allocating different weight of different blocks, the tracker can reduce the influence of occlusion.

C. Weight

Intuitively, the ω^k is used to indicate the distinctiveness of different blocks. If the k th block is not occluded, ω^k should be bigger. Besides, in order to exclude the influence of occluded blocks when describing the target object, the blocks which are well represented by background should be allocated a small weight. Especially, the weight is set close to zero when the block is occlusion completely. With the discussion above, we define a reliability factor q^k as

$$q^k = \exp\left(\mu \frac{\left\|\alpha_+^K\right\|_2^2}{\left\|\alpha^K\right\|_2^2}\right) / \exp\left(\gamma \frac{\left\|\alpha_-^K\right\|_2^2}{\left\|\alpha^K\right\|_2^2}\right) \quad (6)$$

where μ and γ are positive constants to control the tradeoff between foreground sparse coefficients and background sparse coefficients, α_+^K and α_-^K means the positive and negative sparse coding of the k th block. The function of reliability factor is to underline the importance of the positive sparse coding and enhance the discriminative of different block. Then the weight of k th block is

$$\omega^k = \exp(\tau q^k), \quad (7)$$

where the τ is a constant controlling the value of weight. For tracking at time t, every block of each candidate can obtain a weight. Through the similarity measurement, it can improve the robustness of the tracker.

D. Confidence Measure

The proposed algorithm is developed based on the assumption that the target can be better represented by the linear combination of positive templates while the background can be better represented by the span of negative templates. Given the candidate, it is represented by the multiple dictionary set with the coefficients α computed by (2) and (3). The reconstruction error of the candidate x^k with the foreground template set D_p^k is

$$\varepsilon_f^k = \left\|x^k - \mathrm{D}_p^k \alpha_+^k\right\|_2^2. \quad (8)$$

Similarly, the reconstruction error using background template set D_n^k is

$$\varepsilon_f^k = \left\|x^k - \mathrm{D}_p^k \alpha_+^k\right\|_2^2. \quad (9)$$

A candidate with smaller reconstruction error using the foreground dictionary indicates it is more likely to be target object, and vice versa. Thus, we formulate the confidence by:

$$s = \exp\left(-\left(\sum_{k=1}^{K} w^k(\varepsilon_f^k - \varepsilon_b^k)\right) / \sigma\right), \quad (10)$$

where the variable σ is a small constant to balances the confidence values. The tracking result is the candidate with the highest probability.

V. Proposed Tracking Algorithm

A. Two Complementary Part of the Model

The proposed algorithm composes of two parts within the particle filter framework. In our tracking algorithm, the confidence value is based on the classification score and similarity function of multi-block sparse coding. The first part employs the sparse coding and linear classifier to capture the local feature of the target. All of these local features are concatenated together to represent the object target, therefore it models the global information of the target region actually. Besides, the second part introduces the local block information which shares the idea as the part-based tracking method that can handle the partial occlusion. The confidence value of the candidate is computes by

$$P = \frac{1}{1 + e^{-w^T \alpha}} \exp\left(-\left(\sum_{k=1}^{K} w^k(\varepsilon_f^k - \varepsilon_b^k)\right) / \sigma\right) \quad (11)$$

The tracking result is the candidate with the largest confidence value.

B. Update Scheme

There is no need to update negative templates every frame

due to the background usually changes little. In our experiments we update the negative templates every 5 frames from image region away from the current tracking result. For positive templates, we should judge the object whether occlusion occurs or not at first. With the discussion of occlusion and weight allocation we can find that sparse coefficient of block scattered in the region of the entire foreground and background may indicate occlusion. We set a threshold q_0. For the classifier model, we update the dictionary and classifier when block is occluded which can capture the new appearance and is adapted to the altered environment. For the multi-block part, the positive block templates remain the same in the entire sequence to keep the initial value correct and distinct.

VI. Experimental Results

In our experiment, the object target is initialized according to the first frame in ground truth. We evaluate the performance of the model through conducting experiments on nine challenging sequences. These sequences cover the most challenge situations in object tracking: heavy occlusion, motion blur, in-plane and out-of-plane rotation, large illumination change, scale variation and complex background. We will compare our tracking experiment result with five algorithms: Frag tracker [11], L1 tracker [5], MIL tracer [8], and ODLSR tracker [16]. The tracking results of the compared methods were obtained by running the source code or binaries provided by their authors using the same initial positions in the first frame.

The parameters in our experiment are presented follow. The numbers of cluster center of the foreground dictionary J and background dictionary L are both 50, the train numbers of foreground templates M_{pos} and background templates M_{neg} are both 30. The sparse parameter λ_1 and λ_3 are fixed to be 0.01. The numbers of block dictionary set N_p and N_p are both 200. The variables μ, γ and τ in our experiment is fixed to be 2.0, 0.1 and 0.5, respectively. The weight balance factor σ in equation (10) is 0.4.

A. Quantitative Comparison

In this section, we evaluate the proposed method using the average center location errors. The pixel error is every frame is defined as shown in Table I:

$$error = \sqrt{(x' - x)^2 + (y' - y)^2} \qquad (12)$$

where (x', y') represents the object position obtained from the tracker and (x, y) is the ground truth. The quantitative results are summarized in Table I. The number marked with red indicates the best tracker in the test sequence and the blue is the second one. The table shows the effectiveness and robustness of our method.

B. Qualitative Comparison

Occlusion: The faceocc2 caviar and girl sequence are designed to test whether a tracking algorithm can handle partial occlusion and pose change well. In fact, occlusion is one of the most general problems in object tracking. Frag tracker [11], L1 tracker [5], ODLSR tracker and the propose method are develop to solve the problem. Fig. 3 shows some result of the comparison tracker. Frag tracker uses the static part-based to handle the problem. L1 tracker and ODLSR tracker update the dictionary to solve the problem. However, both the Frag tracker and L1 tracker perform poorly. In our multi-block sparse coding part, we estimate the possible occluded block via reliability factor. We allocate small weight for possible occlusion block. At frame 158 and 819 of the faceocc2 sequence, the face is heavy occlusion. The method effectively uses the information of target and alleviates the influence of occlusions. Besides, the first classifier part will update classifier parameter when no occlusion, which ensure the correctness of the object appearance. Our method can accurately keep tracking of the target object as the two part both exploit the spatial information of the local patch.

TABLE I: Average Center Location Error

video	algorithm				
	Frag	*L1*	*MIL*	*ODLSR*	*OUR*
animal	95.71	116.47	89.56	18.59	19.94
board	56.28	216.33	83.52	28.83	54.40
car11	31.34	1.53	74.77	31.91	18.69
caviar	23.47	63.75	90.81	6.38	5.59
faceocc2	16.72	19.49	6.47	7.97	3.45
girl	25.20	122.29	34.76	42.19	12.86
jumping	29.87	42.52	10.84	17.96	8.07
shaking	119.04	118.83	18.08	134.68	14.10
Singer1	22.25	4.07	18.64	33.95	4.26

Similarly, in the caviar sequence, the object target is occluded by the person with similar color and shape. The other tracker is easy to drift due to heavy occlusion and the background clutter. In our experiment, our tracker achieves stable performance when there is large occlusion and pose change.

Rotation: In girl sequence, the major challenges include in-plane and out-of-plane rotations. The L1 tracker fails when the girl rotates her head in the frame 111. For Frag tracker, it could not update in the online fashion, so it may lose tracking when the object appearance changes drastically. For L1 tracker, it is easy to introduce the background information into the template set which may cause the drifting.

In contrast, the MIL tracker, ODLSR tracker and our propose tracker can handle the pose changing problem through the capability of online updating. However, the girl sequence also has other challenges. The MIL tracker fails to handle the combination of all the challenges and tend to drift.

Motion blur: The sequence animal consists some difficulties include fast motion of the target object which leads to blurred image and is hard for tracker to solve. Fig. 3 shows the different results on the animal sequence. Most tracking algorithms fail to follow the target right after the frame 24 which motion fast. The L1 tracker fails and locates a similar object instead of the correct target as the true target is blurred and the background has the similar object. The propose method is using the foreground information and

background information at the same time. The second part of the model select the discriminative features to better separate the target from the background which is robust to blur target.

Fig. 3. Sample tracking results of evaluated algorithm on nine challenging image sequences.

Illumination change: In the singer1 sequence, the object target is a singer with dramatic illumination changes. The stage light changes drastically from frame 121 and frame 321. The L1 tracker is not able to track the object steady. Since we update the classifier and dictionary of first part, our tracker can adapt the change of the appearance.

VII. Conclusion

In summary, we proposed a method based on sparse coding, classifier and multi-block. The tracker combines sparse coding and classifier to encode the appearance information of both object target and the background with update the dictionary and classifier parameter which is robust the appearance change and complex background. The using of the multi-block considers the spatial information among different patches with the occlusion handing method through allocating weight for different block. It can improve the performance of the proposed tracker. By applying several difficult benchmark videos, the experimental results demonstrate the robustness of our tracker.

References

[1] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. van den Hengel, "A survey of appearance models in visual object tracking," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 4, pp. 58.1-58.48, 2013.

[2] X. Li, W. Hu, and Z. Zhang, "Robust visual tracking based on incremental tensor subspace learning," in *Proc. 11th International Conference on Computer Vision*, 2007, pp. 1-8.

[3] Y. Wu, J. Cheng, J. Wang, and H. Lu, "Real-time visual tracking via incremental covariance tensor learning," in *Proc. 12th International Conference on Computer Vision*, 2009, pp. 1631-1638.

[4] J. Lim, D. Ross, R. S. Lin, and M. H. Yang, "Incremental learning for visual tracking," *Advances in Neural Information Processing Systems*, 2004, pp. 142-149.

[5] M. Xue and L. Haibin, "Robust visual tracking using l1 minimization," in *Proc. 12th International Conference on Computer Vision*, 2009, pp. 1436-1443.

[6] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1830-1837.

[7] X. Jia *et al*., "Visual tracking via adaptive structural local sparse appearance model," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1822-1829.

[8] B. Babenko, M. H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619-1632, 2011.

[9] H. Grabner and H. Bischof, "Online boosting and vision," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 260-267.

[10] B. Liu, L. Yang, J. Huang, P. Meer, L. Gong, and C. Kulikowski, "Robust and fast collaborative tracking with two stage sparse optimization," in *Proc. 11th European Conference on Computer Vision(ECCV)*, 2010, pp. 624-637.

[11] A. Adam, E. Rivlin, and I. Shimshoni, "Robust fragments-based tracking using the integral histogram," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 798-805.

[12] J. Wright, Y. Ma, J. Maral, G. Sapiro, T. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," in *Proc. the IEEE*, vol. 98, no. 6, pp. 1031-1044, 2010.

[13] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861-2873, 2010.

[14] L. Zhang, M. Yang, and X. C. Feng, "Sparse representation or collaborative representation which helps face recognition?" in *Proc. IEEE International Conference on Computer Vision*, 2011, pp. 471-478.

[15] W. Zhong, H. Lu, and M. H. Yang, "Robust object tracking via sparsity-based collaborative model," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1838-1845.

[16] Q. Wang, F. Chen, W. Xu and M. H. Yang, "Online discriminative object tracking with local sparse representation," in *Proc. IEEE Workshop on Applications of Computer Vision (WACV)*, 2012, pp. 425-432.

[17] F. Yang, H. Lu, and Y. W. Chen, "Bag of features tracking," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 153-156.

Wu Wang was born in Anhui, China in 1990. He received the B.Sc. degree in information engineering from University of Jinan, Jinan, China, in 2008 and he is currently working towards his M.Sc. degree in signal and information processing from the University of Science and Technology of China, Hefei, China.

His research interests include signal processing, image processing, and pattern recognition.

Jinxu Tao received the M.Sc. degree in automatic engineering from Hefei University of Technology, Hefei, China, in 1988, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China, in 1995.

He was a postdoctoral researcher of advanced material characterization at Windsor University, Windsor, Canada from 1990 to 2000. Currently, he is an associate professor at the Department of Electronic Engineering, University of Science and Technology of China. His research interests include nondestructive tests by ultrasonic methods, scanning microscopy, underwater acoustics as well as linear and nonlinear properties of acoustic waves.

Yongjun Jiang received the B.Sc. degree in Hefei University of Technology, Hefei, China, in 2008 and he is currently working towards his M.Sc. degree in signal and information processing from the University of Science and Technology ofChina, Hefei, China.

Zhongfu Ye received the B.Sc. degree in Hefei University of Technology, Hefei, China, in 1982, the M.Sc. degree from Hefei University of Technology, Hefei, China, in 1986, the Ph.D. degree in signal and information processing from the University of Science and Technology of China, in 1995. Currently, he is a professor at the Department of Electronic Engineering, University of Science and Technology of China.