Patchwise Tracking via Spatio-Temporal Constraint-Based Sparse Representation and Multiple-Instance Learning-Based SVM

Yuxia Wang(✉) and Qingjie Zhao

Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, People's Republic of China
yuxiawang2006@163.com, zhaoqj@bit.edu.cn

Abstract. This paper proposes a patch-based tracking algorithm via a hybrid generative-discriminative appearance model. For establishing the generative appearance model, we present a spatio-temporal constraint-based sparse representation (STSR), which not only exploits the intrinsic relationship among the target candidates and the spatial layout of the patches inside each candidate, but also preserves the temporal similarity in consecutive frames. To construct the discriminative appearance model, we utilize the multiple-instance learning-based support vector machine (MIL&SVM), which is robust to occlusion and alleviates the drifting problem. According to the classification result, the occlusion state can be predicted, and it is further used in the templates updating, making the templates more efficient both for the generative and discriminative model. Finally, we incorporate the hybrid appearance model into a particle filter framework. Experimental results on six challenging sequences demonstrate that our tracker is robust in dealing with occlusion.

Keywords: Patchwise tracking · Hybrid generative-discriminative appearance model · MIL&SVM · Spatio-temporal constraint · Sparse representation

1 Introduction

Visual tracking is an active field of research in computer vision. While numerous tracking methods have been proposed with demonstrated success in recent years, designing a robust tracking method is still an open problem, due to factors such as scale and pose change, illumination variation, occlusion, etc. Especially, occlusion is a core issue. One of the main reasons is the lack of the effective object appearance models, which play a significant role in visual tracking.

For designing a robust tracker, most tracking algorithms employ generative learning or discriminative learning based appearance models. Generative learning based appearance models mainly concentrate on how to fit the data accurately from the object class using generative methods. Among them, sparse representation is a widely used generative method. Jia et al. [3] developed a

S. Arik et al. (Eds.): ICONIP 2015, Part I, LNCS 9489, pp. 264–271, 2015.
DOI: 10.1007/978-3-319-26532-2_29

local appearance model by utilizing the sparse representation of the overlapped patches. Zhang et al. [10] proposed a structural sparse tracking algorithm to exploit the relationship among the target candidates and spatial layout of the patches inside each candidate. Zarezade et al. [9] presented a joint sparse tracker by assuming that the target and the previous best candidates have a common sparsity pattern. Although these methods achieve convincing performance, they either lack of the description of the target spatial layout or ignore the temporal consistency constraint of successive frames. In this paper, we propose a spatio-temporal constraint-based sparse representation (STSR), which not only exploits the spatial layout of the local patches inside each candidate and the intrinsic relationship among the candidates and their local patches, but also preserves the temporal consistency of the sparsity pattern in consecutive frames.

In comparison, discriminative appearance models pose visual tracking as a binary classification issue, aiming to maximize the inter-class separability between the object and non-object regions via discriminative learning techniques. Babenko et al. [2] introduced the multiple-instance learning technique into online object tracking where training samples can be labeled more precisely. In [4], Kalal et al. proposed to train a binary classifier using the P-N learning algorithm with both labeled and unlabeled samples. Despite the convincing performance, most of these methods use holistic representation to represent the object and hence do not handle occlusion well. In this paper, we utilize the patch-based discriminative appearance model proposed by [6] to locate the target from the background, in which the multiple-instance learning-based support vector machine (MIL&SVM) is used as the classifier and it can predict the occlusion state and alleviate the drifting problem. According to the occlusion state, we update the template set as mentioned in [6], making the templates more effective both for the generative and discriminative appearance model.

2 Patchwise Tracking via a Hybrid Generative-Discriminative Appearance Model

In our tracker, we utilize $\mathbf{s}_t$ to denote the object state at time t, and construct our tracker in the particle filter framework (PF). For the dynamic model of PF, $p(\mathbf{s}_t|\mathbf{s}_{t-1})$, we assume a Gaussian distributed model. For the appearance model in PF, $p(\mathbf{y}_t|\mathbf{s}_t)$, we use our patch-based hybrid generative-discriminative appearance model, which will be introduced below.

2.1 Generative Appearance Model Based on STSR

Given the image set of the target templates $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_m]$, where m is the number of target templates, we sample K overlapped local patches inside each target region. The sampled patches are used to form a dictionary $\mathbf{D} = [\mathbf{d}_1^{(1)}, ..., \mathbf{d}_m^{(1)}, ..., \mathbf{d}_1^{(K)}, ..., \mathbf{d}_m^{(K)}]$, each column in $\mathbf{D}$ is obtained by ℓ_2 normalization on the vectorized gray scale image observations extracted from $\mathbf{T}$.

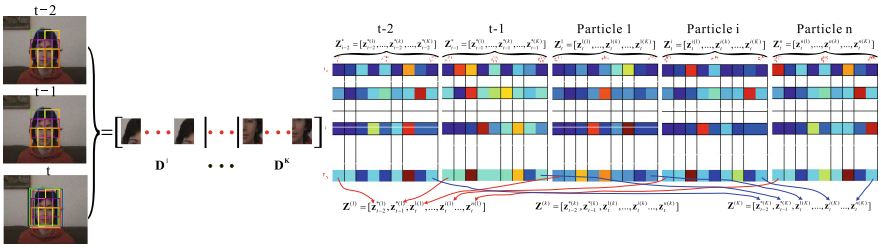

Fig. 1. Spatio-temporal constraint-based sparse representation

Let $\{\mathbf{x}_{t-i}^*\}_{i=1}^N$ and $\{\mathbf{x}_t^i\}_{i=1}^n$ represent the best candidates obtained in the previous tracking and particles from the current frame respectively. For $\{\mathbf{x}_{t-i}^*\}_{i=1}^N$ and $\{\mathbf{x}_t^i\}_{i=1}^n$, we also sample K overlapped local patches as done in the template set and denote $\mathbf{x}_{t-i}^* = [\mathbf{x}_{t-i}^{*(1)}, ..., \mathbf{x}_{t-i}^{*(K)}]$ and $\mathbf{x}_t^i = [\mathbf{x}_t^{i(1)}, ..., \mathbf{x}_t^{i(K)}]$. Let $\mathbf{X}_t^{(k)} = [\mathbf{x}_t^{1(k)}, ..., \mathbf{x}_t^{n(k)}]$ denote the k-th local patches of n particles at time t. In order to represent this observations matrix $\mathbf{X}_t^{(k)}$, we not only consider the spatial constraint of the particles and local patches, but also utilize the temporal constraint in consecutive frames.

Spatio-Temporal Constraint. Based on the fact that n particles at current frame are densely sampled at and around the target of the previous frame and the target's appearance changes smoothly, it is reasonable to assume that these particles are likely to be similar and they have the similar sparse pattern with previous tracking results over a period of time. Thus the k-th image patches of n particles and previous tracking results are expected to be similar. In addition, for patches extracted from a candidate particle or a previous tracking result, their spatial layout should be preserved.

Spatio-Temporal Constraint-Based Sparse Representation (STSR). Based on the above observations, we use $\mathbf{X}^{(k)} = [\mathbf{x}_{t-i}^{*(k)}, ..., \mathbf{x}_{t-1}^{*(k)}, \mathbf{x}_t^{1(k)}, ..., \mathbf{x}_t^{n(k)}]$ to represent the k-th local patches of previous tracking results and n particles in current frame, $\mathbf{D}^{(k)} = [\mathbf{d}_1^{(k)}, \mathbf{d}_2^{(k)}, ..., \mathbf{d}_m^{(k)}]$ to express the k-th patches of m templates, and $\mathbf{Z}^{(k)} = [\mathbf{z}_{t-i}^{*(k)}, ..., \mathbf{z}_{t-1}^{*(k)}, \mathbf{z}_t^{1(k)}, ..., \mathbf{z}_t^{n(k)}]$ to denote the representations of the k-th local patch observations of $\mathbf{X}^{(k)}$ with respect to $\mathbf{D}^{(k)}$. Then the joint sparse appearance model for the object tracking under the spatio-temporal constraint can be obtained by using the $\ell_{2,1}$ mixed norm as

$$\min_{\mathbf{Z}} \frac{1}{2} \sum_{k=1}^{K} ||\mathbf{X}^{(k)} - \mathbf{D}^{(k)}\mathbf{Z}^{(k)}||_F^2 + \lambda ||\mathbf{Z}||_{2,1} \tag{1}$$

where, $\mathbf{Z} = [\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}, ..., \mathbf{Z}^{(K)}]$, $||\cdot||_F$ denotes the Frobenius norm, λ is a regularization parameter which balances reconstruction error with model complexity, $||\mathbf{Z}||_{2,1} = \sum_i (\sum_j |[\mathbf{Z}]_{ij}|^2)^{\frac{1}{2}}$ and $[\mathbf{Z}]_{ij}$ denotes the entry at the i-th row and j-th column of $\mathbf{Z}$. The $\ell_{2,1}$ mixed norm regularizer is optimized using an

Accelerated Proximal Gradient (APG) method. The illustration of the spatio-temporal constraint-based sparse representation is shown in Fig. 1.

Generative Appearance Model Based on STSR. After learning the $\mathbf{Z}$, the observation likelihood of the tracking candidate i is defined as

$$p_g(\mathbf{y}_t|\mathbf{s}_t) = \frac{1}{\beta}\exp(-\alpha\sum_{k=1}^{K}||\mathbf{x}_t^{i(k)} - \mathbf{D}^k\mathbf{z}_t^{i(k)}||_F^2) \tag{2}$$

where, $\mathbf{z}_t^{i(k)}$ is the coefficient of the k-th image patch of the i-th particle corresponding to the target templates, and α and β are normalization parameters.

2.2 Discriminative Appearance Model Based on MIL&SVM

Despite the robust performance of the generative appearance model achieved, it is not effective in dealing with the background distractions. Therefore, we introduce a discriminative appearance model based on MIL&SVM to improve the performance of our tracker.

We denote the overlapped image patches extracted from the target templates as the positive pathes p^+, and the overlapped patches extracted from the background (which is an annular region and the distance from the center-point of the target object to the edge of the negative patch sampling area is set to R) are denoted as negative patches p^-. As we all known, some positive patches obtained above may contain some noisy pixels from background because the bounding box is rectangular whereas the shape of the target may not be a standard rectangle. In order to deal with this problem, we adopt the patch-based MIL&SVM to train a robust classifier. In the training procedure, a row of patches are defined as a positive bag b^+ if they extracted from the target templates, or negative bag b^- if they come from background. The training procedure is illustrated in Fig. 2.

With this classifier, we can classify each patch of a candidate object at time t. For a candidate, we use r^+ to denote the local patches which are classified as positive and use r^- to denote patches classified as negative. Then the probability of a candidate being the tracking result can be defined as

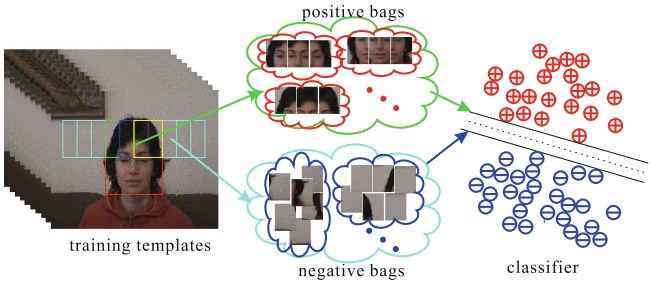

Fig. 2. Illustration for the patch-based MIL&SVM

$$p_d(\mathbf{y}_t|\mathbf{s}_t) = \frac{|r^+|}{|r^-| + |r^+|} \tag{3}$$

where $|r^+|$ and $|r^-|$ are the number of positive patches and negative patches.

Furthermore, according to the classification result, the occlusion state of a candidate can be obtained as

$$O = \frac{|r^-|}{|r^-| + |r^+|} \tag{4}$$

2.3 Adaptive Hybrid Generative-Discriminative Appearance Model

Based on the likelihood obtained from the spatio-temporal constraint-based sparse representation and the probability got via multiple-instance learning-based SVM, we construct our final observation model as:

$$p(\mathbf{y}_t|\mathbf{s}_t) = \eta p_g(\mathbf{y}_t|\mathbf{s}_t) + (1-\eta)p_d(\mathbf{y}_t|\mathbf{s}_t) \tag{5}$$

where $\eta \in [0,1]$ is a control parameter, which can adjust weights of the two methods according to the occlusion state and can be defined as $\eta = \frac{1}{2}(1+O)$.

In order to deal with appearance variation with time, we need to update our templates. We divide the templates $\mathbf{T}$ into two groups according to the occlusion state. The group without occlusion is denoted as $\mathbf{T}_{unocc} = [\mathbf{T}_1, ..., \mathbf{T}_{m_1}]$, and the occluded template set is denoted as $\mathbf{T}_{occ} = [\mathbf{T}_{m_1+1}, ..., \mathbf{T}_m]$, where m_1 is the number of unoccluded patches. The templates in $\mathbf{T}_{unocc}$ are ordered by time and the templates in $\mathbf{T}_{occ}$ are ordered reversely by time. We use two increasing interval sequences and a random number $r \in [0,2]$ to determine the sequence number of the template needed to be deleted as Eq. 6.

$$f(r) = \begin{cases} i, & r \in [\frac{(i-1)^2+(i-1)}{m_1^2+m_1}, \frac{i^2+i}{m_1^2+m_1}], \quad 0 < r \leq 1 \\ j, & r \in [1+\frac{(j-1)^2+(j-1)}{m_2^2+m_2}, 1+\frac{j^2+j}{m_2^2+m_2}], \quad 1 < r \leq 2 \end{cases} \tag{6}$$

where $m_2 = m - m_1$.

After selecting the template to discard, we use the method mentioned in [3] to update the template. For more detail, please refer [3]. After the templates $\mathbf{T}$ is updated, we retrain the MIL&SVM classifier only with the templates without occlusion or with light occlusion.

3 Experiments

We validate our tracker on six challenging sequences and compare it with six state-of-the-art methods proposed in recent years. All of these sequences are publicly available. The challenges of these sequences include severe occlusion and drastic shape deformation. In order to test the effectiveness and robustness of our

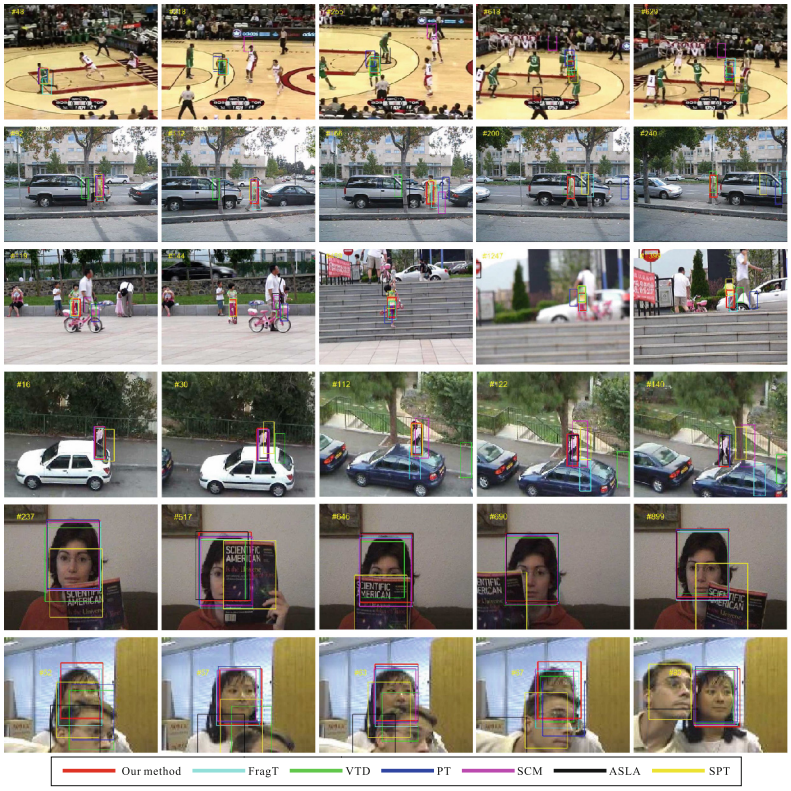

Fig. 3. Comparative experimental tracking results of 7 methods on six sequences, from top to bottom are *basketball*, *DavidOutdoor*, *girl_move*, *woman_sequence*, *face_sequence* and *girl_head*

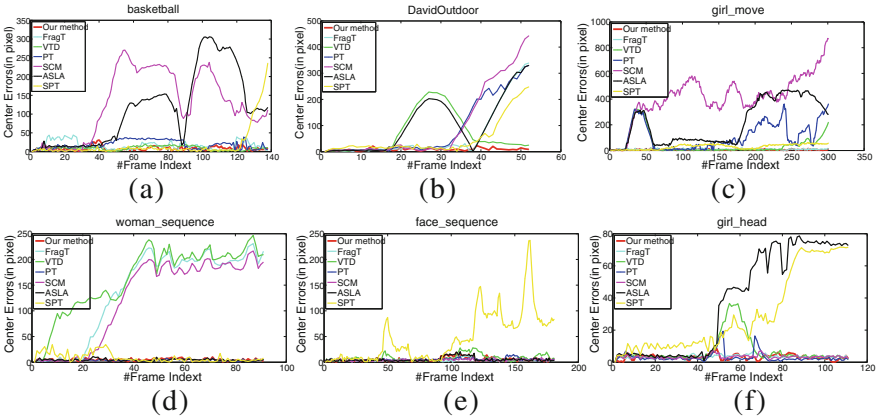

Fig. 4. Center error plots for 7 methods on six video sequences

tracker, we compare it with FragT [1], VTD [5], PT [8], SCM [11], ASLA [3] and SPT [7]. For our tracker, we set the number of templates $m = 10$, the number of local patches $K = 9$, the number of particles $n = 400$, and we use 2 previous tracking results in STSR. We resize all the targets or candidates as (32, 32).The size of the sampling patch is (16,16) and the sampling step is 8 pixels.

Table 1. Location errors (in pixel, the bold font indicates the best performance)

Sequences	FragT	VTD	PT	SCM	ASLA	SPT	Ours
basketball	16.3	9.0	19.1	126.4	112.6	17.4	**8.2**
DavidOutdoor	63.5	70.0	88.0	101.7	105.2	50.0	**8.5**
girl_move	8.9	45.4	110.2	414.8	214.5	30.0	**5.7**
woman_sequence	138.1	163.1	**3.8**	122.9	4.4	9.0	4.7
face_sequence	**4.4**	8.5	5.4	4.5	5.4	48.1	5.5
girl_head	3.6	7.2	3.1	3.3	38.0	30.0	**3.1**
Overall	39.1	50.5	38.3	128.9	80.0	30.8	**6.0**

Comparative tracking results of selected frames are shown in Fig. 3, from which we can find that our proposed tracker performs very well on all these challenging sequences. FragT is designed for dealing with occlusion and performs well in *face_sequence* and *girl_head* when the target is large enough, but it cannot get good results in other sequences when there exists sever occlusion in a small target. VTD adopts multi-trackers to track the target and it achieves satisfactory results in *face_sequence* and *basketball* but also shows less effective in dealing with the situation when there exists both rigid shape deformation and occlusion. PT is a part-based tracker and it performs well in dealing with partial occlusion, but it fails when the target is full occluded. Both SCM and ASLS adopt sparsity-based appearance model and they perform well in dealing with occlusion as shown in *face_sequence*, but cannot get satisfactory performance when there exists rigid shape deformation. SPT achieves good results on *DavidOutdoor* and *girl_move* as shown in Fig. 3, but cannot obtain stable performance in clutter scene or when there exists severe and frequent occlusion as shown in screenshots of sequences *basketball*, *woman_sequence* and *face_sequence*.

We also measure the quantitative tracking error, the Euclidean distance from the tracking center to the ground-truth. The center error plots of 7 methods on 6 sequences are shown in Fig. 4, which demonstrates that our tracker is robust in handling occlusion and shape deformation even in a complex scene. We show the location errors in Table 1, which shows that our tracker achieves the best tracking results on 4 sequences and gives the the best tracking result on average.

4 Conclusion

In this paper, we have proposed a novel patch-based tracking method based on the combination of spatio-temporal constraint-based sparse representation

(STSR) and multiple-instance learning-based SVM (MIL&SVM). By utilizing the STSR, our tracker effectively captures the structure cues of the target and the temporal similarity in consecutive frames. Furthermore, we utilize MIL&SVM as our discriminative appearance model, which is robust in cluttered background and can predict the occlusion state. Based on the occlusion state, we update the template set separately, making the generative method obtain more precise templates and the discriminative method maintain correctness. Qualitative and quantitative experimental results on different challenging sequences demonstrate that our tracker is very robust to the occlusion.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61175096 and 61273273), Specialized Fund for Joint Building Program of Beijing municipal Education Commission.

References

1. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 798–805, IEEE (2006)
2. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 983–990, IEEE (2009)
3. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1822–1829, IEEE (2012)
4. Kalal, Z., Matas, J., Mikolajczyk, K.: PN learning: bootstrapping binary classifiers by structural constraints. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 49–56, IEEE (2010)
5. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1269–1276, IEEE (2010)
6. Li, X., He, Z., You, X., Chen, C.P.: A novel joint tracker based on occlusion detection. Knowl. Based Syst. **71**, 409–418 (2014)
7. Yang, F., Lu, H., Yang, M.H.: Robust superpixel tracking. IEEE Trans. Image Process. **23**(4), 1639–1651 (2014)
8. Yao, R., Shi, Q., Shen, C., Zhang, Y., van den Hengel, A.: Part-based visual tracking with online latent structural learning. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2363–2370, IEEE (2013)
9. Zarezade, A., Rabiee, H., Soltani-Farani, A., et al.: Patchwise joint sparse tracking with occlusion detection. IEEE Trans. Image Process. **23**(10), 4496–4510 (2014)
10. Zhang, T., Liu, S., Xu, C., Yan, S., Ghanem, B., Ahuja, N., Yang, M.H.: Structural sparse tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 150–158 (2015)
11. Zhong, W., Yang, M., et al.: Robust object tracking via sparse collaborative appearance model. IEEE Trans. Image Process. **23**(5), 2356–2368 (2014)