ADAPTIVE VISUAL TARGET TRACKING BASED ON LABEL CONSISTENT K-SVD SPARSE CODING AND KERNEL PARTICLE FILTER

Jinlong Yang[1], Xiaoping Chen[1], Yu Hen Hu[2], Jianjun Liu[1]

[1]School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China
[2]Department of Electrical and Computer Engineering, University of Wisconsin–Madison, WI 53706, USA

ABSTRACT

We propose an adaptive visual target tracking algorithm based on Label-Consistent K-Singular Value Decomposition (LC-KSVD) dictionary learning. To construct target templates, local patch features are sampled from foreground and background of the target. LC-KSVD then is applied to these local patches to simultaneously estimate a set of low-dimension dictionary and classification parameters (CP). To track the target over time, a kernel particle filter (KPF) is proposed that integrates both local and global motion information of the target. An adaptive template updating scheme is also developed to improve the robustness of the tracker. Experimental results demonstrate superior performance of the proposed algorithm over state-of-art visual target tracking algorithms in scenarios that include occlusion, background clutter, illumination change, target rotation and scale changes.

Index Terms—Visual target tracking, label consistent K-SVD, sparse coding, dictionary learning, particle filter

1. INTRODUCTION

Visual target tracking is a key enabling technology for numerous emerging computer vision applications including video surveillance, navigation, human-computer interactions, augmented reality, higher level scene understanding and action recognition among many others [1-4]. It is a challenging task because the visual observations often suffer from interference due to occlusion, scale and shape variation, illumination variation, background clutter, and related factors.

Current visual target tracking algorithms may be categorized into two families: discriminative algorithms [5,6] versus generative algorithms [7,8]. Recently, tracking algorithms based on the sparse model have attracted great interests. Mei et al [9, 10] formulated visual target tracking as a sparse approximation problem in the particle filtering (PF) framework [11, 12]. The target can be represented as a weighted linear combination of very few (hence sparse representation) image templates in the dictionary. The sparse representation can be estimated by solving an l_1-norm regularized least squares (LS) problem. In [13], a real-time robust l_1 tracker is proposed by adding an l_2-norm regularization to the coefficients associated with the trivial templates, and an accelerated proximal gradient (APG) method is employed to speed up the problem solving. Multi-task tracking (MTT) is proposed [14] as a multi-task sparse learning problem in a PF framework. In [5], an adaptive structural local sparse appearance model is proposed to locate the target more accurately by considering the spatial information of the target based on an alignment-pooling method. A collaborative model is proposed [15] that combines a sparsity-based discriminative classifier and a sparsity-based generative model.

However, most of these template update schemes cannot adapt to the changes of the foreground and the background of the target and often lack the ability of discrimination. To address these concerns, in this work, we propose an adaptive visual target tracking algorithm that extends the Label-Consistent K-Singular Value Decomposition (LC-KSVD) [16,17] approach to train the low dimensional dictionary and classification parameters (CP). An adaptive template update algorithm is also developed to update the dictionary. Finally, a kernel particle filter (KPF) is implemented to track the target by merging the Gaussian kernel density (GKD) and the CP, as well as the sparse coefficient information of each patch. Experimental results demonstrate superior performance of the proposed algorithm over state-of-art visual target tracking algorithms in scenarios that include occlusion, background clutter, illumination change, target rotation and scale changes.

In Section 2, we summarize the details of the proposed algorithm. Experiment results comparing against existing visual target trackers are reported in Section 3. Conclusions are presented in Section 4.

2. LC-KSVD BASED VISUAL TARGET TRACKING

A block diagram of the proposed LC-KSVD tracking algorithm is shown in Fig. 1. Image patches of both the foreground and background image will be extracted during initiation. These patches will be fed into the LC-KSVD dictionary learning [16,17] algorithm to yield (i) a low dimension dictionary D, (ii) sparse coefficients X and (iii) classification parameters W. Meanwhile, patches will also be extracted from the candidate target image and then a

kernel density based particle filter (KPF) is applied to deduce a sparse representation using the low dimension dictionary D. To account for potential occlusion of the target, we introduce a detection scheme of sparse coefficient histogram matrix (SCHM), which is the concatenation of the sparse coefficient vector after occlusion detection [15]. The occlusion degree of the target is calculated to design the adaptive parameter model for the proposed template update scheme, which can improve the robustness of the tracker.

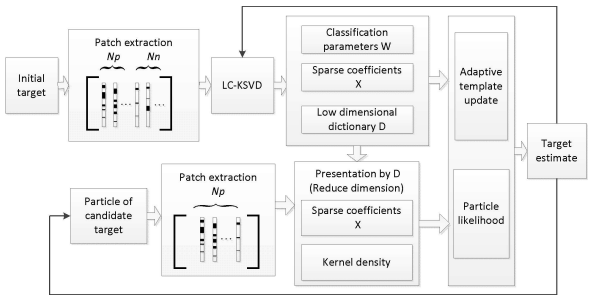


Fig. 1. LC-KSVD tracking algorithm.

2.1. Initialization of the target template

The target template will be normalized to 32×32 pixels. Overlapping 6 × 6 patches then will be extracted within this template to give N_p=196 positive (foreground) patches. N_n=196 negative (background) patches will be extracted along the template border, including partial foreground features.

2.2. Label-Consistent K-Singular Value Deposition

The LC-KSVD dictionary learning algorithm [16,17] can simultaneously train an over-complete dictionary and a linear classifier. The objective function is

$$\langle D,W,A,X\rangle = \arg\min_{D,W,A,X} \|Y-DX\|_2^2 + \alpha\|Q-AX\|_2^2 + \beta\|H-WX\|_2^2,\ s.t.\ \forall i, \|x_i\| \le T \quad (1)$$

Where Y is the observation matrix, D is the dictionary matrix, X is the sparse coefficient matrix, Q is the sparse codes with discriminative power of Y for classification. The matrix A transforms the original sparse codes to be most discriminative in the sparse feature space. H is the class label of input samples. Using KSVD [18], one may update the dictionary, the sparse representation, as well as estimating corresponding A and W matrices. In this work, the low dimensional dictionary consists of 50 positive and 50 negative patches.

2.3. Kernel Particle Filter

Target motion between consecutive frames is modeled using an six-dimensional affine transformation $\boldsymbol{x}_k$ = {x_k, y_k, θ_k, s_k, α_k, β_k}, where (x_k, y_k) is the target position, θ_k, s_k, α_k, and β_k are respectively the rotation angle, the scaling factor, the aspect ratio and the angle of inclination. During the implementation process of the KPF, 100 candidate particles are sampled, and 196 positive patches of each particle are used to learn and produce candidate target sparse coefficient. The dynamics of the KPF can be modeled as $p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}) = N(\boldsymbol{x}_k; \boldsymbol{x}_{k-1}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix whose diagonal consists of variances of elements of $\boldsymbol{x}_k$.

The likelihood of the lth candidate particle is

$$p_l = \sum_{i=1}^{N_p} k\left(\left\|\left(y_k^l - c_i\right)/h\right\|^2\right) M_{k,i}^l L_{k,i}^l \quad (2)$$

where $k\left(\left\|\left(y_k^l - c_i\right)/h\right\|^2\right)$ is an isotropic Gaussian kernel density function. In eq. (2), y_k^l is the center of the lth candidate particle, c_i is the center of the ith patch, $M_{k,i}^l = \cos\left\langle W\varphi_{k,i}^l, \Gamma\right\rangle$ is the likelihood of classification, and $L_{k,i}^l = \sum \min\left(\varphi_{k,i}^l, \psi^i\right)$ is the similarity function [15] of the target sparse coefficients between the candidate and the template. Here $\Gamma = [1,0]^{\mathrm{T}}$ is the base vector of target classification, φ_c^i and ψ^i are the sparse coefficient histogram matrices (SCHM) of the candidate target and the target template, respectively.

It is noted that the larger number of patches belong to the candidate particle is, the more accurate the target appearance may be described. Because the selected patches may be from target templates or background templates. Therefore, if the patch belongs to the target, we should give it a larger weight than that belong to the background according to the likelihood of classification.

2.4. Adaptive Template Update

The template is updated every 5 frames using a linear combination of the SCHM ψ of the first frame and the latest estimated SCHM $\hat{\varphi}_n$:

$$\hat{\psi}_n = \begin{cases} \mu\psi + (1-\mu)\hat{\varphi}_n, & O_n < O_0 \\ \hat{\psi}_{n-1}, & otherwise \end{cases} \quad (3)$$

where μ=$\exp\left((O_n/O_o)-1\right)$ is an adaptive weight parameter. $O_n = (\#\text{ occluded patches})/(\#\text{all patches})$ and O_0 is a threshold of the occlusion degree.

Remark. With increasing number of occluded patches, O_n and hence μ will increase so the latest template $\hat{\varphi}_n$ will be given smaller weight because it is less reliable.

3. EXPERIMENT RESULTS

Eight challenging video sequences drawn from the public domain video target tracking datasets [1,19] are used to examine the performance of the proposed adaptive LC-KSVD visual target tracking algorithm. They are shown in Fig. 2. Major challenges include occlusion, background

clutter, illumination change, target rotation, scale change, etc. In sequences (a) FaceOcc2 and (b) Woman, the targets are heavily occluded or partial occluded for long period of time. In sequence (c) Singer1, there is large illumination variation. In sequences (d)~(h), there are significant background clutter, target rotation and fast motion.

We apply the proposed LC-KSVD tracking algorithm according to the experimenting protocol specified in [1]. We then compare the results against those using five state-of-the-art benchmark tracking algorithms, including robust FRAG [7], IVT [8], L1APG [13], MTT [14] and LSK [20]. The experiments are implemented on computer with Intel Core 2.4 GHz, i7-4700HQ processor with 8GB RAM. The threshold of the occlusion degree is set as $O_0 = 0.8$.

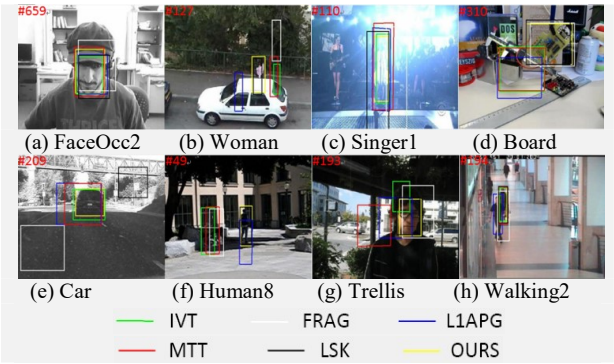
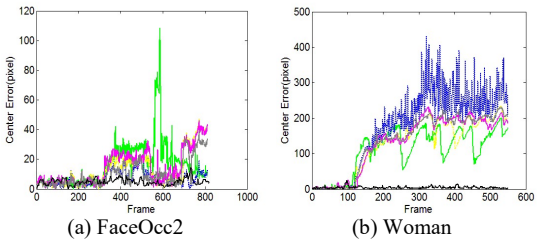


Fig.2. Tracking results of different algorithms.

Two evaluation criteria are employed to quantitatively assess the performance of the proposed algorithm. One is average center location error (ACLE), and another is tracking success rate (SR) [5]. Fig. 3 shows the relative position error (in pixels) between the center and the tracking results. ACLE is defined as the average relative position error. Assume the tracking result is R_r, and the ground truth is R_g, then SR is defined as $\Upsilon = (R_r \cap R_g)/R_r \cup R_g$. Tables 1 and 2 give values of ACLE and SR for different tracking algorithms.

As can be seen from Fig. 3, the LC-KSVD tracking algorithm exhibits much better performance than those of the benchmark algorithms. The tracking result of each frame is accurate, and the error curve remains low without big changing.



(a) FaceOcc2 (b) Woman

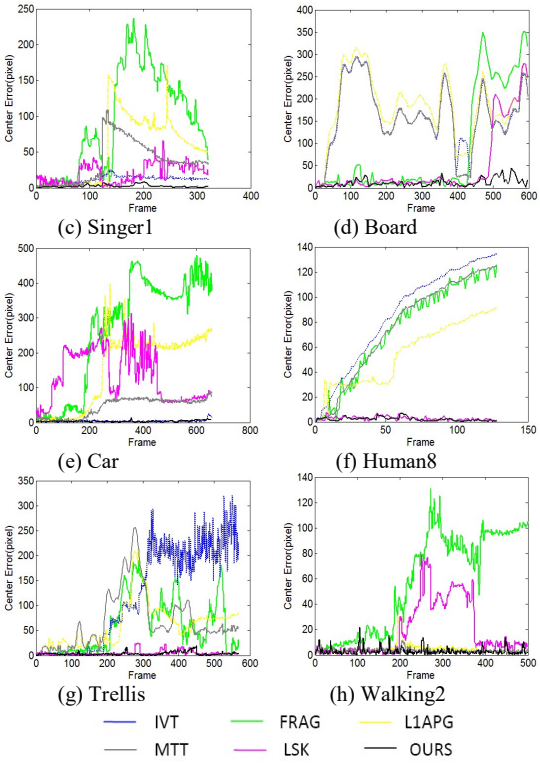


Fig.3. Position errors (in pixel) .

From Tables 1 and 2, clearly the proposed algorithm can adapt to most of the video sequences with the highest accuracy, which is attributed to the detailed description of the local patches by the LC-KSVD dictionary learning and adaptive template update scheme. Moreover, the kernel density and classification information are considered in the proposed KPF, improving the tracking performance.

Table 1. Average center location error (in pixel). The best and second best results are shown in bold and underline.

	IVT	FRAG	L1APG	MTT	LSK	OURS
FaceOcc2	<u>6.9</u>	15.7	12.9	10.2	14.7	**4.88**
Woman	172.6	<u>109.7</u>	126.7	134.8	131.6	**4.43**
Singer1	<u>11.5</u>	91.5	53.1	35.9	21.2	**2.48**
Board	162.2	84.5	184.4	159.2	<u>45.4</u>	**12.08**
Car4	<u>4.08</u>	263.1	153.98	45.25	133.23	**3.89**
Human8	85.96	74.83	54.17	76.42	<u>2.74</u>	**2.18**
Trellis	119.57	59.51	62.30	68.99	<u>4.70</u>	**3.85**
Walking2	<u>3.04</u>	57.53	4.52	3.48	18.95	**2.84**

To verify the effectiveness of the adaptive template update scheme with the adaptive parameter μ in eq.(3), two special challenging sequences with big appearance changes are chosen, the first 200 frames of FaceOcce2 and the first 170 frames of Woman. The results with different constant values (e.g. 0.1, 0.4, 0.7 and 0.9) of μ are compared against

those with adaptive parameter value, and there are demonstrated in Table 3.

Table 2. Success rate. The best and second best results are shown in bold and underline.

	IVT	FRAG	L1APG	MTT	LSK	OURS
FaceOcc2	0.73	0.66	0.68	<u>0.75</u>	0.64	**0.82**
Woman	0.16	0.16	0.17	<u>0.18</u>	0.17	**0.74**
Singer1	<u>0.59</u>	0.23	0.32	0.37	0.37	**0.87**
Board	0.15	0.55	0.11	0.16	<u>0.65</u>	**0.83**
Car4	<u>0.88</u>	0.19	0.26	0.45	0.15	**0.89**
Human8	0.06	0.10	0.16	0.10	<u>0.69</u>	**0.74**
Trellis	0.25	0.29	0.20	0.22	<u>0.66</u>	**0.71**
Walking2	0.76	0.28	<u>0.78</u>	**0.81**	0.47	0.75

As can be seen from Table 3, there are different results of ACLEs and SRs by choosing different values of μ, and smaller value of μ (e.g., 0.1) gets higher accuracy for the FaceOcce2 sequences, while bigger value of μ (e.g., 0.9) gets higher accuracy for the Woman sequences. The reason is that the variations of target appearances are small between 1st and 140th frames of FaceOcce2 sequences, the updated templates mainly rely on the latest templates during this time. But the target appearances are severely occluded between 141st and 190th fames, the updated templates more rely on the template of the first frame. Therefore, it is noted that the differences of the tracking accuracy are small with different values of μ for this sequence. But for the Woman sequences, the target appearances are slightly disturbed by the background clutters between 36th and 170th frames, and there exists partial occlusion between 106th and 165th frames. Therefore, most of the updated templates mainly rely on the latest frame templates, and the bigger value of μ gets better results. While for the proposed algorithm with adaptive parameter μ, it can obtain an ideal tracking result without manually setting the parameter values.

Table 3. Discussion of constant and adaptive parameter μ.

Dataset	Criteria	μ				
		0.1	0.4	0.7	0.9	Adptive
FaceOcc2 1~200(f)	ACLE	**4.48**	4.53	4.65	4.81	4.59
	SR	**0.85**	0.85	0.84	0.84	0.84
Woman 1~170(f)	ACLE	15.66	5.80	4.39	4.32	**2.87**
	SR	0.58	0.80	0.82	0.82	**0.84**

4. CONCLUSION

In this paper, we present a high performance visual tracking algorithm. The template sets constructed by the local patch features from both foreground and background of the target are used to learn the low dimensional dictionary and classification parameters, which are considered to propose an effective KPF to extract target. Moreover, to robustly decide the final tracking states, an adaptive template update scheme is designed. The effectiveness of the proposed algorithm is experimentally demonstrated by comparing against several state-of-the-art trackers on challenging video sequences, and experimental results show that the proposed algorithm has better tracking performance than some benchmark methods in the scenarios with the interference of occlusion, background clutter, illumination change, etc.

5. ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China (No. 61305017) and the Natural Science Foundation of Jiangsu Province (No. BK20130154). Y. H. Hu is supported by a grant by the U.S. National Science Foundation through the Cyber-Physical Systems program (CNS 1329481).

6. REFERENCES

[1]Y. Wu, J. Lim, M. and H. Yang, "Object Tracking Benchmark," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1834-1848, 2015.

[2]M. Kristan, J. Matas, and A. Leonardis, et al., "The visual object tracking vot2015 challenge results," in *Proc. of the IEEE conf. on computer vision workshops*, pp. 1-23, Dec. 2015.

[3]A. W. M. Smeulders, D. M. Chu, and R. Cucchiara, et al., "Visual Tracking: An Experimental Survey," *IEEE Trans. on Pattern Analysis & Machine Intelligence*, vol. 36, no. 7, pp.1442-1468, 2014.

[4]A. Mazzù, P. Morerio, L. Marcenaro, et al., "A Cognitive Control-Inspired Approach to Object Tracking," *IEEE Trans. on Image Process*, vol. 25, no. 6, pp. 2697-2711, 2016.

[5]X. Jia, H. C. Lu, M. and H. Yang, "Visual tracking via adaptive structural local sparse appearance model," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 1822-1829, June 2012.

[6]B. Babenko, M. H. Yang, and S. Belongie, "Visual tracking with online Multiple Instance Learning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 983-990, June 2009.

[7]A. Adam, E. Rivlin, and I. Shimshoni, "Robust Fragments-based Tracking using the Integral Histogram," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 798-805, June 2006.

[8]D. A. Ross, J. Lim, R. S. Lin, et al., "Incremental Learning for Robust Visual Tracking," *International Journal of Computer Vision*, 2008, vol. 77, no. 1-3, pp. 125-141, 2008.

[9]X. Mei, and H. B. Ling, "Robust visual tracking using L1 minimization," in *Proc. of IEEE 12th International Conference on Computer Vision*, pp. 1436-1443, Sep. 2009.

[10] X. Mei, H. B. Ling, and Y. Wu, et al., "Minimum error bounded efficient L1 tracker with occlusion detection," in *Proc. of*

IEEE Conference on Computer Vision and Pattern Recognition, pp. 1257-1264, June 2011.

[11] M. S. Arulampalam, S. Maskell, and N. Gordon, *et al*., "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174-188, 2002.

[12] S. P. Zhang, H. X. Yao, and X. Sun, et al., "Sparse coding based visual tracking: Review and experimental comparison," *Pattern Recognition*, vol. 46, no. 7, pp. 1772-1788, 2013.

[13] C. L. Bao, Y. Wu, and H. B. Ling, et al., "Real time robust L1 tracker using accelerated proximal gradient approach," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1830-1837, Jun 2012.

[14] T. Z. Zhang, B. Ghanem, and S. Liu, et al., "Robust visual tracking via multi-task sparse learning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2042-2049, June 2012.

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

[16] Z. L. Jiang, Z. Lin, and L. S. Davis, "Learning a Discriminative Dictionary for Sparse Coding via label consistent k-svd," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1697-1704, 2011.

[17] Z. L. Jiang, Z. Lin, and L. S. Davis, "Label Consistent K-SVD: Learning A Discriminative Dictionary for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2651-2664, 2013.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311-4322, 2006.

[19] Datasets are available: http://www.visual-tracking.net.

[20] B. Y. Liu, J. Z. Huang, and L. Yang, et al., "Robust tracking using local sparse appearance model and K-selection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1313-1320, June 2011.