A unified online dictionary learning framwork with label information for robust object tracking

*Baojie Fan**
College of Automation
Nanjing University of Posts and Telecommunications
Nanjing, China
jobfbj@gmail.com

Jing Sun, Yang Cong, Yingkui Du
State Key Laboratory of Robotics,
Shenyang Institute Automation, Chinese Academy of Sciences
Shenyang, China

***Abstract*— In this paper, a supervised approach to online learn a structured sparse and discriminative representation for object tracking is presented. Label information from training data is incorporated into the dictionary learning process to construct a robust and discriminative dictionary. This is accomplished by adding an ideal-code regularization term and classification error term to the unified objective function. By minimizing the unified objective function we learn the high quality dictionary and optimal linear multi-classifier jointly. Combined with robust sparse coding, the learned classifier is employed directly to separate the object from background. As the tracking continues, the proposed algorithm alternates between robust sparse coding and dictionary updating. Experimental evaluations on the challenging sequences show that the proposed algorithm performs favorably against state-of-the-art methods in terms of effectiveness, accuracy and robustness.**

Keywords—Label information; the unified objective function for online dictionary learning; optimal linear multi-classifier;

I. INTRODUCTION

Given the initialized position and size of a target in the first frame (or former frames) of a video, the goal of visual tracking is to estimate the states of the moving target in the subsequent frames. Visual tracking is a well-known topic in computer vision with many applications such as robot navigation, automated surveillance, medical imaging, traffic monitoring, human computer interaction, etc. Despite that much progress has been made in recent years [1-20], it is a challenging task to develop a robust tracking algorithm due to numerous factors: dynamic appearance changes, illumination, occlusions, background clutters, abrupt motion, pose variation and shape deformation.

Inspired by the success of sparse representation-based face recognition [28], Mei and Ling [7] proposed the L1 tracker for robust visual tracking under the particle filter framework based on the sparse coding technique. In detail, the target templates are used to describe the tracked object and trivial templates are used to deal with outliers. This representative scheme and its extensions are robust to a wide range of image corruptions, especially moderate occlusions. Some extensions [8-14] are developed to improve the L1 tracker in terms of both speed and accuracy. In [9], APG based solution is used to improve the L1 tracker. Liu et al. [10] also develop a tracking algorithm based on local sparse model which employs histograms of sparse coefficients and the mean-shift algorithm for object tracking. Wei et al.[11] propose a robust object tracking algorithm using a collaborative model that combine a sparsity-based discriminative classifier (SDC) and a sparsity-based generative model (SGM), but it adopt the naive model updating strategy and similar metric measure, this will affect the performance of the tracker. Xu et al.[12] develop a simple yet robust tracking method based on the structural local sparse appearance model. Its representation exploits both partial information and spatial information of the target based on a novel alignment-pooling method In Zhang et al. [13], low-rank sparse learning is adopted to consider the correlations among particles for robust tracking. Inspired by these works, he develops the Multi-Task Tracking (MTT) algorithm [14]. However, the dictionary still include the trivial templates, they will degrade the efficiency and effectiveness of the tracker. In Wang et al [15], online robust non-negative dictionary learning method is developed for visual tracking, a new particle representation formulation using the Huber loss function is proposed to estimate the robust object templates. Many above trackers only use the samples from target and background as the dictionary atom, the constructed dictionary is not compact. Besides, less work focus on the quality of the dictionary during the tracking process. Most existing online dictionary learning methods do not use the robust function in the data fitting term and might be vulnerable to large outliers.

In this paper, we formulate object tracking in a particle filter framework as a binary classification problem. The priori information from training data is exploited effectively to online learn a discriminative and reconstructive dictionary. Specifically, the class label information is incorporated into the dictionary learning process as the classification error term and idea coding regularization term respectively. Combined with the traditional reconstruction error, a total objective function for dictionary learning is constructed. By minimizing the total object function, we can obtain a high quality dictionary and optimal linear classifier jointly using iterative reweighed least squares algorithm. With the help of robust sparse coding, the optimal classifier can separate the tracker object from background effectively.

The main contributions of this paper are:

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.401

(1) The priori information from the training samples is exploited to construct a compact and discriminative dictionary. It is a critical factor for the object tracker based sparse representation. The learned dictionary encourages samples from the same class to have similar representations.
(2) Learning a robust dictionary and optimal linear classifier are accomplished simultaneously by iterative reweighed least squares algorithm.
(3) Experiments show the proposed tracker outperforms some state-of-the-art methods on challenging sequences with heavy occlusion, drastic illumination changes, and large pose variations.

II. RELATED WORK

In this section, we briefly review nominal tracking methods and those that are the most related to our tracker. We focus specifically on the representative tracking methods that use particle filters, sparse representation and dictionary learning. For a more thorough survey of tracking methods, we refer the readers to [1-4]. Existing tracking algorithms can be roughly categorized as either generative or discriminative.

Generative tracking methods learn an appearance model to describe the target observations, and the aim is to search for the target location that has the most similar appearance to the model. These methods are based on either templates [6, 8, 9, 11] or subspace models [5, 8, 12]. Popular generative trackers include eigentracker [16], mean shift tracker [17], incremental tracker (IVT) [5], and VTD tracker [6]. The mean shift tracker [17] is a popular mode-finding method, which successfully copes with camera motion, partial occlusions, clutter, and target scale variations. Ross et al. [5] learn an adaptive linear subspace online for modeling target appearance and implement tracking with a particle filter. However, IVT is less effective in handling heavy occlusion or non-rigid distortion. Kwon *et al.* [6] extend the classic particle filter framework with multiple dynamic observation models to account for appearance and motion variation.

Discriminative methods cast the tracking as a binary classification problem that distinguishes the tracked targets from their surrounding backgrounds. The trained classifier is online updated during the tracking procedure. Examples of discriminative methods are MIL [18], PN [19], CT [21] and Struck [20].

Babenko et al. [18] introduce multiple instance learning into online tracking where samples are considered within positive and negative bags or sets. Kalal et al. [19] propose the pn learning algorithm to exploit the underlying structure of positive and negative samples to learn effective classifiers for object tracking. The Struck [20] ranks top in the recent benchmark [3], and it learns a kernelized structured output support vector machine online. An efficient tracking algorithm [21] based on compressive sensing theory [22] is proposed, which demonstrates that the low dimensional features randomly extracted from the high dimensional multi-scale image feature space can preserve the discriminative capability, thereby facilitating object tracking.

Besides the above trackers, we also focus on the dictionary learning, because dictionary quality is a critical factor for sparse representations. Some approaches [23-30] are developed to learn the compact and discriminative dictionaries. The representative work on dictionary learning methods includes K-SVD [23] and online dictionary learning [25]. K-SVD focuses on the representational power of the learned dictionary, but does not consider the discrimination capability of the dictionary. In order to overcome this drawback, some algorithms attempt to incorporate discriminative terms into the objective function during training have been described in [24-31]. The discrimination criteria include softmax discriminative cost function [27], Fisher discrimination criterion [30], linear predictive classification error [31]. Recently, a label consistent K-SVD (LC-KSVD) algorithm [24] is proposed to learn a compact and discriminative dictionary for sparse coding. However, the corrupted training data will influence the performance of LC-KSVD.

III. BACKGROUND

In this section, we briefly introduce the particle filter to facilitate the presentation of our model in the next section.

A. Particle Filter

Particle filters is a popular tracking framework due to its excellent performance in the presence of nonlinear target motion and to flexibility to different object representations. It can be considered as a Bayesian inference task in a Markov model with hidden state variables, which recursively approximates the posterior distribution using a finite set of weighted samples. It consists of two steps: prediction and update.

Specially, at the frame, let affine parameters $X=(x,y,s,r,\theta,\lambda)$ represent the target state, where x and y are the image coordinates, s and r are the scale and the aspect, θ is the rotation angle, λ is the skew. $Y_{1:t-1}=\{Y_1,Y_2,...,Y_{t-1}\}$ denotes the observation of the target from the first frame to the frame $t-1$. Particle filters tracking estimates and propagates the probability by recursively performing prediction

$$p(X_t \mid Y_{1:t-1}) = \int p(X_t \mid X_{t-1}) p(X_{t-1} \mid Y_{1:t-1}) dX_{t-1} \quad (1)$$

and updating

$$p(X_t \mid Y_{1:t}) = \frac{p(Y_t \mid X_t) p(X_t \mid Y_{1:t-1})}{p(Y_t \mid Y_{1:t-1})} \quad (2).$$

The optimal state for the frame t is obtained according to the maximal approximate posterior probability

$$X_t^* = \arg\max_X p(X \mid Y_{1:t}) \quad (3)$$

This inference is governed by the model $p(X_t \mid X_{t-1})$, which describes the temporal correlation of the tracking results in consecutive frames, and it is modeled to be Gaussian with the dimensions of X_t assumed independent. The observation model $p(Y_t \mid X_t)$ reflects the similarity between a target candidate and dictionary templates. In this paper, $p(Y_t \mid X_t)$ is proportional to the classifier scores.

IV. UNIFIED FRAMWOR FOR ONLINE DICTIONARY LEARNING

Inspired by the above work, an approach to online learn a structured sparse and discriminative representation for object tracking is presented in this section. We develop a supervised dictionary. The class label and structure information among samples are incorporated into the dictionary learning process as the discriminative term and structured regularization term respectively. Combined with the traditional reconstruction error term, a unified objective function for object tracking can be obtained. In this way, the dictionary and the classifier are learned jointly. With the high quality dictionary, structured sparsity based discriminative classifier can be directly used for object tracking.

A. the Unified Object Function

To be concrete, the objective function for our object tracking is defined as

$$\begin{aligned} &< D^*, A^*, X^* >= \arg\min_{D,A,X} \|Y - DX\|_p^p \\ &+\lambda_1 \|H - AX\|_p^p + \lambda_2 \|Q - X\|_p^p + \lambda_3 \|X\|_{p,1} \end{aligned} \quad (5).$$

where parameters $\lambda_1, \lambda_2, \lambda_3$ control the relative weight of three terms: reconstruction error term, classification error, idea coding regularization term, mixed norm regularization term.

Reconstruction Error term $\|Y - DX\|_1^1$: This data fitting term is widely used in sparse representation based tracking [7] and dictionary learning. Its value reflects the presence of occlusion and whether a candidate particle is sampled from the background. We compute the reconstruction errors of all the particles with the learned dictionary items at the same time.

Ideal structured regularization term $\|Q - X\|_1^1$ **:** This term includes the structured and discriminative information from training samples. $Q = [q_i, q_2, ...q_M] \in R^{K \times M}$ is the idea representation for Y, M is the number of training samples. We hope that X is very close to Q, and force the samples from the same class to have similar discriminative sparse representation without losing structure information. q_i is the sparse code corresponding to an input signal y_i with the form $q_i = [q_i^1, q_i^2, ...q_i^K]^{'} = [0,...1,1,1,...]^{'} \in R^K$. We cast the object tracking can be viewed as a binary classification problem: object (class T) and background (class B). If the training data is sampled from the tracked object region, the coefficients in q_i for class T are all 1s, while the others are all 0s. For example, the training samples $Y = [y_1, y_2, y_3, y_4]$ include two classes: y_1, y_2 belong to object T and y_3, y_4 are from background B, the ideal representation Q for Y is as follows:

$$Y = DQ = D * \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad (6)$$

Classification error term $\|H - AX\|_1^1$ **:** the term measures the classification error, and it supports learning an optimal classifier. A simple linear classifier $f(A; X) = AX$ is adopted, where $H = [h_1, h_2, ...h_N] \in R^{2 \times N}$ are the class labels of training data Y. A is the classifier parameters. $h_i = [1,0]^t$ is the corresponding label vector of y_i, and the non-zero position indicates the class label of y_i.

L1 norm regularization term $\|X\|_1$ **:** By adding a sparseness criterion into the objective function (5), we are able to learn a sparse and structural representation with the learned high-quality dictionary D_t. The proposed tracker is under the particle filter framework. The candidate particles are densely sampled around the current tracking target and their representations will be sparse and similar with respect to the given dictionary D_t. In other words, a few items in D_t are required to represent all the particles.

B. Optimization Procedure

The choice of norm p is restricted in this paper. Among its convex options, two popular and widely studied norms (L1, L2) are selected. The solution of Eq. (5) for the choices of p is described in the following.

For $p = 2$, to solve optimization problem in the equation (5), the proposed algorithm alternates between robust sparse coding and dictionary updating. We rewrite the proposed object function as two steps.

Dictionary learning:

$$< D^*, A^* >= \arg\min_{D,A} \|Y_{L1} - D_{L1}X\|_2^2 + \lambda_2 \|Q - X\|_2^2 + \lambda_3 \|X\|_{2,1} \quad (7)$$

where $Y_{L1} = [Y, \sqrt{\lambda_1}H]^{'}$, $D_{L1} = [D, \sqrt{\lambda_1}A]^{'}$.

Sparse coding:

$$X^* = \arg\min_X \|Y_{L2} - D_{L2}X\|_2^2 + \lambda_3 \|X\|_{p,1} \quad (8)$$

where $Y_{L2} = [Y, \sqrt{\lambda_1}Q, \sqrt{\lambda_2}H]^{'}$ $D_{L2} = [D, \sqrt{\lambda_1}I, \sqrt{\lambda_2}A]^{'}$. The equations (7-8) can be transformed as the classical dictionary learning problem that K-SVD [26] and online dictionary learning for sparse coding [32] can both obtain the satisfied solution.

Let $E_i = Y_L - \sum_{j \neq i} d_{new}^j x_j^R$, where d_{new}^j is the dictionary item and x_j^R is the *j-th* row in its corresponding coefficients. Discarding the zero atoms in E_i and x_j^R, we are able to obtain $\bar{E}_i$ and $\bar{x}_j^R$ respectively. $d_i, \bar{x}_j^R$ are computed by minimizing the following object function:

$$< d_i, \bar{x}_j^R >= \arg\min_{d_i, \bar{x}_j^R} \left\| \bar{E}_i - d_i \bar{x}_j^R \right\|_F^2 \quad (9)$$

By performing the SVD operation $U\Sigma V^{'} = SVD(\bar{E}_i)$, the solution of equation (9) is as follows:

$$d_i = U(:,1), \quad \bar{x}_j^R = \Sigma(1,1)V(:,1)$$

For $p = 1$, the data fitting term with L1 norm replaces the L2 norm data fitting term in the traditional sparse coding framework. This operation can improve the robustness for the

dictionary and sparse coding, and make the total object function less vulnerable to large outliers. For example, given a simple L2 norm optimization problem

If $Y=[2,4,6,7,8,9,11,13,15,18]$, the optimal solution for $\mu=\arg\min_{\mu}\sum_{i}(y_i-\mu)^2$ is $\mu_1=\frac{1}{10}\sum_i y_i=9.3$. All the items in Y can be viewed as inliers. In the following, a large value $y_i=75000$ is added into Y, and the corresponding solution for changed Y is $\mu_2=\frac{1}{11}\sum_i y_i=6826.7$. With the simple example, we can see that L2 norm data fitting is not a robust metric when the outlier exits ($y_i=75000$).

During the process of object tracking, the challenging factors such as occlusion, illumination changes, abrupt motion, background clutters are usually regarded as the outlier. If the L2 norm data fitting is adopted for sparse representation based tracker, the drift will cumulative and result in tracking failure. But, L1 fitting functions can overcome above problem and make tracking reliable. So, the equations (7-8) are transformed as

$$\begin{aligned}&<D^*,A^*,X^*>=\arg\min_{D,A,X}\|Y-DX\|_1^1\\&+\lambda_1\|H-AX\|_1^1+\lambda_2\|Q-X\|_1^1+\lambda_3\|X\|_{1,1}\end{aligned}\tag{10}$$

As in [29], iterative reweighed least squares algorithm (IRLS) is used to obtain the optimal solution. It solves the following t wo problems in each iteration until convergence.

$$D_{L1}(j,:)=\arg\min_{d_{L1}}\frac{1}{n}\sum_{i=1}^{n}w_i^j(Y_{L1}(i,j)-d_{L1}X_i)^2\tag{11}$$

$$C^j=D(j,:)M^j\tag{12}$$

where $w_i^j=1\Big/\sqrt{(Y_{L1}(i,j)-d_{L1}X_i)^2+\delta}$, δ is a small positive value . $C^j=\sum_{i=1}^{n}w_i^jY_{L1}(i,j)\beta_i^T$, $M^j=\sum_{i=1}^{n}w_i^j\beta_i\beta_i^T$. As the tracking continues, the dictionary update with the coming data, the online versions of C^j, M^j is as follows:

$$C_N^j=C_n^j+\sum_{i=n+1}^{N}w_i^jY_{L1}(i,j)\beta_i^T\tag{13}$$

$$M_N^j=M_n^j+\sum_{i=n+1}^{N}w_i^j\beta_i\beta_i^T\tag{14}$$

V. EXPERIMENTS

In this section, we demonstrate the merits of the proposed algorithm with extensive experimental results. Our trackers (with different data fitting term: our_L1 and our_L2) are evaluated on 9 challenging tracking sequences (e.g. car11, cliffbar, singer1 and woman sequences) that are publicly available online. We evaluate the proposed tracker against ten state-of-the-art visual tracking algorithms including: ONND [15], LSST [8], MTT [14], CT [21], VTD [6], MIL [18], PN[19], IVT[5], and L1[7]. These trackers are implemented using publicly available source codes or binaries provided by the authors. They are initialized using their default parameters.

A. Parameter Setting

The proposed algorithm is implemented in MATLAB R2011b on a Pentium 2.3 GHz Dual Core laptop with 2GB memory. For each sequence, the location of the target object is manually labeled in the first frame. Each image sample from the target and background is normalized to a 32 × 32 or 48 × 16 patch. We set the parameters $\lambda_1,\lambda_2,\lambda_3$ in Eq.(4) are 4, 2, 0.01 respectively.

B. Qualitative Comparison

In the Car11 sequence, a car is driven into a very dark environment. The contrast between the tracked target and its surrounding background is low, and the ambient light changes significantly. Furthermore, the low image resolution of the target object makes tracking difficult. The tracking results are illustrated in Figure 1(a). Due to changes in lighting, MIL algorithms start to drift around frame 60. L1 method starts to fail in frame 250. IVT, LSST, MTT and our trackers algorithms perform well as in the whole sequence. However, the accuracy and robustness of these methods are less than our proposed algorithm. Whereas the other methods drift away when drastic illumination variation occurs or when similar objects appear in the scene (e.g., #0305), especially the car makes a turn at about frame 280.

The tracking object in the girl sequence undergoes occlusion (complete occlusion of the girl's face as she swivels in the chair), large pose change, and scale variation with in-plane and out-of-plane rotations (from large to small, and from small to large). The tracking results are shown in Fig. 3(b). The experimental results demonstrate that our methods achieve the best performance in this sequence. Other trackers experience drift at different instances: IVT at frame 436, and VTD at frame 477.

In the Cliffbar video, the background has similar texture to the target. Moreover, the target undergoes scale variance, in-plane rotation, and abrupt motion as shown in Figure 3(d). The L1, IVT, CT, MIL, LSST, our_L2 methods drift to the cluttered background, while our proposed tracker has the best performance on this sequence, it can adapt the scale and rotation change of the target, and overcome the influence of similar background and motion blur.

In the skating sequence, there are abrupt object motion, occlusions, severe illumination and scale changes, viewpoint changes, which lead most of the trackers to fail. Only VTD, and our_L1 trackers can handle these changes well and track the target throughout the sequence, as shown in Figure 3(e).

The singer sequence contains abrupt object motion with significant illumination and scale changes, especially, from frame 121 and frame 321, the stage light changes drastically, which is challenging for most of the trackers as shown in Fig. 3(f). Our trackers perform well in the whole sequence. The center error and overlap rate in Table 1 and 2 have verified that our proposed trackers are better than other methods.

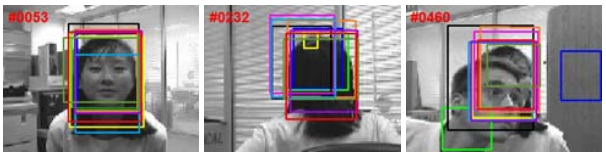
In the caviar sequence, the target is occluded by two people at times and one of them is similar in color and shape to the target. Numerous methods fail to track the target because there are similar objects around it when heavy

occlusion occurs. In contrast, our_L1 algorithm achieves stable performance in the entire sequence when there is a large scale change with heavy occlusion at frame 500.

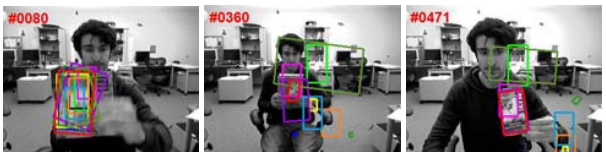
The football sequence is challenging due to the cluttered background, because there are many football players with the similar helmets in appearance to the tracked object in this scene. When the tracked target approaches other football players, some trackers are not robust and begin to drift, as shown in Fig.1 (i). Especially, when the two football players collide at frame 290, most tracking methods cannot locate the target correctly. Only our trackers, CT, VTD overcome this problem and successfully locate the correct object in the whole sequence. The accuracy of our method is the highest.

For the faceocc2 sequence in Figure 1(i), most trackers start drifting from the man's face when it is almost fully occluded by the book. The proposed algorithm performs well especially when partial occlusion or in-plane rotation occurs.

(a) car11

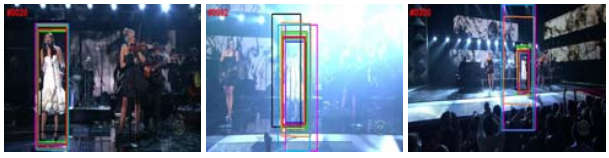


(b) girl



(c) cliffbar

(d) skating1



(e) singer1

(f) caviar

(g) football

(i) faceocc2

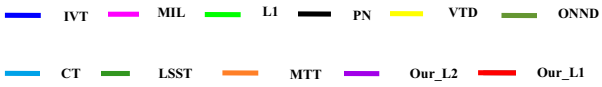


Figure 1 Comparison of 9 trackers on 8 video sequences in terms of bounding box reported

C. Quantitative Comparison

To give a quantitative comparison between the 11 methods, two popular evaluation criteria are used, namely, center location error (CLE) and tracking success rate (TSR). The CLE is computed as the distance between the predicted center position and the ground truth center position. Table.1 summarizes the average center location errors in pixels. The TSR is computed as the ratio of the number of frames the target is successfully tracked to the total frames in the sequence. To define whether the target is successfully tracked at a frame, we use the score in the PASCAL VOC challenge [31], which can be computed as

$$score = area(R_T \cap R_G) \Big/ area(R_T \cup R_G) \quad (15).$$

where R_T is the current the tracking result and R_G is the ground truth. Table.2 gives the average tracking success rate.

D. Discussion

From the above experiments, we can see that the proposed trackers perform well against some state-of-the-art algorithms. Our_L1 has the lowest center location error and highest tracking success rate for the tested image sequences. It can overcome the interference of outlier during the tracking process. But it costs most computing time. The reason is that the IRLS algorithm obtains one column value in dictionary learning in each iteration. Besides, experimental results show that the label information in training data can improve the performance of trackers.

TABLE I. AVERAGE CENTER LOCATION ERROR (IN PIXEL). THE BEST TWO RESULTS ARE SHOWN IN RED AND BLUE FONTS

	car11	gir	cliff	skating	singer1	caviar	football	faceocc2
ivt	2.106	48.47	24.81	11.72	8.483	65.96	13.61	10.21
l1	33.25	62.43	49.60	163.7	4.570	65.67	18.17	11.12
pn	25.11	23.15	11.25	55.68	32.69	44.45	13.54	18.59
vtd	27.05	21.44	34.56	13.32	4.057	58.20	4.300	10.41
mil	43.47	32.21	13.35	161.7	15.17	100.2	13.66	14.06
mtt	2.802	23.89	46.17	195.6	16.62	64.99	9.842	8.65
ct	8.352	32.93	23.42	186.5	13.26	35.79	8.138	22.17
lsst	1.870	73.11	23.31	120.0	3.506	**3.073**	7.574	**3.70**
onnd	**1.742**	27.88	29.61	**7.303**	12.34	63.34	20.37	4.260
Our_l2	2.758	**10.0**	**8.366**	20.89	**3.50**	3.198	**4.199**	**3.920**
Our_l1	**1.692**	**10.73**	**2.630**	12.42	**3.236**	**2.548**	**3.892**	4.281

TABLE II. AVERAGE TRACKING SUCCESS RATE. THE BEST TWO RESULTS ARE SHOWN IN RED AND BLUE FONTS

	car11	gir	cliff	skating	singer1	caviar	football	faceocc2
ivt	0.81	0.42	0.56	**0.68**	0.66	0.14	0.55	0.59
l1	0.44	0.33	0.19	0.09	0.70	0.13	0.57	0.67
pn	0.38	0.58	0.38	0.12	0.41	0.16	0.50	0.49
vtd	0.43	0.51	0.33	0.57	0.79	0.15	0.61	0.59
mil	0.17	0.52	0.46	0.12	0.34	0.13	0.57	0.61
mtt	0.81	0.63	0.31	0.09	0.42	0.14	0.66	0.72
ct	0.53	0.51	0.39	0.05	0.34	0.17	0.69	0.54
lsst	0.81	0.12	0.56	0.13	0.79	**0.85**	0.68	0.80
onnd	**0.82**	0.42	0.35	**0.63**	0.20	0.05	0.41	0.79
Our_l2	0.69	**0.68**	**0.63**	0.44	**0.85**	**0.83**	**0.82**	**0.83**
Our_l1	**0.83**	**0.69**	**0.79**	0.56	**0.86**	**0.83**	**0.83**	**0.82**

VI. CONCLUSIONS

In this paper, by exploiting the strength of the prior information in the training data, a unified object function is constructed to online learn and update a structured sparse and discriminative representation for object tracking. The approach encourages samples from the same class to have similar representations. Experimental results on challenging image sequences demonstrate that the proposed trackers perform favorably against some state-of the- art algorithms. Possible future work is to develop the structured and low-rank representation for robust object tracking.

Acknowledge

This work has been funded by the Natural Science Foundation of China (grant no. 61203270), 2012 State Key Laboratory of Robotics Open Project and the Introduction Foundation for the Talent of Nanjing University of Tele. and Com(no. NY212028).

REFERENCES

[1] A. Yilmaz, O. Javed, M. Shah, "Object tracking: a survey," *ACM Computing Surveys*, vol.38, no.4, 2006.

[2] X. Li, W. Hu, C.Shen, Z.Zhang, A.Dick, A.van den Hengel, "A survey of appearance model in visual object tracking," *ACM Transactions on Intelligent Systems and Technology*, 2013.

[3] Y. Wu, Jongwoo Lim, and Ming-Hsuan yang, "Online object tracking: a Benchmark," In *CVPR*, 2013.

[4] S., Zhang, H., Yao, Xin Sun, X., Lu, "Sparse coding based visual tracking: Review and experimental comparison," *Pattern Recognition*, vol.46, no.7, pp. 1772-1788, 2012.

[5] D. Ross, J. Lim, R. Lin, and M. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no.1, pp.125–141, 2008.

[6] J. Kwon and K. Lee, "Visual tracking decomposition," In *CVPR*, pp. 1269–1276, 2010.

[7] X. Mei and H. Ling, "Robust visual tracking using l1 minimization," In *ICCV*, pp. 1436–1443, 2009.

[8] D. Wang, H. Lu, and M. H. Yang, "Least soft-threshold squares tracking," In CVPR, pp. 2371-2378, 2013.

[9] C. Bao, Y.Wu, H. Ling, and H. Ji. Real time robust ℓ1 tracker using accelerated proximal gradient approach. In *CVPR*, pp.1830–1837, 2012.

[10] B. Liu, J. Huang, L. Yang, & C. Kulikowski, "Robust visual trackingwith local sparse appearance model and k-selection," In *CVPR*, pp. 1–8, 2011.

[11] W. Zhong, H. Lu, and M.-H. Yang, "Robust object tracking via sparsity-based collaborative model," In *CVPR*, pp. 1838–1845, 2012.

[12] X. Jia, H. Lu, and M.-H. Yang. "Visual tracking via adaptive structural local sparse appearance model." In CVPR, pp.1822–1829, 2012.

[13] T. Zhang, B. Ghanem, S. Liu, & N. Ahuja, "Low-rank sparse learning for robust visual tracking," In *ECCV*, pp. 1–8, 2012.

[14] T. Zhang, B. Ghanem, S. Liu, & N. Ahuja, "Robust visual tracking via multi-task sparse learning," In *CVPR*, pp. 1-8, 2012.

[15] N. Wang, J. Wang, D. Yeung, "Online Robust Non-negative Dictionary Learning for Visual Tracking," In *ICCV*, 2013.

[16] M. Black, A. Jepson. "Eigentracking: robust matching and tracking of articu- lated objects using a view-based representation," in: European Conference on Computer Vision, pp. 329–342, 1996.

[17] D. Comaniciu, V. Ramesh, P. Meer. " Kernel-based object tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, 25: 564–575, 2003.

[18] 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.

[19] Z. Kalal, J. Matas, K. Mikolajczyk, "pn learning: boot-strapping binary classifiers by structural constraints," In *CVPR*, pp. 49–56, 2010.

[20] S. Hare, A. Saffari, and P. H. Torr, "Struck: Structured output tracking with kernels," In *ICCV*, pp. 263–270, 2011.

[21] K. Zhang, L. Zhang, M. H. Yang, "Real-Time Compressive Tracking," In *ECCV,* pp. 864-877, 2012.

[22] E. Candes and T. Tao, "Near optimal signal recovery from random projections and universal encoding strategies," *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406–5425, 2006.

[23] 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.1, pp. 4311-4322, 2006.

[24] Z. Jiang, Z. Lin, Larry 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.

[25] J. Marial, F. Bach, J. Ponce, G. Sapiro. Online dictionary learning for sparse coding, In *ICML*, 2009.

[26] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Signal Processing,* 54(1):4311-4322, 2006.

[27] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *CVPR*, pages 1-8, 2008.

[28] J. Wright, A. Ganesh, S. Rao, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[29] C. Lu, J. Shi, and J. Jia. Online robust dictionary learning. In *CVPR*, 2013.

[30] K. Huang and S. Aviyente. Sparse representation for signal classification, In *NIPS*, pages 609-616, 2007.

[31] H. Zhang, A. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition, In *CVPR*, pages 2126-2136, 2006.

[32] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object class (voc) challenge," *Int. J. Comput. Vision.*,vol. 88, pp.303–338, 2010