Discriminative Multi-Task Sparse Learning for Robust Visual Tracking Using Conditional Random Field

Behzad Bozorgtabar[1]
[1]Vision & Sensing, HCC Lab, ESTeM
University of Canberra
Email: Behzad.Bozorgtabar@canberra.edu.au

Roland Goecke[1,2]
[2]IHCC, RSCS, CECS
Australian National University
Email: roland.goecke@ieee.org

***Abstract*—In this paper, we propose a discriminative multi-task sparse learning scheme for object tracking in a particle filter framework. By representing each particle as a linear combination of adaptive dictionary templates, we utilise the correlations among different particles (tasks) to obtain a better representation and a more efficient scheme than learning each task individually. However, this model is completely generative and the designed tracker may not be robust enough to prevent the drifting problem in the presence of rapid appearance changes. In this paper, we use a Conditional Random Field along with a multi-task sparse model to extend our scheme to distinguish the object candidate from the background particle candidate. By this way, the number of particle samples is reduced significantly, while we make the tracker more robust. The proposed algorithm is evaluated on 10 challenging sequences. The results confirm the effectiveness of the approach, which significantly outperforms state-of-the-art trackers in terms of accuracy measures including the centre location error and the overlap ratio, respectively.**

I. Introduction

Visual tracking plays an important role in numerous vision applications such as activity recognition and visual surveillance. While much progress has been made in recent years, it is still a challenging task to develop a robust method for complex scenes due to large appearance changes caused by camera or object motion, varying illumination, occlusions, shape deformation and pose variation. To overcome these challenges, a representation should be strong enough to identify the object and verify predictions in each video frame.

Recently, sparse representation has been introduced for visual tracking [1], [2], where a candidate can be sparsely modelled as a linear combination of the dictionary templates. Treating the candidates' representations individually in a particle filter framework causes a computationally expensive l_1 minimisation. Zhang *et al.* [**?**] proposed multi-task sparse learning for the particles by introducing a mixed norm $l_{p,q}$ to keep the coefficients' similarity level in accordance with the candidates' similarities, thereby making the tracker more efficient. However, from a discriminative point of view, each candidate can be viewed as either foreground or as part of the background. Ignoring this aspect of the particles may cause the tracker to drift in the case of appearance changes, for example due to background clutter.

The Conditional Random Field (CRF) is a well-known probabilistic framework for structured prediction, which is often used to solve the image labelling problem in computer vision, such that each pixel of an image corresponds to a node on the CRF graph [**?**]. In this paper, we utilise CRF as an additional step to separate foreground candidates from the background.

II. Related Work

Tracking methods can be classified as being either generative or discriminative. *Discriminative* models formulate tracking as a classification problem to distinguish the target from the background. In these models, usually a dynamic classifier is trained via different methods, e.g. boosting [3]. Some examples of discriminative methods are online multiple instance learning tracking [3], [4], co-training tracking [5], ensemble tracking [6].

Generative methods formulate the tracking problem as searching for the regions most coherent to the target proposal. They usually construct robust object representations using particular features including subspace representation [7], fragment-based representation [8] and local descriptors [**?**]. Some other examples of generative trackers are the mean-shift tracker [9] and VTD tracker [10].

Recently, the sparse learning method has attracted considerable attention in visual tracking. In [1], a tracking candidate is represented as a sparse linear combination of object templates and trivial templates. Although this method leads to a good performance, it comes at the computational expense of considering multiple independent particles. Zhang *et al.* [**?**] utilised a mixed norm to consider the dependencies among particles for the robust tracking.

In this paper, we take the advantages of both generative and discriminative representations based on sparse learning for the proposed tracker. Using joint sparse learning for the particles, we first extract the discriminative representation of the candidates with respect to the positive and negative adaptive dictionary templates, then we utilise a CRF to extract binary labels for the particles.

III. System Overview

A. Bayesian Filtering Framework

We address visual tracking as a sparse representation problem in the Bayesian filtering framework. In this paper, we utilise the particle filter as an effective realisation of Bayesian filtering, where the posterior distribution $p\left(s_t \left| z_{1:t} \right.\right)$ of the target is recursively approximated by a set of weighted particles $\left\{ s_t^{(i)}, \pi_t^{(i)} \right\}_{i=1}^{N}$, where $z_{1:t}$ denotes the set of observations up to and including the time step t and each particle represents a possible state s_t and a weight π_t associated with it. Considering the Bayesian estimation scheme, the distribution can be recursively updated as

$$p\left(s_t \left| z_{1:t-1} \right.\right) = \int p\left(s_t \left| s_{t-1} \right.\right) p\left(s_{t-1} \left| z_{1:t-1} \right.\right) ds_{t-1} \qquad (1)$$

$$p\left(s_t \left| z_{1:t} \right.\right) \propto p\left(z_t \left| s_t \right.\right) p\left(s_t \left| z_{1:t-1} \right.\right) \qquad (2)$$

In this framework, new particles are drawn by sampling from a known proposal function $q\left(s_t \left| s_{0:t-1}^{(i)} \right., z_{1:t}\right)$ where the simplest choice for the proposal function is the state evolution model $p\left(s_t \left| s_{t-1} \right.\right)$ itself for sampling. Further, the optimal state is obtained by the maximum a posteriori (MAP) estimation over a set of N samples. We model the motion of a target object between two consecutive frames with a six-dimensional affine transformation s_t. The transformation of each parameter is modelled independently by a scalar Gaussian distribution. Then, the dynamic model $p\left(s_t \left| s_{t-1} \right.\right)$ can be represented by a Gaussian distribution. The likelihood (observation) model $p\left(z_t \left| s_t \right.\right)$ reflects the similarity measure for the tracking target.

In the t^{th} frame, we consider N particle samples, whose observations (Gray scale values) are denoted in matrix form as $Y = [y_1, \ldots, y_N] \in R^{m \times N}$. Given a dictionary $D_t = [d_1, \ldots, d_K] \in R^{m \times K}$, where d_i is the i^{th} dictionary item, we seek sparse representation of the particles over the dictionary templates. We denote D_t with a subscript because the dictionary templates will be progressively updated to consider variations in object appearance due to changes in illumination, viewpoint, etc.

B. Multi-Task Sparse Representation

Since in particle filter based trackers, most particle samples are densely sampled around the current target state, their sparse representations with respect to the given dictionary are not independent. Therefore, a global structure exists among particle samples (e.g. the sparse representation of closely located particles is more likely to be similar with regard to the dictionary). To explore this underlying structure between particles, [?] recently presented a tracker based on the group sparse learning. By inducing the mixed norms on the sparse representation matrix $\|W\|_{p,q}$,[1] we can extract the underlying structure between particles as

$$W^* = \arg\min_{W} \frac{1}{2} \|Y - D_t W\|_2^2 + \lambda \sum_i \|W_i\|_p \qquad (3)$$

where λ is a sparsity constraint factor and $W = [w_1, \ldots, w_N] \in R^{K \times N}$ are sparse representations of Y.

C. Solving the Optimisation Problem

The formulated problem in Eq. 5 is a convex optimisation problem with a non-smooth objective function due to the non-negativity constraint assumption for the particles representation matrix W. We seek to solve the above optimisation problem using the accelerated proximal method (APM) [**?**] due to its ability of optimal convergence compared to other first-order techniques. APM iterates between two sequences of variables: **(1)** an attainable solution (updating the current representation matrix) W_k and **(2)** an aggregation matrix R_k.

1) *Proximal Mapping:* At each iteration, the representation matrix W_k can be updated by the generalised proximal mapping as the following problem

$$\min_{W} \frac{1}{2} \|W - H\|_2^2 + \widetilde{\lambda} \|W\|_{p,1} \qquad (4)$$

where $\widetilde{\lambda} = \frac{\lambda}{\gamma_k}$, $H = R_k - \frac{1}{\gamma_k} \nabla^k$ and γ_k denotes the step size. ∇^k is computed as

$$\nabla^k = 2 D_t^T \left(D_t R_k - Y\right) \qquad (5)$$

2) *Aggregation:* At the k^{th} iteration of APM, the aggregation matrix is updated by a linear combination of W_k and W_{k-1} from previous iterations [2]

$$\hat{R}_{k+1} = W_{k+1} + \frac{\mu_{k+1}\left(1 - \mu_k\right)}{\mu_k} \left(W_{k+1} - W_k\right) \qquad (6)$$

where μ_k is conventionally set to $\frac{2}{k+1}$.

D. Conditional Random Field

The tracker designed by multi-task sparse learning is a purely generative one till now. However, in the case of object appearance changes, for example due to occlusion and background clutter, we face the drifting problem. In this paper, we utilise a CRF [**?**] to return a discriminative representation of the target candidates using their geometrical structure (pairwise potential) and prior discriminative representation over the adaptive dictionary (unary potential). By this way, we learn the conditional distribution over the class labelling (foreground and background) (see Fig. 1). Modelling the geometrical structure between candidates, the particles can be treated as a graph $G = (V, E)$, where $V = \{y_1, y_2, \cdots, y_n\}$ denotes the vertex set composed of n particles and E the set of edges, respectively. In the CRF, we aim to minimise the following objective function

$$-log\left(Pr\left(c \mid G; \omega\right)\right) = \sum_{y_i} \Psi\left(c_i \mid y_i\right) + \qquad (7)$$

$$\omega \sum_{y_i, y_j \in Edge} \Phi\left(c_i, c_j \mid y_i, y_j\right) (8)$$

where $Pr\left(c \mid G; \omega\right)$ is the conditional probability of the class label assignments c given the proposed graph and a weight ω, which indicates the trade-off between spatial regularisation and our confidence in the classification. Ψ and Φ are unary and pairwise potentials, respectively. Below, we describe how to obtain these terms in detail.

[1] Where $\|W\|_{p,q} = \left(\sum_{i=1}^{K} \|W_i\|_p^q\right)^{\frac{1}{q}}$, $\|W_i\|_p$ is the l_p norm of W_i, (i^{th} row of matrix W). Here, we set $q = 1$.

[2] μ_k is conventionally set to $\frac{2}{k+1}$.

a) Unary Potential:: Given the sparse representation of the particles W, it is obvious that the target can be better represented by the linear combination of positive templates , while the background can be better described by the range of negative templates. Therefore, we are able to calculate the prior confidence score of each particle belonging to the target as

$$Pr\left(c_i \mid y_i\right) = exp\left(-\frac{\left(\varepsilon_{pos} - \varepsilon_{neg}\right)}{\sigma}\right) \tag{9}$$

where $\varepsilon_{pos} = \left\|y_i - D_t^{(pos)} W^{(pos)}\right\|$ is the reconstruction error of the i^{th} particle y_i candidate using the positive (foreground) template set $D_t^{(pos)}$, and $W^{(pos)}$ is the corresponding sparse representation matrix. Similarly, $\varepsilon_{neg} = \left\|y_i - D_t^{(neg)} W^{(neg)}\right\|$ is the reconstruction error of the candidate y_i with respect to the negative (background) template set $D_t^{(neg)}$ and its related sparse coefficient $W^{(neg)}$. The variable σ is fixed to be a small constant balances two reconstruction errors . Finally, the unary potential is calculated by the log likelihood of the probability obtained by the confidence score as

$$\Psi\left(c_i \mid y_i\right) = -log\left(Pr\left(c_i \mid y_i\right)\right), \quad c_i \in \{1, -1\} \tag{10}$$

b) Pairwise Potential:: Using multi-task sparse learning for the particles' representation, we achieve a common structure among tasks (particle candidates). However, the particles may exhibit a more sophisticated structure (e.g. the sparse representation of closely located particles is more likely to be similar rather than those from distant spatial locations). In our proposed graph structure, the set of edges E explores the mutual dependencies for the particles (here, the pairwise similarity in terms of spatial distance and appearance likeness can be modelled by a Markov Random Field (MRF) framework). The pairwise potential can be modelled as

$$\Phi\left(c_i, c_j \mid y_i, y_j\right) = \left(\frac{1}{1 + \|y_i - y_j\|}\right) \quad i \neq j \tag{11}$$

where $\|y_i - y_j\|$ is the l_2-norm of gray scale difference between particle candidates. For the pairwise potential, we utilise the homogeneous Ising model. In our similarity graph, the particles are connected if and only if they are among the k nearest neighbours of each other. After the CRF process, we will obtain a binary labels for the particles on which the target and background candidates are distinctly distinguished.

IV. Observation Model Using Binary Labels

We propose a discriminative confidence score for the particles, which can naturally integrate the particle candidate importance into the tracking result. In fact, we introduce another aspect of the particle representation where the binary label obtained by the CRF indicates whether a certain particle is supposed to be a foreground object or a part of background. Therefore, we reduce the computational complexity significantly caused by the large number of candidates.

In order to obtain the observation model for the tracking result, we compute the candidate similarity with respect to the positive templates and negative templates by its corresponding foreground $W^{(pos)}$ and background $W^{(neg)}$ representations, respectively. Finally, the scores of the particles are determined by the difference in contribution of these two parts and the tracking result z_t at time instance t is the particle y_i such that

$$i = \arg\max y_{i=1,\ldots,N} \quad L_i \odot \left(\left\|W^{(pos)}\right\|_1 - \left\|W^{(neg)}\right\|_1\right) \tag{12}$$

where L_i is the binary label of the i^{th} candidate obtained by CRF and $\odot$ is an element-wise product. This likelihood function not only encourages the tracking result to be represented well by the object and *not* the background templates, it also ignores those particles, which are supposed to be background.

V. Dictionary Update

Since dictionary templates should handle appearance changes of the target during tracking, e.g. illumination variation, updating target object templates is a vital part of object tracking. A steady appearance model is not reliable for a long period tracking to handle appearance changes of the target. On the other hand, if templates are updated frequently, the tracker will degrade. Here, our initial dictionary comprising n_p positive templates and n_n negative templates are obtained by drawing sample images around the target location.

First, for the background (negative) templates $D_t^{(neg)}$, since the surrounding image region of two consecutive frames are similar, we only update the negative templates from the surrounding of the current tracking result (from image regions away (more than 6 pixels)).

On the other hand, for the object (positive) templates $D_t^{(pos)}$, we allocate a similarity measure that demonstrates how similar the template to the tracking result is. Similar to [**?**], in each frame, we measure the similarity between the current tracking result and the object templates (Euclidean distance). Then, we compare the maximum similarity value with a predefined threshold ($\theta = 0.6$). We substitute the corresponding object template, which has the maximum similarity with the new target appearance, if the similarity is higher than the threshold. By this way, we prevent updating an occluded target object, which causes drifting problem. Fig. 2 shows the effect of using binary labels obtained by CRF in our proposed framework.

VI. Experimental Results

In this paper, we compare the proposed tracker with nine state-of-the-art trackers. For a fair comparison, we use the default parameters for these trackers as they were reported. These trackers can be divided into different categories

- On one side, generative trackers such as the incremental subspace-based IVT [7] and two channels blurring approaches DFT [11].
- On the other side, discriminative heuristic trackers including the STRUCK [12] and the multiple instance learning-based tracker MIL [13].
- Trackers with local target descriptors such as TLD [14], which estimate the new target location by combining the local motion estimates with discriminative learning of the patches and Frag Track [8], in which the target object is represented by multiple image fragments or patches.

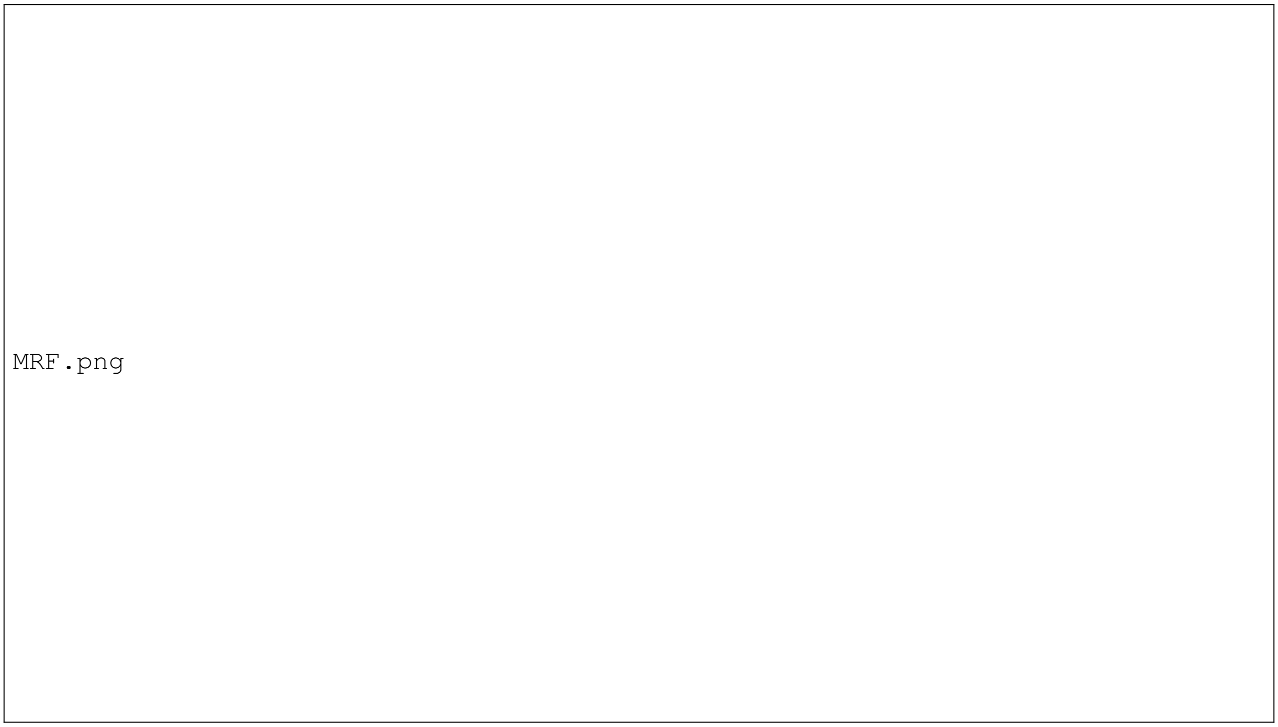

Fig. 1: An illustration of using a CRF in the proposed particle filter based tracker, the algorithm aims to segment foreground candidates (*red* circles) from the background candidates (*blue* circles) using their mutual dependencies.

- In addition, sparse representations based trackers including MTT [**?**] and L1APG [15] are used. In these trackers, the target is represented by holistic templates.
- Finally, the VTD tracker is included, which adapts mixture models based on sparse principal component analysis.

A. Parameters Setting

The parameters of the proposed tracking algorithm are fixed in all experiments. The number of positive templates n_p and negative templates n_n are 8 and 150, respectively. The sparsity factor in Eq. 3 is set to 0.5. In Eq. (4), we set $\widetilde{\lambda}$ (by cross-validation) to 0.005 and γ_k to 1/0.01, respectively. The maximum iteration of objective function in Eq. 5 is set to 15, and 630 particles are chosen as candidate samples in each frame. An observed target image patch is partitioned into non-overlapping local fragments (image patches) of size 8×8 pixels, each of which is independently represented in gray scale values, vectorised and normalised to be a vector with unit l_2 norm. Then, we concatenate these local feature vectors so that the global structural information is maintained. This representation allows us to handle partial occlusion. ω in Eq. 7 is set to 0.5.

B. Challenging Sequences

We evaluate the effectiveness of the proposed tracker on 11 challenging video sequences. Each of the video sequence was labelled with different attributes such as abrupt motion, illumination change, occlusion and scale change. Their ground truths are provided. Figure 3 shows some sample tracking results for different video sequences. In the *Dudek* sequence, where the person displays partial occlusion and a variety of scale rotations (in-plane rotation and out-of-plane rotation), our method outperforms other trackers, e.g. the MIL and L1APG methods. In this sequence, in the presence of rotation and when the pose of the object changes gradually, some trackers drift from the target (e.g. Frag in frame **967**). On the contrary, the IVT and DFT methods perform well on this sequence and comparable to the results of our tracker. Regarding the proposed tracker, due to the discriminative aspect of the proposed scheme, it is able to maximally capture the appearance change information and accurately distinguish the target from the background when the target faces with heavy occlusion in the *Woman* sequence.

C. Performance Measure

For the purpose of measuring the performance of the proposed tracker, two metrics including the centre location error and the overlap ratio are employed. It should be noted that a smaller average error or a bigger overlap rate means a more accurate result. The tracker's overlap rate in each frame is defined as area $\frac{area(BB_T \cap BB_G)}{area(BB_T \cup BB_G)}$, where BB_G and BB_T denote the bounding box obtained by the ground truth and a tracker, respectively. One of the important advantages of the overlap ratio is that it accounts for both position and size of the predicted and proposal bounding boxes simultaneously, and does not lead to arbitrary large errors at tracking failures. As shown in Tables I and II, the results of our tracker outperforms other trackers. Due to space limitations, the centre error and the overlap ratio plots of the two sequences are included in Fig. 4.

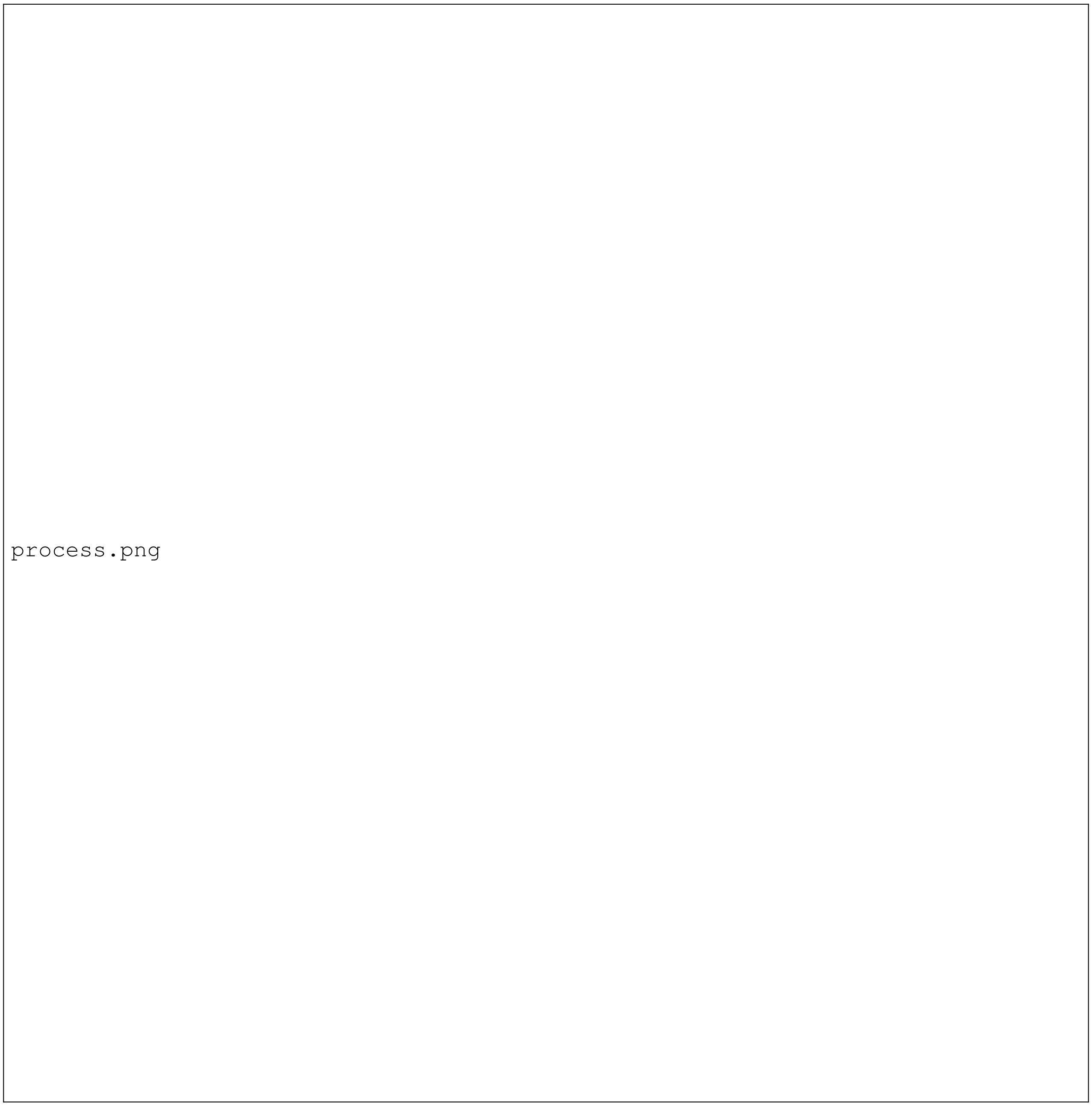

Fig. 2: This figure demonstrates the advantage of using binary labels to indicate of good or bad particle sample. Here, a sample good particle sample and a bad candidate are chosen from *Dudek* sequence. At the right side, their corresponding sparse representation with respect to the dictionary are illustrated before and after using binary labels obtained by CRF. As it can be seen, after using a binary representation of the candidates, the bad candidate is removed as it is considered as part of the background.

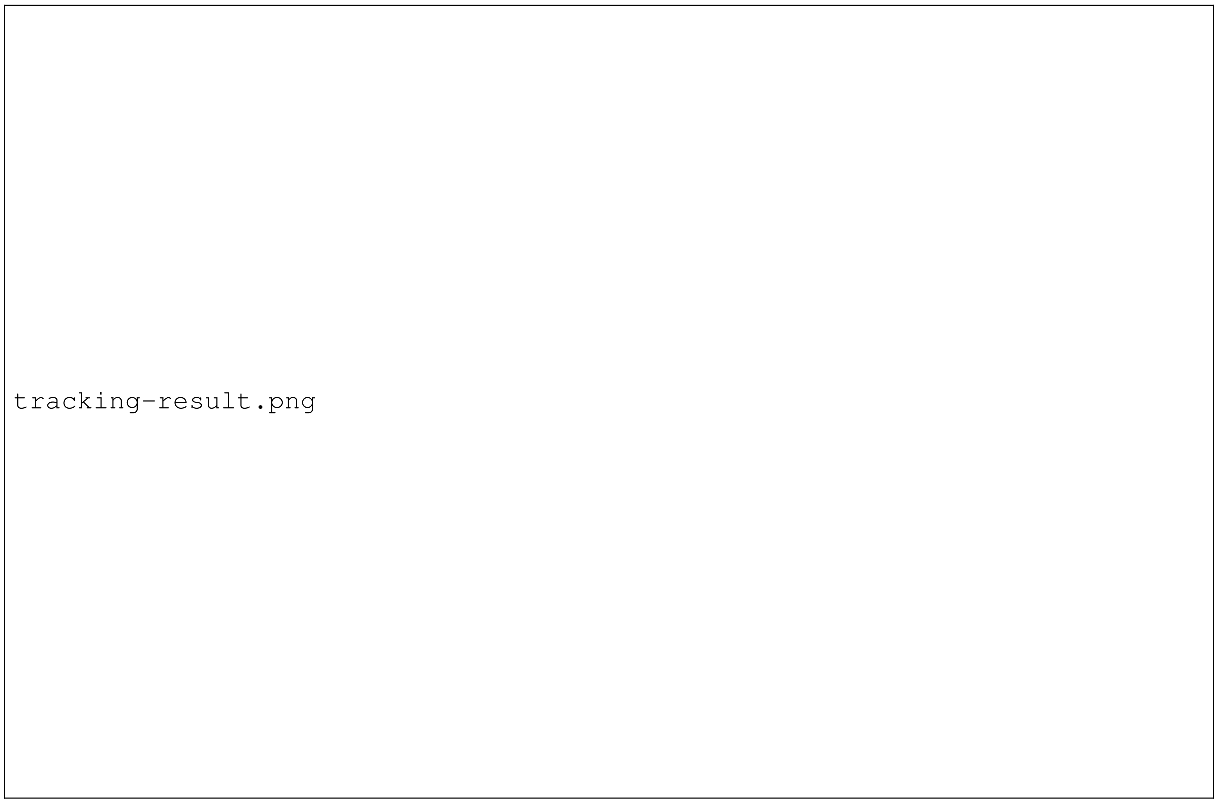

Fig. 3: Sample tracking results obtained by our tracker compared with the benchmark results in challenging sequences with heavy occlusion (*Woman sequence*) and in-plane and out-of-plane rotations in the (*Dudek sequence*).

TABLE I: Average centre location errors (in pixels). The best three results are shown in Red, Blue, and Green fonts.

Video Clip	Frag	IVT	MIL	APG	VTD	MTT	TLD	Struck	DFT	Ours
Singer1	22.0	8.5	15.2	3.2	4.1	41.2	11.6	12.6	10.4	3.7
Girl	18.0	48.6	32.2	62.4	21.4	23.9	20.3	10.0	21.5	9.6
Car11	63.8	2.1	43.5	1.7	27.1	1.8	25.1	1.9	2.2	1.9
Face	48.8	69.7	134.6	57.7	140.9	127.2	67.5	25.0	26.8	20.2
David	76.7	3.6	16.1	14.3	13.6	124.2	16.3	3.1	10.2	3.1
Dudek	61.5	8.8	20.3	70.6	66.0	53.8	10.5	11.5	9.5	8.7
Woman	113.6	167.4	122.3	118.5	136.6	127.2	110.4	10.1	15.3	8.9
Bolt	240.1	170.6	163.9	225.5	22.3	106.0	34.5	98.5	102.3	17.8
Jumping	58.6	36.7	10.2	9.1	63.2	19.3	8.0	42.0	39.5	6.9
Mountain	141.6	33.2	128.3	130.2	7.5	11.3	96.5	10.5	122.4	3.3
Tiger1	39.5	158.7	14.2	21.5	28.9	30.9	13.9	12.2	10.0	9.2

TABLE II: Average overlap rate (in pixels). The best three results are shown in Red, Blue, and Green fonts.

Video Clip	Frag	IVT	MIL	APG	VTD	MTT	TLD	Struck	DFT	Ours
Singer1	0.34	0.66	0.33	0.83	0.79	0.32	0.65	0.59	0.72	0.85
Girl	0.69	0.43	0.51	0.33	0.52	0.63	0.57	0.94	0.56	0.95
Car11	0.09	0.81	0.17	0.83	0.43	0.58	0.38	0.86	0.63	0.88
Face	0.39	0.44	0.15	0.35	0.24	0.26	0.46	0.78	0.75	0.83
David	0.19	0.71	0.45	0.57	0.53	0.28	0.44	0.79	0.61	0.81
Dudek	0.46	0.81	0.64	0.61	0.46	0.36	0.71	0.75	0.68	0.84
Woman	0.20	0.18	0.16	0.06	0.15	0.17	0.07	0.86	0.74	0.81
Bolt	0.07	0.13	0.16	0.10	0.82	0.19	0.77	0.15	0.11	0.89
Jumping	0.13	0.29	0.54	0.57	0.09	0.31	0.98	0.18	0.20	0.96
Mountain	0.06	0.66	0.14	0.11	0.89	0.81	0.25	0.87	0.10	0.93
Tiger1	0.19	0.71	0.39	0.15	0.73	0.75	0.65	0.73	0.89	0.92

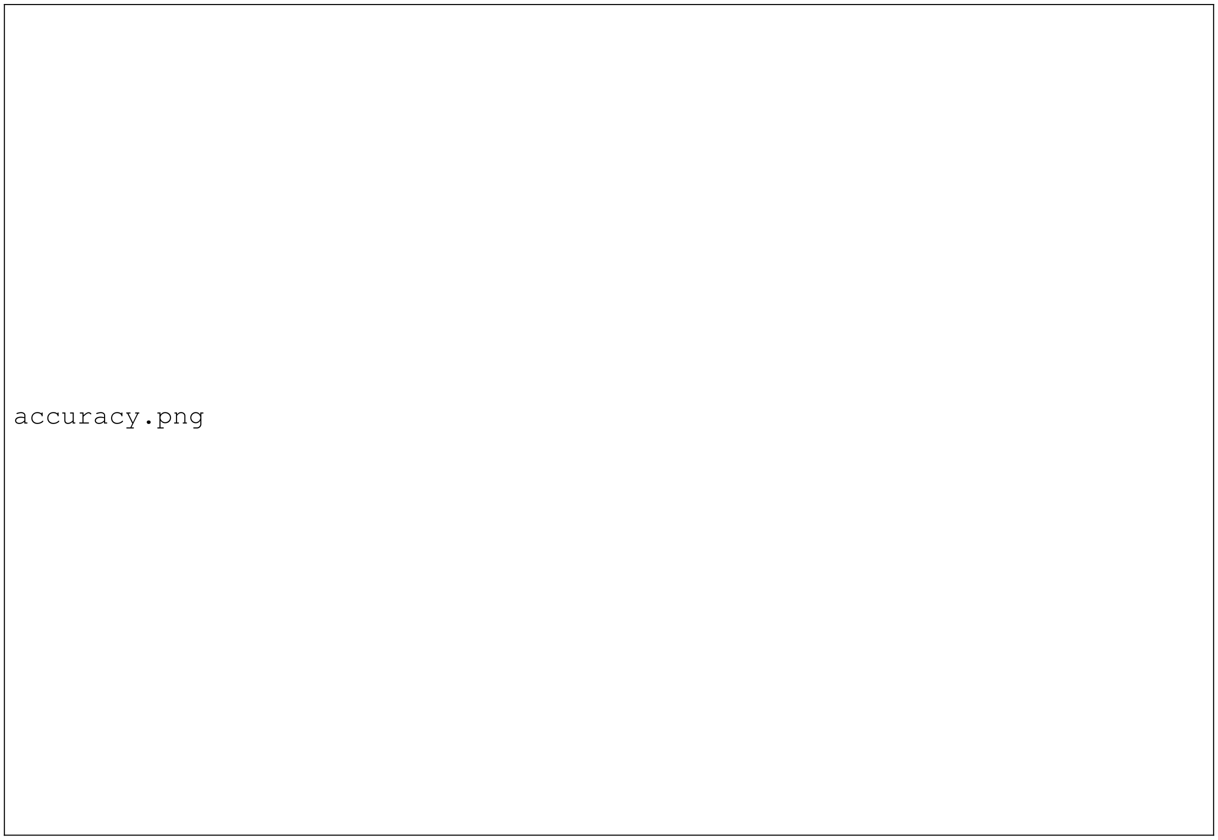

Fig. 4: Evaluation of different trackers by centre error rate and overlap ration on the sample sequences (*Mountain bike* and *Dudek* sequences). It should be noted that a smaller error for centre locations means a more accurate result, while a larger value for overlap ratio indicates a better result.

VII. Conclusions

In this paper, we have formulated particle filter based tracking as a discriminative multi-task sparse learning problem, where the dependencies between particles are explored by the learned mixed sparsity norm. In addition, we consider another aspect of the particles and induce binary labels obtained by the CRF, which helps to improve the accuracy and robustness of our tracker. We extensively evaluate the performance of our tracker on 11 challenging videos and show its superior performance in both devised accuracy measures.

References

[1] X. Mei and H. Ling, "Robust visual tracking and vehicle classification via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 11, pp. 2259–2272, 2011.

[2] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai, "Minimum error bounded efficient ? 1 tracker with occlusion detection," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1257–1264.

[3] H. Grabner, M. Grabner, and H. Bischof, "Real-Time Tracking via Online Boosting," in *British Machine Vision Conference (BMVC)*, 2006, pp. 47–56.

[4] B. Babenko, M. Yang, and S. Belongie, "Visual Tracking with Online Multiple Instance Learning," in *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] R. Liu, J. Cheng, and H. Lu, "A robust boosting tracker with minimum error bound in a co-training framework," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1459–1466.

[6] S. Avidan, "Ensemble tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 2, pp. 261–271, 2007.

[7] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 125–141, 2008.

[8] A. Adam and E. Rivlin, "Robust Fragments-based Tracking Using the Integral Histogram," in *Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 798–805.

[9] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 5, pp. 564–577, 2003.

[10] J. Kwon and K. M. Lee, "Visual tracking decomposition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1269–1276.

[11] L. Sevilla-Lara and E. Learned-Miller, "Distribution fields for tracking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1910–1917.

[12] S. Hare, A. Saffari, and P. H. Torr, "Struck: Structured output tracking with kernels," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 263–270.

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

[14] Z. Kalal, J. Matas, and K. Mikolajczyk, "Pn learning: Bootstrapping binary classifiers by structural constraints," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 49–56.

[15] C. Bao, Y. Wu, H. Ling, and H. Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," in *Computer Vision and*

Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 1830–1837.