Coupling Semi-supervised Learning and Example Selection for Online Object Tracking

Min Yang, Yuwei Wu, Mingtao Pei, Bo Ma and Yunde Jia

Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, Beijing 100081, China

Abstract. Training example collection is of great importance for discriminative trackers. Most existing algorithms use a sampling-and-labeling strategy, and treat the training example collection as a task that is independent of classifier learning. However, the examples collected directly by sampling are not intended to be useful for classifier learning. Updating the classifier with these examples might introduce ambiguity to the tracker. In this paper, we introduce an active example selection stage between sampling and labeling, and propose a novel online object tracking algorithm which explicitly couples the objectives of semi-supervised learning and example selection. Our method uses Laplacian Regularized Least Squares (LapRLS) to learn a robust classifier that can sufficiently exploit unlabeled data and preserve the local geometrical structure of feature space. To ensure the high classification confidence of the classifier, we propose an active example selection approach to automatically select the most informative examples for LapRLS. Part of the selected examples that satisfy strict constraints are labeled to enhance the adaptivity of our tracker, which actually provides robust supervisory information to guide semi-supervised learning. With active example selection, we are able to avoid the ambiguity introduced by an independent example collection strategy, and to alleviate the drift problem caused by misaligned examples. Comparison with the state-of-the-art trackers on the comprehensive benchmark demonstrates that our tracking algorithm is more effective and accurate.

1 Introduction

Object tracking aims to estimate the trajectory of an object automatically in a video sequence. Although the task is easily fulfilled by human vision system, designing a robust online tracker remains a very challenging problem due to significant appearance variations caused by factors such as object deformation, illumination change, occlusion, and background clutters.

Numerous tracking algorithms have been proposed to address appearance variations, and most of them fall into two categories: generative methods and discriminative methods. Generative methods represent an object in a particular feature space, and then find the best candidate with maximal matching score. Some popular generative trackers include incremental visual tracking [1],

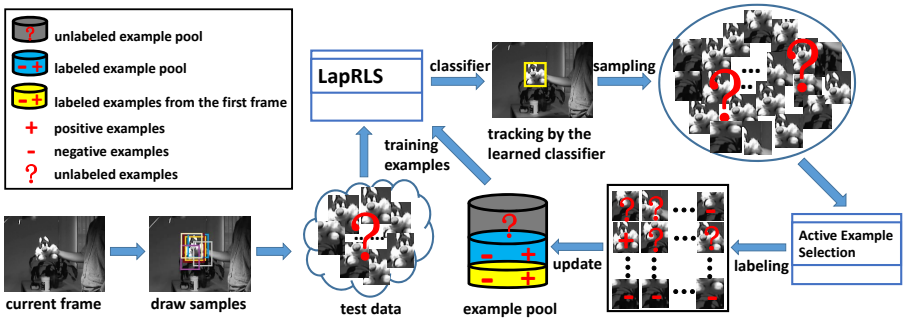


Fig. 1. Overview of our tracker. LapRLS is used to learn a robust classifier which is able to exploit both labeled and unlabeled data during tracking. An active example selection stage is introduced between sampling and labeling, which couples the objectives of semi-supervised learning and example selection. The figure is best viewed in color.

visual tracking decomposition [2], sparse representation based tracking [3–7], and least soft-threshold squares tracking [8]. Discriminative methods cast tracking as a binary classification problem that distinguishes the object from the background [9–13]. Benefiting from the explicit consideration of background information, discriminative trackers usually are more robust against appearance variations under complex environments. In this paper, we focus on learning an online classifier which is able to capture appearance changes adaptively for object tracking.

The performance of discriminative trackers largely depends on the training examples used for classifier learning. Existing algorithms often collect training examples via a two-stage strategy [9]: sampling and labeling. The sampling process generates a set of examples around the current tracking result, and the labeling process estimates the labels of these examples using heuristic approach that depends on the current tracking result (*e.g.*, examples with small distance to the current track are labeled as positive, and examples far away from the current track are negative).

This widely used example collection strategy raises several issues. Firstly, the objective of the sampling process may not be consistent with the objective for the classifier, which makes the example collection strategy independent of classifier learning. The examples collected directly by sampling are neither necessarily informative nor intended to be useful for the classifier learning, and might introduce ambiguity to the tracker. Secondly, assigning labels estimated by the current tracking result to unlabeled examples may easily cause drift [14, 15, 9]. Slight inaccuracy of tracking results can lead to incorrectly labeled examples, and consequently degrades the classifier. State-of-the-art discriminative trackers mainly focus on learning a classifier that is robust to poorly labeled examples (*e.g.*, semi-supervised learning [14, 16–18], P-N learning [19], multiple instance learning [15] and self-paced learning [20]). However, the first issue is rarely mentioned in the literature of object tracking.

In this paper, we propose an online object tracking algorithm which explicitly couples the objectives of semi-supervised learning and example selection. The overview of our tracker is shown in Fig. 1. We use a manifold regularized semi-supervised learning method, *i.e.*, Laplacian Regularized Least Squares (LapRLS) [21], to learn a robust classifier for object tracking. We show that it is crucial to exploit the abundant unlabeled data which can be easily collected during tracking to improve the classifier and alleviate the drift problem caused by label noisy. To avoid the ambiguity introduced by an independent example collection strategy, an *active example selection* stage is introduced between sampling and labeling to select the examples that are useful for LapRLS. The active example selection approach is designed to maximize the classification confidence of the classifier using the formalism of active learning [22, 23], thus guarantees the consistency between classifier learning and example selection in a principled manner. Our experiments suggest that coupling semi-supervised learning and example selection leads to significant improvement on tracking performance. To make the classifier more adaptive to appearance changes, part of the selected examples that satisfy strict constraints are labeled, and the rest are considered as unlabeled data. According to the stability-plasticity dilemma [24], the additional labels provide reliable supervisory information to guide semi-supervised learning during tracking, and hence increases the plasticity of the tracker, which is validated in our experiments.

Semi-supervised tracking: Semi-supervised approaches have been previously used in tracking. Grabner *et al.* [14] proposed an online semi-supervised boosting tracker to avoid self-learning as only the examples in the first frame are considered as labeled. Saffari *et al.* [16] proposed a multi-view boosting algorithm which considers the given priors as a regularization component over the unlabeled data, and validate its robustness for object tracking. Kalal *et al.* [19] presented a P-N learning algorithm to bootstrap a prior classifier by iteratively labeling unlabeled examples via structural constraints. Gao *et al.* [18] employed the cluster assumption to exploit unlabeled data to encode most of the discriminant information of their tensor representation, and showed great improvement on tracking performance.

The methods mentioned above actually determine the "pseudo-label" of the unlabeled data, and do not discover the intrinsic geometrical structure of the feature space. In contrast, the LapRLS algorithm employed in our algorithm learns a classifier that predicts similar labels for similar data points by constructing a data adjacency graph. We show that it is crucial to consider the similarity in terms of label prediction during tracking. Bai and Tang [17] introduced a similar algorithm, *i.e.*, Laplacian ranking SVM, for object tracking. However, they adopt a handcrafted example collection strategy to obtain the labeled and unlabeled data, which limits the performance of their tracking method.

Active learning: Active learning, also referred to as experimental design in statistics, aims to determine which unlabeled examples would be the most informative (*i.e.*, improve the classifier the most if they were labeled and used as training data) [22, 23], and has been well applied in text categorization [25]

and image retrieval [26, 27]. In this work, we propose an active example selection approach to couple semi-supervised learning and example selection by using the framework of active learning, in which the task is to select the examples that improve the prediction accuracy of LapRLS the most.

We show that the active example selection introduces several advantages for object tracking over existing methods. Firstly, it guarantees the consistency between classifier learning and example collection in a principled way. That is, the selected examples are meaningful for LapRLS, which can improve the classification performance. Secondly, the active example selection tends to choose the representative examples, which reduces the amount of training data without performance loss. Thirdly, assigning labels to the selected examples alleviates the drift problem caused by label noise. According to the theory of active learning, the examples, that minimize the predictive variance when they are used for training, will be selected. Thus misaligned examples are intended to be rejected by the active example selection.

2 The Proposed Tracking Algorithm

Our tracker operates by alternately performing two stages: classifier learning with LapRLS, and training example collection with active example selection. After describing these two stages in Sec. 2.1 and Sec. 2.2, respectively, we formulate object tracking in a Bayesian inference framework and summarize our tracking algorithm in Sec. 2.3.

2.1 Classifier Learning with LapRLS

Given a set of l labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$, and a set of u unlabeled examples $\{x_i\}_{i=l+1}^{l+u}$, the LapRLS algorithm seeks for a real valued function $f : \mathcal{X} \to \mathbb{R}$ by solving the following optimization problem [21]:

$$f^* = \arg\min_{f \in \mathcal{H}_K} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \lambda_1 \|f\|_K^2 + \frac{\lambda_2}{2} \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 W_{ij}, \quad (1)$$

where $\mathcal{H}_K$ is a Reproducing Kernel Hilbert Space (RKHS) which is associated with a positive definite Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\|\cdot\|_K$ is the norm defined in $\mathcal{H}_K$, and W is a $(l+u) \times (l+u)$ similarity matrix with entries W_{ij} indicating the adjacency weights between data points x_i and x_j. The last term in Eq.(1) is an approximated manifold regularizer that preserves the local geometrical structure represented by a weighted adjacency graph with similarity matrix W. It actually respects a smoothness assumption, that is, data points which are closed to each other in a high-density region should share similar measurements (or labels) given by trained function. According to the spectral graph theory, this regularized term can be rewritten as

$$\frac{1}{2} \sum_{i,j=1}^{l+u} (f(x_i) - f(x_j))^2 W_{ij} = \boldsymbol{f}^\top L \boldsymbol{f}, \quad (2)$$

where $\boldsymbol{f} = [f(x_1), f(x_2), \ldots, f(x_{l+u})]^\top$, and L is the graph Laplacian given by $L = D - W$. Here, D is a diagonal matrix defined as $D_{ii} = \sum_{j=1}^{l+u} W_{ij}$. We adopt the local scaling method [28] to define the similarity matrix,

$$W_{ij} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|_2^2}{\sigma_i \sigma_j}\right), \text{ if } i \in N_k^j \text{ or } j \in N_k^i, \\ 0, \text{otherwise}, \end{cases} \tag{3}$$

where N_k^i indicates the index set of the k nearest neighbors of x_i in $\{x_i\}_{i=l}^{l+u}$, $\sigma_i = \|x_i - x_i^{(k)}\|_2$, and $x_i^{(k)}$ is the k-th nearest neighbor of x_i in $\{x_i\}_{i=l}^{l+u}$.

The Representer Theorem (see details in [21]) shows that the solution of Eq.(1) is an expansion of kernel functions over both labeled and unlabeled data,

$$f^*(x) = \sum_{i=1}^{l+u} \omega_i^* K(x, x_i). \tag{4}$$

By substituting this form into Eq.(1), we get a convex differentiable objective function of the $(l+u)$-dimensional vector $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_{l+u}]^\top$,

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega} \in \mathbb{R}^{l+u}}{\arg\min} \|\tilde{\boldsymbol{y}} - \Lambda K \boldsymbol{\omega}\|^2 + \lambda_1 \boldsymbol{\omega}^\top K \boldsymbol{\omega} + \lambda_2 \boldsymbol{\omega}^\top K L K \boldsymbol{\omega}, \tag{5}$$

where K is the $(l+u) \times (l+u)$ Gram matrix with entries $K_{ij} = K(x_i, x_j)$, $\tilde{\boldsymbol{y}}$ is the augmented label vector given by $\tilde{\boldsymbol{y}} = [y_1, \ldots, y_l, 0, \ldots, 0]^\top$, and Λ is an $(l+u) \times (l+u)$ diagonal matrix with the first l diagonal entries being 1 and the rest 0, *i.e.*, $\Lambda = diag(1, \ldots, 1, 0, \ldots, 0)$.

The solution of Eq.(5) can be acquired by setting the gradient *w.r.t* $\boldsymbol{\omega}$ to zero,

$$\boldsymbol{\omega}^* = (\Lambda K + \lambda_1 I + \lambda_2 L K)^{-1} \tilde{\boldsymbol{y}}, \tag{6}$$

where I is an $(l+u) \times (l+u)$ identity matrix. Obviously, the prediction function can be efficiently obtained by solving a single system of linear equations described in Eq.(6), and then the predicted label of a test data x is given by Eq.(4).

2.2 Training Example Collection with Active Example Selection

Given the object location at the current frame, a large set of unlabeled examples is generated by random sampling around the object location, denoted as $P = \{p_i\}_{i=1}^{N_p}$, where N_p is the number of examples. Existing tracking algorithms directly employ a labeling process on this example set, and ignore the correlation between example collection and classifier learning. In this work, we propose an active example selection approach using the formulism of active learning to automatically select the most informative examples among P for LapRLS.

Now we consider the example selection problem from the perspective of active learning. Given the candidate set $V = \{v_i\}_{i=1}^{n}$, the task is to find a set of examples $Z = \{z_i\}_{i=1}^{m}$ that together are maximally informative [22]. Suppose

that we can observe the labels of z_i by a measurement process $c_i = f(z_i) + \epsilon_i$, where c_i is the observed label of example z_i, f is the underlying label prediction function and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is measurement noise. Using Z as labeled data and the rest in V as unlabeled data, the estimate of f, denoted as $\hat{f}$, can be obtained by using LapRLS,

$$\hat{f}(x) = K_{x,V}\hat{\boldsymbol{\omega}}, \tag{7}$$

$$\hat{\boldsymbol{\omega}} = (K_{VZ}K_{ZV} + \lambda_1 K + \lambda_2 KLK)^{-1} K_{VZ}\boldsymbol{c}, \tag{8}$$

where $(K_{x,V})_{1j} = K(x, v_j)$, $(K_{VZ})_{ij} = K(v_i, z_j)$, $(K_{ZV})_{ij} = K(z_i, v_j)$, $(K)_{ij} = K(v_i, v_j)$, and $\boldsymbol{c} = [c_1, \ldots, c_m]^\top$. Note that Eq.(7) and Eq.(8) can be easily derived from Eq.(4) and Eq.(6), respectively.

Denote $H = K_{VZ}K_{ZV} + \lambda_1 K + \lambda_2 KLK$ and $\Delta = \lambda_1 K + \lambda_2 KLK$, the covariance matrix of $\hat{\boldsymbol{\omega}}$ can be expressed as

$$\begin{aligned} Cov(\hat{\boldsymbol{\omega}}) &= Cov(H^{-1}K_{VZ}\boldsymbol{c}) \\ &= H^{-1}K_{VZ}Cov(\boldsymbol{c})K_{ZV}H^{-1} \\ &= \sigma^2(H^{-1} - H^{-1}\Delta H^{-1}), \end{aligned} \tag{9}$$

where the third equation uses the assumption $Cov(\boldsymbol{c}) = \sigma^2 I$. The covariance matrix $Cov(\hat{\boldsymbol{\omega}})$ characterizes the confidence of the estimation, or the informativeness of the selected examples [23]. Different criteria can be applied to the covariance matrix to obtain different active learning algorithms for LapRLS. He [27] used the D-optimality criterion that minimizes the determinant of $Cov(\hat{\boldsymbol{\omega}})$ to design an active learning method for image retrieval. However, the criteria does not directly consider the quality of predictions on test data.

Inspired by the work in [25], we design the objective of our active example selection approach in a transductive setting. Let $\mathbf{f}_V = [f(v_1), \ldots, f(v_n)]^\top$ be the true labels of all examples in V given by the underlying label prediction function f, and $\hat{\mathbf{f}}_V = [\hat{f}(v_1), \ldots, \hat{f}(v_n)]^\top$ be the predictions on V given by the estimator $\hat{f}$, then the covariance matrix of the predictive error $\boldsymbol{f}_V - \hat{\boldsymbol{f}}_V$ is given by

$$\begin{aligned} Cov(\boldsymbol{f}_V - \hat{\boldsymbol{f}}_V) &= KCov(\hat{\boldsymbol{\omega}})K \\ &= \sigma^2 K(H^{-1} - H^{-1}\Delta H^{-1})K. \end{aligned} \tag{10}$$

We aim to select m examples Z from V such that the average predictive variance $\frac{1}{n}\mathrm{Tr}(Cov(\boldsymbol{f}_V - \hat{\boldsymbol{f}}_V))$ is minimized, *i.e.*, a high confidence of predictions on V is ensured. Since the regularization parameters (*i.e.*, λ_1 and λ_2) are usually very small, we have

$$\mathrm{Tr}(K(H^{-1} - H^{-1}\Delta H^{-1})K) \approx \mathrm{Tr}(KH^{-1}K). \tag{11}$$

Therefore, the formulation of our active example selection approach can be expressed as

$$\begin{aligned} &\max_{Z} \ \mathrm{Tr}\big(K(K_{VZ}K_{ZV} + \lambda_1 K + \lambda_2 KLK)^{-1}K\big) \\ &s.t. \quad Z \subset V, \ |Z| = m \end{aligned} . \tag{12}$$

Algorithm 1 Sequential Active Example Selection

Initialize: $M = K(\lambda_1 K + \lambda_2 KLK)^{-1}K$; Z'; $Z = \emptyset$
$M \leftarrow M - M_{VZ'}(M_{Z'Z'} + I)^{-1}M_{Z'V}$
while $|Z| < m$ **do**
  select z according to Eq.(15);
  $Z' = Z' \cup \{z\}$, $Z = Z \cup \{z\}$;
  $M \leftarrow M - M_{V,z}M_{z,V}/(1 + M_{z,z})$;
end while
return Z

Note that the example selection itself is independent of the observed labels $\mathbf{c}$.

Let Δ^{-1} be the Moore-Penrose inverse of Δ, we can get the following equations by applying Woodbury matrix identity,

$$\begin{aligned} KH^{-1}K &= K(K_{VZ}K_{ZV} + \Delta)^{-1}K \\ &= K\Delta^{-1}K - K\Delta^{-1}K_{VZ}(K_{ZV}\Delta^{-1}K_{VZ} + I)^{-1}K_{ZV}\Delta^{-1}K, \end{aligned} \tag{13}$$

where I is an $m \times m$ identity matrix. We define a new kernel matrix $M = K\Delta^{-1}K$, and rewrite Eq.(12) into a much simple form,

$$\begin{aligned} &\max_{Z} \ \mathrm{Tr}\left(M_{VZ}(M_{ZZ} + I)^{-1}M_{ZV}\right) \\ &s.t. \quad Z \subset V, \ |Z| = m \end{aligned} \tag{14}$$

The problem of Eq.(14) is actually a combinatorial optimization problem which is NP-hard. We present a sequential greedy optimization approach to solve Eq.(14). The rational is two-fold. First, a sequential assumption greatly simplifies the problem and ensures the efficiency of our tracker. Second, it is straightforward to incorporate the current set of labeled examples in an incremental way. Considering the current set of labeled examples during example selection ensures the representativeness of the selected examples.

The sequential approach selects just one example in each iteration until m examples have been selected. Denote the selected examples in the previous iterations as Z', the task of each iteration is to seek for a new example $z \in V - Z'$ by solving Eq.(14). Denote $\tilde{\Delta} = K_{VZ'}K_{Z'V} + \Delta$, Eq.(14) can be rewritten into a canonical form,

$$\begin{aligned} &\max_{z} \ \|\tilde{M}_{V,z}\|^2/(1 + \tilde{M}_{z,z}) \\ &s.t. \quad z \in V - Z' \end{aligned} \tag{15}$$

where $\tilde{M} = K\tilde{\Delta}^{-1}K = M - M_{VZ'}(M_{Z'Z'} + I)^{-1}M_{Z'V}$, $\tilde{M}_{V,z}$ and $\tilde{M}_{z,z}$ are z's column and diagonal entry in $\tilde{M}$, respectively. Eq.(15) can be easily solved by directly selecting $z \in V - Z'$ with the highest $\|\tilde{M}_{V,z}\|^2/(1 + \tilde{M}_{z,z})$.

Starting from a set Z' and $M = K(\lambda_1 K + \lambda_2 KLK)^{-1}K$, m most informative examples can be selected sequentially. We summarize our active example

selection approach in Algorithm 1. Note that there is no need for matrix inverse at each iterative step.

Recall that we employ active example selection to automatically select informative examples from the set P generated by sampling. In addition, we intend to incorporate the current set of labeled examples into the example selection problem to ensures the representativeness of the selected examples. Hence, we set Z' as the current set of labeled examples and construct the candidate set as $V = Z' \cup P$ before we perform Algorithm 1 to select useful examples.

2.3 Bayesian inference framework

In this paper, we cast object tracking as a Bayesian inference task with a hidden Markov model. Given the observed image set $\mathcal{O}_{1:t} = \{\boldsymbol{o}_1, \cdots, \boldsymbol{o}_t\}$ up to time t, the optimal state $\boldsymbol{s}_t$ of an object can be estimated by Bayesian theorem,

$$p\big(\boldsymbol{s}_t\big|\mathcal{O}_{1:t}\big) \propto p(\boldsymbol{o}_t|\boldsymbol{s}_t) \int p\big(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}\big) p\big(\boldsymbol{s}_{t-1}\big|\mathcal{O}_{1:t-1}\big)\ d\boldsymbol{s}_{t-1}, \tag{16}$$

where $p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1})$ is the motion model that predicts the next state $\boldsymbol{s}_t$ from the previous state $\boldsymbol{s}_{t-1}$, and $p(\boldsymbol{o}_t|\boldsymbol{s}_t)$ is the observation model that estimates the likelihood of the observation $\boldsymbol{o}_t$ at the state $\boldsymbol{s}_t$ belonging to the object class. In practice, a particle filter [29] is used to approximate the posterior $p(\boldsymbol{s}_t\big|\mathcal{O}_{1:t})$ by a finite set of N_s samples $\{\boldsymbol{s}_t^i\}_{i=1}^{N_s}$ with importance weights $\{\pi_t^i\}_{i=1}^{N_s}$. The samples $\boldsymbol{s}_t^i$ are drawn from the motion model and the corresponding weights are given by the observation likelihood $p(\boldsymbol{o}_t|\boldsymbol{s}_t^i)$.

Motion model: We apply the affine transformation with six parameters to model the object motion. Formally, $\boldsymbol{s}_t = (x_t, y_t, \sigma_t, \alpha_t, \theta_t, \phi_t)$ where (x_t, y_t) denote translation, $\sigma_t, \alpha_t, \theta_t, \phi_t$ are scale, aspect ratio, rotation angle, and skew direction at time t, respectively. The motion model is formulated as Brownian motion, *i.e.*, $p(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) = \mathcal{N}(\boldsymbol{s}_t; \boldsymbol{s}_{t-1}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix which indicates the variances of affine parameters.

Observation model: For the tracking at time t, we first generate N_s samples $\{\boldsymbol{s}_t^i\}_{i=1}^{N_s}$ from the previous state $\boldsymbol{s}_{t-1}$. Then the corresponding image regions can be cropped from the observed image $\boldsymbol{o}_t$ by applying affine transformations using $\boldsymbol{s}_t^i$ as parameters. After feature extraction, we can obtain a set of test data, denoted as $\{b_t^i\}_{i=1}^{N_s}$. Integrating this newly test data into the current set of unlabeled examples, denoted as U, together with the current set of labeled examples, denoted as L, an adaptive prediction function f_t can be learned with LapRLS. The observation likelihood of the sample $\boldsymbol{s}_t^i$ is given by

$$p\big(\boldsymbol{o}_t|\boldsymbol{s}_t^i\big) \propto exp\big(-\|1 - f_t(b_t^i)\|^2\big). \tag{17}$$

Here we assume that positive examples are labeled with 1, and negative examples are labeled with 0. At each time stamp, the sample with the maximum observation likelihood is chosen as the tracking result.

Model update: Once the object is located, we sample a large set of unlabeled examples P, and employ active example selection to select a set of informative examples Z. To make the trained classifier more adaptive to appearance

changes, we assign labels to part of the set Z according to the following constraints: the distances between positive examples and the current track should be smaller than a threshold τ, and negative examples should not overlap the current track. The rest examples that do not satisfy the constraints are considered as unlabeled data. Then the informative examples Z are used to update the current set of labeled examples L and the current set of unlabeled examples U, where random replacement happens once the number of examples in L or U reaches the maximum values $|L|$ or $|U|$.

3 Experimental Results

We evaluate our tracker with 10 state-of-the-art methods on a recent benchmark [30], where each tracker is tested on 51 challenging videos. The state-of-the-art trackers include TLD [19], MIL [15], VTD [2], Struck [9], SCM [4], CT [10], SPT [12], LSST [8], RET [13] and ONNDL [6]. We use the source codes publicly available on the benchmark (except that the source codes of SPT, LSST, RET and ONNDL are provided by the authors) with the same initialization and their default parameters. Since the trackers involve randomness, we run them 5 times and report the average result for each sequence.

3.1 Implementation Details

We normalize the object region to 32×32 pixels, and extract 9 overlapped 18×18 local patches within the region by sliding windows with 7 pixels as step length. Each patch is represented as a 32-dimensional HOG feature [31], and these features are grouped into a 288-dimensional feature vector. For LapRLS and active example selection, we apply linear kernel and empirically set the regularization parameters λ_1 and λ_2 to be 0.001 and 0.1, respectively. The parameter k in Eq.(4) is empirically chosen as 7 according to [28]. In the first frame, 20 positive examples, 80 negative examples and 300 unlabeled examples are used to initialize the classifier. The example set capacity $|L| = 200$ and $|U| = 600$. Given the object location at the current frame, $N_p = 1200$ unlabeled examples are generated by random sampling and 20 informative examples are selected by active example selection. We set the labeling constraint parameters τ to be 3 pixels. For particle filter, the number of samples $N_s = 600$, and the state transition matrix $\boldsymbol{\Sigma} = diag(8, 8, 0.01, 0, 0, 0)$. Note that the parameters are fixed throughout the experiments in this section. Our tracker is implemented in MATLAB, which runs at 2 fps on an Intel Core i7 3.5 GHz PC with 16 GB memory.

3.2 Quantitative Evaluation

We use the center location error as well as the overlap rate for quantitative evaluations. Center location error is the per frame distance (in pixels) between the center of the tracking result and that of ground truth. Overlap rate is defined as $\frac{area(R_T \cap R_G)}{area(R_T \cup R_G)}$, where R_T is the bounding box of tracking result and R_G denotes

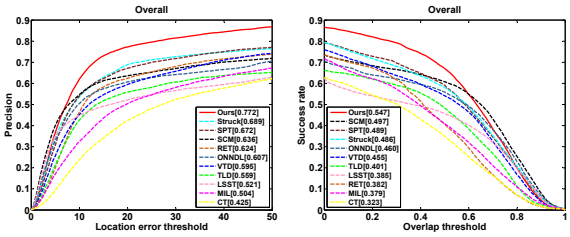


Fig. 2. Overall performance of the competing trackers on 51 video sequences. The precision plot and the success plot are used, and the performance score for each tracker is shown in the legend.

the ground truth. We employ precision plot and success plot [30] to evaluate the robustness of trackers, rather than directly using the average center location error and the average overlap rate over all frames of one video sequence to indicate the overall performance. The precision plot indicates the percentage of frames whose estimated location is within the given threshold distance of the ground truth, and the success plot shows the ratios of successful frames whose overlap rate is larger than the given threshold.

The overall performance of the competing trackers on the 51 sequences is illustrated by the precision plot and the success plot as shown in Fig. 2. For the precision plot, the results at error threshold of 20 pixels are used for ranking, while for the success plot we use area under curve (AUC) scores to summarize and rank the trackers.

We can observe from Fig. 2 that both our tracker and the SCM, SPT and Struck methods achieve good tracking performance. In the precision plot, our tracker performs 8.3% better than the Struck, 10% better than the SPT, and 13.6% better than the SCM. In the success plot, our tracker performs 5% better than the SCM, 5.8% better than the SPT, and 6.1% better than the Struck. We also observe that the SCM method provides higher precision and success rate when the error threshold is relatively small (*e.g.*, 5 pixels in the precision plot, and 80% in the success rate). It owes to the fact that the SCM method exploits both holistic and local representation approaches based on sparse coding to handle appearance variations.

We also utilize the attribute based performance analysis approach [30] to demonstrate the robustness of our tracker. The video sequences used in the benchmark are annotated with 11 attributes which can be considered as different factors that may affect the tracking performance. One sequence can be annotated with several attributes. By putting the sequences that share a common attribute into a subset, we can analyze the performance of trackers to handle a specific challenging condition. Fig. 3 illustrates the success plots of the competing trackers for these 11 attributes (arranged in ascending order of the number of video sequences in each subset), and the precision plots can be found in the *supplementary material*. As indicated in Fig. 3, our method provides the best tracking performance in 7 of the 11 video subsets and also performs well in the

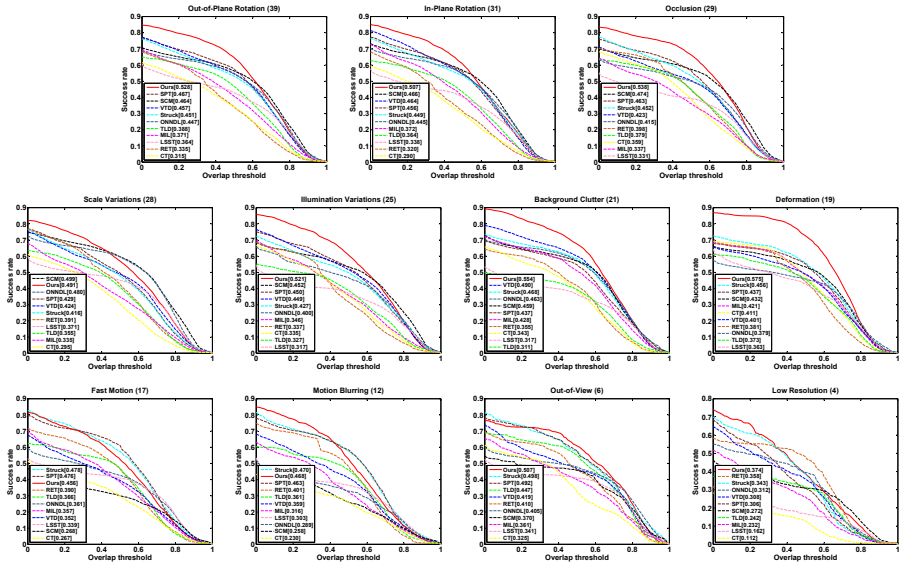


Fig. 3. Attribute based performance analysis using success plot. The number of video sequences in each subset is shown in the title. Best viewed on high-resolution display.

other 4 subsets, which demonstrates that the proposed algorithm is robust to appearance variations caused by a set of factors.

Overall, our tracker performs favorably against the state-of-the-art algorithms in terms of location accuracy and robustness. It can be attribute to the facts that LapRLS is effective for learning a robust classifier for object tracking, and that the proposed training example collection strategy which includes active example selection and the conservative labeling stage makes the classifier robust and adaptive to appearance changes. Our experimental results validate these claims in the following sections.

3.3 Diagnostic Analysis

As previously mentioned, our tracking method chooses the most informative examples for classifier learning via active example selection, leading to a significant improvement on tracking performance. In addition, we assign labels to part of the selected examples that satisfy strict constraints, which can increase the adaptivity of the classifier. To demonstrate the effectiveness of the active example selection approach and the conservative labeling strategy, we build three baseline algorithms to do validation and analyze various aspects of our method.

We begin with a "naive" tracker based on a classifier learned with LapRLS, denoted as BaseLine1. The BaseLine1 only exploits the labeled examples from the first frame, and collects unlabeled examples using random sampling. We add the active example selection stage after the sampling process to select informative examples for LapRLS, resulting in another baseline, denoted as BaseLine2.

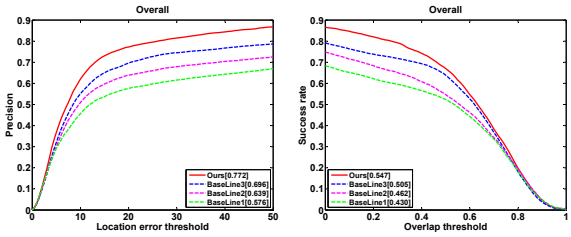


Fig. 4. Diagnostic Analysis. The overall performance of three baseline algorithms and our method on the 51 video sequences is presented for comparison in terms of precision and success rate.

Both the BaseLine1 and the BaseLine2 are stable versions, since no supervisory information is added during tracking, *i.e.*, the training examples are collected without the labeling stage. We get the last baseline by allowing the BaseLine1 to assign labels to part of the unlabeled examples that satisfy the strict constraints described in Sec. 2.3, denoted as BaseLine3. Note that adding supervisory information to the BaseLine2 leads to the proposed method.

The overall tracking performance of these baseline algorithms and our method is presented in Fig. 4. Surprisingly, even without additional example selection and labeling process, the BaseLine1 produces good performance in terms of precision and robustness, outperforming the CT, MIL, LSST and TLD trackers and being comparable to the VTD. It demonstrates the effectiveness of LapRLS which can sufficiently exploit unlabeled data and preserve the local geometrical structure of feature space. The performance of our method and the Baseline3 is obviously better than the BaseLine1 and the BaseLine2, which demonstrates that the additional supervisory information is significant for object tracking. The conservative labeling strategy used in our tracking method achieves a suitable trade-off between stability and plasticity in terms of capturing appearance variations. The performance of our method is significantly better than BaseLine3, and the BaseLine2 outperforms the BaseLine1. It validates the effectiveness of selecting informative examples for classifier learning. The active example selection guarantees the consistency between example collection and classifier learning, and thus improves the tracking performance. Furthermore, assigning labels to examples selected by active example selection alleviates the drift problem caused by label noise, since misaligned examples will be rejected to ensure the high prediction confidence of the classifier.

3.4 Qualitative Evaluation

We present a qualitative evaluation of the tracking results in this section. 12 representative sequences are chose from the subsets of four dominant attributes, *i.e.*, occlusion, illumination variations, background clutter and deformation. Several screenshots of the tracking results on these 12 sequences are illustrated in Fig. 5. We mainly discuss the four dominant challenges in the following.

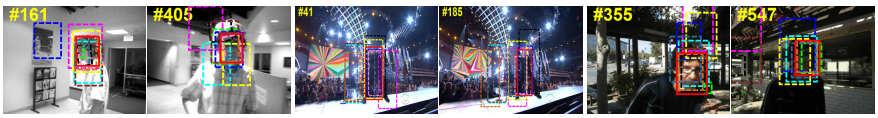
Occlusion is one of the most general yet crucial problems in object tracking, as shown in Fig. 5(a). In the *David3* sequence, the person suffers from partial occlusion as well as drastic pose variations (*e.g.*, #249). Only the LSST, RET, ONNDL and our method success in this sequence. In the *Jogging2* sequence, there is a short-term complete occlusion for the tracked object (*e.g.*, #60). The TLD, SCM, ONNDL and our method are able to reacquire the object and provide satisfactory tracks. Note that the TLD method employs a detector to reacquire the object and the SCM and ONNDL trackers involve occlusion resolving scheme based on sparse representation. In the *Woman* sequence, only the Struck, SPT and our method are able to track the object when the long-term occlusion happens (*e.g.*, #134). Most of the trackers lock onto a wrong object with similar appearances after occlusion. Our method selects informative examples for classifier learning via active example selection, and thus alleviates the drift problem caused by misaligned examples in handling occlusions.

The tracked objects in the *David1*, *Singer2* and *Trellis* sequences undergo significant illumination changes and pose variations, as shown in Fig. 5(b). Most of the trackers can not handle the appearance variations caused by illumination changes together with pose variations (*e.g.*, *David1* #161, *Singer2* #185 and *Trellis* #355), whereas the VTD and SCM methods perform better. In contrast, our method achieve stable performance in the entire sequences. In the *Singer2* sequence, the contrast between the foreground and the background is very low. Our method tracks the object accurately, but most trackers drift away at the beginning of the sequence (*e.g.*, #41). The robustness of our tracker against illumination variations comes from the fact that the adopted HOG feature has been proved to be invariant to illumination changes.

In the *Football*, *Lemming* and *Subway* sequences, the objects appear in background clutters, as shown in Fig. 5(c). Most trackers drift away from the objects as there exists the interference of similar appearances in the background (*e.g.*, *Football* #312, *Lemming* #545, *Subway* #46). Our method learns an online classifier that takes the background information into account, and thus can achieve robust performance under complex environments.

In the *Basketball*, *Bolt* and *Skating1* sequences, the object appearances change drastically due to significant non-rigid object deformation, such as viewpoint changes and pose variations, as shown in Fig. 5(d). We can see that only our method tracks the objects successfully in all these three sequences. In the *Basketball* sequence, the person changes his pose frequently and often partially occluded by other players. Only the VTD and our method can keep track all the time. In the *Bolt* sequences, there exist significant pose variations of the person, together with the viewpoint change. The trackers except the ONNDL and our method fail when the viewpoint start to change (*e.g.*, #107). In the *Skating1* sequence, all of the methods except our tracker gradually drift away when there is severe occlusion and large scale change of the object (*e.g.*, #178). We show that our method adaptively copes with appearance variations through online update with the selected informative examples, thus provides more accurate and consistent tracking results.

(a) Tracking results on sequence *David3*, *Jogging2* and *Woman* with occlusions.



(b) Tracking results on sequence *David*, *Singer2* and *Trellis* with illumination changes.

(c) Tracking results on sequence *Football*, *Lemming* and *Subway* with clustered background.

(d) Tracking results on sequence *Basketball*, *Bolt* and *Skating1* with object deformation.

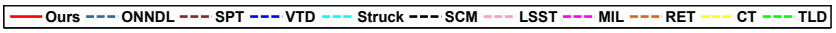


Fig. 5. Sample tracking results of the competing trackers on 12 representative video sequences.

4 Conclusion

In this paper, we have presented a novel online object tracking algorithm that explicitly couples the objectives of semi-supervised learning and example selection in a principled manner. We have shown that selecting informative examples for classifier learning results in more robust tracking, and have proposed an active example selection approach using the formulism of active learning. We have also shown that assigning labels to part of the selected examples achieves a suitable trade-off between stability and plasticity in terms of capturing appearance variations. Both quantitative and qualitative evaluations compared with state-of-the-art trackers on a comprehensive benchmark demonstrate the effectiveness and robustness of our tracker.

Acknowledgement. This work was supported in part by the Natural Science Foundation of China (NSFC) under grant NO. 61203291, the 973 Program of China under grant NO. 2012CB720000, the Specialized Research Fund for the Doctoral Program of Higher Education of China (20121101120029), and the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission.

References

1. Ross, D., Lim, J., Lin, R., Yang, M.H.: Incremental learning for robust visual tracking. International Journal of Computer Vision **77** (2008) 125–141
2. Kwon, J., Lee, K.: Visual tracking decomposition. In: CVPR. (2010) 1269–1276
3. Mei, X., Ling, H.: Robust visual tracking using $\ell 1$ minimization. In: ICCV. (2009) 1–8
4. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: CVPR. (2012) 1838–1845
5. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: CVPR. (2012) 1822–1829
6. Wang, N., Wang, J., Yeung, D.Y.: Online robust non-negative dictionary learning for visual tracking. In: ICCV. (2013) 657–664
7. Wu, Y., Ma, B., Yang, M., Zhang, J., Jia, Y.: Metric learning based structural appearance model for robust visual tracking. IEEE Transactions on Circuits and Systems for Video Technology **24** (2014) 865–877
8. Wang, D., Lu, H., Yang, M.H.: Least soft-thresold squares tracking. In: CVPR. (2013) 2371–2378
9. Hare, S., Saffari, A., Torr, P.H.: Struck: Structured output tracking with kernels. In: ICCV. (2011) 263–270
10. Zhang, K., Zhang, L., Yang, M.H.: Real-time compressive tracking. In: ECCV. (2012) 864–877
11. Li, X., Shen, C., Dick, A.R., van den Hengel, A.: Learning compact binary codes for visual tracking. In: CVPR. (2013) 2419–2426
12. Yao, R., Shi, Q., Shen, C., Zhang, Y., van den Hengel, A.: Part-based visual tracking with online latent structural learning. In: CVPR. (2013) 2363–2370
13. Bai, Q., Wu, Z., Sclaroff, S., Betke, M., Monnier, C.: Randomized ensemble tracking. In: ICCV. (2013) 2040–2047
14. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: ECCV. (2008) 234–247
15. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. & Mach Intell. **33** (2011) 1619–1632
16. Saffari, A., Leistner, C., Godec, M., Bischof, H.: Robust multi-view boosting with priors. In: ECCV. (2010) 776–789
17. Bai, Y., Tang, M.: Robust tracking via weakly supervised ranking svm. In: CVPR. (2012) 1854–1861
18. Gao, J., Xing, J., Hu, W., Maybank, S.: Discriminant tracking using tensor representation with semi-supervised improvement. In: ICCV. (2013)
19. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N learning: Bootstrapping binary classifiers by structural constraints. In: CVPR. (2010) 49–56
20. Supancic III, J.S., Ramanan, D.: Self-paced learning for long-term tracking. In: CVPR. (2013) 2379–2386
21. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research **7** (2006) 2399–2434
22. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of Artificial Intelligence Research **4** (1996) 129–145
23. Atkinson, A.C., Donev, A.N.: Optimum experimental designs. Oxford University Press (2002)

24. Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H.: PROST: Parallel robust online simple tracking. In: CVPR. (2010) 723–730
25. Yu, K., Bi, J., Tresp, V.: Active learning via transductive experimental design. In: ICML. (2006) 1081–1088
26. He, X., Min, W., Cai, D., Zhou, K.: Laplacian optimal design for image retrieval. In: ACM SIGIR. (2007) 119–126
27. He, X.: Laplacian regularized d-optimal design for active learning and its application to image retrieval. IEEE Trans. Image Processing **19** (2010) 254–263
28. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: NIPS. (2004) 1601–1608
29. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. International Journal of Computer Vision **29** (1998) 5–28
30. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR. (2013) 2411–2418
31. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005) 886–893