Graph Embedding Based Semi-Supervised Discriminative Tracker

Jin Gao[1], Junliang Xing[1], Weiming Hu[1], and Xiaoqin Zhang[2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, China
[2]Institute of Intelligent System and Decision, Wenzhou University, China
{jgao10, jlxing, wmhu}@nlpr.ia.ac.cn, zhangxiaoqinnan@gmail.com

Abstract

Recently, constructing a good graph to represent data structures is widely used in machine learning based applications. Some existing trackers have adopted graph construction based classifiers for tracking. However, their graph structures are not effective to characterize the inter-class separability and multi-model sample distribution, both of which are very important to successful tracking. In this paper, we propose to use a new graph structure to improve tracking performance without the assistance of learning object subspace generatively as previous work did. Meanwhile, considering the test samples deviate from the distribution of the training samples in tracking applications, we formulate the discriminative learning process, to avoid overfitting, in a semi-supervised fashion as ℓ_1-graph based regularizer. In addition, a non-linear variant is extended to adapt to multi-modal sample distribution. Experimental results demonstrate the superior properties of the proposed tracker.

1. Introduction

Visual object tracking is an essential component of several practical vision applications such as automated surveillance, vehicle navigation, traffic monitoring, and so on. Generally speaking, a typical tracking system consists of three components: 1) an adaptive appearance model, which can evaluate the likelihoods of the candidate object regions; 2) a motion model, which relates the locations of the candidates over time; and 3) a search strategy for finding the most likely location in the current frame. We refer the readers to [22] for a thorough review of these components. Note that this paper is focused on dealing with the first component.

Current tracking algorithms use two typical techniques to learn the varieties of appearance models, *i.e*., either generative or discriminative approaches. Generative methods mainly concentrates on how to construct robust object representation in specified feature spaces, such as gray-scale image vector subspace learning [18], log-Euclidean Riemannian subspace learning [16, 9], multiple patch votes with each patch represented by gray-scale histogram features [1] or multi-cue integration [6], sparse principal component analysis (SPCA) of a set of feature templates (*e.g*. hue, saturation, intensity, and edge) [11, 12], and so on. These methods have achieved great success in the tracking literature, however they suffer a problem that the information for classification in the background is discarded. Discriminative methods formulate visual tracking as a binary classification problem to separate the object from the background, such as graph embedding based classifiers [26, 14], feature selection based boosting classifiers [2, 7, 8, 13], graph mode-based SVM classifier [15], and so on. Among them, one kind of promising methods is based on graph embedding [26, 14], which capture the underlying geometry of collected object/background samples based on the manifold assumption. This novel trial has demonstrated good performance of graph embedding in tracking applications thanks to the introduction of local structure preservation property, however more aspects need to be fully exploited as follows.

(a) *david* with out-of-plane rotation

(b) *trellis* with dramatic illumination change and out-of-plane rotation

Figure 1: Collected object samples in the *david* and *trellis* sequences.

Motivations. Zhang *et al*. [26] use a PCA graph and a local-geometry-preserved graph to construct an object graph and a background graph respectively. Meanwhile, they use inter-class margin [4, 24] to characterize the separability of different classes. All of the above result in a discriminative subspace. They also use incremental PCA subspace as the object subspace to assist the discriminative subspace. A combination of generative and discriminative

models based on the object subspace and the discriminative subspace respectively forms the final appearance model. There are several shortcomings existing in this work. First, due to the variations of collected object samples (see Fig. 1), it is not appropriate to approximate the object sample distribution by a Gaussian and construct a PCA graph as the object graph for discriminative learning. Second, it is not appropriate to assume that maximization of the inter-class margin only based on the marginal samples is equivalent to maximization of the inter-class separability, especially when the marginal samples are selected by the kNN (k nearest neighbors) method in a high-dimensional manifold. This selection process is easily corrupted by noise and outliers. That is why Zhang's work needs assistance of learning object subspace generatively to track objects accurately. However, when the object undergoes severe appearance variations, the generatively learned object subspace can not capture these variations, and the straightforward combination of generative and discriminative models may degrade the superiority of discriminative learning. We plug Zhang's graph into our system and make a direct comparison in Section 3.1. Third, since the tracking environment severely varies from frame-to-frame, the test samples collected from the current frame deviate from the distribution of the training samples. Semi-supervised learning can avoid overfitting under such circumstances while Zhang's work does't take it into account. Last but not least, the graph embedding with each vertex represented by image-as-vector (a 32×32 image patch results in a 1024-dimensional vector) is faced with the curse of dimensionality problem, which is an ill-conditioned problem when the dimension of the original data space is larger than the number of the training samples. Li *et al*. [14] inherit Zhang's work and introduce a novel Volterra kernel for non-linear embedding, however they do not pay more attention to the aforementioned shortcomings.

Several years have witnessed great developments in graph construction [4, 20, 24, 3, 10, 23, 5, 21, 27], especially in face recognition and image classification applications. Experimental results in these applications demonstrate that graph embedding based machine learning applications have two key premises: i) constructing a good graph to represent data structures; ii) making use of the unlabeled samples appropriately. Inspired by above insights, we propose a new graph embedding based semi-supervised discriminative tracker (GSDT). The main contribution is three-fold. First, in order to get rid of the assistance of learning object subspace generatively, we add an edge between each pair of samples from different classes to characterize the separability of different classes instead of only between the marginal samples. It is robust to noise and outlier corruptions. Second, considering that the test samples deviate from the distribution of the training samples in the real-world complex tracking environment, we formulate the discriminative learning process, to avoid overfitting, in a novel semi-supervised fashion as ℓ_1-graph [23] based regularizer. It imposes a cluster assumption based regularizer to our graph embedding framework. Meanwhile, the ℓ_1-graph explores higher order relationships among more data points, and hence is more powerful to model the neighborhood relationship than the custom kNN method. Third, a non-linear variant of our semi-supervised discriminant learning method is extended to adapt to multi-modal sample distribution. We also introduce the well-known covariance matrix descriptor measured under log-Euclidean Riemannian metric for feature extraction to reduce feature dimension and avoid encountering the curse of dimensionality problem.

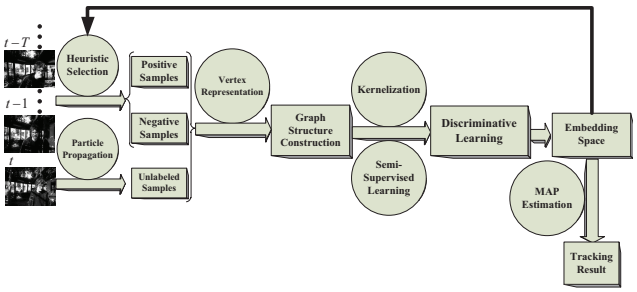


Figure 2: The architecture of the tracking framework.

Semi-supervised learning has recently been introduced for tracking in [8] and later extended by "Covariate Shift" in [13]. These two studies both add cluster assumption based loss function terms to the original feature selection based boosting classifiers using a "SemiBoost" technique. Li *et al*. [15] construct an adjacency graph using all the samples to capture the useful contextual information, and hence develop a new contextual kernel for SVM tracking. The similarity measures in these work only characterize the pairwise relationships of samples, however do not explore higher order relationships among all the samples. The covariance matrix descriptor measured under log-Euclidean Riemannian metric has widely used for tracking (*e.g*. [16, 9, 14]). However, they only use it for constructing generative learning based appearance models, and hence do not explore its discriminant capability when used in the discriminative learning based appearance model. A direct comparison of results between the methods of [8, 9] and ours is given in Section 3.2.

2. The Proposed Approach

Fig. 2 is an overview of the proposed GSDT. For better illustration, we elaborate the important components of the proposed approach in this section, mainly including feature extraction for vertex representation, graph structure construction, and the formulations of semi-supervised fashion and non-linear extension.

2.1. Feature Extraction

By applying affine transformations, we crop the candidate object regions from the current frame (gray-scale im-

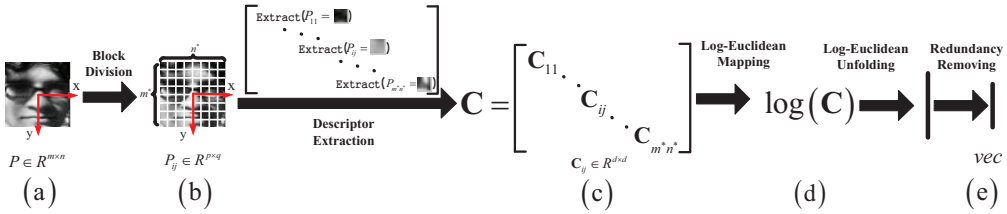


Figure 3: Representation of the graph vertex. A sample image patch P is shown in (a). The results of appearance block division are exhibited in (b), where m^* is denoted as $[m/p]$, n^* is denoted as $[n/q]$, and $[\cdot]$ is denoted as the rounding operator. The block-division based covariance matrix descriptor $\mathbf{C}$ is shown in (c). The Log-Euclidean mapping and unfolding are displayed in (d) and (e).

age) and normalize each of them to a patch of size $m \times n$, which is then divided into several blocks, each of size $p \times q$, as illustrated in Fig. 3(a) and (b). For each patch P, its blocks are denoted by $P_{ij} \in R^{p \times q}$, and the covariance matrix is extracted for representing P_{ij} as follows:

$$\mathbf{C}_{ij} = \frac{1}{L-1} \sum_{i=1}^{L} (f_i - \mu)(f_i - \mu)^T \tag{1}$$

where μ is the mean of $\{f_i\}_{i=1,\ldots,L}$, L is the number of pixels in the block P_{ij}, and f_i is the d-dimensional image feature vector defined at the pixel coordinate (x, y) (referred to [17] for more details). Block-division based appearance representation has been used in [9]. However, they do log-Euclidean mapping and construct appearance model for each block individually, and combine them to a holistic appearance model using novel filtering. In this paper, we modify the block-division based appearance representation, and develop a holistic block-division based covariance matrix descriptor $\mathbf{C}$ as illustrated in Fig. 3(c), which is used to represent patch P. By log-Euclidean mapping under log-Euclidean Riemannian metric, as illustrated in Fig. 3(d), the descriptor $\mathbf{C}$ is converted into a new one $\log(\mathbf{C})$. Due to the vector space structure of $\log(\mathbf{C})$ (referred to [16] for more details), it can be unfolded into a vector to represent the graph vertex. Because $\log(\mathbf{C})$ is also a symmetric matrix and has many zero entries, the redundancy entries of the unfolded vector should be removed (see Fig. 3(e)). This representation method can reduce the dimension of the graph vertex representation effectively, and hence avoid encountering the curse of dimensionality problem while more information is retained. For a patch of size 32×32 ($m = n = 32$), we divide it into 16 blocks and each block has 8×8 pixels ($p = q = 8$), resulting in a 336-dimensional vector when $d = 6$, which is much lower than the original 1024-dimensional vector.

2.2. Graph Structure for Embedding

In real-world scenarios, objects undergo many kinds of appearance changes even within a short period of time, the distributions of the object and background samples are both multi-modal (not Gaussian). We construct two new graphs specially designed to model the local geometrical and discriminative structure of the training samples according to the graph embedding framework [24, 20]. The intrinsic graph consists of an object graph and a background graph, each of which has local-geometry-preserved property. The penalty graph is constructed by adding an edge between each pair of samples from different classes to characterize the separability of different classes instead of only between the marginal samples, so that it can be robust to noise.

Let $\mathbf{x}_i \in \mathbb{R}^D (i = 1, 2, \ldots, l)$ be D-dimensional vectors to represent the graph vertices corresponding to the labeled samples, and $y_i \in \{1, 2, \ldots, C\}$ be associated class labels, where l is the number of the labeled samples and C is the number of classes. In this paper, $y_i = 1$ indicates the object, $y_i = 2$ indicates the background, and $C = 2$. Let n_c be the number of samples in the class c: $\sum_{c=1}^{C} n_c = l$. Using the information about the labeled samples, we aim to find a discriminative embedding space and map $\mathbf{X} \equiv (\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_l) \in \mathbb{R}^{D \times l} \mapsto \mathbf{Z} \equiv (\mathbf{z}_1|\mathbf{z}_2|\cdots|\mathbf{z}_l) \in \mathbb{R}^{R \times l}$, where $R < D$ and is the dimension of the embedding space, such that in the embedding space unlabeled samples are more reliably to be labeled by the nearest neighbor rule, owing to the locally discriminative nature. To achieve this goal, we need to construct two graphs: an intrinsic graph $G = \{\mathbf{X}, \mathbf{W}\}$ and a penalty graph $G^p = \{\mathbf{X}, \mathbf{W}^p\}$ where $\mathbf{W}$ and $\mathbf{W}^p$ are edge weight matrices, and minimize the *graph-preserving criterion* as follows:

$$\begin{aligned} \mathbf{Z}^* &= \underset{\mathrm{tr}(\mathbf{Z}\mathbf{L}^p\mathbf{Z}^T)=R}{\operatorname{argmin}} \sum_{i,j} \|\mathbf{z}_i - \mathbf{z}_j\|^2 w_{ij} \\ &= \underset{\mathrm{tr}(\mathbf{Z}\mathbf{L}^p\mathbf{Z}^T)=R}{\operatorname{argmin}} 2\,\mathrm{tr}\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^T\right) \end{aligned} \tag{2}$$

where the Laplacian matrices $\mathbf{L}$ and $\mathbf{L}^p$ of G and G^p are defined by the diagonal matrices $\mathbf{D}$ and $\mathbf{D}^p$ as follows:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \quad D_{ii} = \sum_{j \neq i} w_{ij}, \quad \forall i \tag{3}$$

$$\mathbf{L}^p = \mathbf{D}^p - \mathbf{W}^p, \quad D^p_{ii} = \sum_{j \neq i} w^p_{ij}, \quad \forall i\,. \tag{4}$$

Construct the intrinsic graph $G = \{\mathbf{X}, \mathbf{W}\}$. Some edges are added between some vertex pairs in G to characterize the similarity relationships between them. Each element w_{ij} of the edge weight matrix $\mathbf{W}$ refers to the weight of the edge between one vertex pair:

$$w_{ij} = \begin{cases} A_{ij}/n_1, & \text{if } y_i = y_j = 1, \\ A_{ij}/n_2, & \text{if } y_i = y_j = 2, \\ 0, & \text{if } y_i \neq y_j, \end{cases} \tag{5}$$

where $n_1 + n_2 = l$, and the affinity A_{ij} is defined by the local scaling method in [25]. Without loss of generality, we assume that the data points in $\{\mathbf{x}_i\}_{i=1}^{l}$ are ordered according to their labels $y_i \in \{1, 2\}$. When $y_i = y_j = 1$,

$$A_{ij} = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/(\sigma_i\sigma_j)\right), \tag{6}$$

where $\sigma_i = \left\|\mathbf{x}_i - \mathbf{x}_i^{(k)}\right\|$, and $\mathbf{x}_i^{(k)}$ is the kth nearest neighbor of $\mathbf{x}_i$ among $\{\mathbf{x}_j\}_{j=1}^{n_1}$. When $y_i = y_j = 2$,

$$A_{ij}=\begin{cases} \exp(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\sigma_i\sigma_j}), \text{if } i\in N_k^+(j) \text{ or } j\in N_k^+(i), \\ 0, \text{otherwise}, \end{cases} \tag{7}$$

where $N_k^+(i)$ indicates the index set of the k nearest neighbors of the vertex $\mathbf{x}_i$ among $\{\mathbf{x}_j\}_{j=n_1+1}^{l}$, $\sigma_i = \left\|\mathbf{x}_i - \mathbf{x}_i^{(k)}\right\|$, and $\mathbf{x}_i^{(k)}$ is the kth nearest neighbor of $\mathbf{x}_i$ among $\{\mathbf{x}_j\}_{j=n_1+1}^{l}$. The parameter k above is empirically chosen as 7 based on [25].

Construct the penalty graph $G^p = \{\mathbf{X}, \mathbf{W}^p\}$. In G^p, each element w_{ij}^p of $\mathbf{W}^p$ is defined as follows:

$$w_{ij}^p = \begin{cases} A_{ij}\left(1/l - 1/n_1\right), & \text{if } y_i = y_j = 1, \\ A_{ij}\left(1/l - 1/n_2\right), & \text{if } y_i = y_j = 2, \\ 1/l, & \text{if } y_i \neq y_j, \end{cases} \tag{8}$$

where A_{ij} has the same definition as in G.

Linear discriminative learning. Assuming that the low-dimensional vector representations of the vertices can be obtained from a linear projection as $\mathbf{Z} = \mathbf{P}^T\mathbf{X}$, where $\mathbf{P}$ is a $D \times R$ transformation matrix, the objective function in Eq. (2) becomes

$$\begin{aligned} \mathbf{P}^* &= \underset{\mathrm{tr}(\mathbf{P}^T\mathbf{X}\mathbf{L}^p\mathbf{X}^T\mathbf{P})=R}{\operatorname{argmin}} 2\,\mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}\right) \\ &= \underset{\mathbf{P}\in R^{D\times R}}{\operatorname{argmin}} \mathrm{tr}\left(\frac{\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}}{\mathbf{P}^T\mathbf{X}\mathbf{L}^p\mathbf{X}^T\mathbf{P}}\right), \end{aligned} \tag{9}$$

where the analytic form of $\mathbf{P}^*$ is obtained by solving a generalized eigenvalue problem as follows:

$$\mathbf{P}^T\mathbf{X}\mathbf{L}^p\mathbf{X}^T\mathbf{P}\varphi = \lambda\mathbf{P}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{P}\varphi\,. \tag{10}$$

Denoting R principal generalized eigenvectors corresponding to the R largest eigenvalues of Eq. (10) as $\{\varphi_r\}_{r=1}^{R}$, we obtain the discriminative projection $\mathbf{P}^* = (\varphi_1|\varphi_2|\cdots|\varphi_R)$. An efficient implementation of this process is proposed in [20].

Discussion. In the intrinsic graph G, we construct the background graph as same as in Zhang's work [26], while construct object graph also with local geometry preserved due to the variations of the collected object samples (see Fig. 1). In the penalty graph G^p, we add an edge between each pair of samples from different classes ($w_{ij}^p = 1/l$, if $y_i \neq y_j$) to characterize the separability of different classes. This can make the samples from different classes apart even when they are corrupted by noise and outliers, and avoid kNN selection (easily corrupted) of the marginal samples in a high-dimensional manifold. Actually, our proposed graph structure is reduced to the graph structure of linear discriminant analysis (LDA) when all the affinities are set to 1.

2.3. ℓ_1-Graph based Semi-Supervised Regularizer

Considering that the unlabeled candidates deviate from the distribution of the labeled samples when the appearances of the object and the background undergo severe changes, we formulate the discriminative learning process, to avoid overfitting, in a novel semi-supervised fashion as ℓ_1-graph [23] based regularizer as follows.

Let $\{\mathbf{x}_i\}_{i=l+1}^{l+u}$ be D-dimensional vectors to represent the graph vertices corresponding to u unlabeled samples. Denote $\mathbf{X}^* \in \mathbb{R}^{D\times(l+u)}$ as a matrix of all the input samples $(\mathbf{x}_1|\mathbf{x}_2|\cdots|\mathbf{x}_{l+u})$, and $\mathbf{Z}^* \in \mathbb{R}^{R\times(l+u)}$ as a matrix of all the embeddings $(\mathbf{z}_1|\mathbf{z}_2|\cdots|\mathbf{z}_{l+u})$. We define $n = l + u$. In the cluster assumption, the "similar" samples may have nearby embeddings (low-dimensional representations). This assumption allows the unlabeled samples to regularize the decision boundary. A popular way to define the inconsistency between the embeddings $\{\mathbf{z}_i\}_{i=1}^{n}$ of the samples $\{\mathbf{x}_i\}_{i=1}^{n}$ is the *quadratic criterion*:

$$\mathcal{F}(\mathbf{Z}^*) = \frac{1}{2}\sum_{i,j=1}^{n} S_{ij}\left(\mathbf{z}_i - \mathbf{z}_j\right)^2 = \mathrm{tr}(\mathbf{Z}^*\mathbf{L}^r\mathbf{Z}^{*T})\,, \tag{11}$$

where S_{ij} is the pairwise similarity, and the graph Laplacian matrix $\mathbf{L}^r$ is defined by the diagonal matrix $\mathbf{D}^r$ as follows:

$$\mathbf{L}^r = \mathbf{D}^r - \mathbf{S}, \quad D_{ii}^r = \sum_{j\neq i} S_{ij}, \quad \forall i\,. \tag{12}$$

We define $\mathbf{S}$ by adding an edge between samples $\mathbf{x}_i$ and $\mathbf{x}_j$ if they are "similar". This concept "similar" is defined based on the ℓ_1 directed graph construction process [23]. If a directed edge is placed from node i to j ($a_i^j \neq 0$), or from j to i ($a_j^i \neq 0$), we assume these two samples are "similar". Specifically,

$$S_{ij} = \begin{cases} 1/n, & \text{if } a_i^j \neq 0 \text{ or } a_j^i \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

In general, the optimal $\mathbf{P}^*$ in Eq. (9) should also minimize $\mathcal{F}(\mathbf{Z}^*) = \mathrm{tr}(\mathbf{Z}^*\mathbf{L}^r\mathbf{Z}^{*T})$. Thus, a natural regularizer can be defined as follows:

$$\mathcal{J}(\mathbf{P}) = \mathrm{tr}(\mathbf{Z}^*\mathbf{L}^r\mathbf{Z}^{*T}) = \mathrm{tr}(\mathbf{P}^T\mathbf{X}^*\mathbf{L}^r\mathbf{X}^{*T}\mathbf{P}). \tag{14}$$

With this ℓ_1-graph based regularizer, we get the objective function of our semi-supervised formulation of Eq. (9) as

$$\mathbf{P}^* = \underset{\mathbf{P}\in R^{D\times R}}{\operatorname{argmin}} \mathrm{tr}\left(\frac{\mathbf{P}^T\left(\mathbf{X}\mathbf{L}\mathbf{X}^T + \beta\mathbf{X}^*\mathbf{L}^r\mathbf{X}^{*T}\right)\mathbf{P}}{\mathbf{P}^T\mathbf{X}\mathbf{L}^p\mathbf{X}^T\mathbf{P}}\right) \tag{15}$$

where $\beta \geq 0$ is a trade-off parameter. The regularizer imposes a cluster assumption based regularizer to our original graph embedding framework. Meanwhile, the ℓ_1-graph explores higher order relationships among all the samples, and hence is more powerful to model the neighborhood relationship than the custom kNN method.

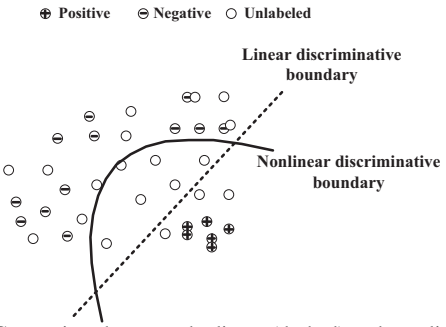


Figure 4: Comparison between the linear (dashed) and non-linear (solid) discriminative boundary.

2.4. Non-Linear Extension

Recall that the sample distributions of the object and background samples are often multi-modal due to the drastic appearance and background changes. In order to find a more discriminative embedding space, we need to adopt a non-linear projection, because non-linear discriminative boundary tends to provide a more reasonable solution space than linear one, as illustrated in Fig. 4.

We present the non-linear extension of our semi-supervised discriminative learning method using the *kernel trick* under a graph view [19]. Let $\phi : \mathbf{x} \mapsto \mathcal{H}$ be a function mapping the points in the input space to a high-dimensional Hilbert space. For a proper chosen ϕ, we replace the explicit mapping with the inner product $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. In this paper, we use the Gaussian kernel to define this product: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2\right)$. For convenience, we rewrite the vertex matrix in the Hilbert space as $\mathbf{X}^{\phi} \equiv (\phi(\mathbf{x}_1)|\phi(\mathbf{x}_2)|\cdots|\phi(\mathbf{x}_l))$ and $\mathbf{X}^{\phi*} \equiv (\phi(\mathbf{x}_1)|\cdots|\phi(\mathbf{x}_n))$. According to Representer Theorem, the optimal $\mathbf{P}^*$ of Eq. (15) in the Hilbert space is given by

$$p_j^{\phi*} = \sum_{i=1}^{n} \alpha_{ij}^* \phi(\mathbf{x}_i), \quad j = 1, 2, \cdots, R \tag{16}$$

where α_{ij}^* is the weight that defines how $p_j^{\phi*}$ is represented in the space spanned by a set of over-complete bases $\{\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \cdots, \phi(\mathbf{x}_n)\}$. We rewrite Eq. (15) in the Hilbert space as follows:

$$\mathbf{P}^{\phi*} = \underset{\mathbf{P}^{\phi} = \mathbf{X}^{\phi*}\boldsymbol{\alpha}}{\operatorname{argmin}} \operatorname{tr} \left(\frac{\mathbf{P}^{\phi T} \left(\mathbf{X}^{\phi} \mathbf{L} \mathbf{X}^{\phi T} + \beta \mathbf{X}^{\phi*} \mathbf{L}^r \mathbf{X}^{\phi*T} \right) \mathbf{P}^{\phi}}{\mathbf{P}^{\phi T} \mathbf{X}^{\phi} \mathbf{L}^p \mathbf{X}^{\phi T} \mathbf{P}^{\phi}} \right) \tag{17}$$

where $\mathbf{L}$ and $\mathbf{L}^p$ are calculated in the Hilbert space. When $\mathbf{K}^* = \mathbf{X}^{\phi*T}\mathbf{X}^{\phi*}$ and $\mathbf{K} = \mathbf{X}^{\phi T}\mathbf{X}^{\phi*}$ are calculated, we can further rewrite Eq. (17) as follows:

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \operatorname{tr} \left(\frac{\boldsymbol{\alpha}^T \left(\mathbf{K}^T \mathbf{L} \mathbf{K} + \beta \mathbf{K}^* \mathbf{L}^r \mathbf{K}^* \right) \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \mathbf{K}^T \mathbf{L}^p \mathbf{K} \boldsymbol{\alpha}} \right), \tag{18}$$

and then the optimal solution can be obtained as $\mathbf{P}^{\phi*} = \mathbf{X}^{\phi*}\boldsymbol{\alpha}^*$. A data point in the Hilbert space can be embedded into a R-dimensional subspace by: $\phi(\mathbf{x}) \mapsto \mathbf{z} = \mathbf{P}^{\phi*T}\phi(\mathbf{x}) = \boldsymbol{\alpha}^{*T}\mathbf{X}^{\phi*T}\phi(\mathbf{x}) = \boldsymbol{\alpha}^{*T}K(:, \mathbf{x})$, where $K(:, \mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x})|\cdots|K(\mathbf{x}_n, \mathbf{x}))^T$.

2.5. Heuristic Selection of Training Samples

In our proposed tracker, we take a heuristic strategy for training sample selection from the propagated candidates generated from the particle filter (see [18]). First, we assign a reliable confidence to each graph vertex $\mathbf{x}_i$ corresponding to each propagated candidate, which reflects the probability that the candidate belongs to the object. Its confidence can be defined as follows:

$$p(\mathbf{x}_i | \mathbf{z}^+, \boldsymbol{\alpha}^*) \propto \exp\left(-\|\mathbf{z}^+ - \boldsymbol{\alpha}^{*T} K(:, \mathbf{x}_i)\|\right) \tag{19}$$

where $\mathbf{z}^+$ represents the center of the graph vertices corresponding to the positive samples in the embedding space. Hence, we can determine the optimal object region from the candidate regions by the MAP (maximum a posterior) estimation in the Bayesian inference framework, where the observation model is defined by Eq. (19). Second, we make a descending sort for the candidates according to Eq. (19), resulting in a sorted vertex set. By selecting the top one of the vertices from it, we add it to a positive buffer set $\mathbb{N}^+$ with buffer size set to $\mathbb{T}^+$; by selecting the bottom $1/3$ of the vertices from it, we add them to a negative buffer set $\mathbb{N}^-$ with buffer size set to $\mathbb{T}^-$.

3. Experimental Results and Analysis

In this section, we present experimental results that validate the superior properties of our new graph embedding based discriminative tracker with the ℓ_1-graph based semi-supervised regularizer (GSDT). First, to demonstrate the effectiveness of the proposed tracking approach, we evaluate each component of our tracker individually, such as the new graph structure, the ℓ_1-graph based semi-supervised regularizer, and non-linear extension. Second, we compare our tracker with six trackers on various test videos and prove that our tracker tracks objects robustly and accurately.

Implementation details. All our experiments are done using MATLAB R2008b on a 2.83GHz Intel Core2 Quad PC with 4GB RAM. As shown in Fig. 2, we use the particle filter to draw unlabeled samples from frame I_t, and set the number of particles to 300 ($u = 300$). The parameters $\mathbb{T}^+$ and $\mathbb{T}^-$ are both set to 50 and 300 respectively, indicating that $n_1 = 50$, $n_2 = 300$, $l = 350$, and $n = 650$. For feature vector f_i in Eq. (1), we only consider the coordinate (x, y), the intensity value $I(x, y)$, the first order intensity derivatives $I_x(x, y)$ and $I_y(x, y)$, and $\sqrt{(I_x(x,y))^2 + (I_y(x,y))^2}$, resulting in a 6-dimensional feature vector. All the cropped image patches are normalized to size 32×32 ($m = n = 32$), and $p = q = 8$ in Fig. 3. Sample representation used in the ℓ_1 directed graph construction process [23] is based on the image-as-vector representation method, where each sample corresponds to a

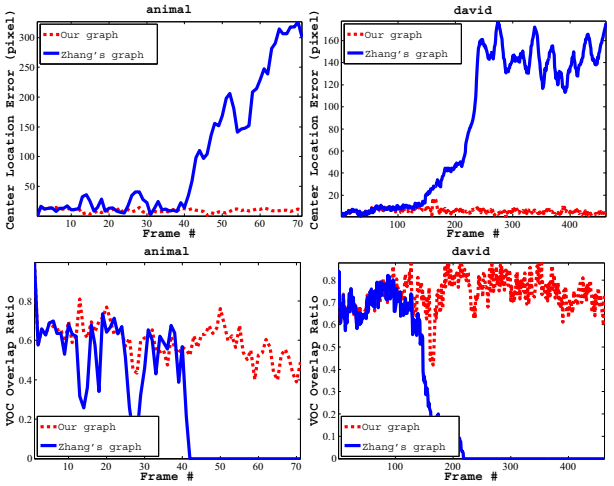


Figure 5: Quantitative comparison of the proposed tracker with our new graph and Zhang's graph [26] on two videos. The left two subfigures are associated with the tracking performance in CLE and VOR on the *animal* video, respectively; the right two subfigures correspond to the tracking performance in CLE and VOR on the *david* video, respectively.

normalized 10×10 image template. The dimension of the embedding space is empirically set to $R = 1$, because the largest eigenvalue is more than ten times larger than the second largest one. The trade-off parameter β is set to 0.5. In Section 2.4, σ in the Gaussian kernel is empirically set to 7. The above settings remain the same in all the experiments.

3.1. The Effectiveness of Our Tracker

To evaluate each component of our tracker GSDT individually, such as the new graph structure, the ℓ_1-graph based semi-supervised regularizer, and non-linear extension, we conduct a set of experiments on four challenging video sequences with only $2D$ translation and scale tracked. For quantitative comparison, two evaluation criteria are introduced, namely, the center location error (CLE) and the VOC overlap ratio (VOR) between the predicted bounding box B_p and the ground truth B_{gt} such that $\text{VOR} = \frac{\text{area}(B_p \bigcap B_{gt})}{\text{area}(B_p \bigcup B_{gt})}$. If VOR is larger than 0.5, then the target is considered to be successfully tracked.

Table 1: Comparison of the proposed tracker with/without semi-supervised learning. ACLE: the average CLE; AVOR: the average VOR; TSR: the tracking success rate.

	david			skating1			trellis*		
	ACLE	*AVOR*	*TSR*	*ACLE*	*AVOR*	*TSR*	*ACLE*	*AVOR*	*TSR*
Semi	3.6	0.74	0.99	6.7	0.67	0.95	4.8	0.72	1.00
non-Semi	6.5	0.50	0.40	36.8	0.41	0.47	20.3	0.57	0.75

Evaluation of our graph structure. To validate the fact that our proposed new graph structure is more effective to separate the object from the background than Zhang's graph [26], we plug Zhang's graph into our system for comparison. Fig. 5 shows the corresponding experimental results of the proposed tracker with our new graph and Zhang's graph on the *animal* and *david* videos. In the *animal* video, fast motion and blur indicate large appearance

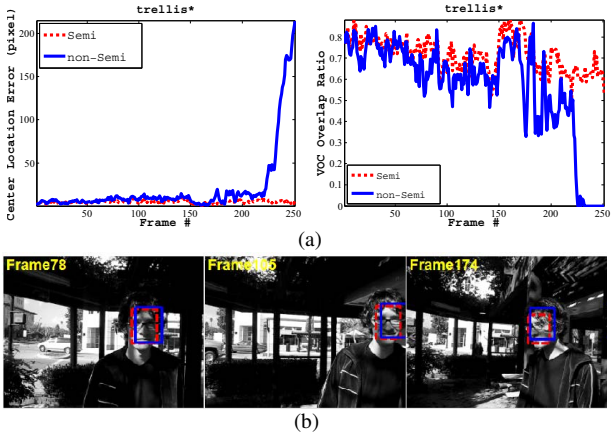


Figure 6: Comparison of the proposed tracker with/without semi-supervised learning on the *trellis** video. (a) shows tracking performance in CLE and VOR; (b) shows some screenshots, red dashed line indicates scenario with semi-supervised learning, blue line indicates without.

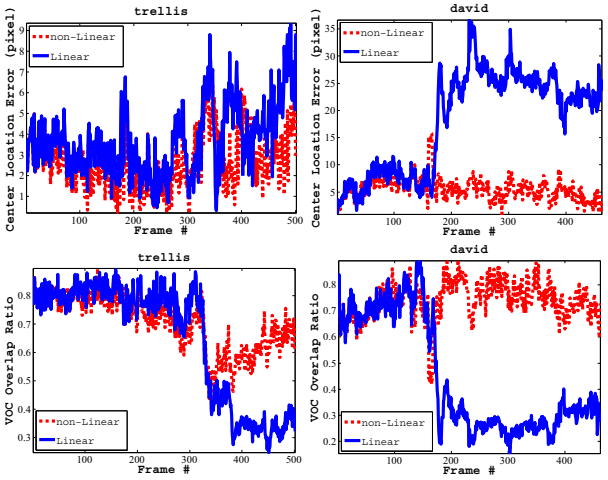


Figure 7: Quantitative comparison of the proposed tracker with/without non-linear extension on two videos. The left two subfigures are associated with the tracking performance in CLE and VOR on the *trellis* video, respectively; the right two subfigures correspond to the tracking performance in CLE and VOR on the *david* video, respectively.

variations in short time; in the *david* video, significant out-of-plane rotation appears around frame #150. From Fig. 5, we can see that Zhang's graph easily impose the tracker drift away from the target due to the shortcoming of separating different classes only based on marginal samples.

Performance with and without regularizer. To validate the effectiveness of the ℓ_1-graph based semi-supervised regularizer, we extract odd numbered frames of the *trellis* video (501 frames in total), and make a new video *trellis** (251 frames in total with low frame rate). In this converted video, the position and background of the object are drastically changed. Meanwhile, severe illumination change and out-of-plane rotation translate the appearance of the object into different one. Fig. 6 shows that the proposed tracker with the ℓ_1-graph based semi-supervised regularizer can adapt to these changes and reliably track the object. More quantitative comparison results are reported in

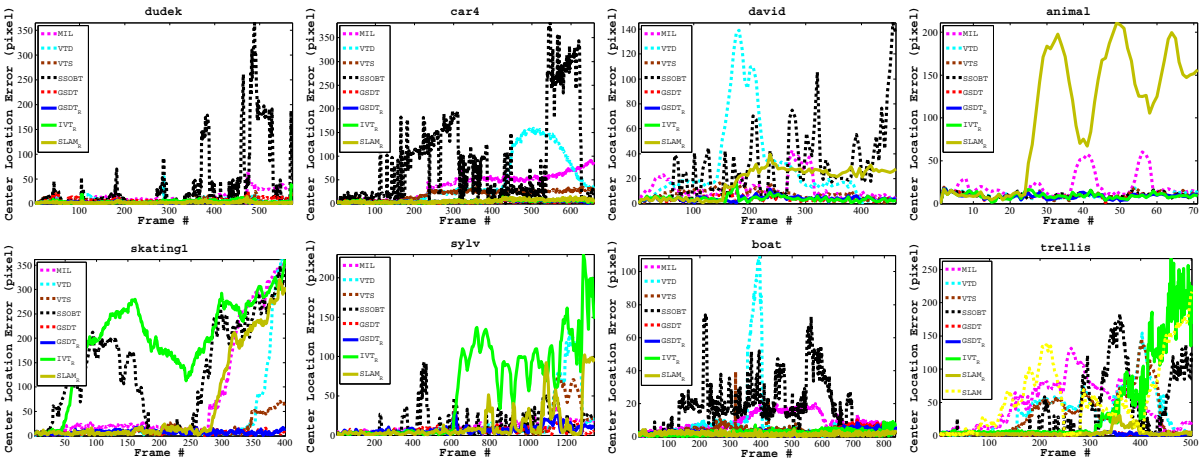


Figure 8: Quantitative comparison of different trackers in terms of CLE on eight videos.

Table 1, which demonstrates that the proposed regularizer gives the plausible boost.

Evaluation of non-linear extension. We claim that the non-linear discriminative boundary tends to provide a more reasonable solution space than linear one, especially in the complex tracking environment. Experimental results on the *trellis* and *david* videos (see Fig. 7) show that, after the objects have undergone out-of-plane rotations (after frame #350 in the *trellis* video, and after frame #170 in the *david* video), the proposed tracker with non-linear extension can more reliably to capture the original appearances of the objects than the one without non-linear extension.

3.2. Comparison with competing trackers

To show the superiority of GSDT over other competing trackers, we perform experiments using SLAM[9], IVT[18], VTD[11], VTS[12], MIL[2] and SSOBT[8] on eight videos. SLAM also introduces the covariance matrix descriptor measured under log-Euclidean Riemannian metric. By comparing GSDT with SLAM, we can intuitively find how our graph embedding based semi-supervised discriminative appearance model explores the descriptor's discriminant capability and achieves more robust and accurate tracking results than SLAM. We implement these trackers using publicly available source codes or binaries provided by the authors. For fair evaluation, each tracker is run with appropriately adjusted parameters. Because IVT and SLAM achieve their best tracking results when their motion models are set to be affine transform (with 2D translation, scale, and in-plane rotation tracked), we consider setting the parameters in our tracker such that all these affine parameters are tracked. This is indicated by the subscript R. Considering the specificities of VTD, VTS, MIL and SSOBT, we set the parameters in our tracker such that only 2D translation and scale are tracked for comparing with them. This is indicated by no-subscript.

To quantitatively evaluate the tracking performances (robustness and accuracy) of the seven trackers under challenging scenarios, we have manually labeled the ground truth of the *boat* and *trellis* videos, and downloaded others from the websites[1] of the video providers. The tracking error evaluation is based on center location error (CLE) between the center of the tracking result and that of the ground truth. Fig. 8 plots the center location error plots (highlighted in different colors) obtained by the seven trackers in the eight experiments. Further, we also compute the average of the center location errors and report the results in Table 2. From Fig. 8 and Table 2, we can see that our proposed GSDT outperforms the others in terms of tracking accuracy and robustness. Some qualitative tracking results especially in Fig. 9(e), Fig. 9(f) and Fig. 9(h) also show that the proposed GSDT can handle pose variation and drastic illumination more robustly than all others.

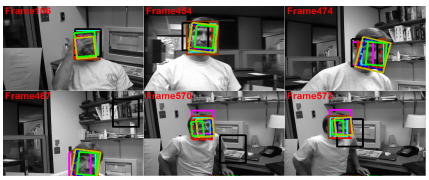
Table 2: Tracking object location: average center location errors (pixels). **Bold green** font indicates best performance with rotation tracked, **Bold red** font indicates best performance without rotation tracked.

Videos	MIL	VTD	VTS	SSOBT	GSDT	GSDT$_R$	IVT$_R$	SLAM$_R$
dudek	12.9	6.6	**7.0**	37.8	7.3	2.4	4.0	**2.2**
car4	38.2	47.5	19.6	87.4	**4.1**	**4.6**	5.0	5.5
david	11.7	28.3	7.8	29.5	**3.6**	**4.0**	**4.0**	18.0
animal	19.9	9.7	10.0	10.1	**9.3**	**8.8**	9.2	100.9
skating1	79.8	32.0	14.3	133.4	**6.7**	**7.7**	192.4	65.5
sylv	9.4	15.7	11.4	14.2	**6.6**	**7.4**	58.3	13.4
boat	9.0	8.4	3.5	17.0	**3.2**	2.8	2.6	**1.7**
trellis	46.0	35.7	33.2	36.4	**2.7**	**2.8**	44.0	4.4

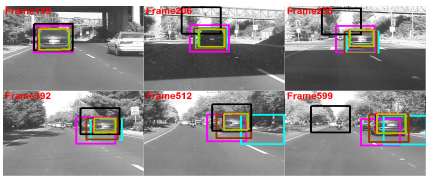
4. Conclusion

In this paper, we have proposed an effective and robust new graph embedding based discriminative tracker with the ℓ_1-graph based semi-supervised regularizer. The superiority of our approach can be attributed to: 1) two new specially designed graphs for modeling the local geometrical and dis-

[1] *dudek*: `http://www.cs.toronto.edu/~dross/ivt/`; *car4*: `http://www.dabi.temple.edu/~hbling/code_data.htm`; *david* and *sylv*: `http://vision.ucsd.edu/~bbabenko/project_miltrack.shtml`; *animal* and *skating1*: `http://cv.snu.ac.kr/research/~vtd/`.

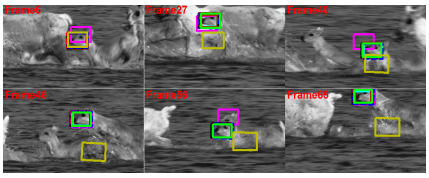


(a) *dudek*, occlusion, scale change and out-of-plane rotation

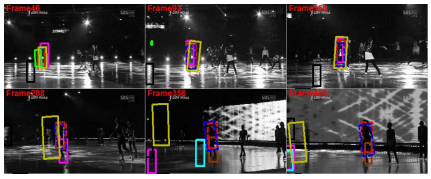


(b) *car4*, scale and illumination changes

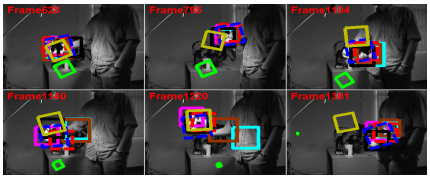
(c) *david*, scale variation and out-of-plane rotation



(d) *animal*, fast motion and blur



(e) *skating1*, pose variation, occlusion and illumination change



(f) *sylv*, in-plane and out-of-plane rotations

(g) *boat*, out-of-plane rotation and background clutter

(h) *trellis*, drastic illumination change and out-of-plane rotation

Figure 9: Screenshots of some sampled tracking results of evaluated approaches on eight challenging videos.

criminative structure of the samples, especially in characterizing the separability of different classes; 2) a novel semi-supervised formulation of the discriminative learning process using the ℓ_1-graph based regularizer, which explores higher order relationships among all the samples and hence is more powerful to model the neighborhood relationship in the cluster assumption for regularizing the decision boundary; 3) a non-linear variant of our semi-supervised discriminative learning method is extended to adapt to multi-modal sample distribution. Experimental results compared with several state-of-the-art trackers on challenging videos demonstrate the effectiveness and robustness of the proposed tracker.

Acknowledgment. This work is partly supported by NSFC (Grant No. 60935002), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504), the Natural Science Foundation of Beijing (Grant No. 4121003), and The Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081).

References

[1] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. *In CVPR*, 2006.

[2] B. Babenko, M. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *Trans. on PAMI*, 33(8):1619–1632, 2011.

[3] D. Cai, X. He, and J. Han. Semi-supervised discriminant analysis. *In ICCV*, 2007.

[4] H. Chen, H. Chang, and T. Liu. Local discriminant embedding and its variants. *In CVPR*, 2005.

[5] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. Huang. Learning with ℓ_1-graph for image analysis. *Trans. on IP*, 19(4):858–866, 2010.

[6] E. Erdem, S. Dubuisson, and I. Bloch. Fragments based tracking with adaptive cue integration. *Computer Vision and Image Understanding*, 116(7):827–841, 2012.

[7] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. *In BMVC*, 2006.

[8] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. *In ECCV*, 2008.

[9] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang. Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model. *Trans. on PAMI*, 34(12):2420–2440, 2012.

[10] T. Jebara, J. Wang, and S. Chang. Graph construction and b-matching for semi-supervised learning. *In ICML*, 2009.

[11] J. Kwon and K. Lee. Visual tracking decomposition. *In CVPR*, 2010.

[12] J. Kwon and K. Lee. Tracking by sampling trackers. *In ICCV*, 2011.

[13] G. Li, L. Qin, Q. Huang, J. Pang, and S. Jiang. Treat samples differently: object tracking with semi-supervised online covboost. *In ICCV*, 2011.

[14] W. Li, X. Zhang, W. Luo, W. Hu, H. Ling, and O. Wu. Robust object tracking with boosted discriminative model via graph embedding. *In ICCV Workshops*, 2011.

[15] X. Li, A. Dick, H. Wang, C. Shen, and A. van den Hengel. Graph mode-based contextual kernels for robust svm tracking. *In ICCV*, 2011.

[16] X. Li, W. Hu, Z. Zhang, M. Zhu, and J. Cheng. Visual tracking via incremental log-euclidean riemannian subspace learning. *In CVPR*, 2008.

[17] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on lie algebra. *In CVPR*, 2006.

[18] D. A. Ross, J. Lim, R. Lin, and M. Yang. Incremental learning for robust visual tracking. *Int. J. Comp. Vis.*, 77(1):125–141, 2008.

[19] Y. Song, F. Nie, C. Zhang, and S. Xiang. A unified framework for semi-supervised dimensionality reduction. *Journal of Pattern Recognition*, 41(9):2789–2799, 2008.

[20] M. Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research*, 8:1027–1061, 2007.

[21] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by $k$nn-sparse graph-based label propagation over noisily-tagged web images. *ACM Trans. on Intelligent Systems and Technology*, 2(2), 2011.

[22] Q. Wang, F. Chen, W. Xu, and M. Yang. An experimental comparison of online object tracking algorithms. *In Proceedings of Image and Signal Processing*, 2011.

[23] S. Yan and H. Wang. Semi-supervised learning by sparse representation. *In SIAM Int'l Conf. on Data Mining*, 2009.

[24] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *Trans. on PAMI*, 29:40–51, 2007.

[25] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. *In NIPS*, 2005.

[26] X. Zhang, W. Hu, S. Maybank, and X. Li. Graph based discriminative learning for robust and efficient object tracking. *In ICCV*, 2007.

[27] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu. Non-negative low rank and sparse graph for semi-supervised learning. *In CVPR*, 2012.