Visual Object Tracking via a Unified Graph Optimization and Labeling Model

No Author Given

No Institute Given

Abstract. Recently, superpixel based object segmentation and tracking methods have been usually developed for segmentation-based tracking. In this paper, we propose a novel unified graph optimization and labeling model for superpixel based object tracking. The main benefits of the proposed method have two main aspects. First, it provides an effective way to exploit both spatial and temporal consistency constraint for target object segmentation. Second, it pursues to learn a multiple connected components graph to better capture the latent relationship among nodes. Extensive experiments demonstrate that our method obtains better performance against the state-of-the-art trackers.

Keywords: visual object tracking · object segmentation · spatial-temporal constraint · graph learning.

1 Introduction

Visual object tracking is an active research problem in computer vision and multimedia area. It has been widely used in many computer vision applications, such as action recognition, video surveillance, augmented reality, etc.

In the past decade, many methods have been proposed for object tracking problem[4, 9, 5, 24, 2, 6, 8, 20, 10, 19, 22, 23]. In general, most of these methods can categorize into two branches, i.e., tracking-by-detection and tracking-by-segmentation. Tracking-by-detection methods generally aim to estimate the locations of the target object in the video sequences by using a rectangle bounding box surrounding the target object while tracking-by-segmentation methods first segment the object from background and then track it based on segmented results. For tracking-by-segmentation methods, early works generally develop some pixel-level based object tracking and segmentation methods [2, 8, 6]. For example, Chad et al. [2] propose a probabilistic framework for joint segmentation and tracking by employing a probabilistic principal component analysis model. Godec et al.[8] propose to employ a generalized Hough-transform and GrabCut[18] method for tracking. Duffner et al.[6] introduce to integrate both generalized Hough transform and probabilistic segmentation approach together for pixel-based object tracking. Obviously, the above pixel-level based tracking methods mainly only explore the feature information of pixels which fail to consider semantic-level and structure information of target object.

In this paper, we propose a novel unified optimization model for superpixel based object tracking, named Spatial-Temporal via Graph Learning(STGL) algorithm. The proposed model provides a general optimization framework which integrates the cues of i) label linear prediction, ii) spatial-temporal consistent constraint and iii) graph learning together to obtain an accurate target object segmentation. The main contributions of this paper are summarized as follows.

- We present a unified optimization framework to integrate multiple cues together for superpixel based object tracking by exploring both local appearance and global structure information of superpixels simultaneously.
- The proposed unified optimization model explores both spatial-temporal consistency constraint and graph learning in visual object segmentation and tracking.
- An effective algorithm is developed to find the global optimal solution for our unified model.

Comprehensive experiments on several widely used benchmark datasets demonstrate the advantages of proposed tracking approach against the state-of-the-art trackers.

2 Methodology

In this section, we first propose the unified optimization model step by step. Then, we derive an simple algorithm to optimize the proposed model.

2.1 Spatial-Temporal Constraint

Given one candidate region in current frame, we first partition it into n non-overlapping superpixels $\{p_1, p_2, \ldots, p_n\}$ via SLIC approach [1]. Normally, let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n] \in \mathbb{R}^{d\times n}$ denotes the collection of some kind of superpixel feature descriptors, where d means the feature dimension. Define $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^{n\times 1}$ to represent indicator vector. We aim to assign each superpixel p_i with an object indicative value $y_i \in \{0, 1\}$ to indicate whether the superpixel p_i belongs to the target object, i.e., $y_i = 1$ indicates that superpixel p_i belongs to the target object, and $y_i = 0$ otherwise. Let $\mathbf{q} = [q_1, q_2, \ldots, q_n]^T \in \mathbb{R}^{n\times 1}$ denotes some kind of foreground prior (or measurement) for superpixels, in which larger q_i indicates the more likely that superpixel p_i belongs to the target object.

Linear Regression. Classically, we utilize linear regression function to measure the latent relationship between superpixel feature and it's corresponding indicator value. Formally, the linear regression function can be written as

$$\min_{\mathbf{y},\mathbf{w},b} \quad \|\mathbf{X}^T\mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^{d\times 1}$ and b is a scalar. $\|\cdot\|_2$ denotes ℓ_2-norm operation. Vector $\mathbf{1} = [1, \ldots, 1]^T \in \mathbb{R}^{n\times 1}$.

Spatial Consistency. Intuitively, neighboring superpixels with similar visual appearance are likely to belong to the same object and thus share similar indicative value. In order to incorporate this spatial consistency constraint, we first construct a neighbor graph $G(V, E)$ whose nodes V represent superpixels and edges E denote the spatial relationship among superpixels. Let $\mathbf{S} \in \mathbb{R}^{n\times n}$ be the corresponding edge weight matrix of graph G. Then, we can incorporate the spatial consistency constraint by adding a graph Laplacian regularization term into Eq.(1):

$$\min_{\mathbf{y},\mathbf{w},b} \|\mathbf{X}^T\mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \alpha\mathbf{y}^T\mathbf{L}_S\mathbf{y} \quad s.t. \;\; y_i \in \{0, 1\} \tag{2}$$

where $\mathbf{L}_S = \mathbf{D}_S - (\mathbf{S}^T + \mathbf{S})/2$ is Laplacian matrix and $\mathbf{D}_S = \mathrm{diag}(d_{11}, d_{22}, \ldots, d_{nn})$, and $d_{ii} = \sum_{j=1}^{n}(\mathbf{S}_{ij} + \mathbf{S}_{ji})/2$. $\mathbf{y}^T\mathbf{L}_S\mathbf{y}$ is computed as

$$\mathbf{y}^T\mathbf{L}_S\mathbf{y} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{S}_{ij}(y_i - y_j)^2 \tag{3}$$

Temporal Consistency. The above model conducts object segmentation on each frame separately, which obviously ignores the temporal correlation between current frame and previous frames. When the superpixel features $\mathbf{X}$ in current frame are partially contaminated or corrupted, the above separate learning model $\mathbf{w}, b$ may be less effective. To overcome this problem, we propose to further incorporate the temporal consistency by considering the consensus between consecutive frames.

$$\min_{\mathbf{y},\mathbf{w},b} \|\mathbf{X}^T\mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \alpha\mathbf{y}^T\mathbf{L}_S\mathbf{y} + \xi\|\widetilde{\mathbf{X}}^T\mathbf{w} + \mathbf{1}b - \widetilde{\mathbf{y}}\|_2^2 \;\; s.t. \;\; y_i \in \{0, 1\} \tag{4}$$

where $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{y}}$ denote the feature collection and indicative vector in the previous frame, respectively. Note that, in the above model, the previous frame features $\widetilde{\mathbf{X}}$ and the already obtained $\widetilde{\mathbf{y}}$ are employed to guide the superpixel indicator vector $\mathbf{y}$ in current frame via the common linear regression parameter $\mathbf{w}, b$.

Additionally, considering that the final indicative values should not be far from preacquired prior values, we add a label propagation term $\|\mathbf{y} - \mathbf{q}\|_2^2$ to minimize the distance between final indicative values and preacquired prior values. We also add regularization term to avoid over-fitting. Ultimately, the spatial-temporal consistency constraint model becomes

$$\begin{aligned}\min_{\mathbf{y},\mathbf{w},b} \mathcal{J}(\mathbf{y}, \mathbf{w}, b) &= \|\mathbf{X}^T\mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \xi\|\widetilde{\mathbf{X}}^T\mathbf{w} + \mathbf{1}b - \widetilde{\mathbf{y}}\|_2^2 \\ &+ \alpha\mathbf{y}^T\mathbf{L}_S\mathbf{y} + \beta\|\mathbf{y} - \mathbf{q}\|_2^2 + \gamma\|\mathbf{w}\|_2^2 \qquad s.t. \quad y_i \in \{0, 1\}\end{aligned} \tag{5}$$

where the parameters $\alpha, \xi, \beta, \gamma$ are used to balance the weight of each term.

2.2 Graph Learning and unified model

Due to implicit relationship and latent structures hidden in data, graph learning has conspicuous significance in the area of machine learning and computer vision applications. Formally, given the some kind of feature descriptors of m-th superpixel sample $\mathbf{X}_m$, the feature data is preprocessed by $\mathbf{X}_m \leftarrow (\mathbf{X}_m - \overline{\mathbf{X}})/\sigma(\mathbf{X})$, where $\overline{\mathbf{X}}$ and $\sigma(\mathbf{X})$ represent the average value and standard deviation of all samples $\mathbf{X}$, respectively. Similar to works [15, 16], graph learning model is represented as

$$\min_{\mathbf{S}} \varPhi(\mathbf{S}) = \sum_{i,j}^{n} \|\mathbf{X}_i - \mathbf{X}_j\|_2^2 \mathbf{S}_{ij} + \varphi\|\mathbf{S}\|_F^2 \quad s.t. \quad \mathbf{S}_i\mathbf{1} = 1,\ 0 \le \mathbf{S}_{ij} \le 1 \tag{6}$$

where $\mathbf{S}_i \in \mathbb{R}^{1\times n}$ is a vector of i-th row of $\mathbf{S}$. $\|\cdot\|_F$ denotes Frobenius norm function. Reasonably, $\mathbf{S}_{ij}$ can be seen as the probability(similarity) between $\mathbf{X}_i$ and $\mathbf{X}_j$ for the reason that the value of $\mathbf{S}_{ij}$ belongs to 0 to 1. Obviously, farther samples should have smaller probability while closer samples should have larger probability. The second term is regularization term which is used to avoid the trivial solution.

We combine spatial-temporal consistency constraint and graph learning to one unified energy optimization model, which iteratively updates weight matrix and indicator vector simultaneously. Thus, the Spatial-Temporal via Graph Learning(STGL) algorithm can be given as

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{y},\mathbf{w},b} \mathcal{J}(\mathbf{y},\mathbf{w},b) + \varPhi(\mathbf{S}) = \|\mathbf{X}^T\mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \xi\|\widetilde{\mathbf{X}}^T\mathbf{w} + \mathbf{1}b - \widetilde{\mathbf{y}}\|_2^2 \\ &+ \alpha\mathbf{y}^T\mathbf{L}_S\mathbf{y} + \beta\|\mathbf{y} - \mathbf{q}\|_2^2 + \gamma\|\mathbf{w}\|_2^2 + \sum_{i,j}^{n} \|\mathbf{X}_i - \mathbf{X}_j\|_2^2\mathbf{S}_{ij} + \varphi\|\mathbf{S}\|_F^2 \\ &s.t. \quad y_i \in \{0,1\},\ \mathbf{S}_i\mathbf{1} = 1,\ 0 \le \mathbf{S}_{ij} \le 1 \end{aligned} \tag{7}$$

3 Optimization

In this section, we propose an effective optimization algorithm to solve the proposed model. Then, the convergence analysis of the proposed algorithm is given. Since Eq.(7) is a convex problem, the proposed algorithm can obtain the global optimal solution for it. First, it is difficult to enforce the integral constraint $\mathbf{y}_i \in \{0,1\}$ in model optimization. One popular way is to relax this integral constraint $\mathbf{y}_i \in \{0,1\}$ to the nonnegative domain $\mathbf{y}_i \ge 0$ and propose to solve the following convex problem as

$$\begin{aligned} &\min_{\mathbf{S},\mathbf{y},\mathbf{w},b} \mathcal{J}(\mathbf{y},\mathbf{w},b) + \varPhi(\mathbf{S}) = \|\mathbf{X}^T\mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \xi\|\widetilde{\mathbf{X}}^T\mathbf{w} + \mathbf{1}b - \widetilde{\mathbf{y}}\|_2^2 \\ &+ \alpha\mathbf{y}^T\mathbf{L}_S\mathbf{y} + \beta\|\mathbf{y} - \mathbf{q}\|_2^2 + \gamma\|\mathbf{w}\|_2^2 + \sum_{i,j}^{n} \|\mathbf{X}_i - \mathbf{X}_j\|_2^2\mathbf{S}_{ij} + \varphi\|\mathbf{S}\|_F^2 \\ &s.t. \quad y_i \ge 0,\ \mathbf{S}_i\mathbf{1} = 1,\ 0 \le \mathbf{S}_{ij} \le 1 \end{aligned} \tag{8}$$

Then, we derive an effective update algorithm which updates the variables $\mathbf{S}, \mathbf{w}, b, \mathbf{y}$ alternatively until convergence.
Step 1: Fix $\mathbf{y}, \mathbf{w}, b$, update $\mathbf{S}$.
The problem becomes

$$\begin{aligned} &\min_{\mathbf{S}} \alpha \mathbf{y}^T \mathbf{L}_S \mathbf{y} + \sum_{i,j}^{n} \|\mathbf{X}_i - \mathbf{X}_j\|_2^2 \mathbf{S}_{ij} + \varphi \|\mathbf{S}\|_F^2 \\ &s.t. \quad y_i \geq 0,\ \mathbf{S}_i \mathbf{1} = 1,\ 0 \leq \mathbf{S}_{ij} \leq 1 \end{aligned} \tag{9}$$

For clarity, we denote $d_{ij}^y = \mathbf{y}^T \mathbf{L}_S \mathbf{y}$ and $d_{ij}^x = \|\mathbf{X}_i - \mathbf{X}_j\|_2^2$. Note that each row of $\mathbf{S}$ is solved independently. Thus, we handle each $\mathbf{S}_i$ individually by

$$\min_{\mathbf{S}_i} \sum_{j=1}^{n} (\alpha d_{ij}^y \mathbf{S}_{ij} + d_{ij}^x \mathbf{S}_{ij} + \varphi \mathbf{S}_{ij}^2) \qquad s.t.\ \ y_i \geq 0, \mathbf{S}_i \mathbf{1} = 1,\ 0 \leq \mathbf{S}_{ij} \leq 1 \tag{10}$$

$\mathbf{S}_i$ can be obtained by setting the first order derivative of Eq.(10) with respect to variables $\mathbf{S}_i$ to zeros. Let d_{ij} denotes $\alpha d_{ij}^y + d_{ij}^x$, the optimal $\mathbf{S}_i$ is given as

$$\mathbf{S}_i^* = -\frac{1}{2\varphi} d_{ij} \tag{11}$$

Step 2: Fix $\mathbf{S}, \mathbf{y}, b$, update $\mathbf{w}$.
The problem becomes

$$\min_{\mathbf{w}} \|\mathbf{X}^T \mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \xi \|\widetilde{\mathbf{X}}^T \mathbf{w} + \mathbf{1}b - \widetilde{\mathbf{y}}\|_2^2 + \gamma \|\mathbf{w}\|_2^2 \quad s.t.\ \ y_i \geq 0 \tag{12}$$

Similarity, $\mathbf{w}$ can be acquired by setting the first order derivative of Eq.(12) with respect to variables $\mathbf{w}$ to zeros. Then, the optimal $\mathbf{w}$ is given as

$$\mathbf{w}^* = \big(\mathbf{X}\mathbf{X}^T + \xi \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T + \gamma \mathbf{I}\big)^{-1} \big(\mathbf{X}(\mathbf{y} - \mathbf{1}b) + \xi \widetilde{\mathbf{X}}(\widetilde{\mathbf{y}} - \mathbf{1}b)\big) \tag{13}$$

Step 3: Fix $\mathbf{S}, \mathbf{y}, \mathbf{w}$, update b.
The problem becomes

$$\min_{b} \|\mathbf{X}^T \mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \xi \|\widetilde{\mathbf{X}}^T \mathbf{w} + \mathbf{1}b - \widetilde{\mathbf{y}}\|_2^2 \quad s.t.\ \ y_i \geq 0 \tag{14}$$

Then, the optimal b is given as

$$b^* = \frac{1}{(1+\xi)n} \mathbf{1}^T \big(\mathbf{y} - \mathbf{X}^T \mathbf{w} + \xi(\widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}^T \mathbf{w})\big) \tag{15}$$

Step 4: Fix $\mathbf{S}, \mathbf{w}, b$, update $\mathbf{y}$.
The problem becomes

$$\begin{aligned} &\min_{\mathbf{y}} \|\mathbf{X}^T \mathbf{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \alpha \mathbf{y}^T \mathbf{L}_S \mathbf{y} + \beta \|\mathbf{y} - \mathbf{q}\|_2^2 \\ &s.t.\ \ y_i \geq 0,\ \mathbf{S}_i \mathbf{1} = 1,\ 0 \leq \mathbf{S}_{ij} \leq 1 \end{aligned} \tag{16}$$

which is simply rewritten as

$$\min_{\mathbf{y}} \|\mathbf{y}-\mathbf{u}\|_2^2 + \alpha \mathbf{y}^T \mathbf{L}_S \mathbf{y} \quad s.t. \;\; \mathbf{y}_i \geq 0 \tag{17}$$

here $\mathbf{u} = \mathbf{X}^T\mathbf{w} + \mathbf{1}b + \beta\mathbf{q}$. We can simply prove that Eq.(17) is equivalent to

$$\min_{\mathbf{y}} \|\mathbf{y}-\check{\mathbf{u}}\|_2^2 + \alpha \mathbf{y}^T \mathbf{L}_S \mathbf{y} \quad s.t. \;\; \mathbf{y}_i \geq 0 \tag{18}$$

where $\check{\mathbf{u}}_i = \max\{\mathbf{u}_i, 0\}$. It is known that the optimal solution of problem Eq.(18) is naturally nonnegative. Thus, the optimal solution can be obtained by setting the first derivative of Eq.(18) with respect to variable $\mathbf{y}$ to zero. Then, we get the optimal $\mathbf{y}$

$$\mathbf{y}^* = (\mathbf{I} + \alpha \mathbf{L}_S)^{-1}\check{\mathbf{u}} \tag{19}$$

4 Segmentation and Tracking

In each current frame, we first obtain a candidate Region of Interest (RoI), which is identified based on an extended bounding box surrounding the target object established in the previous frame by using an optical flow algorithm [3]. We use SLIC algorithm [1] to segment the RoI into a set of non-overlapping superpixels $\{p_1, p_2, \ldots, p_n\}$. For each superpixel p_i, we extract visual feature descriptor $\mathbf{X}_i$ (color, edge and texture) for it. Then, we construct a regular graph $G(V, E)$ whose nodes V represent superpixels and edges E contain the spatial relationships among superpixels. The edges connect superpixels within 2 hops, i.e., each node connects to its 2 nearest neighbor nodes. Similar to work [23], we adopt a support vector regression to learn a score r_i for each node by projecting the original feature $\mathbf{X}_i$. Then, we compute the weight of each edge as,

$$\mathbf{S}_{ij} = \exp(-\frac{\|r_i - r_j\|}{\sigma}), \tag{20}$$

where σ is a constant, r_i and r_j denote the scores of node v_i and v_j, respectively.

Based on the feature $\mathbf{X}$ of current RoI, feature $\widetilde{\mathbf{X}}$ of RoI in the previous frame, graph $\mathbf{S}$ of superpixels in current RoI and foreground prior $\mathbf{q}$, we first use the proposed model to compute an optimal confidence of target object $\mathbf{y}$ for superpixels in current frame. In this paper, we compute the foreground prior $\mathbf{q}$ by using an absorbing markov chain model and use the normalized absorbing time as the foreground prior measurement [23]. Then, we obtain the final segmentation mask of target object by selecting the superpixels that have high confidence value (higher than mean value of $\mathbf{y}$). Finally, we cover the segmentation mask with a minimum rectangular bounding box and achieve object tracking.

5 Experiment

We evaluate the proposed method on five widely used benchmark datasets including DAVIS [17], GBS [11, 14, 13], ST2 [12], NR [19] and VS dataset [7]. We use the two common used evaluation metrics to measure the quantitative performances of different tracker methods, i.e., precision rate and success rate [21].

Table 1. Average overlap ratio of segmentation masks for tracking-by-segmentation algorithms. The best results are in bold fonts.

	STGL	AMCT	OGBDT	HT	SPT	PT
DAVIS	**62.2**	59.2	44.9	33.1	27.1	26.1
GBS	**75.7**	74.8	59.7	40.4	45.9	35.3
ST2	**62.7**	58.8	47.6	43.0	26.3	21.2
NR	**58.7**	58.6	53.3	41.1	29.7	28.3
VS	**87.5**	84.1	79.8	51.2	61.0	73.9

Table 2. Average overlap ratio of bounding boxes for tracking algorithms. The best results are in bold fonts.

	tracking-by-segmentation algorithms						tracking-by-detection algorithms			
	STGL	AMCT	OGBDT	HT	SPT	PT	ECO	MUSTer	DSST	MEEM
DAVIS	**61.9**	60.9	50.0	35.8	43.2	41.6	55.1	25.9	58.4	52.7
GBS	**81.3**	80.0	61.2	43.0	55.2	44.7	68.1	59.4	62.9	52.6
ST2	**68.0**	64.8	50.2	44.9	53.5	32.2	59.6	58.8	62.0	59.5
NR	**67.2**	66.9	60.8	40.9	35.7	16.1	38.1	36.2	35.4	33.1
VS	**89.5**	88.2	78.8	57.6	61.5	51.9	71.1	64.1	66.9	60.3

5.1 Implementation Details

We set parameters $\alpha, \beta, \xi, \gamma, \varphi$ to $0.001, 50, 0.01, 1, 0.1$, respectively. The objective function difference value $minError$ between two consecutive iterations and max iteration times $maxIter$ are set to 0.1 and 50 respectively.

5.2 Comparative Results

We compare our method with some other state-of-the-art related segmentation-based tracking methods include Absorbing Markov Chain Tracking(AMCT) [23], Online Gradient Boosting Decision Tree Tracker (OGBDT) [19], Hough-based Tracking (HT) [8], Superpixel Tracker (SPT) [20], PixelTrack (PT) [6]. We also compare our method with some other detection-based tracking algorithms including ECO [4], MUSTer [9], DSST [5], and MEEM [24].

Table 1 shows the average overlap ratio between tracked segmentation results and manually labeled ground-truth masks. The first column represents the proposed Spatial-Temporal via Graph Learning (STGL) model. Since tracking-by-detection methods do not return the binary segmentation masks, they are not compared in Table 1. Note that, our method yields the best performance on all five datasets. Table 2 demonstrates the average overlap ratio of bounding boxes for tracking-by-detection and tracking-by-segmentation algorithms. Here, we can note that, our method comprehensively performs better than the state-of-the-art methods and achieves the best performance on all datasets.

Figure 1 demonstrates the average success and precision curves on all five datasets. Figure 1 (a) and (b) show the success curves in terms of bounding box

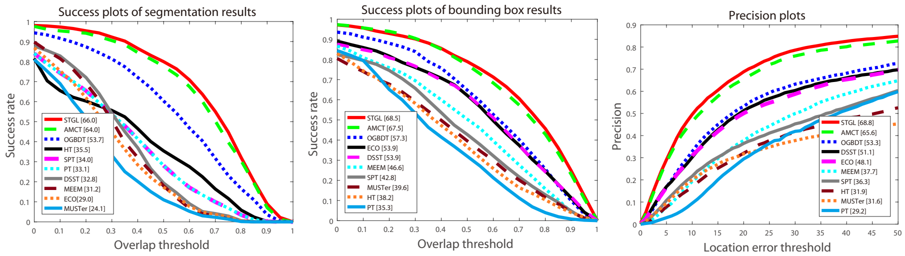

Fig. 1. Success and precision plots in overall five datasets: (a) success plots in terms of segmentation overlap ratio (b) success plots in terms of bounding box overlap ratio (c) precision plots.

Table 3. Ablation study of each module in our algorithm on overall datasets.

	Success-Seg	Success-Box	Precision
baseline (Eq.(1))	64.4	67.0	67.3
+S (Eq.(2))	65.0	67.5	67.8
+ST (Eq.(5))	65.3	67.7	67.9
+STGL (Eq.(7))	**66.0**	**68.5**	**68.8**

and segmentation mask, respectively. For tracking-by-detection methods, we regard the all pixels in rectangular bounding boxes directly as segmentation masks to compute the success rates, as used in work [23]. Figure 1(c) demonstrates the precision curve in terms of bounding box. Here, one can note that the proposed tracker yields the best average results on all five datasets in terms of both object tracking and segmentation, which further validates the effectiveness of the proposed method. Qualitative comparisons of our approach with segmentation-based tracking methods are shown in Figure 2.

5.3 Ablation Study

In this section, we investigate the effects of each component in our unified model. Table 3 summarizes how performance gets improved by adding each module step-by-step into our model on overall datasets. We implement three special variants of our model, i.e., baseline, baseline+S, baseline+ST, and the complete model baseline+STGL. (1) baseline as our baseline model only uses the linear regression for superpixel classification, as shown in Eq.(1). (2) baseline+S adds spatial consistency constraint, as shown in Eq.(2). (3) baseline+ST denotes the spatial-temporal consistency constraint model, as shown in Eq.(5). (4) baseline+STGL denotes the complete Spatial-Temporal via Graph Learning(STGL) model, as shown in Eq.(7). Here, one can note that, after incorporating the spatial-temporal consistency constraint and graph learning modular into baseline model, our model improves Success-Seg, Success-Box and Precision by 1.6%,1.5% and 1.5% respectively, which shows that the proposed unified op-

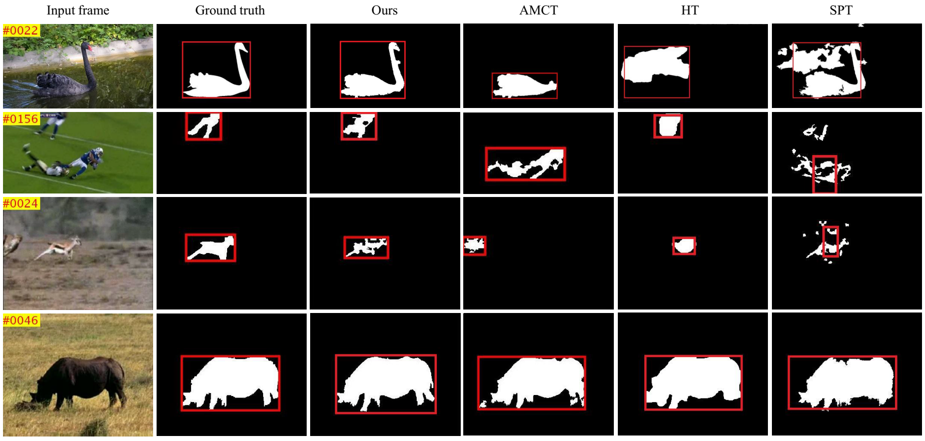

Fig. 2. Qualitative performance evaluation results. From top to bottoms, we show segmentation and tracking results of *blackswan* in DAVIS, *NFL2* in GBS, *cheetah-1* in ST2 and *Rhino* in VS. White regions and red boxes depict segmentation masks and bounding boxes, respectively. The numbers at the top-left corners in first column are frame indices.

timization model is essential and effective for non-rigid and articulated motion problems.

6 Conclusion

We have proposed an effective and efficient segmentation-based via superpixel tracking algorithm using a unified optimization model, named STGL algorithm. Our model incorporates spatial-temporal consistent constraint and graph learning into a formulation to tackle object tracking involved non-rigid and articulated motions. We compare our algorithm with the state-of-the-art segmentation-based trackers and detection-based trackers in multiple challenging datasets. Experiments demonstrate that our algorithm achieves the state-of-the-art performance.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE transactions on pattern analysis and machine intelligence **34**(11), 2274–2282 (2012)
2. Aeschliman, C., Park, J., Kak, A.C.: A probabilistic framework for joint segmentation and tracking. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1371–1378. IEEE (2010)
3. Bao, L., Yang, Q., Jin, H.: Fast edge-preserving patchmatch for large displacement optical flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3534–3541 (2014)

4. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: Eco: efficient convolution operators for tracking pp. 6638–6646 (2017)
5. Danelljan, M., Häger, G., Khan, F., Felsberg, M.: Accurate scale estimation for robust visual tracking (2014)
6. Duffner, S., Garcia, C.: Pixeltrack: a fast adaptive algorithm for tracking non-rigid objects. In: Proceedings of the IEEE international conference on computer vision. pp. 2480–2487 (2013)
7. Fukuchi, K., Miyazato, K., Kimura, A., Takagi, S., Yamato, J.: Saliency-based video segmentation with graph cuts and sequentially updated priors. In: 2009 IEEE International Conference on Multimedia and Expo. pp. 638–641. IEEE (2009)
8. Godec, M., Roth, P.M., Bischof, H.: Hough-based tracking of non-rigid objects. vol. 117, pp. 1245–1256. Elsevier (2013)
9. Hong, Z., Chen, Z., Wang, C., Mei, X., Prokhorov, D., Tao, D.: Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 749–758 (2015)
10. Hong, Z., Wang, C., Mei, X., Prokhorov, D., Tao, D.: Tracking using multilevel quantizations (2014)
11. Kwak, S., Lim, T., Nam, W., Han, B., Han, J.H.: Generalized background subtraction based on hybrid inference by belief propagation and bayesian filtering. In: 2011 International Conference on Computer Vision. pp. 2174–2181. IEEE (2011)
12. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2192–2199 (2013)
13. Lim, J., Han, B.: Generalized background subtraction using superpixels with label integrated motion estimation pp. 173–187 (2014)
14. Lim, T., Hong, S., Han, B., Hee Han, J.: Joint segmentation and pose tracking of human in natural videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 833–840 (2013)
15. Nie, F., Wang, X., Huang, H.: Clustering and projected clustering with adaptive neighbors. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 977–986. ACM (2014)
16. Nie, F., Zhu, W., Li, X.: Unsupervised feature selection with structured graph optimization. In: Thirtieth AAAI conference on artificial intelligence (2016)
17. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 724–732 (2016)
18. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts **23**(3), 309–314 (2004)
19. Son, J., Jung, I., Park, K., Han, B.: Tracking-by-segmentation with online gradient boosting decision tree. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3056–3064 (2015)
20. Wang, S., Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: 2011 International Conference on Computer Vision. pp. 1323–1330. IEEE (2011)
21. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2411–2418 (2013)
22. Xiao, J., Stolkin, R., Leonardis, A.: Single target tracking using adaptive clustered decision trees and dynamic multi-level appearance models. In: Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. pp. 4978–4987 (2015)

23. Yeo, D., Son, J., Han, B., Hee Han, J.: Superpixel-based tracking-by-segmentation using markov chains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1812–1821 (2017)
24. Zhang, J., Ma, S., Sclaroff, S.: Meem: robust tracking via multiple experts using entropy minimization. In: European conference on computer vision. pp. 188–203. Springer (2014)