An Approach to Streaming Video Segmentation With Sub-Optimal Low-Rank Decomposition

Chenglong Li, Liang Lin, Wangmeng Zuo, Wenzhong Wang, and Jin Tang

***Abstract*— This paper investigates how to perform robust and efficient video segmentation while suppressing the effects of data noises and/or corruptions, and an effective approach is introduced to this end. First, a general algorithm, called sub-optimal low-rank decomposition (SOLD), is proposed to pursue the low-rank representation for video segmentation. Given the data matrix formed by supervoxel features of an observed video sequence, SOLD seeks a sub-optimal solution by making the matrix rank explicitly determined. In particular, the representation coefficient matrix with the fixed rank can be decomposed into two sub-matrices of low rank, and then we iteratively optimize them with closed-form solutions. Moreover, we incorporate a discriminative replication prior into SOLD based on the observation that small-size video patterns tend to recur frequently within the same object. Second, based on SOLD, we present an efficient inference algorithm to perform streaming video segmentation in both unsupervised and interactive scenarios. More specifically, the constrained normalized-cut algorithm is adopted by incorporating the low-rank representation with other low level cues and temporal consistent constraints for spatio-temporal segmentation. Extensive experiments on two public challenging data sets VSB100 and SegTrack suggest that our approach outperforms other video segmentation approaches in both accuracy and efficiency.**

***Index Terms*— Video processing, streaming segmentation, low-rank representation, spectral clustering.**

I. Introduction

VIDEO segmentation is to partition the video into several semantically consistent spatio-temporal regions. It is a fundamental computer vision problem in many applications, such as object tracking, activity recognition, video analytics, summarization and indexing. However, it is still a challenging research area due to its computational complexity and inherent difficulties, like the large intra-category variations and the large inter-category similarities.

According to the amount of manual annotation, recent video segmentation algorithms can be categorized into four groups, *i.e.*, unsupervised, interactive, semi-supervised and supervised. 1) The unsupervised methods produce coherent spatial-temporal regions from the bottom-up fashion, and have been introduced ranging from mean-shift [1], spectral clustering [2], [3], graph-based processing [4], [5] and superpixel tracking [6]. Besides, some benchmarks [7], [8] have also been provided to evaluate existing methods and help further study. 2) A small amount of human at the start frame or frames is required in the interactive approaches to segment the foreground from the background [9]–[13]. Some of these approaches [9]–[11] are strongly interactive that allow the user to correct any mistakes in the loop if needed. 3) The semi-supervised foreground propagation approaches accept a frame labeled manually with the foreground region and propagate it to the remaining frames [14]. 4) Methods for the supervised setting attempt to segment the same object or object category of interest as foreground by learning an object model from labeled exemplars [15].

In this paper, we investigate the problem of streaming video segmentation under the Low-Rank Representation (LRR) framework.[1] Although LRR had been very successful in image segmentation [16]–[18], there exists several remaining issues for applying LRR to video segmentation. First, most LRR algorithms relax the rank constraint with the nuclear norm to make the objective tractable. The relaxed objective usually is optimized by ALM method [19] which converges slowly, making it computationally inefficient for video segmentation. Second, it is shown that internal video statistics is helpful to improve segmentation performance, but it remains not well studied for incorporating internal video statistics into LRR. Finally, to cope with arbitrarily long video, temporally consistent constraints is indispensable for streaming video segmentation.

Aimed at addressing these issues and motivated by the advances in subspace clustering [20], [21], especially LRR methods for image segmentation [16]–[18], we propose an effective approach for streaming video segmentation with a *Sub-Optimal Low-rank Decomposition* (SOLD) algorithm, which pursues the low-rank representation by exploiting the low-rank structure of low-level supervoxel features. It is well known that the rank constraint can suppress the effects of

Manuscript received June 19, 2015; revised December 3, 2015, January 21, 2016, and February 24, 2016; accepted February 28, 2016. Date of publication March 2, 2016; date of current version March 18, 2016. This work was supported in part by the 973 Program of China under Grant 2015CB351705, in part by the National Natural Science Foundation of China under Grant 61472002, in part by the Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase), in part by the Guangdong Natural Science Foundation under Grant S2013050014548 and Grant 2014A030313201, and in part by the Program of Guangzhou Zhujiang Star of Science and Technology under Grant 2013J2200067. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Janusz Konrad. *(Corresponding author: Liang Lin.)*

C. Li, W. Wang, and J. Tang are with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com; wenzhong@ahu.edu.cn; ahhftang@gmail.com).

L. Lin is with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China (e-mail: linliang@ieee.org).

W. Zuo is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: cswmzuo@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2537211

[1] Project webpage: http://vision.sysu.edu.cn/projects/sold/.

1057-7149 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

severe noises and/or corruptions, which is important for robust video segmentation.

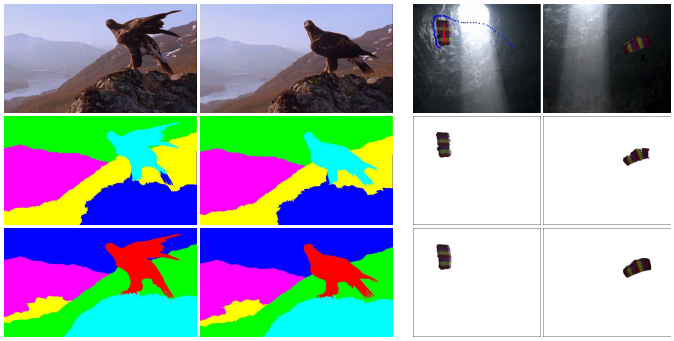
Instead of pixels or superpixels in previous works like [3], [4], we take supervoxels as graph nodes to infer their optimal affinities. Supervoxels can preserve local spatio-temporal coherence as well as good boundaries, and substantially improve segmentation efficiency. We assume that *the intra-class supervoxels are drawn from one identical low-rank feature subspace, and all supervoxels in a temporal window lie on a union of multiple subspaces*. Herein, a temporal window is defined as a number of adjacent frames. Thus, we can represent each supervoxel descriptor as a linear combination of other supervoxel descriptors, and seek for the low-rank representation of all supervoxels in a joint fashion. Moreover, we also integrate *discriminative replication prior* in the formulation for enhancing its discriminative ability. This prior, *local small-size video cubes (e.g.,* $6 \times 6 \times 6$ *voxels) with certain appearance patterns tend to recur frequently within the semantic region, but may not appear in the different semantic regions*, exploits the small non-local recurring regions [22] to refine affinities among supervoxels. Herein, a semantic region is defined as a set of spatio-temporal pixels of the same object. It also can be viewed as the extension of internal image statistics [23] for video data, but can substantially reduce the computational complexity.

Unlike relaxing the rank minimization to the nuclear norm minimization in other works [16], [17], the rank of the representation coefficient matrix in SOLD is explicitly determined for better representation. In particular, the representation coefficient matrix with the fixed rank can be decomposed into two low rank sub-matrices. Thus, we efficiently optimize the low-rank representation by iteratively solving several sub-problems with closed-form solutions. The optimization solution is then employed to define affinities among supervoxels.

Based on SOLD, two special tasks, unsupervised and interactive video segmentation, are addressed in our framework. First, we combine the low-rank representation matrix with other low-level cues to define the affinity matrix. Then, we directly apply constrained NCut algorithm [24] on the defined affinity matrix to achieve the unsupervised segmentation. In interactive task, we define the appearance models of foreground and background by user interactions, and combine with the low-rank representation and the spatio-temporal smoothness constraints to accurately segment the target object. We formulate it as the Markov Random Field (MRF) problem, which can be efficiently solved by the Primal-Dual method [25]. Fig. 1 illustrates the unsupervised and interactive segmentation results of our approach.

This paper makes the following contributions to video processing and related applications.

- It presents an effective approach for segmenting videos into consistent spatio-temporal regions, which pursues the low-rank representation of the video supervoxel feature matrix. Our approach is able to deal with both unsupervised and interactive scenarios and outperforms other video segmentation methods on the standard benchmarks.



(a) (b)

Fig. 1. The unsupervised and interactive segmentation results of our approach are shown in (a) and (b), respectively. The different colors indicate the different regions in (a). The first row shows the quintessential frames in video sequences, and the dash lines in (b) indicate the user interactions, in which the red denotes the foreground and the blue denotes the background. Our results and the corresponding ground truth are shown in the middle and last row, respectively.

- It presents a novel low-rank decomposition method with the fixed-rank representation coefficient matrix, achieving a very efficient sub-optimal solution by iteratively solving three closed-form sub-problems. This proposed method can be extended to other similar tasks for pursuing low-rank representations.
- It utilizes an internal replication prior for enhancing discriminative ability between supervoxels, which is naturally incorporated into SOLD. Moreover, we utilize several temporal consistent constraints during the inference of streaming video segmentation, effectively improving the robustness.

The rest of this paper is organized as follows. In Sect. II, the relevant existing unsupervised and interactive video segmentation methods are introduced. In Sect. III, we describe the details of our approach. The experimental results on two public challenging datasets are shown in Sect. V. The final Sect. VI concludes this paper.

II. Literature Review

Some of the relevant state-of-the-art methods on the unsupervised and interactive video segmentation are reviewed in this section.

Recent advances in hierarchical methods [4], [26], [27], streaming methods [5], [28] and related benchmarks [7], [8] have shown that unsupervised supervoxel segmentation has gained potential as a first step in early video processing. Hierarchical video segmentation provides a rich multiscale decomposition of a given video. Grundmann *et al.* [4] proposed Hierarchical Graph-Based video segmentation (HGB) algorithm based on local properties. It iteratively merged nodes in a region graph to produce a hierarchical segmentation. To process arbitrary long video, Xu *et al.* [5] proposed a streaming hierarchical video segmentation framework and instantiated HGB within this framework (SHGB). This method enforced a Markov assumption on the video stream, which leveraged ideas from data streams. Galasso *et al.* [28] proposed a spectral graph reduction algorithm for efficient streaming video segmentation. In this method, the reduced superpixel graph was reweighted such that the resulting segmentation

was equivalent to the full graph under certain assumptions. Xu and Corso [7] presented a thorough evaluation of five supervoxel methods on a suite of suitable metrics designed to access supervoxel desiderata. A united video segmentation benchmark was provided by Galasso *et al.* [8] to evaluate effectively over- and under-segmentation of current video segmentation methods. These benchmarks not only allow to analyze the current state-of-the-art in video segmentation, but encourage the progress on new aspects of the video segmentation methods.

Recent works on video segmentation focus only on salient moving objects by analyzing point trajectories, while taking background as a single cluster [2], [29]. Some other works [3], [6] over-segment frames into superpixels, and partition them spatially and match them temporally. These methods provide a desirable computational reduction and powerful within-frame representation [30]. For instance, Galasso *et al.* [3] proposed a robust Video Segmentation approach with Superpixels (VSS) to explore various within- and between-frame affinities. In addition, Tarabalka *et al.* [31] presented a more efficient method for joint segmentation of monotonously growing or shrinking shapes in a time sequence of noisy images, and this method was applied to three practical problems to validate its performance and practicality.

Different from unsupervised video segmentation, interactive video segmentation focused on extracting foreground object in clutter background with simple user interventions (often just one scribble for the object and one for the background). Wang *et al.* [9] introduced a hierarchical mean-shift preprocess to reduce the number of nodes for efficient computation, and extended 2D alpha matting scheme to 3D video volumes. Bai *et al.* [10] presented an interactive framework for soft segmentation and matting of natural images and videos. The proposed technique was based on weighted geodesics distance functions, which can be solved in computationally optimal linear time. It also allowed additional constraints into the distance definition to efficiently handle occlusions. A learning-based method was proposed by Price *et al.* [11] to automatically weighted multiple features by learning from the previous implicitly-validated frame or the user corrections required in the previous frame. The above methods segmented or matted object frame by frame, and may require additional supervision in more complex videos. The long video intervals (up to 100 frames) were considered by Dondera *et al.* [12] on the basis of occlusion and long term spatio-temporal structure cues. Their system obtained good results quickly by running spectral clustering on superpixels.

III. Sub-Optimal Low-Rank Decomposition

Given an arbitrarily long input video, we adopt the overlapping sliding temporal window approach to save memory and space. In this section, we focus on the proposed model to obtain the low rank coefficient matrix $\mathbf{Z}$ of the supervoxel feature matrix of a temporal window.

A. Formulation

The proposed low rank decomposition model is imposed on the supervoxels for better tradeoff of efficiency and accuracy.

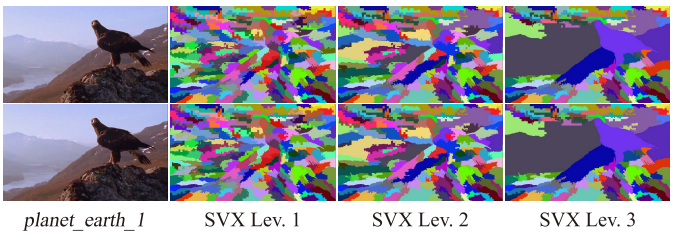


Fig. 2. Sample supervoxels at level 1 (200, where 200 indicates the number of supervoxels), 2 (150) and 3 (100) extracted from a hierarchical video segmentation [4]. The different colors indicate the different supervoxels.

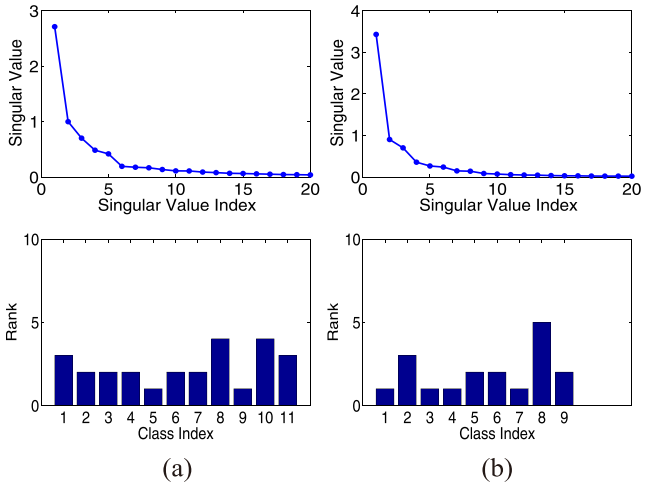


Fig. 3. Illustration of low-rank assumption in our framework. The first row indicates the singular value plots of supervoxel feature matrix in one temporal window, and the second one denotes the rank of each semantic class in this temporal window. These two sequences are randomly selected in VSB100 dataset. (a) Arctic_kayak. (b) Palm_tree.

We over-segment a temporal window into supervoxels by employing unsupervised video segmentation method [4], where each supervoxel comprises an ensemble of voxels that are coherent both spatially and temporally, and perceptually similar with respect to certain appearance features ($e.g.$ color). Generally speaking, multi-level supervoxel representation can provide more appearance and motion features. However, as shown in Fig. 2, the finest-level supervoxels have good spatio-temporal coherence and boundaries whilst the coarse-level supervoxels usually introduce large under-segmentation errors. Therefore, our model is formulated in the finest-level supervoxels to avoid error propagation.

Each temporal window of the video is segmented into n supervoxels. For each supervoxel, a set of appearance and motion features are extracted and combined into one single d-dimensional feature vector $\mathbf{x}_i$ for supervoxel representation. Then, all the feature vectors of the n supervoxels form the data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$.

We assume that supervoxels belonging to the same semantic region are all drawn from the same low-rank subspace, and all supervoxels in one temporal window lie on a union of multiple subspaces. As illustrated in Fig. 3, the supervoxel feature matrix can be well approximated by a matrix with rank less than 10, and the rank of each semantic class is less than 5, which justifies the low rank

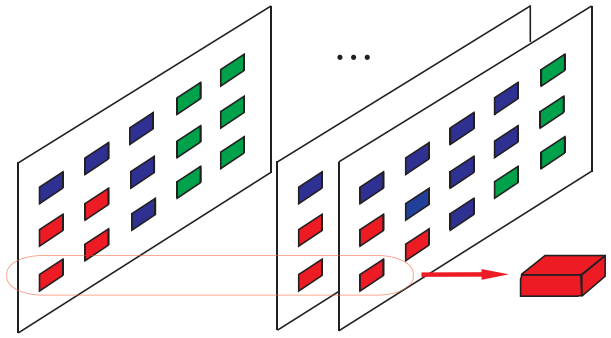


Fig. 4. Illustration of discriminative replication prior in SOLD. A video cube consists of a set of spatially overlapped patches, where repeatedly occurred patches are identified with the same color. One red cube is highlighted for clarity.

representation assumption. Here, each supervoxel descriptor can be represented as the linear combination of the supervoxel descriptors, the low-rank representation (LRR) of all supervoxels can then be pursued in a joint fashion, *i.e.*, $\mathbf{X} = \mathbf{XZ}$, where $\mathbf{Z}$ is the desired low-rank representation coefficient matrix. Thanks to the low-rank constraint, the solution of LRR can better capture the global structure of the matrix $\mathbf{X}$ than sparse coding [21], and benefit subspace segmentation [20]. Since the supervoxel feature matrix is often noisy or grossly corrupted, the low-rank representation can be solved by the following program,

$$\mathbf{X} = \mathbf{XZ} + \mathbf{E} + \epsilon, \quad s.t.\ rank(\mathbf{Z}) \leq r, \tag{1}$$

where r is the desired rank, and $r \ll n$. $\mathbf{Z} \in \mathbb{R}^{n\times n}$ is the desired low-rank representation coefficient matrix, and $\mathbf{E} \in \mathbb{R}^{d\times n}$ and $\epsilon \in \mathbb{R}^{d\times n}$ denote the sparse corrupted noise and the dense Gaussian noise, respectively. Thus, the low-rank representation model can be formulated as

$$\min_{\mathbf{Z},\mathbf{E}} \frac{1}{2}\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 + \lambda\|\mathbf{E}\|_1, \quad s.t.\ rank(\mathbf{Z}) \leq r, \tag{2}$$

where λ denotes the regularization parameter. $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius norm and the ℓ_1-norm of a matrix, respectively.

To enhance the discriminative ability of the low-rank representation coefficient matrix, we further integrate into the model in Eq. (2) the discriminative replication prior based on internal video statistics. Discriminative replication prior exploits the small non-local recurring regions [22] to refine the affinity between supervoxels. Similar work on image segmentation was proposed in [23]. In our work, we assume that local small-size cubes (*e.g.*, $6 \times 6 \times 6$ voxels) tend to recur frequently within the same object, yet less frequently within the different objects. Further, the prior to video also benefits the preservation of temporal coherence and improvement on computational efficiency, as shown in Fig. 4.

We utilize the cube recurrence density to quantify the discriminative replication prior. Let Λ denote the spatio-temporal subregion. We first define the empirical density of small-size cube indexed by p with respect to Λ by Parzen window method:

$$D(p,\ \Lambda) = \frac{1}{|\Lambda|}\sum_{q\in\Lambda} \delta_\zeta(\kappa(\|\mathbf{x}_p - \mathbf{x}_q\|)), \tag{3}$$

where q indexes the small-size cubes, $\mathbf{x}_p$ and $\mathbf{x}_q$ are the features extracted from p and q, respectively, κ is a Gaussian kernel, and the function $\delta_\zeta(a)$ denotes the hard-threshold operator,

$$\delta_\zeta(a) = aI(|a| > \zeta), \tag{4}$$

where $I(\cdot)$ is the indicator function, and the threshold ζ is fixed to be 0.4 in this work. Parzen window method does not distinguish between the smaller number of perfectly similar patches and the larger number of partially similar patches. Therefore, introducing the hard-thresholding operator can depress the effects of partially similar patches.

Next, we define the discriminative replication prior to measure how likely two supervoxels belong to different semantic region. The discriminative replication prior matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is defined as follows:

$$Q_{ij} = e^{-(\frac{1}{|\Lambda_i|}\sum_{p\in\Lambda_i} D(p,\ \Lambda_j) + \frac{1}{|\Lambda_j|}\sum_{q\in\Lambda_j} D(q,\ \Lambda_i))}, \tag{5}$$

where Λ_i denotes the spatio-temporal subregion covered by supervoxel i, and $|\Lambda_i|$ indicates the number of cubes within Λ_i.

Then, we incorporate the discriminative replication prior into the model in Eq. (2):

$$\min_{\mathbf{Z},\mathbf{E}} \frac{1}{2}\|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 + \lambda\|\mathbf{E}\|_1 + \gamma\, tr(\mathbf{Z}^T\mathbf{Q}),$$
$$s.t.\ rank(\mathbf{Z}) \leq r, \tag{6}$$

where $tr(\cdot)$ returns the matrix trace, and γ is a tuning parameter. Note that larger Q_{ij} indicates that the supervoxel i and j belong to different semantic spatio-temporal regions with higher probability, and will encourage smaller Z_{ij} by minimizing the last term $tr(\mathbf{Z}^T\mathbf{Q})$. Therefore, minimizing $tr(\mathbf{Z}^T\mathbf{Q})$ prefers to enforce the coefficient matrix $\mathbf{Z}$ to be block diagonal, where $\mathbf{Z}_{ij}$ is zero if the supervoxel i and j are from different semantic regions, and vice versa. It is known that block-diagonal structure is critical for accurate subspace segmentation [32]. In this way, high-level semantic internal statistics can be incorporated as a soft constraint to enhance the discriminative ability.

The model in Eq. (6) is nonconvex due to the rank constraint, which is usually relaxed to a convex problem, *i.e.*, minimizing nuclear norm of $\mathbf{Z}$. In this way, model optimization can be performed using the Augmented Lagrangian Method (ALM) [19] or linearized ALM [33]. However, in many applications it is easier to explicitly determine the desired rank rather than implicitly tuning the tradeoff parameter of nuclear norm [34]. For example, rigid Structure From Motion (SFM) can be formulated as a rank-3 matrix factorization problem [35], [36], while nonrigid SFM can be formulated as a rank-3k matrix factorization, where k is the number of shape basis for depicting nonrigid deformation [37]. Moreover, as demonstrated in [38]–[40], the incorporation of explicit rank constraint may result in more efficient optimization algorithm. Therefore, unlike using the nuclear-norm regularizer in conventional LRR models [16]–[18], [20], [23], we explicitly impose the fixed-rank constraint on $\mathbf{Z}$.

We decompose the representation coefficient matrix as $\mathbf{Z} = \mathbf{AB}$, where $\mathbf{A} \in \mathbb{R}^{n\times r}$, $\mathbf{B} \in \mathbb{R}^{r\times n}$. By replacing $\mathbf{Z}$ with $\mathbf{AB}$, the Sub-Optimal Low-rank Decomposition (SOLD) model is then formulated as,

$$\min_{\mathbf{A},\mathbf{B},\mathbf{E}} \frac{1}{2}\|\mathbf{X}-\mathbf{XAB}-\mathbf{E}\|_F^2 + \lambda\|\mathbf{E}\|_1 + \frac{\beta}{2}\|\mathbf{AB}\|_F^2 + \gamma\ tr((\mathbf{AB})^T\mathbf{Q}), \quad (7)$$

where β is a regularization parameter that controls overfitting. Even SOLD is nonconvex and sub-optimal, as demonstrated in our experiments, such formulation can deliver both efficient algorithms and promising video segmentation accuracy.

B. Optimization

To optimize Eq. (7), we adopt the alternating optimization method, and denote

$$J(\mathbf{A},\mathbf{B},\mathbf{E}) = \frac{1}{2}\|\mathbf{X}-\mathbf{XAB}-\mathbf{E}\|_F^2 + \lambda\|\mathbf{E}\|_1 + \frac{\beta}{2}\|\mathbf{AB}\|_F^2 + \gamma\ tr((\mathbf{AB})^T\mathbf{Q}). \quad (8)$$

Given $\mathbf{E}$, taking the derivative of $J(\mathbf{A},\mathbf{B},\mathbf{E})$ w.r.t. $\mathbf{B}$, and setting it to zero, we obtain

$$\mathbf{B} = (\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2, \quad (9)$$

where

$$\mathbf{S}_1 = \mathbf{X}^T\mathbf{X} + \beta\mathbf{I},$$
$$\mathbf{S}_2 = (\mathbf{X}^T(\mathbf{X}-\mathbf{E}) - \gamma\mathbf{Q}). \quad (10)$$

By substituting Eq. (9) back into Eq. (7), the subproblem on $\mathbf{A}$ becomes

$$\mathbf{A}^* = \arg\max_{\mathbf{A}} tr\{(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\mathbf{S}_2^T\mathbf{A}\}. \quad (11)$$

Eq. (11) can be transformed to a generalized eigen-problem, where its global optimal solution is the top r eigenvectors of $\mathbf{S}_1^\dagger\mathbf{S}_2\mathbf{S}_2^T$ corresponding to the nonzero eigenvalues, where $\mathbf{S}_1^\dagger$ denotes the pseudo-inverse of $\mathbf{S}_1$.

Given $\mathbf{A}$ and $\mathbf{B}$, the noise matrix $\mathbf{E}$ can be solved by the soft-thresholding (or shrinkage) operator in [19]:

$$\mathbf{E}^* = \arg\min_{\mathbf{E}} \lambda\|\mathbf{E}\|_1 + \frac{1}{2}\|\mathbf{E}-(\mathbf{X}-\mathbf{XAB})\|_F^2. \quad (12)$$

Please refer to Appendix A for the detailed derivations of above equations. A sub-optimal solution can be obtained by alternating between the updating of $\{\mathbf{A}, \mathbf{B}\}$ and the updating of $\mathbf{E}$, and the algorithm is summarized in Alg. 1.

Although Alg. 1 is an iterative algorithm, we can guarantee its convergence to a stationary point. Note that both the $\mathbf{E}$ subproblem and the $\{\mathbf{A}, \mathbf{B}\}$ subproblem have unique closed-form solutions. Therefore, the generated sequence is monotone, *i.e.*, $J(\mathbf{A}^t,\mathbf{B}^t,\mathbf{E}^t) \geq J(\mathbf{A}^{t+1},\mathbf{B}^{t+1},\mathbf{E}^t) \geq J(\mathbf{A}^{t+1},\mathbf{B}^{t+1},\mathbf{E}^{t+1})$. Moreover, the sequence of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{E}$ in each iteration are bounded, see Appendix B for more details. As shown in of [41] ($Proposition$ 2.7.1 in page 268), if the solutions to each subproblem is unique, the accumulation point of the sequence generated by alternating minimization is a stationary point.

Algorithm 1 Optimization Procedure to Eq. (7)

Input: The supervoxel feature matrix $\mathbf{X}$, the discriminative replication prior matrix $\mathbf{Q}$, the low-rank r, the parameter λ, β and γ;
Set $\mathbf{E} = 0$; $\varepsilon = 10^{-8}$, $maxIter = 500$.
Output: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{E}$.
1: **while** not converged **do**
2: Update $\mathbf{A}$ by Eq. (11);
3: Update $\mathbf{B}$ by Eq. (9);
4: Update $\mathbf{E}$ by Eq. (12);
5: Check the convergence condition: the maximum element change of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{E}$ between two consecutive iterations is less than ε or the maximum number of iterations reaches $maxIter$.
6: **end while**

Our optimization delivers a more efficient algorithm. Some matrices (computing $\mathbf{S}_1$ and $\mathbf{X}^T\mathbf{X} - \gamma\mathbf{Q}$) in our algorithm can be pre-computed, and we require compute the top r generalization eigenvectors, where r is the desired rank. Note that r is generally much smaller than the size n of coefficient matrix $\mathbf{Z}$, making our algorithm more efficient to be optimized.

It should be noted that, both [38] and our SOLD adopt the alternating minimization algorithm, but the algorithm in [38] alternates between updating $\mathbf{A}$ and $\mathbf{B}$ while our SOLD alternates between updating $\mathbf{E}$ and $\{\mathbf{A}, \mathbf{B}\}$. Moreover, even for our specific subproblem on $\{\mathbf{A}, \mathbf{B}\}$, instead of the AltMin algorithm by [38], we suggest a generalized eigenvalue decomposition algorithm which can directly obtain the closed-form solutions to $\mathbf{A}$ and $\mathbf{B}$.

The low-rank coefficient matrix can be obtained by $\mathbf{Z} = \mathbf{AB}$, and will be combined with other low-level cues (*e.g.*, edge strength and spatial smoothness) to define the affinity between supervoxels in Sect. IV. Therefore, the optimized $\mathbf{Z}$ can be utilized to suppress the effects of data noise and/or corruption in video segmentation.

C. Implementation

Some important implementation details are briefly introduced. In this work, we utilize the hierarchical graph-based method (HGB) [4] to generate one layer supervoxels. HGB performs well on all the metrics of the unified video segmentation benchmarks [7], [8], and only involves one input parameter, *i.e.*, the total number n of supervoxels. Note that the supervoxel number n should not be set too small (large under-segmentation errors) or too large (heavy computational cost). On one hand, as shown in Fig. 2, the supervoxel segmentation result with $n = 100$ usually introduce large under-segmentation errors. On the other hand, as demonstrated in [7], when the supervoxel number n is between 200 and 900, the 3D under-segmentation error of HGB on the SegTrack [42] dataset only changes a little, and so do the other performance metrics including boundary recall, segmentation accuracy and explained variation. Therefore, it is reasonable to set $n \geq 200$. Moreover, considering that the computational complexity of SOLD (two SVD operations in each iteration) is $O(n^3)$,

we should let n as small as possible, and thus set $n = 200$ to balance the accuracy-efficiency tradeoff.

For robust supervoxel description, four low-level features are extracted from supervoxels and normalized with unit ℓ_2 norm. These feature vectors, including 12-dimension color histogram in each channel of RGB, 58-dimension Local Binary Pattern (LBP), 31-dimension Histogram of Oriented Gradient (HOG) and 18-dimension Histogram of Optical Flow (HOF), are concatenated into a single descriptor vector. To reduce computational complexity, we perform PCA [43] on the feature matrix $\mathbf{X}$ to remove insignificant components. The same operation is employed on computing the discriminative replication prior matrix $\mathbf{Q}$.

IV. Streaming Video Segmentation

In this section, we will deploy the optimized low-rank representation in Sect. III to perform streaming video segmentation in both unsupervised and interaction scenarios. First, the coefficient matrix $\mathbf{Z}$ is combined with other low-level cues to define the affinity matrix. Then, we apply NCut with temporal consistent constraints for clustering supervoxels. Finally, both unsupervised and interactive video segmentation can be conducted based on the supervoxel clustering results.

An effective streaming (sometimes called online as a synonym) algorithm can enable us to process an arbitrary-long video with limited memory and computational resources. Thus, it is essential to perform video segmentation in a streaming way. To this end, we segment the video in overlapping sliding windows. In particular, we consider both the temporal consistent constraints and low-rank representations to improve the long-range consistency and segmentation accuracy of the inference algorithm.

A. Affinity Definition

We define the affinity between two supervoxels as a linear combination of three cues:

$$\mathbf{W}_{ij} = \sum_{m=1}^{3} \omega^m \phi_{ij}^m, \tag{13}$$

where ϕ^m is the m-th affinity value in the feature space, and ω^m is the linear combination weight. In this work, ϕ^1 is the intervening contours kernel, defined as

$$\phi_{ij}^1 = e^{-\alpha^1 max_{x \in Lines(i,j)} \|Edge(x)\|}, \tag{14}$$

where $Lines(i, j)$ is a straight line set, in which each line joins centers of one within-frame superpixel-pair. Herein, supervoxel-pair (i, j) is decomposed into the within-frame superpixel-pairs. The $Edge(x)$ is the edge strength computed by gradient at location x and α^1 is a tuning parameter. ϕ^2 is the smoothness kernel defined as

$$\phi_{ij}^2 = e^{-\alpha^2 \|c_i - c_j\|^2}, \tag{15}$$

where c_i represents the centroid of the supervoxel i, and α^2 is the tuning parameter. And the third kernel ϕ^3 is defined as

$$\phi_{ij}^3 = e^{-\alpha^3 e^{\frac{-(|\mathbf{Z}_{ij}| + |\mathbf{Z}_{ji}|)/2}{2\sigma^2}}}, \tag{16}$$

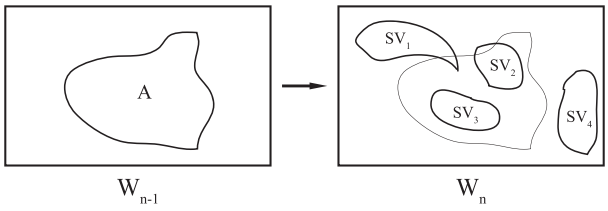


Fig. 5. The generation of the temporal consistent constraints between two neighboring sliding windows W_{n-1} and W_n. A denotes one segmentation region in W_{n-1}, and provides some constraints to the segmentation of W_n. For clarity, four typical supervoxels are shown here, which stand for four typical supervoxel types based on their relationship to the region A: complete (SV_3), almost (SV_2), part (SV_1) and none (SV_4). Thus, only SV_2 and SV_3 compose a partial grouping supervoxel set and generate a constraint due to A.

where $\mathbf{Z}_{ij}$ indicates the (i, j)-th element of the optimized lowest-rank representation in Sect. III, α^3 is the tuning parameter, and σ is the Gaussian parameter. The settings of all the parameters are described in Sect. V-A.

B. NCut With Temporal Consistent Constraints

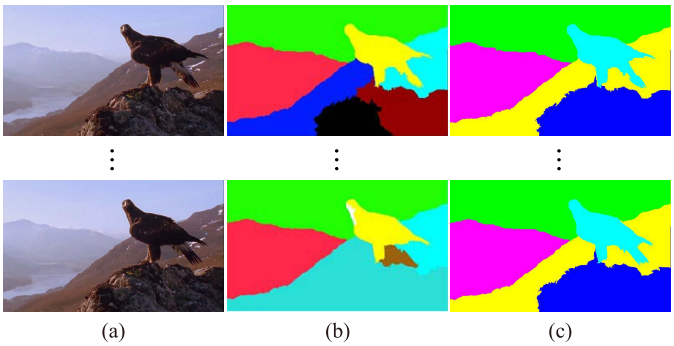
The temporal consistent constraints are introduced to properly propagate solutions between neighboring windows. We utilize some reasonable constraints to propagate the segmentation labels, while avoiding some bad results should not affect the quality of segmentation in the future frames. To this end, we divide the supervoxels into two categories as follows. Given segmentation labels of the current window, the supervoxels in the next are divided into the deterministic supervoxels and the non-deterministic supervoxels. More specifically, the deterministic supervoxel is defined as completely or almost (over 90% in this paper) belonging to one specific label, and the non-deterministic supervoxel is defined as partly belonging to some label. Then, the partial grouping supervoxel set is composed by only the deterministic supervoxels. Fig. 5 shows this process.

Given the partial grouping supervoxel set $\mathscr{U}_t$, we can obtain $|\mathscr{U}_t| - 1$ independent constraints, where $|\cdot|$ denotes the size of a set, and $t \in T$ indicates the label index. Then, the temporal consistent constraint matrix $\mathbf{U}$ is computed as follows: For each row k, there is two nonzero elements $\mathbf{U}_k(i) = 1$ and $\mathbf{U}_k(j) = -1$, where $i, j \in \mathscr{U}_t$ and $k \in [\sum_{t=1}^{T}(|\mathscr{U}_t| - 1)]$, $[n]$ indicates the set of integers between 1 and n: $[n] = \{1, 2, \ldots, n\}$. Alg. 2 summarizes this procedure, and Fig. 6 demonstrate its effectiveness.

Generally, conventional methods [16], [20], [23] directly employ NCut on affinity matrix to perform subspace segmentation. In our work, we aim to integrate the temporal consistent constraints in segmentation for improving temporal segmentation accuracy. To this end, we apply the constrained NCut method [24] on $\mathbf{W}$ to achieve the supervoxel-level segmentation. The tractable K-ways normalized segmentation criterion with temporal consistent constraints is formulated as

$$\begin{aligned} &\max_{\mathbf{G}} \ \frac{1}{K} tr(\mathbf{G}^T \mathbf{W} \mathbf{G}) \\ &s.t. \ \mathbf{U}\mathbf{G} = \mathbf{0}, \quad \mathbf{G}^T \mathbf{D} \mathbf{G} = \mathbf{I}_K, \end{aligned} \tag{17}$$

where $\mathbf{D} = \mathbf{W}\mathbf{1}_N$ is the degree matrix, and $\mathbf{G} = \mathbf{M}(\mathbf{M}^T \mathbf{D} \mathbf{M})^{-\frac{1}{2}}$ is the scaled partition matrix. N is total number



(a) (b) (c)

Fig. 6. Illustrations of the temporal consistent constraints for temporal consistency. Frame 41 to 61 of the video sequence *"planet_earth_1"* in the dataset VSB100 [8] are shown in (a), and the segmentation results without and with the temporal consistent constraints are shown in (b) and (c), respectively. The different colors indicate the different segmentation labels.

Algorithm 2 Temporal Consistent Constraint Matrix Computation Between Two Neighboring Windows

Input: Label set $\mathscr{T}$ from the previous window; Supervoxel set $\mathscr{S}$ in the current window.

Output: Temporal consistent constraint matrix $\mathbf{U}$.

```
for t = 1 : |T| do
    Find the deterministic supervoxel set U_t (⊆ S) for the label T(t) according to the overlap ratio of overlapping frame(s);
    k = 0;
    for s = 1 : |U_t| − 1 do
        k = k + 1;
        U(k, U_t(s)) = 1;
        U(k, U_t(s + 1)) = −1.
    end for
end for
```

of supervoxels. $\mathbf{1}$ and $\mathbf{I}$ denote all ones vector and identity matrix, respectively. $\mathbf{M} \in \{0,1\}^{N\times K}$ is the partition matrix. The optimization of Eq. (17) has been addressed in [24], and the main results are as follows. Let $\mathbf{P}$ be the row-normalized weight matrix and $\mathbf{H}$ be a projector onto the feasible solution space:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}, \quad \mathbf{H} = \mathbf{I} - \mathbf{U}^{-1}\mathbf{U}^{T}(\mathbf{U}\mathbf{D}^{-1}\mathbf{U}^{T})^{-1}\mathbf{U}. \tag{18}$$

Let $\mathbf{V}_{[K]}$ be the first K eigenvectors of the matrix $\mathbf{HPH}$, then the solutions of Eq. (17) are $\hat{\mathbf{V}}_{[K]} = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}_{[K]}$, which indicates the desired segments. Next, we apply $\hat{\mathbf{V}}_{[K]}$ to two video segmentation tasks.

C. Unsupervised Video Segmentation

The eigenvectors $\hat{\mathbf{V}}_{[K]}$ can be discretized by spectral rotation [44] or k-means (spectral rotation in this paper) to obtain the discrete solutions of graph partition.

In order to create new labels or remove old labels when the objects enter or leave the camera view, we utilize a reasonable strategy to determine the label mapping by their spatial overlap [6]. An overlap of one frame between neighboring windows is used to determine whether current labels are new ones or mapped from previous ones. For simplicity, the overlaps between new labels (from the current processing window) and old labels (from the preceding processing window) are measured by their Dice coefficients. For a current label l, it is mapped from previous one if it significantly overlaps with some previous label p, but barely overlaps with any other previous label q. Otherwise, it is considered as new one s, *i.e.*, a new object:

$$l = \begin{cases} p, & if\ o(l,\ p) > o_1\ and\ o(l,\ q) < o_2 \\ s, & else, \end{cases} \tag{19}$$

where $o(\cdot,\ \cdot)$ denotes the Dice coefficient in overlap between two labels, and o_1, o_2 are fixed parameters, which is set to be 0.8 and 0.2, respectively.

D. Interactive Object Segmentation

For the applications which utilize priors from user interactions, we employ an energy minimization approach to achieve interactive object segmentation. Since objects are spatially compact and temporally consistent, we integrate the appearance model of foreground and background by user interactions, the spatio-temporal smoothness constraints and the low-rank representation into our framework to accurately segment the target object. To this end, we formulate it as the MRF model, and the energy function is defined as:

$$\begin{aligned} \min_{\pi} & \sum_{k=0}^{1}\sum_{i=1}^{n}\bar{\delta}(k,\pi_i)\bar{A}_k(i) + \xi_1\sum_{k=0}^{1}\sum_{i\in C_k}\bar{\delta}(k,\pi_i) \\ & + \xi_2\sum_{k=0}^{1}\sum_{i=1}^{n}\bar{\delta}(k,\pi_i)\bar{d}_k(i) + \xi_3\sum_{i=1}^{n}\sum_{j=\mathscr{N}(i)}^{n}\bar{\delta}(\pi_i,\pi_j)\bar{d}(i,j), \end{aligned} \tag{20}$$

where $\pi \in \{0,1\}$ and $\bar{\delta}$ are the assignment function and the Dirac delta function, respectively. $\mathscr{N}(i)$ denotes the neighboring node set of node i and C_k indicates the penalty node set of segment k. ξ_1, ξ_2 and ξ_3 are the weighted parameters. The appearance model $\bar{A}$ consists of two Gaussian Mixture Models (GMMs) over RGB values, one for the foreground and one for the background. The parameters of GMMs are estimated from manually labeled pixels in first window. Since the appearance of the foreground and background typically changes smoothly over time, we update these models over time in later windows by employing the segmentation results. $\bar{d}_k(i)$ and $\bar{d}(i,j)$ are a unary potential function indicating the cost of node i belonging to segment k and a pairwise potential function denoting the cost of node i and j belonging to a same segment, and defined as:

$$\bar{d}_k(i) = \max(\hat{\mathbf{V}}_{i,:}) - \hat{\mathbf{V}}_{i,k}, \tag{21}$$

$$\bar{d}(i,j) = \frac{1}{\|\hat{\mathbf{V}}_{i,:} - \hat{\mathbf{V}}_{j,:}\|_2^{2\eta}}, \tag{22}$$

where η is a parameter to control the penalty gap between large and small $\hat{\mathbf{V}}$ differences, and fixed to be 2 in our experiments. In Eq. (20), the first term evaluates how likely a supervoxel is to be foreground or background according to the appearance model; the second term reinforces that the labeled

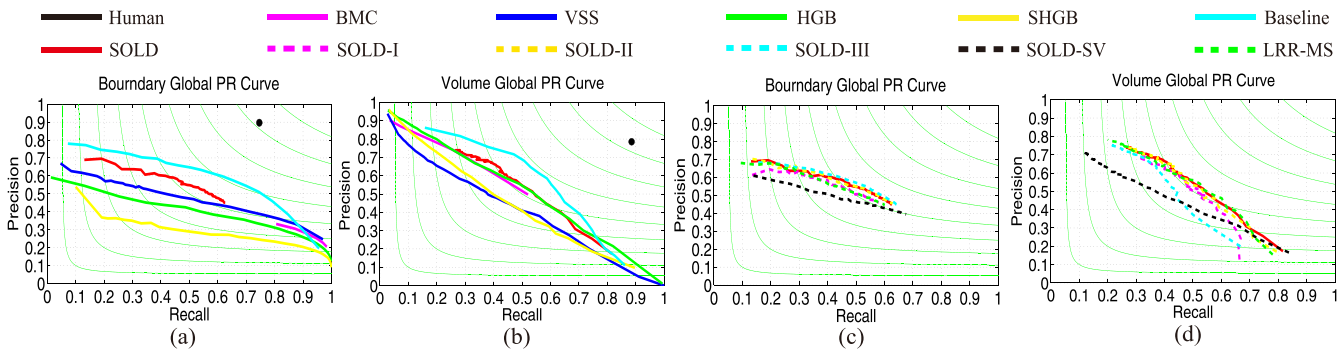


Fig. 7. Comparison curves of SOLD with its variants and other video segmentation approaches, including BMC [26], VSS [3], HGB [4], SHGB [5] and Baseline [8]. The first two subfigures are the comparison curves for comparing SOLD with previous works, and the last two subfigures are the comparison curves for comparing SOLD with its variants. (a) and (b) show Boundary Precision-Recall (BPR) and Volume Precision-Recall (VPR) curves, respectively, on VSB100 dataset. (c) and (d) show the component evaluations of our approach. See text for more details.

supervoxels should be assigned with correct labels; the third term represents the total distance between the supervoxels and their corresponding segments, and the fourth term encodes the spatial non-smoothness [45]. We adopt the Primal-Dual solver in MRF framework introduced in [25] to optimize it due to its high accuracy and efficiency.

For obtaining reliable interactive segmentation in later sliding windows, we propagate the user interactions over time by supervoxel propagation in overlapping frame(s), instead of optical flow propagation because of its incorrect estimation. Furthermore, we divide the penalty node set C_k into the interactive node set (reliable) and the propagated node set (less reliable) from preceding segmentation results, and empirically set ξ_1 to be 10^{10} and 10^2, respectively. Herein, the process of supervoxel propagation is the same as generation of the temporal consistent constraints.

V. Experimental Results

In this section, we evaluate our approach on two challenging datasets VSB100 [8] and SegTrack [42], and compare with other video segmentation methods. Then, we further analyze the effectiveness of the main components of our approach. At last, the efficiency analysis is discussed.

A. Evaluation Settings

To make the comparison comprehensive, we employ the segment number set $\{2, 3, \ldots, 51\}$ to produce multilevel segmentation results in unsupervised settings, and fix all parameters in all evaluations: we empirically set $\{\lambda, \beta, \gamma\} = \{0.5, 0.5, 0.05\}$ in optimization, and $\{\omega^1, \omega^2, \omega^3, \alpha^1, \alpha^2, \alpha^3, \sigma\} = \{0.4, 0.3, 0.3, 30, 0.6, 10, 0.12\}$ in affinity definition [18]. Followed by [45], we set $\{\xi_1, \xi_2, \xi_3\} = \{10^{10}(10^2), 10, \frac{10^{-3}}{n^2\sqrt{n}}\}$ in MRF (see III-C for details of setting ξ_1), where n denotes the number of supervoxels. In addition, the number of frames per window is set to be 6, and one frame is overlapped between neighboring windows. Generally, the rank r of representation coefficient matrix is in the range of $[k/2, k)$, where k is the number of classes [39]. In our experiments, the segmentation performance is slightly different with respect to different r in $[6, 12)$, where 12 indicates the average number of classes on the VSB100 [8]. Therefore, we simply fixed r to be 10 in unsupervised settings, and 2 in interactive settings.

TABLE I

The Aggregation Measures of Boundary Precision-Recall (BPR) and Volume Precision-Recall (VPR) for Comparing Previous Works With SOLD on Dataset VSB100 [8]. (*) Denotes Evaluated on Video Frames Resized by 0.5 Due to Large Computational Demands and the Italic Denotes the Streaming Method. The Bold Fonts Indicate the Best Performance

	BPR			VPR		
Algorithm	**ODS**	**OSS**	**AP**	**ODS**	**OSS**	**AP**
*BMC [26]	0.47	0.48	0.32	0.51	0.52	0.38
*VSS [3]	0.51	0.56	**0.45**	0.45	0.51	0.42
*HGB [4]	0.47	0.54	0.41	0.52	0.55	**0.52**
SHGB [5]	0.38	0.46	0.32	0.45	0.48	0.44
**SOLD*	**0.54**	**0.57**	0.39	**0.53**	**0.59**	0.47
Human	0.81	0.81	0.67	0.83	0.83	0.70
Baseline [8]	0.61	0.65	0.59	0.59	0.62	0.56

B. Exp-I: Unsupervised Video Segmentation

The selected VSB100 [8] for empirical evaluation is very challenging. It is the largest video segmentation dataset with high definition frames, and consists of four difficult sub-datasets: general, motion segmentation, non-rigid motion segmentation and camera motion segmentation. Using the same setting as [8], we regard the general sub-dataset (60 video sequences) as our test set for all the approaches.

1) Comparison Results: We compare our method [46] with four unsupervised video segmentation methods, including BMC [26], VSS [3], HGB [4] and SHGB [5]. The first two subfigures of Fig. 7 illustrate the Boundary Precision-Recall (BPR) and Volume Precision-Recall (VPR) curves of the comparisons on the VSB100 dataset. Tab. I gives a summary of the aggregation performance evaluations, which includes Optimal Dataset Scale (ODS), Optimal Segmentation Scale (OSS) and Average Precision (AP) of BPR and VPR. Herein, the baseline introduced by [8] is the extension of a state-of-the-art image segmentation method [47] by propagating the segmentation [47] of the central frame to the other frames with optical flow [48] and labelling the image segments (across the hierarchy) with maximum voting. This baseline

Fig. 8. Qualitative comparisons of SOLD against other video segmentation methods HGB [4] and SHGB [5]. For clarity, the last two rows show one frame results of two challenging samples. We can see that our method qualitatively improves on HGB, and substantially outperforms SHGB. (a) Image. (b) HGB. (c) SHGB. (d) SOLD. (e) GT.

adopted more complex image features, and was introduced to develop video segmentation methods referring to image segmentation methods and exploiting additional cues like motion.

From Fig. 7 and Tab. I, we can conclude that our approach achieves comparable performance against other compared approaches in both BPR and VPR on the VSB100 dataset. Specifically, our approach achieves best ODS and OSS values in both BPR and VPR. Though exploiting more informative cues as VSS, our approach performs better for its insensitivity to noise. This owes to the proposed sub-optimal low-rank decomposition of representation coefficient matrix of supervoxel features. Besides, the temporal consistent constraints adopted by our method bring better performance than other methods in VPR. It is also worth noting that SHGB is also a streaming mode. These superior performances demonstrate that our approach can not only effectively infer the affinities between supervoxels, but also preserve the longer-range temporal consistency in a streaming mode. In addition, the qualitative comparisons of our approach and previous works are shown in Fig. 8 to demonstrate the superior performance of our framework.

Though our approach has achieved superior performance, its AP in both BPR and VPR is lower than some of the state-of-the-arts (VSS and HGB). This is due to the low recall caused by the small maximum supervoxel number for over-segmentation. As a matter of fact, we can alleviate it by simply increasing the supervoxel number. However, to balance the accuracy-efficiency trade-off, we will develop an adaptive version of SOLD in our future work.

2) Component Analysis: To justify the significance of the main components of our approach, we implement three special versions and two variants of our approach for empirical analysis. They are: 1) *SOLD-I*, that sets $\omega^3 = 0$ to remove

TABLE II

THE AGGREGATION MEASURES OF BOUNDARY PRECISION-RECALL (BPR) AND VOLUME PRECISION-RECALL (VPR) FOR COMPARING SOLD WITH ITS VARIANTS ON DATASET VSB100 [8]. THE DESCRIPTION OF THIS TABLE IS THE SAME AS TABLE I

	BPR			VPR		
Algorithm	**ODS**	**OSS**	**AP**	**ODS**	**OSS**	**AP**
**SOLD*	**0.54**	**0.57**	0.39	**0.53**	**0.59**	**0.47**
**SOLD-I*	0.51	0.55	0.34	0.51	0.58	0.39
**SOLD-II*	0.53	**0.57**	0.39	0.52	0.58	0.46
**SOLD-III*	**0.54**	**0.57**	**0.41**	0.47	0.54	0.38
**SOLD-SV*	0.51	0.55	0.36	0.45	0.50	0.39
**LRR-MS [18]*	0.52	0.56	0.37	**0.53**	**0.59**	**0.47**

the affinity term inferred by the sub-optimal low-rank decomposition in streaming segmentation approach. 2) *SOLD-II*, that sets $\gamma = 0$ in Eq. (7) to remove the regularization term of the discriminative replication prior in our approach. 3) *SOLD-III*, that sets $\bar{\mathbf{U}} = 0$ to perform segmentation without the temporal consistent constraints. 4) *SOLD-SV*, that substitutes the optimal affinities optimized by the sub-optimal low-rank decomposition with the affinities based on feature descriptors, *i.e.* letting $\phi^3_{ij} = e^{-\alpha^4 \|\mathbf{x}_i - \mathbf{x}_j\|^2}$ in Eq. (16), where α^4 is empirically set to be 0.5 in our implementation. 5) *LRR-MS* [18], that enforces multiscale consistency between multilayers, and is solved by ALM method [19]. Specifically, we implement *LRR-MS* as following three steps. Firstly, three level supervoxel representations of video are generated via HGB [4] with the supervoxel numbers of 200, 150, 100, respectively. Secondly, we introduce both the cross-scale consistent constraint matrix and the discriminative replication prior matrix into the LRR model, where the cross-scale consistent constraint matrix is obtained based on [23] to enforce consistency of representation matrices at different scales, and the discriminative replication prior matrix is obtained based Eq. (5). Thirdly, we integrate it into our streaming framework to facilitate the evaluation.

The last two subfigures of Fig. 7 show the components evaluation of our approach, and corresponding aggregation measures are reported in Tab. II. From Fig. 7 and Tab. II, we can make some observations and conclusions as follows. 1) The complete approach outperforms SOLD-I in both BPR and VPR. This justifies the significance of the low-rank representation optimized by SOLD. 2) Comparing to the complete approach, SOLD-II has a little performances drop in BPR and VPR. This demonstrates the contribution of the discriminative replication prior. 3) Though worse than SOLD-III in BPR, our approach with the temporal consistent constraints substantially improves the performance in VPR, *i.e.,* keeping longer-range temporal consistency. 4) Our approach outperforms SOLD-SV in both BPR and VPR, and it shows that the representations inferred by the sub-optimal low-rank decomposition can alleviate the noises of low-level features effectively. It is worth noting that VPR is greatly affected by noises in our approach. 5) Our approach obtains better results than LRR-MS. This validates that the multiscale consistency constraints may not help to improve the segmentation results due to error propagation as we previously discussed.

TABLE III

QUANTITATIVE EVALUATION OF THE INTERACTIVE VIDEO SEGMENTATION ON THE SEGTRACK DATASET [42]. THE SCORE IS THE AVERAGE LABEL MISMATCH PER FRAME. THE BOLD FONTS INDICATE THE BEST PERFORMANCE

Sequence	**RotoBrush** [10]	**PF-MRF** [49]	**SOLD**
birdfall2	462	405	**291**
cheetah	1553	**1288**	1508
girl	5520	8575	**4981**
monkeydog	**816**	1225	1203
parachute	569	1042	**460**
penguin	708	**482**	1042
Average	1605	2170	**1581**

We also evaluate the streaming settings of our approach. 1) Setting the number of frames in a window as 4 and 6, we find that the results are slightly different with the current setting (The former obtains 0.01 higher in OSS of BPR, and the latter obtains 0.01 lower in ODS of VPR). 2) Setting the number of the overlapped frames as 2, we find that the results are slightly helpful to propagate the segmentation while reducing the values of BPR (0.01 higher in OSS of VPR, and 0.02 lower in ODS of BPR).

C. Exp-II: Interactive Object Segmentation

We further evaluate our approach on the SegTrack [42] dataset under the interactive settings. The SegTrack consists of six challenging video sequences, that were between 21 and 71 frames in length. With respect to the challenging measures (color, motion, and shape), these video sequences can be characterized as: low-low-low (*parachute*), low-low-high (*girl*), low-high-high (*monkeydog*), high-low-low (*penguin*), high-high-low (*birdfall2*), and high-high-high (*cheetah*).

For the evaluation of the interactive video segmentation scenario, we make a few scribbles on the initial frame, and no other user interactions are applied for the other frames. We compare our approach against the high-performing interactive video segmentation method from the literature [10], which is a local classifier based segmentation method and included in Adobe After Effects CS5 as the *roto-brush tool*. We call this method as RotoBrush in this paper. We also compare our approach to the MRF-based algorithm PF-MRF [49], which employs the Pixel-Flow MRF (PF-MRF) for propagation. In this paper, we discard the step of actively selecting frames for labelling in PF-MRF for fair comparison.

Tab. III shows the quantitative results of our approach compared to RotoBrush and PF-MRF. From Tab. III, our approach outperforms them in 3 of the 6 video sequences. Specifically, our approach can propagate the foreground well in *birdfall2*, in which the background is extremely clutter and significantly overlaps appearance of the foreground in many frames. Our approach accurately segments the foreground of *parachute* in spite of its large illumination variation. Though our approach outperforms other methods in *girl*, the segmentation quality of our method is still poor due to under-segmentation in the supervoxels. Fig. 9 shows some representative examples.

Our weaker performance on *cheetah*, *monkeydog* and *penguin* is due to the failure of supervoxel segmentation, which

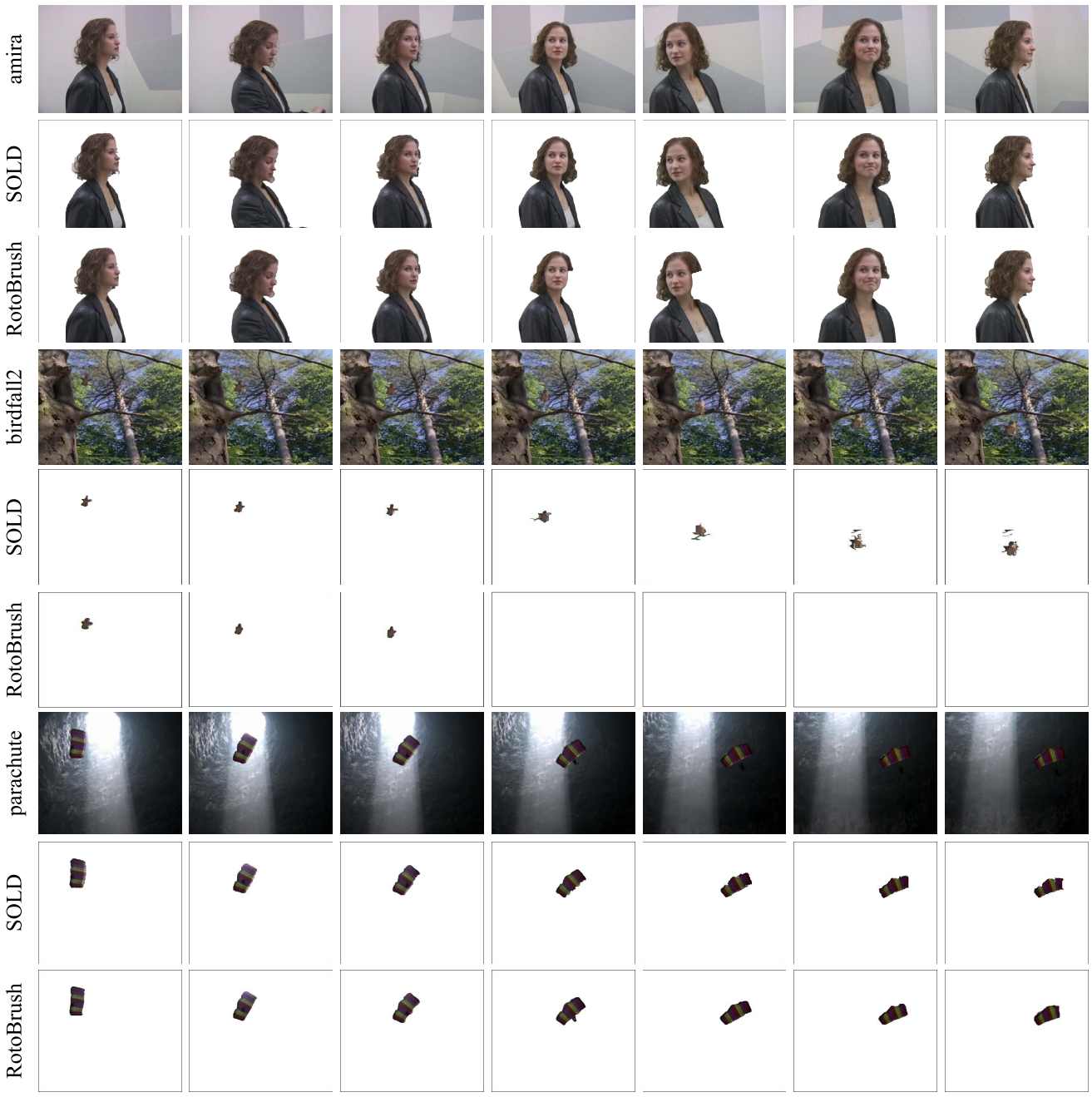


Fig. 9. Qualitative results on three sequences against the state-of-the-art approach RotoBrush [10]. The first sequence is from [9], and the last two sequences are from the SegTrack [42].

may attributes to fast camera motion, strong non-rigid deformations, similar appearance between the foreground and the background or incorrect optical flow estimation.

In addition, we present some unsatisfying results generated by our approach. The graph-based video over-segmentation algorithm [4] adopted by our framework as preprocessing step is sensitive to noises and/or corruptions, and usually introduces under-segmentation errors. In such circumstance, it will lead to the bad segmentation quality of our framework. Fig. 10 illustrates one quintessential case.

D. Efficiency Analysis

Runtime of our approach against other methods is presented in Tab. V. It is worth mentioning that our approach is faster than the original HGB due to two main reasons. First, we employ HGB in a streaming way instead of batch processing. Second, we just generate the fine supervoxels.

To further explore whether the proposed sub-optimal low-rank decomposition is more efficient than the widely used ALM method, we further compare the time efficiency of SOLD with LRR-MS. Herein, LRR-MS and SOLD refer to solving their respective low-rank problems. The experiments are carried out on a desktop with an Intel i7 3.4GHz CPU and 10GB RAM, and implemented on mixing platform of C++ and MATLAB without any optimization. Fig. 11 shows the convergence curves of LRR-MS and SOLD, and Tab. IV reports their average iterations and running time. Thanks to the proposed sub-optimal low-rank decomposition, it only

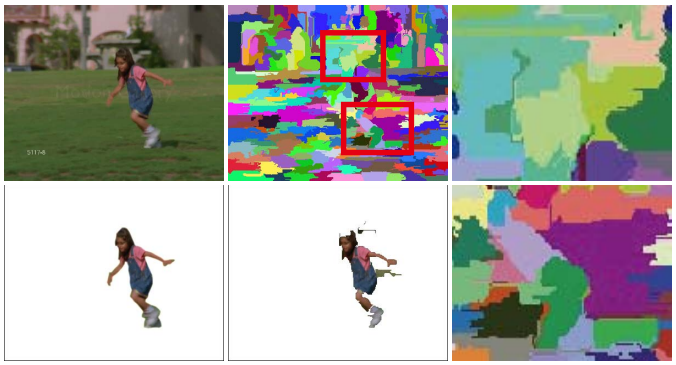


Fig. 10. Illustrated unsatisfying result on the video sequence *girl*. The first column shows one original frame and corresponding ground truth, respectively. And the second column denotes over-segmentation and our result, respectively. The amplified images indexed by the red box in over-segmentation result are presented in last column.

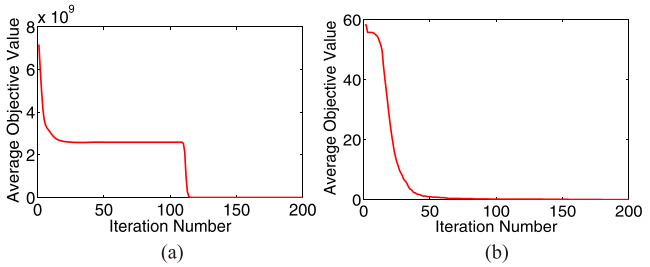


Fig. 11. Convergence curves of LRR-MS and SOLD on dataset VSB100 [8]. (a) LRR-MS. (b) SOLD.

TABLE IV

THE AVERAGE ITERATIONS AND RUNNING TIME (SECONDS PER FRAME) OF LRR-MS AND SOLD

Algorithm	LRR-MS [18]	SOLD
Iteration Number	112	16
Running Time	2.40	0.12

TABLE V

RUNNING TIME (SECONDS PER FRAME) OF OUR FRAMEWORK AGAINST OTHER METHODS

Method	HGB [4]	SHGB [5]	SOLD framework
Running Time	5.43	10.87	3.03

costs 0.12 sec./frame for SOLD, which converges faster than LRR-MS (see Fig. 11), and brings 20-times over it (see Tab. IV).

We also report runtime of other main procedures in SOLD with the typical resolution of 640 × 360 pixels. 1) The features extraction, including the discriminative replication prior matrix $\mathbf{Q}$ calculation, takes approximately 0.35 second per frame. 2) The constrained NCut is efficiently solved within 0.01 second per frame due to the proposed supervoxel-level segmentation. 3) The MRF minimization problem in interactive settings takes about 0.30 second per frame. 4) The graph-based over-segmentation algorithm [4] is mostly time consuming procedure, which costs approximately 2.24 second per frame (about 74%). Hence, through introducing the efficient over-segmentation algorithms, we can achieve much better computation time under our approach.

VI. CONCLUSION

In this paper, we have proposed a general algorithm for low-rank representation pursuit by decomposing the matrix with the fixed rank and proved that a sub-optimal solution can be achieved by alternating closed-form optimization. Based on this algorithm, we have developed an effective and efficient approach that automatically segments streaming videos in both unsupervised and interactive way. In future work, we will improve our video segmentation framework by introducing more robust video features or deep feature learning methods [50]. Our low-rank decomposition algorithm can be also extended to other vision tasks such as multi-object tracking and saliency detection.

APPENDIX A
OPTIMIZATION TO SOLD

Given $\mathbf{E}$, taking the derivative of $J(\mathbf{A},\mathbf{B},\mathbf{E})$ w.r.t. $\mathbf{B}$, and setting it to zero, we obtain

$$-\mathbf{A}^T\mathbf{X}^T(\mathbf{X}-\mathbf{X}\mathbf{A}\mathbf{B}-\mathbf{E})+\beta\mathbf{A}^T\mathbf{A}\mathbf{B}+\gamma\,\mathbf{A}^T\mathbf{Q}=0. \quad (23)$$

According to Eq. (10), Eq. (23) can be rewritten as Eq. (9). By substituting Eq. (9) back into Eq. (7), the subproblem on $\mathbf{A}$ becomes

$$\begin{aligned}\min_{\mathbf{A}}\ &\frac{1}{2}\|(\mathbf{X}-\mathbf{E})-\mathbf{X}\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\|_F^2\\&+\frac{\beta}{2}\|\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\|_F^2\\&+\gamma\ tr(\mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Q}).\end{aligned} \quad (24)$$

Note that $\|\mathbf{x}\|_F^2=tr(\mathbf{x}^T\mathbf{x})$, we have

$$\begin{aligned}\min_{\mathbf{A}}\ &tr((\mathbf{X}-\mathbf{E})^T(\mathbf{X}-\mathbf{E})\\&-2(\mathbf{X}-\mathbf{E})^T\mathbf{X}\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\\&+\mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X}^T\mathbf{X}\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2)\\&+\beta tr(\mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2)\\&+2\gamma\ tr(\mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Q}).\end{aligned} \quad (25)$$

Merging the third and the fourth term, we have

$$\begin{aligned}\min_{\mathbf{A}}\ &tr((\mathbf{X}-\mathbf{E})^T(\mathbf{X}-\mathbf{E})\\&-2(\mathbf{X}-\mathbf{E})^T\mathbf{X}\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\\&+\mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2)\\&+2\gamma\ tr(\mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Q}).\end{aligned} \quad (26)$$

Substituting first $\mathbf{S}_2$ to $\mathbf{X}^T(\mathbf{X}-\mathbf{E})-\gamma\,\mathbf{Q}$ in the third term of Eq. 26 and employing $tr(\mathbf{x}^T)=tr(\mathbf{x})$, we obtain

$$\begin{aligned}\min_{\mathbf{A}}\ &tr((\mathbf{X}-\mathbf{E})^T(\mathbf{X}-\mathbf{E})\\&-(\mathbf{X}-\mathbf{E})^T\mathbf{X}\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\\&-\gamma\,\mathbf{Q}^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2)\\&+2\gamma\ tr(\mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Q}),\end{aligned} \quad (27)$$

and it equals to

$$\begin{aligned}\min_{\mathbf{A}}\ &tr((\mathbf{X}-\mathbf{E})^T(\mathbf{X}-\mathbf{E})\\&-(\mathbf{X}-\mathbf{E})^T\mathbf{X}\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\\&+\gamma\,\mathbf{Q}^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2).\end{aligned} \quad (28)$$

Thus, we have

$$\min_{\mathbf{A}} tr((\mathbf{X}-\mathbf{E})^T(\mathbf{X}-\mathbf{E}) - \mathbf{S}_2^T\mathbf{A}(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2). \quad (29)$$

According to Eq. (29), we utilize the fact that $tr(\mathbf{xy}) = tr(\mathbf{yx})$, and solve $\mathbf{A}$ by Eq. (11).

Eq. (11) can be transformed to a generalized eigen-problem. Its global optimal solution is the top r eigenvectors of $\mathbf{S}_1^\dagger\mathbf{S}_2\mathbf{S}_2^T$ corresponding to the nonzero eigenvalues, where $\mathbf{S}_1^\dagger$ denotes the pseudo-inverse of $\mathbf{S}_1$.

Given $\mathbf{A}$ and $\mathbf{B}$, the optimization of $\mathbf{E}$ is written as Eq. (12), which can be solved by the soft-threshold (or shrinkage) method in [19].

A sub-optimal solution can be obtained by alternating between the updating of $\{\mathbf{A}, \mathbf{B}\}$ and the updating of $\mathbf{E}$.

APPENDIX B
PROOF OF BOUNDEDNESS

From Eq. (12) and Eq. (9), we have

$$\begin{aligned}
||\mathbf{E}_{k+1}||_F &= ||\mathbb{S}_\lambda(\mathbf{X}-\mathbf{X}\mathbf{A}_{k+1}\mathbf{B}_{k+1})||_F \\
&\le ||\mathbf{X}-\mathbf{X}\mathbf{A}_{k+1}\mathbf{B}_{k+1}||_F \\
&= ||\mathbf{X}-\mathbf{X}\mathbf{A}_{k+1}(\mathbf{A}_{k+1}^T\mathbf{S}_1\mathbf{A}_{k+1})^{-1}\mathbf{A}_{k+1}^T\mathbf{S}_{2,k}||_F \\
&= ||\mathbf{X}-\mathbf{X}\mathbf{S}_1^{-1}(\mathbf{X}^T\mathbf{X}-\gamma\mathbf{Q}) + \mathbf{X}\mathbf{S}_1^{-1}\mathbf{X}^T\mathbf{E}_k||_F \\
&= ||\mathbf{K}(\mathbf{I}-\mathbf{L})^{-1}(\mathbf{I}-\mathbf{L}^k) + \mathbf{L}^k\mathbf{E}_1||_F \\
&\le ||\mathbf{K}(\mathbf{I}-\mathbf{L})^{-1}||_F(1+||\mathbf{L}^k||_F) + ||\mathbf{L}^k||_F||\mathbf{E}_1||_F,
\end{aligned} \quad (30)$$

where $\mathbb{S}_\lambda(\cdot)$ denotes the soft-thresholding operation with parameter λ. $\mathbf{K} = \mathbf{X}-\mathbf{X}\mathbf{S}_1^{-1}(\mathbf{X}^T\mathbf{X}-\gamma\mathbf{Q})$ and $\mathbf{L} = \mathbf{X}\mathbf{S}_1^{-1}\mathbf{X}^T$. Since $||\mathbf{L}||_2 < 1$, $||\mathbf{L}^k||_F \to 0$ when $k \to \infty$. Thus, $\{\mathbf{E}_k\}$ is bounded.

Since $\{\mathbf{E}_k\}$ is bounded, $\mathbf{S}_{2,k}$ is bounded. Besides, $\mathbf{S}_1$ is constant, and thus $\mathbf{S}_1^\dagger\mathbf{S}_{2,k}\mathbf{S}_{2,k}^T$ is bounded. Therefore, $\mathbf{A}_{k+1}$, the top r eigenvectors of $\mathbf{S}_1^\dagger\mathbf{S}_{2,k}\mathbf{S}_{2,k}^T$ corresponding to the nonzero eigenvalues, is also bounded. According to Eq. (9), we obtain $\mathbf{B}_{k+1} = \mathbf{A}_{k+1}^T\mathbf{S}_1^{-1}\mathbf{S}_{2,k}$. $\{\mathbf{B}_k\}$ is bounded due to the boundedness of $\{\mathbf{A}_k\}$ and $\mathbf{S}_{2,k}$.

REFERENCES

[1] S. Paris, "Edge-preserving smoothing and mean-shift segmentation of video streams," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 460–473.

[2] K. Fragkiadaki, G. Zhang, and J. Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1846–1853.

[3] F. Galasso, R. Cipolla, and B. Schiele, "Video segmentation with superpixels," in *Proc. Asian Conf. Comput. Vis.*, 2012, pp. 760–774.

[4] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2141–2148.

[5] C. Xu, C. Xiong, and J. J. Corso, "Streaming hierarchical video segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 626–639.

[6] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller, "Multiple hypothesis video segmentation from superpixel flows," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 268–281.

[7] C. Xu and J. J. Corso, "Evaluation of super-voxel methods for early video processing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1202–1209.

[8] F. Galasso, N. S. Nagaraja, T. J. Cardenas, T. Brox, and B. Schiele, "A unified video segmentation benchmark: Annotation, metrics and analysis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3527–3534.

[9] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, and M. F. Cohen, "Interactive video cutout," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 585–594, Jul. 2005.

[10] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video SnapCut: Robust video object cutout using localized classifiers," *ACM Trans. Graph.*, vol. 28, no. 3, Aug. 2009, Art. no. 70.

[11] B. L. Price, B. S. Morse, and S. Cohen, "LIVEcut: Learning-based interactive video segmentation by evaluation of multiple propagated cues," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 779–786.

[12] R. Dondera, V. Morariu, Y. Wang, and L. Davis, "Interactive video segmentation using occlusion boundaries and temporally coherent superpixels," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Mar. 2014, pp. 784–791.

[13] L. Lin, W. Yang, C. Li, J. Tang, and X. Cao, "Inference with collaborative model for interactive tumor segmentation in medical image sequences," *IEEE Trans. Cybern.*, doi: 10.1109/TCYB.2015.2489719, 2016.

[14] V. Badrinarayanan, F. Galasso, and R. Cipolla, "Label propagation in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3265–3272.

[15] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei, "Discriminative segment annotation in weakly labeled video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2483–2490.

[16] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2439–2446.

[17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.

[18] X. Liu, Q. Xu, J. Ma, H. Jin, and Y. Zhang, "MsLRR: A unified multiscale low-rank representation for image segmentation," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2159–2167, May 2014.

[19] Z. Lin, A. Ganesh, J. Wright, M. Chen, L. Wu, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," Univ. Illinois Urbana–Champaign, Champaign, IL, USA, Tech. Rep. UILU-ENG-09-2214, 2009.

[20] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 1–8.

[21] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2790–2797.

[22] A. Faktor and M. Irani, "Video segmentation by non-local consensus voting," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014, pp. 1–12.

[23] X. Liu, L. Lin, and A. L. Yuille, "Robust region grouping via internal patch statistics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1931–1938.

[24] S. X. Yu and J. Shi, "Segmentation given partial grouping constraints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 173–183, Feb. 2004.

[25] N. Komodakis, G. Tziritas, and N. Paragios, "Performance vs computational efficiency for optimizing single and dynamic MRFs: Setting the state of the art with primal-dual strategies," *Comput. Vis. Image Understand.*, vol. 112, no. 1, pp. 14–29, Oct. 2008.

[26] J. J. Corso, E. Sharon, S. Dube, S. El-Saden, U. Sinha, and A. Yuille, "Efficient multilevel brain tumor segmentation with integrated Bayesian model classification," *IEEE Trans. Med. Imag.*, vol. 27, no. 5, pp. 629–640, May 2008.

[27] L. Lin, X. Wang, W. Yang, and J.-H. Lai, "Discriminatively trained And-Or graph models for object shape detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 959–972, May 2015.

[28] F. Galasso, M. Keuper, T. Brox, and B. Schiele, "Spectral graph reduction for efficient image and streaming video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 49–56.

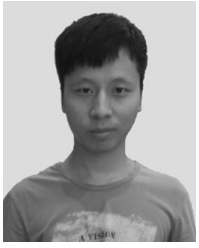
[29] K. Wang, L. Lin, J. Lu, C. Li, and K. Shi, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3019–3033, Oct. 2015.

[30] L. Lin, K. Wang, W. Zuo, M. Wang, J. Luo, and L. Zhang, "A deep structured model with radius–margin bound for 3D human activity recognition," *Int. J. Comput. Vis.*, doi: 10.1109/TCYB.2015.2489719, 2016.

[31] T. Tarabalka, G. Charpiat, L. Brucker, and B. H. Menze, "Spatio-temporal video segmentation with shape growth or shrinkage constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3829–3840, Sep. 2014.

[32] J. Feng, Z. Lin, H. Xu, and S. Yan, "Robust subspace segmentation with block-diagonal prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3818–3825.

[33] J. Yang and X. Yuan, "Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization," *Math. Comput.*, vol. 82, no. 281, pp. 301–329, 2012.
[34] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2117–2130, Sep. 2013.
[35] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, Nov. 1992.
[36] T. Okatani and K. Deguchi, "On the Wiberg algorithm for matrix factorization in the presence of missing components," *Int. J. Comput. Vis.*, vol. 73, no. 3, pp. 329–337, May 2007.
[37] C. Bregler, A. Hertzmann, and H. Biermann, "Recovering non-rigid 3D shape from image streams," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2000, pp. 690–696.
[38] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. Annu. Symp. Theory Comput.*, 2013, pp. 665–674.
[39] X. Cai, C. Ding, F. Nie, and H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1124–1132.
[40] R. Liu, Z. Lin, F. De la Torre, and Z. Su, "Fixed-rank representation for unsupervised visual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 598–605.
[41] B. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.
[42] D. Tsai, M. Flagg, and J. M. Rehg, "Motion coherent tracking with multi-label MRF optimization," in *Proc. Brit. Mach. Vis. Conf.*, 2010.
[43] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[44] S. X. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 313–319.
[45] H. Hu, J. Feng, C. Yu, and J. Zhou, "Multi-class constrained normalized cut with hard, soft, unary and pairwise priors and its applications to object segmentation," *IEEE Trans. Image Process.*, vol. 22, no. 11, pp. 4328–4340, Nov. 2013.
[46] C. Li, L. Lin, W. Zuo, S. Yan, and J. Tang, "SOLD: Sub-optimal low-rank decomposition for efficient video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5519–5527.
[47] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[48] X. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L^1 optical flow," in *Proc. Joint DAGM Symp.*, 2008, pp. 214–223.
[49] S. Vijayanarasimhan and K. Grauman, "Active frame selection for label propagation in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 496–509.
[50] R. Zhang, L. Lin, R. Zhang, W. Zuo, and L. Zhang, "Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4766–4779, Dec. 2015.



Chenglong Li received the B.S. degree in applied mathematics and the M.S. degree in computer science from Anhui University, Hefei, China, in 2010 and 2013, respectively, where he is currently pursuing the Ph.D. degree in computer science.

His current research interests include computer vision, machine learning, and intelligent media technology.

Liang Lin received the B.S. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 1999 and 2008, respectively. From 2006 to 2007, he was a joint Ph.D. Student with the Department of Statistics, University of California, Los Angeles (UCLA), Los Angeles, CA, USA.

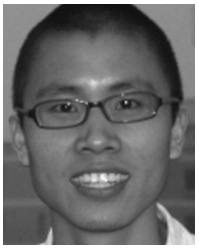
He was a Post-Doctoral Research Fellow with the Center for Vision, Cognition, Learning, and Art, UCLA. He is currently a Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He has authored over 80 papers in top-tier academic journals and conferences. His current research interests include new models, algorithms, and systems for intelligent processing and understanding of visual data, such as images and videos.

Prof. Lin was a recipient of the Best Paper Runners-Up Award in NPAR 2010, the Google Faculty Award in 2012, the Hong Kong Scholars Award 2014, and the Best Student Paper Award in the IEEE ICME 2014. He currently serves as an Associate Editor of the IEEE TRANSACTIONS HUMAN-MACHINE SYSTEMS, *Neurocomputing*, and *The Visual Computer*.

Wangmeng Zuo (M'09–SM'14) received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007.

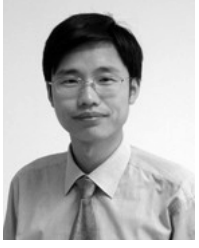
He was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, from 2004 to 2008. From 2009 to 2010, he was a Visiting Professor with Microsoft Research Asia, Beijing, China. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He has authored over 50 papers in the research areas. His current research interests include image modeling and low-level vision, discriminative learning, and biometrics.

Dr. Zuo is an Associate Editor of the *IET Biometrics*.



Wenzhong Wang received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2007.

He is currently a Lecturer with the School of Computer Science and Technology, Anhui University. His research interests include computer vision, computer graphics, and virtual reality.



Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively.

He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, and machine learning.