SOLD: Sub-Optimal Low-rank Decomposition for Efficient Video Segmentation

Chenglong Li[1,2], Liang Lin[2]*, Wangmeng Zuo[3], Shuicheng Yan[4], Jin Tang[1]

[1]School of Computer Science and Technology, Anhui University, Hefei,China

[2]School of Advanced Computing, Sun Yat-sen University, Guangzhou, China

[3]School of Computer Science and Technology, Harbin Institute of Technology, China

[4]Department of ECE, National University of Singapore, Singapore

lcl1314@foxmail.com, linliang@ieee.org, {cswmzuo, ahhftang}@gmail.com, eleyans@nus.edu.sg

Abstract

This paper investigates how to perform robust and efficient unsupervised video segmentation while suppressing the effects of data noises and/or corruptions. We propose a general algorithm, called Sub-Optimal Low-rank Decomposition (SOLD), which pursues the low-rank representation for video segmentation. Given the supervoxels affinity matrix of an observed video sequence, SOLD seeks a sub-optimal solution by making the matrix rank explicitly determined. In particular, the affinity matrix with the rank fixed 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 our framework based on the obervation that small-size video patterns tend to recur frequently within the same object. The video can be segmented into several spatio-temporal regions by applying the Normalized-Cut (NCut) algorithm with the solved low-rank representation. To process the streaming videos, we apply our algorithm sequentially over a batch of frames over time, in which we also develop several temporal consistent constraints improving the robustness. Extensive experiments on the public benchmarks demonstrate superior performance of our framework over other state-of-the-art approaches.

1. 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. Recently, various works on video segmentation have been introduced ranging from mean-shift [14], spectral clustering [11, 16], graph-based processing [2, 12] and superpixel tracking [10, 13]. And some benchmarks [4, 7] have also been provided to evaluate existing methods and help further study. Despite of much progress on video segmentation, there exists a critical limitation, *i.e.*, most of video segmentation methods have worse segmentation quality due to only utilizing low-level features. On one hand, the low-level features are easily contaminated by video noises. On the other hand, the low-level features are usually not powerful enough to differentiate the different semantic regions.

Motivated by the advances in subspace clustering [30], especially the Low-Rank Representation (LRR) methods for image segmentation [6, 22], we propose a *Sub-Optimal Low-rank Decomposition* (SOLD) algorithm, which pursues the low-rank representation for video segmentation. Instead of using superpixels in previous works like [13, 16], we take supervoxels as graph nodes to infer their optimal affinities because they can preserve local spatio-temporal coherence as well as good boundaries. To seek the unbiased and task-independent video segmentation solution, we define our low-rank model based on very generic assumption inspired by [1, 8, 20]. In particular, we assume that the intra-class supervoxels are drawn from one identical low-rank feature subspace, and all supervoxels in a period lie on a union of multiple subspaces, which can be justified by natural statistic and observations of videos. 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 enlarging its discriminative ability. As a natural extension from internal image statistics [1], this prior,

*Corresponding author is Liang Lin. This work was done when the first author worked as a visiting student in Sun Yat-sen University. This work was supported by the 863 Program of China (no. 2013AA013801), Natural Science Foundation of China (no. 61472002), Guangdong Natural Science Foundation (no.S2013050014548), and Program of Guangzhou Zhujiang Star of Science and Technology (no.2013J2200067).

local small-size video cubes ($e.g.$ $6 \times 6 \times 6$ voxels) with certain appearance patterns tend to recur frequently within the same semantic spatio-temporal region, but may not appear in semantically different spatio-temporal regions, can substantially reduce the computational complexity in video segmentation.

Unlike relaxing the rank minimization to the nuclear norm minimization in other works [6, 22], the rank of the affinity matrix in SOLD is explicitly determined for better representation. In particular, the affinity matrix with the rank fixed can be decomposed into two sub-matrices of low rank, and thus we efficiently optimize the low-rank representation by iteratively solving several closed-form subproblems to obtain a sub-optimal solution, which is utilized to infer the affinities between supervoxls.

In the inference of video segmentation, we process the video in the sliding windows instead of in the whole video to facilitate the arbitrarily long video processing in the limited memory and space, and enforce the temporal consistent constraints on the video stream to approximate the full video segmentation. Specifically, one or more frames are overlapped to propagate solutions between neighboring windows. We construct the reasonable temporal consistent constraints by the overlap ratio between temporal supervoxels in overlapping frames, and apply the efficient constrained NCut method [9] to achieve the final supervoxel-level segmentation.

The key contributions of this work are two-folds. First, We propose a general algorithm for Low-Rank Representation pursuit, which decomposes the affinity matrix with the rank fixed into two sub-matrices of low rank and conducts a sub-optimal solution. Second, we develop an effective framework for unsupervised streaming video segmentation, where several informative priors and constraints over video supervoxels are developed. The extensive experiments on the public challenging dataset VSB100 [7] validate superior effectiveness compared to the state-of-the-art approaches and efficiency of our approach.

2. Literature Review

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

Recent advances in hierarchical methods [2, 3, 28], streaming methods [12, 17] and related evaluations [4, 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. [2] 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. [12] proposed a streaming hierarchical video segmentation framework and instantiated HGB within this framework (SHGB). This method enforced a Markovian assumption on the video stream, which leveraged ideas from data streams. Galasso et al. [17] 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 [4] 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. [7] to evaluate effectively over- and under-segmentation of current video segmentation. These works also encourage the progress on new aspects of the video segmentation problem.

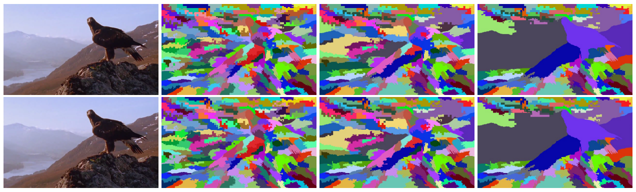


Figure 1. Sample supervoxels at level 1 (200, where 200 indicates the number of supervoxels), 2 (150) and 3 (100) extracted from a hierarchical video segmentation [2]. The different colors indicate the different supervoxels.

Recent works on video segmentation focus only on salient moving objects by analyzing point trajectories, while taking background as a single cluster [11]. Some other works [13, 16] 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. For instance, Galasso et al. [16] proposed a robust Video Segmentation approach with Superpixels (VSS) to explore various within- and between-frame affinities suitable for video segmentation. In addition, Tarabalka et al. [21] 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.

3. SOLD Algorithm

Given an arbitrarily long input video, we adopt the overlapping sliding window approach for saving memory and space, and solve the segmentation of video frames within the observed window.

3.1. Formulation

The proposed low rank decomposition model is imposed on the supervoxels for better tradeoff of efficiency and accuracy. In one temporal window, the supervoxels are generated by unsupervised video segmentation method [2], 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). Although multilevel supervoxel representation can provide more appearance and motion features, as shown in Fig. 1, 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 instead of enforcing multilevel consistency in multilayers to avoid error propagation.

Each temporal window of the video is segmented into n supervoxels. Note that n should not be set too small (large under-segmentation errors) or too large (low computational speed). Empirically, we fix $n = 200$ in this work to balance the accuracy-efficiency trade-off. 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. Based on the low-rank representation assumption, we have

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

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

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

where λ denotes the regularization parameter. $\| \cdot \|_F$ and $\| \cdot \|_1$ denote the Frobenius norm and the ℓ_1-norm of a matrix, respectively. The model in Eq. 2 is nonconvex, and the low rank affinity matrix is usually obtained by solving its convex relaxation problem,

$$\min_{\mathbf{Z}, \mathbf{E}} \frac{1}{2} \|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 + \alpha \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_1, \tag{3}$$

where $\| \cdot \|_*$ denotes the nuclear norm of a matrix, and the parameters α and λ are balance factors of three parts.

To enhance the discriminative ability of the low rank affinity matrix, we further integrate into the model in Eq. 3 the discriminative replication prior based on internal video statistics. Discriminative replication prior was proposed for modeling statistical observation on natural images and was successfully applied to image segmentation [1]. In this work, we extend it to videos in a natural way: local small-size cubes ($e.g.$, $6 \times 6 \times 6$ voxels) tend to recur frequently within the same semantic spatio-temporal region, yet less frequently within semantically different spatio-temporal regions. Further, the extension to video also benefits the preservation of temporal coherence and improvement on computational efficiency, as shown in Fig. 2.

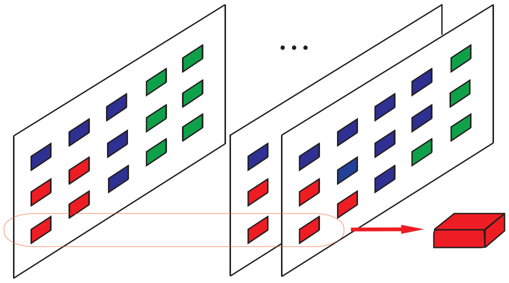


Figure 2. Illustration of the discriminative replication prior. A video cube consists of a set of spatially overlapped patches, where repeatedly occurred patches are identified with the same color. One red cube is highlighted for clarity.

Denote Λ_i by the spatio-temporal subregion covered by supervoxel i, and $\mathbf{Q} \in \mathbb{R}^{n \times n}$ be the discriminative replication prior matrix. Supposing that the number of cubes within Λ_i is $|\Lambda_i|$, we have

$$\begin{aligned} Q_{ij} &= e^{-(\frac{1}{|\Lambda_i|} \sum_{p \in \Lambda_i} D(p,\, \Lambda_j) + \frac{1}{|\Lambda_j|} \sum_{q \in \Lambda_j} D(q,\, \Lambda_i))}, \\ D(p,\, \Lambda) &= \frac{1}{|\Lambda|} \sum_{q \in \Lambda} \delta_\zeta(\kappa(\|\mathbf{x}_p - \mathbf{x}_q\|)), \end{aligned} \tag{4}$$

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

$$\delta_\zeta(a) = aI(|a| > \zeta), \tag{5}$$

where $I(\cdot)$ is the indicator function, and the threshold ζ is fixed to be 0.4 in this work.

From Eq. 4, one can see that larger $\mathbf{Q}_{ij}$ indicates that the supervoxel i and j belong to different semantic spatio-temporal regions with higher probability, and vice versa. Therefore, we incorporate the discriminative replication prior into the model in Eq. 3:

$$\min_{\mathbf{Z}, \mathbf{E}} \frac{1}{2} \|\mathbf{X} - \mathbf{XZ} - \mathbf{E}\|_F^2 + \alpha \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_1 + \gamma tr(\mathbf{Z}^T \mathbf{Q}), \tag{6}$$

where $tr(\cdot)$ returns the matrix trace, and γ is a tuning parameter. Therefore, high-level semantic internal statistics can be incorporated as a soft constraint to enhance the discriminative ability.

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 $\lambda$, $\beta$ and $\gamma$;
    Set $\mathbf{E} = 0$; $\varepsilon = 10^{-8}$, $maxIter = 500$.
Output: $\mathbf{A}$, $\mathbf{B}$, $\mathbf{E}$.
while not converged do
    Update $\mathbf{A}$ by Eq. 11;
    Update $\mathbf{B}$ by Eq. 9;
    Update $\mathbf{E}$ by Eq. 12;
    Check the convergence condition: the maximum element change of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{E}$ between two consecutive iterations is less than $\varepsilon$ or the maximum number of iterations reaches $maxIter$.
end while
```

The low rank representation model in Eq. 6 can be solved using the augmented Lagrangian method (ALM) [20] or linearized ALM [29]. However, in many applications it is easier to explicitly determine the desired rank rather than implicitly tuning the tradeoff parameter α [24]. For example, rigid structure from motion (SFM) can be formulated as a rank-3 matrix factorization problem [27], while nonrigid SFM can be formulated as a rank-$3k$ matrix factorization, where k is the number of shape basis for depicting nonrigid deformation [26]. Moreover, as demonstrated in [19, 23], the incorporation of explicit rank constraint may result in more efficient optimization algorithm. Therefore, we remove the nuclear-norm regularizer in Eq. 6, and explicitly impose the fixed-rank constraint on $\mathbf{Z}$. Supposing the rank of the affinity matrix $\mathbf{Z}$ is r, we have $\mathbf{Z} = \mathbf{AB}$, where $A \in \mathbb{R}^{n \times r}$, $B \in \mathbb{R}^{r \times n}$, and $r < min(n, d)$. By replacing $\mathbf{Z}$ with $\mathbf{AB}$, the Sub-Optimal Low-rank Decomposition (SOLD) method is then formulated as,

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

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.

3.2. Optimization

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

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

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

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

where

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

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

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

Eq. 11 can be transformed to a generalized eigenproblem, 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 noises matrix $\mathbf{E}$ can be solved by the soft-threshold (or shrinkage) method in [20]:

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

Please refer to the supplementary material for detailed derivation of the 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 the global convergence of the algorithm is not proved, we empirically validate its convergence in Sect. 5.4. Finally, the low rank affinity matrix of the supervoxels can be obtained by $\mathbf{Z} = \mathbf{AB}$.

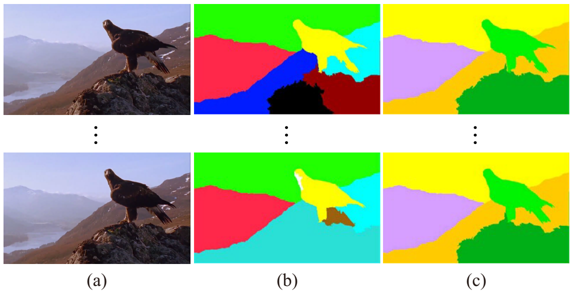

(a) (b) (c)

Figure 3. Illustration of the temporal consistent constraints for temporal consistency. Frame 41 to 61 of the video sequence *"planet_earth_1"* in the dataset VSB100 [7] 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.

3.3. Implementation Details

To make SOLD clear and complete, some important implementation details are briefly introduced.

Since the hierarchical graph-based method [2] performs well on all the metrics of the united video segmentation

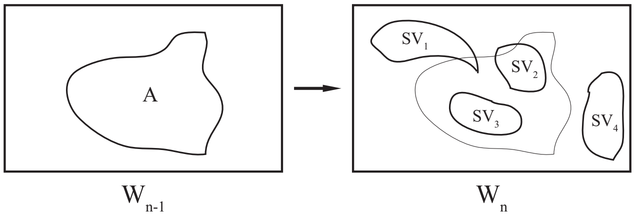


Figure 4. The generation of the temporal consistent constraints between two neighboring sliding windows W_{n-1} and W_n. A denotes one segmentation region in W_{n-1}, and provides some constraints to the segmentation of W_n. For clarity, four typical supervoxels are shown here, which stand for four typical 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.

benchmarks [4, 7], we utilize it to generate one layer supervoxels. Thus, in this work, the hierarchical graph-based method is utilized to generate one layer supervoxels. The only input parameter is the total number of supervoxels, which is fixed to be 200 to balance the accuracy-efficiency trade-off in this work.

To withstand noise and moderate appearance variation, 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) [15], are concatenated into a single descriptor vector.

4. Streaming Video Segmentation

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, and thus is essential in video segmentation. To this end, we segment the video in overlapping sliding windows, and adopt the the supervoxel-level NCut for efficient video segmentation. Besides, both the temporal consistent constraints and low rank affinity are considered to improve the longer-range consistency and segmentation accuracy of the inference algorithm.

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

$$\mathbf{W}_{ij} = \sum_{m=1}^{3} \omega^m \phi^m_{ij}, \qquad (13)$$

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

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

where $Lines(i,j)$ is a straight line set, and each line joins centers of within-frame superpixel-pair, which belongs to supervoxel-pair (i,j). 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

$$\phi^2_{ij} = e^{-\alpha^2\|c_i-c_j\|}, \qquad (15)$$

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

$$\phi^3_{ij} = e^{-\alpha^3 e^{\frac{-|\mathbf{Z}_{ij}|}{2\sigma^2}}}, \qquad (16)$$

where $\mathbf{Z}_{ij}$ indicates the (i,j)-the element of the low rank affinity matrix, α^3 is the tuning parameter, and σ is the Gaussian parameter. $|\cdot|$ indicates the absolute operator. The settings of all the parameters are described in Sect. 5.1.

The temporal consistent constraints are further introduced to properly propagate solutions from current temporal window to the next window. In this way, we can generate some constraints between neighboring windows 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, which completely or almost (over 90% in this paper) belong to one specific label, and non-deterministic supervoxels, which partly belong to some label. Then the partial grouping supervoxel set is composed by only the deterministic supervoxels. Fig. 4 illustrates this process.

Given the partial grouping supervoxel set $\mathbf{U}_t$, we can obtain $|\mathbf{U}_t|-1$ independent constraints, where $|\cdot|$ denotes the size of a set, and $t \in T$ indicates the label index. Then, the temporal consistent constraint matrix $\bar{\mathbf{U}}$ is computed as follows: For each row k, there is two nonzero elements $\bar{\mathbf{U}}_k(i) = 1$ and $\bar{\mathbf{U}}_k(j) = -1$, where $i,j \in \mathbf{U}_t$ and $k \in [\sum_{t=1}^{T}(|\mathbf{U}_t|-1)]$, $[n]$ indicates the set of integers between 1 and n: $[n] = \{1,2,\ldots,n\}$. Alg. 2 summarizes this procedure, and Fig. 3 illustrates its effectiveness.

Finally, we apply the constrained NCut method [9] on $\mathbf{W}$ to achieve the supervoxel-level segmentation. The tractable K-ways normalized segmentation criterion with temporal consistent constraints is formulated as

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

where $\bar{\mathbf{Z}} = \bar{\mathbf{X}}(\bar{\mathbf{X}}^T\bar{\mathbf{D}}\bar{\mathbf{X}})^{-\frac{1}{2}}$, $\bar{\mathbf{D}} = \mathbf{W}\mathbf{1}_N$, $\bar{\mathbf{X}}$ and $\bar{\mathbf{Z}}$ are the partition matrix and the scaled partition matrix, respectively, and $\bar{\mathbf{D}}$ is the degree matrix, $\mathbf{1}$ and $\mathbf{I}$ denote all ones vector and identity matrix, respectively, and N is total number of supervoxels. The optimization of Eq. 17 has been

Algorithm 2 Temporal Consistent Constraint Matrix Computation between Two Neighboring Windows

```
Input: Label set $T$ from the first window; Supervoxel set
    $S$ in the second window.
Output: Temporal consistent constraint matrix $\bar{\mathbf{U}}$.
for $t = 1 : |T|$ do
    Find the deterministic supervoxel set $\mathbf{U}_t$ $(\subseteq S)$ for
    the label $T(t)$ according to the overlap ratio of over-
    lapping frame(s);
    $k = 0$;
    for $s = 1 : |\mathbf{U}_t| - 1$ do
        $k = k + 1$;
        $\bar{\mathbf{U}}(k,\ \mathbf{U}_t(s)) = 1$;
        $\bar{\mathbf{U}}(k,\ \mathbf{U}_t(s+1)) = -1$.
    end for
end for
```

addressed in [9], and the main results are as follows. Let $\bar{\mathbf{P}}$ be the row-normalized weight matrix and $\bar{\mathbf{Q}}$ be a projector onto the feasible solution space:

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

Let $\hat{\mathbf{V}}'_{[K]}$ be the first K eigenvectors of the matrix $\bar{\mathbf{Q}}\bar{\mathbf{P}}\bar{\mathbf{Q}}$, then the solutions of Eq. 17 are $\hat{\mathbf{V}}_{[K]} = \bar{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{V}}'_{[K]}$. The $\hat{\mathbf{V}}_{[K]}$ are discretized by spectral rotation [18] or k-means (spectral rotation in this paper) to obtain the final 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 labels mapping by their spatial overlap [13]. 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.

5. Experiment Results

In this section, we evaluate our video segmentation framework on the standard benchmark VSB100 [7], and compare with other state-of-the-art methods. Then, we further analyze the effectiveness of our main components. At last, the efficiency analysis of our framework is discussed.

5.1. Evaluation Settings

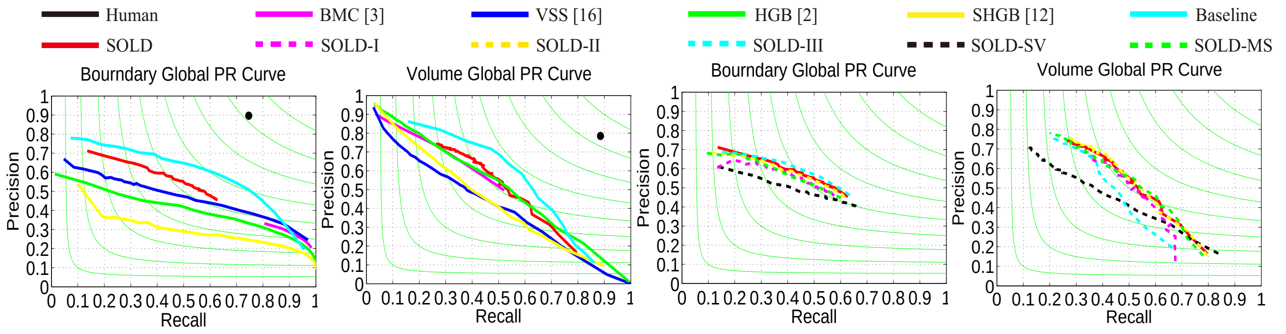
The selected VSB100 [7] for empirical evaluation is very challenging. It is the largest video segmentation dataset with high definition frames, and consists of four difficult sub-datasets: general, motion segmentation, non-rigid motion segmentation and camera motion segmentation. The same setting as [7], we regard the general sub-dataset (60 video sequences) as our test set for all the approaches.

To make the comparison comprehensive, we employ the segment number set {2, 3, ..., 51} to produce multilevel segmentation results, and fix all parameters in all evaluations: we set $\{r,\ \lambda,\ \beta,\ \gamma\}$ = {8, 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 inference. In addition, the number of frames per window is set to be 6, and one frame is overlapped between neighboring windows.

5.2. Comparison Results

We compare our approach with four state-of-the-art video segmentation algorithms, including BMC [3], VSS [16], HGB [2] and SHGB [12]. The first two subfigures of Fig. 5 illustrate the Boundary Precision-Recall (BPR) and Volume Precision-Recall (VPR) curves of the comparisons on the VSB100 dataset. Tab. 1 gives a summary of the aggregate performance evaluations, which includes Optimal Dataset Scale (ODS), Optimal Segmentation Scale (OSS) and Average Precision (AP) of BPR and VPR. Herein, the baseline [7] is extension of [5] by propagating the results [5] of central frame to other frames with optical flow [25] and labeling image segments (across hierarchy) with maximum voting. It adopted more complex image features while exploiting additional cues like motion.

From Fig. 5 and Tab. 1, we can conclude that our approach achieves comparable performance against the state-of-the-art methods in both BPR and VPR on the VSB100 dataset. Specifically, our approach achieves best ODS and OSS values in both BPR and VPR. Though all 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 affinity matrix of supervoxel features. Besides, the temporal consistent constraints adopted by our approach 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 to previous works are shown in Fig. 6 to demonstrate the superior per-



(a) (b) (c) (d)

Figure 5. Comparison curves of our framework with its variants and the state-of-the-art video segmentation approaches [2, 3, 12, 16]. The first two subfigures are the comparison curves for comparing our framework with previous works, and the last two subfigures are the comparison curves for evaluating the variants of our framework.

Table 1. The aggregation measures of Boundary Precision-Recall (BPR) and Volume Precision-Recall (VPR) for comparing with previous works on dataset VSB100 [7]. (*) denotes evaluated on video frames resized by 0.5 due to large computational demands and the *italic* denotes the streaming method. Red fonts indicate the best performance.

	BPR			VPR		
Algorithm	**ODS**	**OSS**	**AP**	**ODS**	**OSS**	**AP**
*BMC [3]	0.47	0.48	0.32	0.51	0.52	0.38
*VSS [16]	0.51	0.56	0.45	0.45	0.51	0.42
*HGB [2]	0.47	0.54	0.41	0.52	0.55	0.52
SHGB [12]	0.38	0.46	0.32	0.45	0.48	0.44
**SOLD*	0.54	0.58	0.40	0.53	0.60	0.46
Human	0.81	0.81	0.67	0.83	0.83	0.70
Baseline [7]	0.61	0.65	0.59	0.59	0.62	0.56

Table 2. The aggregation measures of Boundary Precision-Recall (BPR) and Volume Precision-Recall (VPR) for comparing our approach with its variants on dataset VSB100 [7]. The description of this Table is the same as Table 1.

	BPR			VPR		
Algorithm	**ODS**	**OSS**	**AP**	**ODS**	**OSS**	**AP**
**SOLD*	0.54	0.58	0.40	0.53	0.60	0.46
**SOLD-I*	0.51	0.55	0.34	0.52	0.58	0.41
**SOLD-II*	0.53	0.56	0.38	0.53	0.59	0.47
**SOLD-III*	0.54	0.57	0.41	0.47	0.54	0.39
**SOLD-SV*	0.51	0.55	0.36	0.45	0.50	0.39
**SOLD-MS*	0.52	0.56	0.37	0.53	0.59	0.47

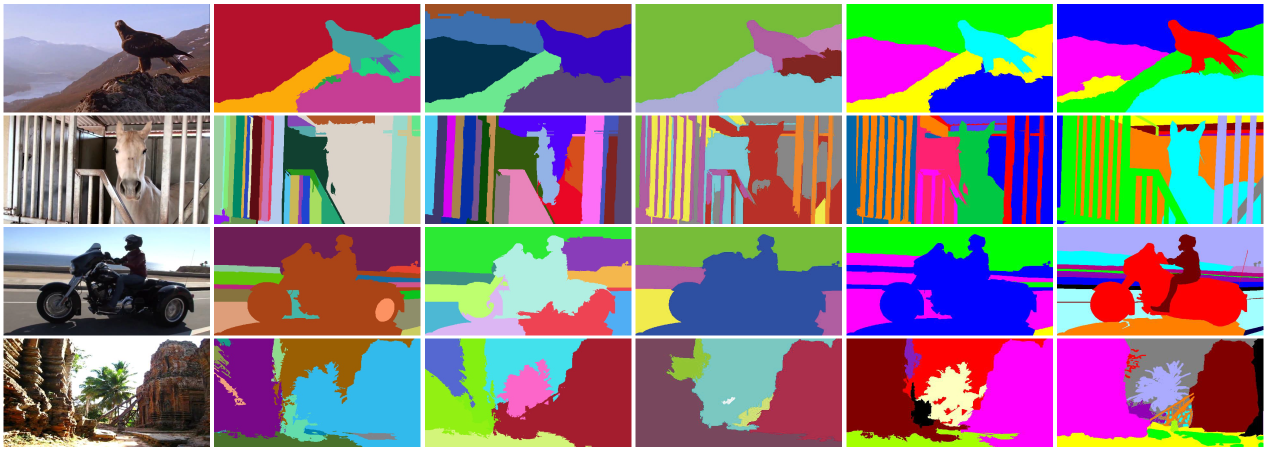
formance of our framework.

Though our framework 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 currently exploit the small number and will develop an adaptive version in our future work.

5.3. Component Analysis

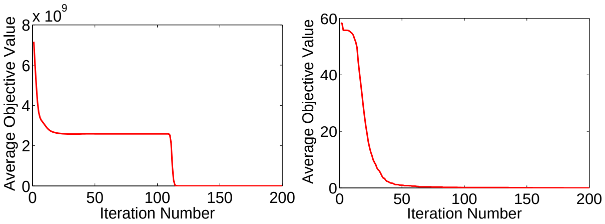
To justify the significance of the main components of our framework, we implement three special versions and two variants for empirical analysis. They are: 1) *SOLD-I*, that sets $\omega^3 = 0$ to remove the affinity term inferred by the sub-optimal low-rank decomposition in streaming segmentation framework. 2) *SOLD-II*, that sets $\gamma = 0$ in Eq. 7 to remove the regularization term of the discriminative replication prior in SOLD. 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) *SOLD-MS*, that enforces multiscale consistency between multilayers, and is solved by ALM method [20]. Specifically, we implement *SOLD-MS* as following three steps. Firstly, three layer supervoxels are generated by hierarchical graph-based video segmentation approach [2] with supervoxel numbers set to be 200, 150, 100, respectively. Secondly, we introduce the multiscale consistent constraint matrix and the discriminative replication prior matrix into Eq. 3, and apply ALM method to solve it. Thirdly, we integrate it into our streaming framework to facilitate the evaluation.

The last two subfigures of Fig. 5 show the components evaluation of our framework, and corresponding aggregation measures are reported in Tab. 2. From Fig. 5 and Tab. 2, we can draw the following conclusions. 1) The complete framework outperforms SOLD-I in both BPR and VPR. This justifies the significance of the optimal affinities optimized by the sub-optimal low-rank decomposition. 2) Comparing to the complete framework, 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, the complete frame-



(a) Image (b) HGB [2] (c) SHGB [12] (d) SOLD-MS (e) SOLD (f) GT

Figure 6. Qualitative comparisons with the state-of-the-art video segmentation methods HGB and SHGB. We can see that our method qualitatively improves on the algorithm of HGB, and substantially outperforms the algorithm of SHGB.



(a) SOLD-MS (b) SOLD

Figure 7. Convergence curves of our approach on dataset VSB100 [7].

work with the temporal consistent constraints substantially improves the performance in VPR, *i.e.,* keeping longer-range temporal consistency. 4) Our framework outperforms SOLD-SV in both BPR and VPR, and it shows that the affinities 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 streaming framework. 5) Our framework obtains better results than SOLD-MS. This validates that the multiscale consistency constraints do not help to improve the segmentation results due to error propagation as we previously discussed.

5.4. Efficiency Analysis

To 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 SOLD-MS. Herein, SOLD-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. 7 shows the convergence curves of SOLD-MS and SOLD, and Tab. 3 reports their average iterations and running time. Thanks to the proposed sub-optimal low-rank decomposition, it only costs 0.12 sec./frame for our SOLD, which converges faster than SOLD-MS (see Fig. 7), and brings 20-times over it (see Tab. 3).

Table 3. The average iterations and running time (seconds per frame) of SOLD-MS and SOLD.

	SOLD-MS	SOLD
Iteration Number	112	16
Running Time	2.40	0.12

Furthermore, constrained NCut is also efficiently solved within 0.01 sec./frame due to the proposed supervoxel-level segmentation. In addition, graph-based over-segmentation and feature extraction are mostly two time consuming procedures, and will be parallelized by GPU in future work.

6. Conclusion

In this paper, we have proposed a general algorithm for low-rank representation pursuit by decomposing the matrix with the rank fixed 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 framework that automatically segments streaming videos in an unsupervised way. Extensive experiments on the standard benchmarks have demonstrated the superior performances of our approach over other state-of-the-art methods. In future work, we will improve our video segmentation framework by introducing more robust video features and over-segmentation methods. Our low-rank decomposition algorithm can be also extended to other vision tasks such as multi-object tracking.

References

[1] X. Liu, L. Lin, and A. Yuille. Robust region grouping via internal patch statistic. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013. 1, 3

[2] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010. 1, 2, 3, 4, 6, 7

[3] 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.*, 27(5): 629–640, May 2008. 2, 6, 7

[4] C. Xu, and J. J. Corso. Evaluation of supervoxel methods for early video processing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012. 1, 2, 5

[5] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5): 898–916, May 2011. 6

[6] B. Chen, G. Liu, Z. Huang, and S. Yan. Multi-task low-rank affinities pursuit for image segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2011. 1, 2

[7] F. Galasso, N. Nagaraja, T. Cardenas, T. Brox, and B. Schiele. A unified video segmentation benchmark: annotation, metrics and analysis. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2013. 1, 2, 4, 5, 6, 7, 8

[8] L. Lin, Y. Xu, X. Liang, and J. Lai. Complex Background Subtraction by Pursuing Dynamic Spatio-Temporal Models. *IEEE Transactions on Image Processing*, 23(7): 3191-3202, 2014. 1, 2

[9] S. X. Yu and J. Shi. Segmentation Given Partial Grouping Constraints. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2): 173–183, Feb. 2004. 2, 5, 6

[10] W. Brendel, and S. Todorovic. Video object segmentation by tracking regions. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2009. 1

[11] K. Fragkiadaki, and J. Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012. 1, 2

[12] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *Proc. Eur. Conf. Comput. Vis.*, 2012. 1, 2, 6, 7

[13] A. Vazquez-Reina, S. Avidan, H. Pfister, and E. Miller. Multiple hypothesis video segmentation from superpixel flows. In *Proc. Eur. Conf. Comput. Vis.*, 2010. 1, 2, 6

[14] S. Paris. Edge-preserving smoothing and mean-shift segmentation of video streams. In *Proc. Eur. Conf. Comput. Vis.*, 2008. 1

[15] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *Proc. Brit. Mach. Vis. Conf.*, 2009. 5

[16] F. Galasso, R. Cipolla, and B. Schiele. Video segmentation with superpixels. In *Proc. Asian Conf. Comput. Vis.*, 2012. 1, 2, 6, 7

[17] 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.*, 2014. 2

[18] S. X. Yu and J. Shi. Multiclass Spectral Clustering. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2003. 6

[19] 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*, 2013. 4

[20] 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. *UIUC Technical Report UILU-ENG-09-2214*, July 2009. 1, 4, 7

[21] T. Tarabalka, G. Charpiat, L. Brucker, and B. Menze. Spatio-temporal video segmentation with shape growth or shrinkage constraint. *IEEE Trans. Image Process.*, 23(9), Sep. 2014. 2

[22] 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.*, 35(1): 171–184, Jan. 2013. 1, 2

[23] R. Liu, Z. Lin, F. Torre, and Z. Su. Fixed-rank representation for unsupervised visual learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012. 4

[24] 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.*, 35(9): 2117–2130, Sep. 2013. 4

[25] X. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *Proc. DAGM*, 2007. 6

[26] L. Lin, X. Liu and S. C. Zhu. Layered Graph Matching with Composite Cluster Sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8): 1426-1442, 2010. 4

[27] T. Okatani, and K. Deguchi. On the wiberg algorithm for matrix factorization in the presence of missing components. *Int. J. Comput.Vis.*, 72(3):329-337, May 2007. 4

[28] L. Lin, T. Wu, J. Porway, and Z. Xu. A Stochastic Graph Grammar for Compositional Object Representation and Recognition. *Pattern Recognition*, 42(7): 1297-1307, 2009. 2

[29] J. Yang, and X. Yuan. Lineared augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of computation*, 82(281): 301-329, Mar. 2012. 4

[30] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *Proc. Int. Conf. Mach. Learn.*, 2010. 1