# SOLD: Sup-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.

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, there exists several remaining issues to be addressed. First, most of video segmentation methods have worse segmentation quality due to only utilizing the low-level features, which are easily contaminated by video noises and usually not powerful enough to differentiate the different semantic regions. Second, exploring the internal video statistics is indispensable to improve the segmentation performance other than employing a large number of related exemplars, which is obviously time-consuming and computationally inefficient. Third, a streaming setting for video segmentation must take into account temporal long-range relationships between voxels.

Motivated by the advances in subspace clustering [4], especially the Low-Rank Representation (LRR) methods for image segmentation [1, 3], we propose a *Sub-Optimal Low-rank Decomposition* (SOLD) algorithm, which pursues the low-rank representation for efficient video segmentation. Instead of using superpixels in previous works like [2], 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 [5]. 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. Based on this assumption, the tractable low-rank representation model can be formulated as

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

where $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n]\in\mathbb{R}^{d\times n}$, denoting the feature matrix of supervoxels. $\mathbf{Z}\in\mathbb{R}^{n\times n}$ and $\mathbf{E}\in\mathbb{R}^{d\times n}$ are the desired low-rank affinity matrix the sparse corrupted noises, respectively. The parameters $\alpha$ and $\lambda$ 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. 1 the discriminative replication prior based on internal video statistics: local small-size cubes (*e.g.*, $6\times6\times6$ voxels) tend to recur frequently within the same semantic spatio-temporal region, yet less frequently within semantically different spatio-temporal regions. We denote $\mathbf{Q}\in\mathbb{R}^{n\times n}$ as the discriminative replication prior matrix, and larger $\mathbf{Q}_{ij}$ indicates that the supervoxel $i$ and $j$ belong to different semantic spatio-temporal regions with higher probability, and vice versa. Thus, we incorporate $\mathbf{Q}$ into the model in Eq. 1:

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

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

The low rank representation model in Eq. 2 can be solved using the augmented Lagrangian method (ALM) [9] or linearized ALM [7]. However, in many applications it is easier to explicitly determine the desired rank rather than implicitly tuning the tradeoff parameter $\alpha$ [6]. Therefore, we remove the nuclear-norm regularizer in Eq. 2, 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,

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{E}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{A}\mathbf{B}-\mathbf{E}\|_F^2+\lambda\|\mathbf{E}\|_1+\\ \frac{\beta}{2}\|\mathbf{A}\mathbf{B}\|_F^2+\gamma\, tr((\mathbf{A}\mathbf{B})^T\mathbf{Q}),\end{aligned} \tag{3}$$

where $\beta$ is a regularization parameter that controls overfitting. We can factorize Eq. 2 into three sub-problems with closed-form solution, *i.e.*

$$\begin{aligned}\mathbf{A}^* &= \arg\max_{\mathbf{A}} tr\{(\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2\mathbf{S}_2^T\mathbf{A}\},\\ \mathbf{B}^* &= (\mathbf{A}^T\mathbf{S}_1\mathbf{A})^{-1}\mathbf{A}^T\mathbf{S}_2,\\ \mathbf{E}^* &= \arg\min_{\mathbf{E}} \lambda\|\mathbf{E}\|_1+\frac{1}{2}\|\mathbf{E}-(\mathbf{X}-\mathbf{X}\mathbf{A}\mathbf{B})\|_F^2,\end{aligned} \tag{4}$$

where $\mathbf{S}_1=\mathbf{X}^T\mathbf{X}+\beta\mathbf{I}$, and $\mathbf{S}_2=(\mathbf{X}^T(\mathbf{X}-\mathbf{E})-\gamma\mathbf{Q})$. Even SOLD is non-convex and sub-optimal, as demonstrated in our experiments, such formulation can deliver both efficient algorithms and promising video segmentation accuracy. 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 details please refer to the supplementary material. Finally, the low rank affinity matrix of the supervoxels can be obtained by $\mathbf{Z}=\mathbf{AB}$.

An effective streaming 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. 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 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. Thus, 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. Then, we apply the constrained NCut method [8] on affinity matrix while incorporating above constraints to achieve the supervoxel-level segmentation.

[1] B. Chen, G. Liu, Z. Huang, and S. Yan. Multi-task low-rank affinities pursuit for image segmentation. In *Proceedings of IEEE International Conference on Computer Vision*, 2011.
[2] R. Cipolla F. Galasso and B. Schiele. Video segmentation with superpixels. In *Proceedings of Asian Conference on Computer Vision*, 2012.
[3] S. Yan J. Sun Y. Yu G. Liu, Z. Lin and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.
[4] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of International Conference on Machine Learning*, 2010.
[5] L. Lin X. Liu and A. Yuille. Robust region grouping via internal patch statistics. In *Proceedings of IEEE Conference on Computer Vision and Computer Vision*, 2013.
[6] J. Ye X. Li Y. Hu, D. Zhang and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2013.
[7] J. Yang and X. Yuan. Lineared augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of computation*, 82(281): 301–329, 2012.
[8] S. X. Yu and J. Shi. Segmentation given partial grouping constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):173–183, 2004.
[9] J. Wright M. Chen L. Wu Z. Lin, A. Ganesh and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *UIUC Technical Report UILU-ENG-09-2214*, 2009.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.