Learning Must-Link Constraints for Video Segmentation based on Spectral Clustering

Anna Khoreva[1], Fabio Galasso[1], Matthias Hein[2], and Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarbrücken, Germany
{khoreva,galasso,schiele}@mpi-inf.mpg.de
[2]Saarland University, Saarbrücken, Germany
hein@cs.uni-saarland.de

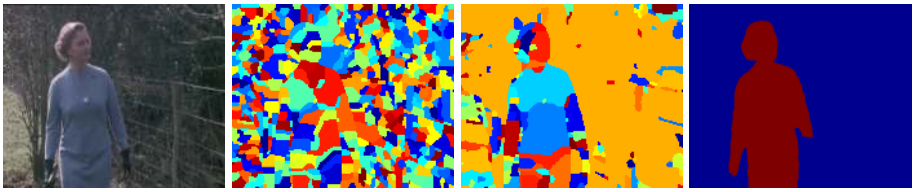
Abstract. In recent years it has been shown that clustering and segmentation methods can greatly benefit from the integration of prior information in terms of must-link constraints. Very recently the use of such constraints has been integrated in a rigorous manner also in graph-based methods such as normalized cut. On the other hand spectral clustering as relaxation of the normalized cut has been shown to be among the best methods for video segmentation. In this paper we merge these two developments and propose to learn must-link constraints for video segmentation with spectral clustering. We show that the integration of learned must-link constraints not only improves the segmentation result but also significantly reduces the required runtime, making the use of costly spectral methods possible for today's high quality video.

1 Introduction

Video segmentation is an open problem in computer vision, which has recently attracted increasing attention. The problem is of high interest due to its potential applications in action recognition, scene classification, 3D reconstruction and video indexing, among others. The literature on the topic has become prolific [7, 43, 2, 28, 27, 19, 11, 10, 4, 29] and a number of techniques have become available, e.g. generative layered models [25, 26], graph-based models [20, 46, 36] and spectral techniques [39, 8, 15, 18, 32, 35, 16].

Spectral methods, stemming from the seminal work of [39, 34], have received much attention from the theoretical viewpoint [31, 9, 21], and currently provide state-of-the-art segmentation performance [3, 40, 18, 41, 35, 42, 32, 16]. Spectral clustering, as a relaxation of the NP-hard normalized cut problem, is suitable due to its ability to include long-range affinities [18, 40] and its global view on the problem [14], providing balanced solutions.

In this paper, we focus on two important limitations of spectral techniques: the *excessive resource requirements* and the *lack of exploiting available training data*. The large demands of spectral techniques [40, 18] are particularly clear in the case of high-quality video datasets [17], limiting their current large-scale applicability. While often a labeled dataset is available, a systematic learning of the affinities used to build the graph for spectral clustering is very difficult.



a. Video sequence b. SPX c. Proposed $\mathcal{M}$ SPX d. Video segm.

Fig. 1. Video segmentation [18] employs fine superpixels (b), resulting in large resource requirements, *esp.* when using spectral methods. We propose learned must-links to merge superpixels into fewer must-link-constrained $\mathcal{M}$ superpixels (c). This reduces runtime and memory consumption and maintains or improves the segmentation (d).

In particular, as the normalized cut itself is a NP-hard problem and even the spectral relaxation is non-convex, the optimization of the minimizer which yields the segmentation is out of reach. Thus in practice one typically validates a few model parameters [8, 18, 32], refraining spectral methods to make use of recently available large training data [17].

We propose to *learn must-link constraints* to overcome both limitations. Recent spectral theory [38, 16] has shown that the integration of must-links (i.e. forcing two vertices to be in the same cluster) allows to reduce the size of the problem, while preserving the original optimization objective for all partitions satisfying the must-links. On the other hand by learning must-link constraints we can leverage the available training data in order to guide spectral clustering towards a desired segmentation. Figure 1 illustrates the advantages of learning must-links: superpixel-based techniques [18] build spectral graphs on fine superpixels, Figure 1(b); by contrast, we propose to build graphs merging superpixels based on learned must-link constraints, Figure 1(c). In particular, specifically training a classifier to minimize the number of false positives allows conservative superpixel merging, which: **i**. reduces the problem size significantly; **ii**. preserves the original optimization problem; and **iii**. improves the video segmentation, Figure 1(d), because *correct* must-links avoid undesired solutions (cf. Section 3).

In the following, we present the integration and learning of must-link constraints in Section 3 and validate them experimentally under various setups in Section 4 on two recent video segmentation datasets [8, 17].

2 Related Work

The usage of must-link constraints, first introduced in [44], is an active area of research in machine learning known as *constrained clustering* (see [5] for an overview). The goal of integrating must-link constraints into spectral clustering has been tried via: **i**. modifying the value of affinities (cf. [24], which first considered constrained spectral clustering); **ii**. modifying the spectral embedding [30]; or **iii**. adding constraints in a post-processing step [49, 13, 48, 45, 33]. Interestingly, none of these methods can guarantee that the must-link constraints

are actually satisfied in the final clustering. By contrast, we employ must-link constraints to reduce the original graph to one of smaller size, thus enforcing the constraints while additionally benefiting runtime and memory consumption.

In particular, [38, 16] have shown that must-link constraints can be used to reduce the graph, based on the corresponding point groupings, and proved equivalence between the reduced and the original graph, respectively in terms of NCut [38] and SC [16], for any clustering satisfying the must-link constraints. We employ these recent advances and propose to learn the must-link constraints in a data-driven discriminative fashion for video segmentation.

Other related work in segmentation have looked at merging superpixels with equivalence [1], but using hand-designed affinities, or learned pair-wise relations between superpixels [23], disregarding equivalence in the agglomerative merging process. This work brings together learning affinities and merging with equivalence guarantees for the first time.

3 Learning spectral must-link constraints

We provide here the steps of a video segmentation framework based on the normalized cut [39, 34, 22] and review the integration of must-link constraints by graph reductions as proposed in [38, 16]. While the idea of learning must-link constraints applies to any segmentation problem, we discuss in detail learning and inference in the specific case of the video segmentation features of [18].

3.1 Segmentation and Must-link Constraints

We represent a video sequence as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$: nodes $i \in \mathcal{V}$ represent superpixels, extracted at each frame of the video sequence with an image segmentation algorithm [3]; edges $e_{ij} \in \mathcal{E}$ between superpixels i and j take non-negative weights w_{ij} and express the similarity (*affinity*) between the superpixels.

A video segmentation can be defined as a partition $S = \{S_1, S_2, \ldots, S_K\}$ of the (superpixel) vertex set $\mathcal{V}$, i.e. $\cup_k S_k = \mathcal{V}$, $S_k \cap S_m = \varnothing \quad \forall \quad k \neq m$. Given $\mathcal{S}$ the set of all partitions, we look for an optimal video segmentation $S^* = \{S_1^*, S_2^*, \ldots, S_N^*\} \in \mathcal{S}$ (where N is the number of visual objects), minimizer of an objective function, implicit [20, 47, 37] or explicit [39, 34, 43, 10].

Must-link constraints alter the video segmentation by reducing the set of feasible partitions $\mathcal{S}$. Given *correct*[1] must-links, a video segmentation algorithm generally improves in performance, since the solver is constrained to disregard non-optimal segmentations *wrt* S^*. Moreover, the integration of must-links leads to reduced runtime and memory load as the recent work [38, 16] suggests.

We are interested in learning a *must-link grouping function* $\mathcal{M}$, which groups *certain*[2] superpixels in the graph, while respecting S^*. $\mathcal{M}$ should *conservatively*

[1] correct refers to the desired ground truth segmentation, which ideally corresponds with the optimal segmentation S^*

[2] certain groupings are the conservative grouping decisions which we propose to learn

associate each node i with a point grouping $I_k \subseteq S_l^*$ (in most uncertain cases a point grouping may only include a single node). More formally:

$$\begin{aligned} &\mathcal{M} : \mathcal{V} \mapsto \mathcal{P}, \quad i \mapsto I_k \\ &\quad s.t. \;\; I_k \subseteq S_l^* \subseteq \mathcal{V}\,, \;\; \cup_k I_k = \mathcal{V}\,, \;\; I_k \cap I_m = \varnothing \;\; \forall \;\; k \neq m\,, \end{aligned} \tag{1}$$

where $\mathcal{P}$ is the set of possible partitions of $\mathcal{V}$.

3.2 Framework

Here we tailor the general theory to a video segmentation framework based on the normalized cut, solved either via the spectral [39, 34] or 1-spectral [9, 21] relaxation. Further, we discuss the integration of learned must-link constraints via graph reduction techniques [38, 16] and learning and inference strategies.

Video segmentation setup. We build upon Galasso et al. [18]. Their constructed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ uses superpixels extracted from the lowest level (level 1) of a hierarchical image segmentation [3]. Edges connect superpixels from spatial and temporal neighbors and are weighted by their pair-wise affinities, computed from motion, appearance and shape features.

We consider six pairwise affinities: spatio-temporal appearance (STA), based on the median CIE Lab color distance; spatio-temporal motion (STM), based on median optical flow distance; across boundary appearance (ABA) and motion (ABM), computed across the common boundary of superpixels; short-term-temporal (STT), measuring shape similarity by the spatial overlap of optical flow-propagated superpixels; long-term-temporal (LTT), given by the fraction of common trajectories between the superpixels. Additionally we consider the number of common intersecting trajectories (IT). We distinguish four types of affinities, depending on whether the related superpixels: **i**. lie within the same frame (STA,STM,ABA,ABM); **ii**. lie on adjacent frames (STA,STM,STT); **iii-iv**. lie on frames at a distance of 2 (STT,LTT,IT) or more frames (LTT,IT) respectively.

Video segmentation objective function. Given a partition of $\mathcal{V}$ into N sets $S_1, \ldots, S_N$, the normalized cut (NCut) is defined [31] as:

$$\mathrm{NCut}(S_1, \ldots, S_N) = \sum_{k=1}^{N} \frac{\mathrm{cut}(S_k, \mathcal{V} \backslash S_k)}{\mathrm{vol}(S_k)}, \tag{2}$$

where $\mathrm{cut}(S_k, \mathcal{V} \backslash S_k) = \sum_{i \in S_k, j \in \mathcal{V} \backslash S_k} w_{ij}$ and $\mathrm{vol}(S_k) = \sum_{i \in S_k, j \in \mathcal{V}} w_{ij}$. The balancing factor prevents trivial solutions and is ideal when unary terms cannot be defined, but is also the reason why minimization of the NCut is NP-Hard.

Spectral relaxations. The most widely adopted relaxation of NCut is spectral clustering (SC) [39, 34, 31], where the solution of the relaxed problem is given by representing the data points with the first few eigenvectors and then clustering them with k-means.

While widely adopted [16, 32, 3, 8, 40, 18, 41], the SC relaxation is known to be *loose*. We therefore additionally consider the 1-spectral clustering (1-SC) [21, 22] - a tight relaxation based on the 1-Laplacian. However, the relaxation is only tight for bi-partitioning, for multi-way partitioning recursive splitting is used as greedy heuristic.

Reducing the original graph size with learned must-link constraints allows to experiment with 1-SC on state-of-the-art video segmentation benchmarks [8, 17], notwithstanding the increased computational costs.

Graph reduction schemes. Given must-link constraints provided as point groupings $\{I_1, I_2, \ldots, I_q\}$ on the original vertex set $I_k \subseteq \mathcal{V}$, recent work [38, 16] shows how to integrate such constraints into the original problem with respectively preserving the NCut and the spectral clustering objective function.

In more detail, integration proceeds by reducing the original graph $\mathcal{G}$ to one of smaller size $\mathcal{G}^M = (\mathcal{V}^M, \mathcal{E}^M)$, whereby the vertex set is given by the point grouping $\mathcal{V}^M = \{I_1, I_2, \ldots, I_q\}$, the edge set $\mathcal{E}^M$ preserves the original node connectivity and weights w_{IJ}^M are estimated so as to preserve the original video segmentation problem in terms of the NCut or spectral clustering objective. In particular, the NCut reduction is given by

$$w_{IJ}^M = \sum_{i \in I} \sum_{j \in J} w_{ij} \tag{3}$$

while the spectral clustering reduction is defined as

$$w_{IJ}^M = \begin{cases} \displaystyle\sum_{i \in I} \sum_{j \in J} w_{ij} & \text{if } I \neq J \\ \displaystyle\frac{1}{|I|} \sum_{i \in I} \sum_{j \in J} w_{ij} - \frac{(|I|-1)}{|I|} \sum_{i \in I} \sum_{j \in \mathcal{V} \setminus I} w_{ij} & \text{if } I = J, \end{cases} \tag{4}$$

provided equal affinities of elements of $\mathcal{G}$ constrained in $\mathcal{G}^M$, cf. [16].

3.3 Learning

An ideal must-link constraining function $\mathcal{M}$ (Eq. 1) should only merge superpixels which are *correct*, i.e. belong to the same set in the optimal segmentation. From an implementation viewpoint, it is convenient to consider instead $\mathcal{M}_{pw}$, defined over the set of edges $\mathcal{E}$ of the graph $\mathcal{G}$ representing the video sequence:

$$\mathcal{M}_{pw} : \mathcal{E} \mapsto \{0, 1\} \tag{5}$$

$\mathcal{M}_{pw}$ casts the must-link constraining problem as a binary classification one, where a TRUE output for an input edge e_{ij} means that i and j belong to the same point grouping, in the must-link constrained graph $\mathcal{G}^M$.

We learn $\mathcal{M}_{pw}$ with Random Forests [6, 12] using as features the affinities of [18] (STA,STM,ABA,ABM,STT,LTT) and the additional IT which we described in Section 3.2. Since different sets of affinities are available depending on whether two superpixels lie on the same or on different frames, we learn 4 different classifiers to match the 4 types of affinities.

We train a set of independent trees by estimating optimal parameters θ_p for the split functions $h(x, \theta_p)$ at each tree node p, as a function of the computed features x. Given a training set $T_p \subset X \times Y$, with X the vector of computed features and $Y = \{0, 1\}$ the corresponding ground truth video annotations, we seek to maximize the information gain I_p:

$$I_p(T_p, T_p^L, T_p^R) = H(T_p) - \frac{|T_p^L|}{|T_p|} H(T_p^L) - \frac{|T_p^R|}{|T_p|} H(T_p^R), \tag{6}$$

with $T_p^L = \{(x, y) \in T_p | h(x, \theta_p) = 0\}$, $T_p^R = T_p \backslash T_p^L$, the Shannon entropy $H(T) = -\sum_{y \in \{0,1\}} p_y \log(p_y)$ and p_y is the pdf of outcome y.

We extend the formulation of (6) to allow for learning must-link constraints on pre-grouped nodes. [16] uses superpixel groupings (larger superpixel named *level 2*, cf. 4). It is important, as we found out, to consider the node multiplicity. We define therefore $|T_p| = \sum_{k \in T_p} m_k$, where $m_k = |I_k| \cdot |J_k|$ is the multiplicity of the edge between superpixel groupings I_k and J_k, thus $p_y = \frac{\sum_y m_y}{\sum_{y \in \{0,1\}} m_y}$.

Must-link constraints have a transitive nature: $\mathcal{M}_{pw}(e_{ij}) = 1$ and $\mathcal{M}_{pw}(e_{ik}) = 1$ imply $\mathcal{M}_{pw}(e_{jk}) = 1$. It is therefore crucial that all decided constraints are correct, as a few wrong ones may result in a larger set of incorrect decisions by transitive closure and potentially spoil the segmentation. Thus we define the hyper-parameters (threshold of the classifier and tree depth) such that $\mathcal{M}_{pw}$ provides the largest number of positive predictions (the must-link decisions), while making zero false positives on the validation set. In such a conservative way we ensure that the resulting classifier makes only a very small number of false positives on unseen data. Although this conservative classifier might imply that in the worst case, no must-link constraints are predicted, it turns out our classifier actually predicts for a large fraction of the edges to be linked and thus leads to a significant reduction in size, while making a few false positives on the unseen data (overall, 1 false positive per 242k true predictions).

3.4 Inference

The learned must-link constraining function $\mathcal{M}_{pw}$ provides must-link decisions for each edge of graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. A further propagation of merge decisions in the graph accounts for the transitivity closure of $\mathcal{M}_{pw}$, consistently with the validation procedure (cf. Section 3.3). Based on the must-link decisions, we use the graph reduction techniques of Section 3.2, which integrate must-link decisions into graph $\mathcal{G}$ by reducing it to the smaller one $\mathcal{G}^M = (\mathcal{V}^M, \mathcal{E}^M)$ based on the determined groupings.

The described framework allows for evaluating different reduction schemes (equivalence in terms of NCut [38] and SC [16]) and various spectral partitioning

functions (1-SC [22] and SC [39, 34]). It further allows to include spatial must-link constraints and use larger superpixels, as done in [16]. We report experimental results on all these combinations in the following section.

4 Experimental validation

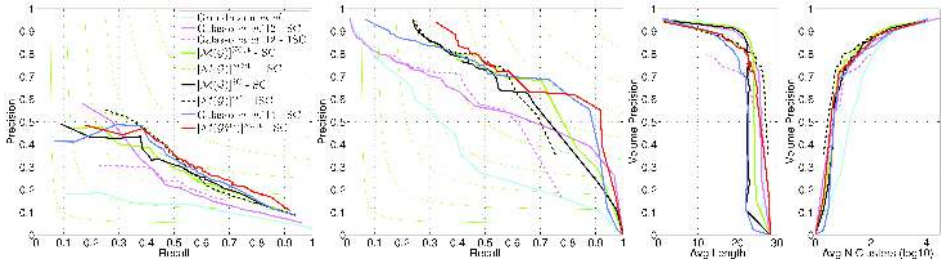
We conduct two sets of experiments to analyze performance and efficiency of must-link constrained graphs $\mathcal{G}^M$. In both cases we adopt the recently proposed benchmark metrics of [17]: the boundary precision-recall (BPR) from [3] and the volume precision-recall (VPR) metric. Besides the PR curves, we report aggregate performance for BPR and VPR: optimal dataset scale [ODS], optimal segmentation scale [OSS], average precision [AP].

In the first set of experiments, we consider the *Berkeley Motion Segmentation Dataset* (BMDS) [8], which consists of 26 VGA-quality video sequences, representing mainly humans and cars, which we arrange into training, validation and test sets (6+4+16). We restrict sequences to the first 30 frames. The ground truth is provided for the 1st, 10th, 20th, 30th frame. We further annotate the 2nd, 9th, 11th frame to learn must-links across 1 and 2 frames (we release the extra annotations).

We compare the baseline of [18] with the proposed variants, $[\mathcal{M}(\mathcal{G})]^{\text{NCut}}$ - SC and $[\mathcal{M}(\mathcal{G})]^{\text{SC}}$ - SC, reducing the original graph $\mathcal{G}$ of [18] with learned must-links to $\mathcal{G}^M$ by using respectively the normalized cut (NCut) and spectral clustering (SC) reductions, and then performing SC. Figure 2 (*plots*) shows that both proposed variants outperform the baseline algorithm [18] both on BPR and VPR. The table shows improvement by 4.7% in BPR and 9% in VPR. Since the average number of superpixels is reduced by 66.7%, the better performance is accompanied by a reduction of 60% in runtime and 90% in memory load.

In Figure 2, we further experiment by adopting 1-spectral clustering (1-SC) [22] for the NCut within the baseline algorithm (Galasso et al. [18] - 1-SC), and we compare this with our proposed variants, $[\mathcal{M}(\mathcal{G})]^{\text{NCut}}$ - 1-SC and $[\mathcal{M}(\mathcal{G})]^{\text{SC}}$ - 1-SC, where we have grouped superpixels according to learned must-links prior to processing (here with 1-SC). Since 1-SC is more costly, the provided computational reduction is even more desirable here. Again, our proposed variants improve in performance, as it appears both in the plots and the tables (average improvement of 12.3% in BPR and 9% in VPR), while significantly reducing runtime (improved by 80%) and memory load (improved by 90%). We note the similar performance of 1-SC for both reduction variants, $[\mathcal{M}(\mathcal{G})]^{\text{NCut}}$ and $[\mathcal{M}(\mathcal{G})]^{\text{SC}}$, which surprises because only the NCut reduction is theoretically justified in combination with 1-SC. Moreover, we observe the better performance of SC over 1-SC. This may indicate that the affinities of [18], designed for SC, do not fit as well the original (but different) NCut problem.

Additionally, we consider the recent work of [16], which uses superpixels extracted from a higher hierarchical level of an image segmentation algorithm [3] (superpixels at level 2), computes affinities between them and re-weights them according to SC, to take the finest superpixels at level 1 into account. Our



BPR VPR Length NCL

	BPR			VPR			Length	NCL
Algorithm	ODS	OSS	AP	ODS	OSS	AP	$\mu(\delta)$	μ
Grundmann *et al.* [20]	0.22	0.25	0.12	0.42	0.44	0.39	26.06(6.34)	13.81
Galasso *et al.* [18] - SC	0.37	0.39	0.24	0.57	0.72	0.59	25.75(6.46)	4.00
$[\mathcal{M}(\mathcal{G})]^{\text{NCut}}$ - SC	0.40	0.45	0.26	**0.69**	**0.77**	**0.69**	24.17(8.57)	6.00
$[\mathcal{M}(\mathcal{G})]^{\text{SC}}$ - SC	**0.41**	**0.46**	**0.27**	0.64	0.75	0.67	22.66(9.55)	6.00
Galasso *et al.* [18] - 1SC	0.34	0.36	0.19	0.56	0.62	0.49	25.99(6.61)	5.00
$[\mathcal{M}(\mathcal{G})]^{\text{NCut}}$ - 1SC	**0.44**	**0.48**	**0.34**	**0.64**	0.70	**0.60**	26.62(5.80)	5.00
$[\mathcal{M}(\mathcal{G})]^{\text{SC}}$ - 1SC	0.43	**0.48**	**0.34**	**0.64**	**0.71**	**0.60**	26.41(5.95)	5.00
Galasso *et al.*'14 [16] - SC	**0.43**	**0.48**	**0.29**	**0.71**	0.79	0.71	22.04(8.92)	7.00
$[\mathcal{M}(\mathcal{G}_2^{SC})]^{\text{NCut}}$ - SC	**0.43**	**0.48**	0.28	**0.71**	**0.80**	**0.75**	24.77(7.49)	5.00

Fig. 2. Comparison of state-of-the-art video segmentation algorithms with the learned must-links, on BMDS (restricted to first 30 frames) [8]. The plots and table show BPR and VPR, aggregate measures ODS, OSS and AP, and length statistics (mean μ, std. δ, no. clusters NCL) [17].

proposed method based on must-links also allows learning constraints on the larger superpixel graph $\mathcal{G}_2$ (the multiplicity of point groupings plays a role in this case, cf. Section 3.3). Figure 2 shows that the reduction $[\mathcal{M}(\mathcal{G}_2^{SC})]^{\text{NCut}}$ - SC leads to the same performance as the original algorithm [16] on BPR and improves on VPR, while reducing the problem size *wrt* [16] (runtime by 30% and memory load by 70%).

Figure 3 qualitatively supports the positive results. Note that the learned must-links respect the GT objects while reducing the number of employed superpixels, $\mathcal{M}$ SPX. Improvements in the video segmentation output ($\mathcal{M}$ Segm Vs. (SPX) Segm.) are more evident for 1-SC. The proposed learned must-links determine merging both in the spatial and temporal dimension. It is interesting to note that for the BMDS [8] most merging comes from the first: it seems easier to make conservative merging assumptions within the frame.

In the second set of experiments we consider the novel benchmark VSB100 [17], which includes 100 HD quality videos [41] arranged into train and test sets (40+60) (we split test – 25 – and validation set – 15). In Figure 4 we compare the proposed method $[\mathcal{M}(\mathcal{G}_2^{SC})]^{\text{NCut}}$ - SC to the baseline [16] and state-of-the-art video segmentation algorithms. Our method maintains the performance of [16] on BPR and slightly improves on VPR. This shows that [16], by jointly lever-

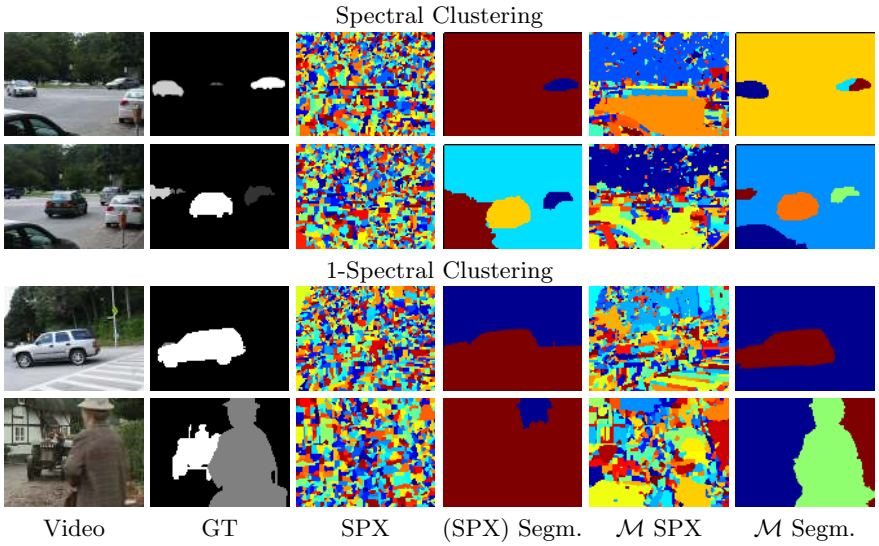
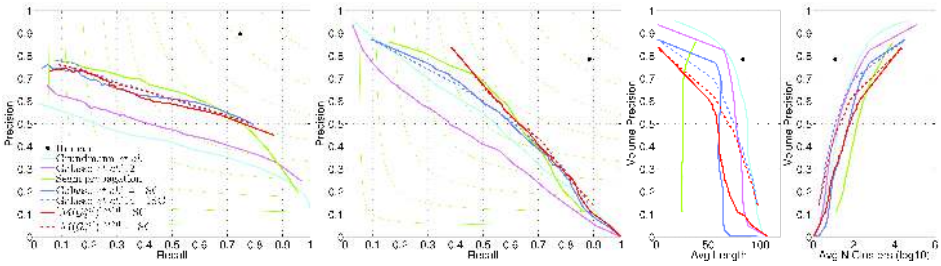


Fig. 3. Sample superpixels (SPX) and segmentation results of [18], compared with the proposed learned must-link variants, both when employing SC and 1-SC (cf. Section 4 for details). The proposed superpixels ($\mathcal{M}$ SPX) respect the video segmentation output while reducing the problem size. Additionally, $\mathcal{M}$ SPX improve results, *esp.* for 1-SC.

aging large powerful superpixels [3], *saturates* the few affinities of [18], which we also use here. Thus learned must-links closely follow the spectral clustering optimization and our proposed method only provides further reduction of the problem size. With similar arguments, as also maintained in [16], the segmentation propagation method of [17] is only partially outperformed, due to its more complex image features e.g. textures. Both observations suggest to use more complex features for learning. With respect to the efficient reduction of [16], we further reduce runtime by 30% and memory load by 65%, while we reduce runtime by 97% and memory load by 87% *wrt* [18].

In addition, we adopt 1-spectral clustering [22] within the baseline (Galasso et al. [16] - 1-SC), and compare this with our proposed method ($[\mathcal{M}(\mathcal{G}_2)]^{\mathrm{NCut}}$ - 1SC). Figure 4 shows that $[\mathcal{M}(\mathcal{G}_2^{SC})]^{\mathrm{NCut}}$ - 1SC results in the same performance on BPR and minor improvement on VPR, while significantly reducing runtime (by 70%) and memory load (by 65%) *wrt* [16].

Implementation details. We use the Random Forests implementation of [12]. The number of features to sample for each node split is set to $\sqrt{\mathrm{F}}$, where F is the dimensionality of the feature space. The averaged prediction of the individual trees is taken for prediction of the ensemble. As weak learners we use linear binary split functions and conic sections, and the forest size is set to 100 trees. The tree depth is varied in the range $[2, 12]$ and validated along with the threshold, which yields the largest number of must-links with zero false positives. Following [18], we extract the first 6 eigenvectors.



BPR VPR Length NCL

	BPR			VPR			Length	NCL
Algorithm	ODS	OSS	AP	ODS	OSS	AP	$\mu(\delta)$	μ
Human	0.81	0.81	0.67	0.83	0.83	0.70	83.24(40.04)	11.90
Grundmann *et al.* [20]	0.47	0.54	0.41	0.52	0.55	0.52	87.69(34.02)	18.83
Galasso *et al.*'12 [18]	0.51	0.56	0.45	0.45	0.51	0.42	80.17(37.56)	8.00
Segm. propagation [17]	0.61	0.65	**0.59**	**0.59**	**0.62**	0.56	25.50(36.48)	258.05
Galasso *et al.*'14 [16] - SC	**0.62**	0.65	0.50	0.55	0.59	0.55	61.25(40.87)	80.00
$[\mathcal{M}(\mathcal{G}_2^{SC})]^{\text{NCut}}$ - SC	0.60	**0.66**	**0.51**	**0.58**	**0.61**	**0.58**	51.72(39.90)	176.65
Galasso *et al.*'14 [16] - 1SC	**0.61**	**0.64**	**0.52**	0.55	0.60	0.54	69.80(42.26)	19.00
$[\mathcal{M}(\mathcal{G}_2^{SC})]^{\text{NCut}}$ - 1SC	**0.61**	**0.64**	0.51	**0.58**	**0.61**	**0.58**	60.48(43.19)	50.00

Fig. 4. Comparison of state-of-the-art video segmentation algorithms with our proposed method based on the learned must-links, on VSB100 [17] (cf. Section 4 for details).

5 Conclusions

We have formalized must-link constraints and proposed the relevant learning and inference algorithms. While this theory is applicable to general clustering and segmentation problems, we have particularly shown the use of learned must-link constraints in conjunction with spectral techniques, whereby recent theoretical advances employ these to reduce the original problem size, hence the runtime and memory requirements. Experimentally, we have shown that learned must-link constraints improve efficiency and, in most cases, performance, as these allow discriminatively training on GT data.

Acknowledgments

The authors would like to thank Syama Sundar Rangapuram for his support on the use of the 1-spectral clustering code.

References

1. Alpert, S., Galun, M., Brandt, A., Basri, R.: Image segmentation by probabilistic bottom-up aggregation and cue integration. TPAMI (2012)
2. Andres, B., Kappes, J.H., Beier, T., Köthe, U., Hamprecht, F.A.: Probabilistic image segmentation with closedness constraints. In: ICCV (2011)

3. Arbeláez, P., Maire, M., Fowlkes, C.C., Malik, J.: Contour detection and hierarchical image segmentation. TPAMI 33(5), 898–916 (2011)
4. Banica, D., Agape, A., Ion, A., Sminchisescu, C.: Video object segmentation by salient segment chain composition. In: ICCV, IPGM Workshop (2013)
5. Basu, S., Davidson, I., Wagstaff, K.: Constrained clustering: Advances in algorithms, theory, and applications (2008)
6. Breiman, L.: Random forests. Machine Learning (2001)
7. Brendel, W., Todorovic, S.: Video object segmentation by tracking regions. In: ICCV (2009)
8. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV (2010)
9. Bühler, T., Hein, M.: Spectral clustering based on the graph p-Laplacian. In: ICML (2009)
10. Chang, J., Wei, D., Fisher, J.W.: A video representation using temporal superpixels. In: CVPR (2013)
11. Cheng, H.T., Ahuja, N.: Exploiting nonlocal spatiotemporal structure for video segmentation. In: CVPR (2012)
12. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. In: Foundations and Trends in Computer Graphics and Vision (2012)
13. Eriksson, A.P., Olsson, C., Kahl, F.: Normalized cuts revisited: A reformulation for segmentation with linear grouping constraints. In: ICCV (2007)
14. Fowlkes, C., Malik, J.: How much does globalization help segmentation? Tech. rep., EECS – UC Berkeley (2004)
15. Fragkiadaki, K., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: CVPR (2012)
16. Galasso, F., Keuper, M., Brox, T., Schiele, B.: Spectral graph reduction for efficient image and streaming video segmentation. In: CVPR (2014)
17. Galasso, F., Nagaraja, N.S., Cardenas, T.Z., Brox, T., Schiele, B.: A unified video segmentation benchmark: Annotation, metrics and analysis. In: ICCV (2013)
18. Galasso, F., Cipolla, R., Schiele, B.: Video segmentation with superpixels. In: ACCV (2012)
19. Galasso, F., Iwasaki, M., Nobori, K., Cipolla, R.: Spatio-temporal clustering of probabilistic region trajectories. In: ICCV (2011)
20. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. In: CVPR (2010)
21. Hein, M., Bühler, T.: An inverse power method for nonlinear eigenproblems with applications in 1spectral clustering and sparse pca. In: NIPS (2010)
22. Hein, M., Setzer, S.: Beyond spectral clustering - tight relaxations of balanced graph cuts. In: NIPS (2011)
23. Jain, V., Turaga, S.C., Briggman, K.L., Helmstaedter, M., Denk, W., Seung, H.S.: Learning to agglomerate superpixel hierarchies. In: NIPS (2011)
24. Kamvar, S.D., Klein, D., Manning, C.D.: Spectral learning. In: IJCAI (2003)
25. Kannan, A., Jojic, N., Frey, B.J.: Generative model for layers of appearance and deformation. In: AISTATS (2005)
26. Kumar, M.P., Torr, P., Zisserman, A.: Learning layered motion segmentations of video. In: IJCV. pp. 301–319. No. 76 (2008)
27. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV (2011)
28. Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the future: Spatio-temporal video segmentation with long-range motion cues. In: CVPR (2011)

29. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV (2013)
30. Li, Z., Liu, J., Tang, X.: Constrained clustering via spectral regularization. In: CVPR (2009)
31. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17(4), 395–416 (December 2007)
32. Maire, M., Yu, S.X.: Progressive multigrid eigensolvers for multiscale spectral segmentation. In: ICCV (2013)
33. Maji, S., Vishnoi, N.K., Malik, J.: Biased normalized cuts. In: CVPR (2011)
34. Ng, A.Y., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS (2001)
35. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. TPAMI (2013)
36. Palou, G., Salembier, P.: Hierarchical video representation with trajectory binary partition tree. In: CVPR (2013)
37. Paris, S.: Edge-preserving smoothing and mean-shift segmentation of video streams. In: ECCV (2008)
38. Rangapuram, S., Hein, M.: Constrained 1-spectral clustering. In: AISTATS (2012)
39. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI (2000)
40. Sundaram, N., Keutzer, K.: Long term video segmentation through pixel level spectral clustering on gpus. In: ICCV Workshops (2011)
41. Sundberg, P., T.Brox, Maire, M., Arbelaez, P., Malik, J.: Occlusion boundary detection and figure/ground assignment from optical flow. In: CVPR (2011)
42. Taylor, C.: Towards fast and accurate segmentation. In: CVPR (2013)
43. Vazquez-Reina, A., Avidan, S., Pfister, H., Miller, E.: Multiple hypothesis video segmentation from superpixel flows. In: ECCV (2010)
44. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: ICML (2001)
45. Wang, X., Davidson, I.: Flexible constrained spectral clustering. In: KDD (2010)
46. Xu, C., Corso, J.J.: Evaluation of super-voxel methods for early video processing. In: CVPR (2012)
47. Xu, C., Xiong, C., Corso, J.J.: Streaming hierarchical video segmentation. In: ECCV (2012)
48. Xu, L., Li, W., Schuurmans, D.: Fast normalized cut with linear constraints. In: CVPR (2009)
49. Yu, S.X., Shi, J.: Grouping with bias. In: NIPS (2001)