Unsupervised Spatio-Temporal Segmentation with Sparse Spectral Clustering

Mahsa Ghafarianzadeh[1]
masa@gwu.edu
Matthew B. Blaschko[2]
matthew.blaschko@inria.fr
Gabe Sibley[1]
gsibley@gwu.edu

[1] Computer Science Department
The George Washington University
Washington DC, USA

[2] École Centrale Paris
INRIA Saclay
Châtenay-Malabry, France

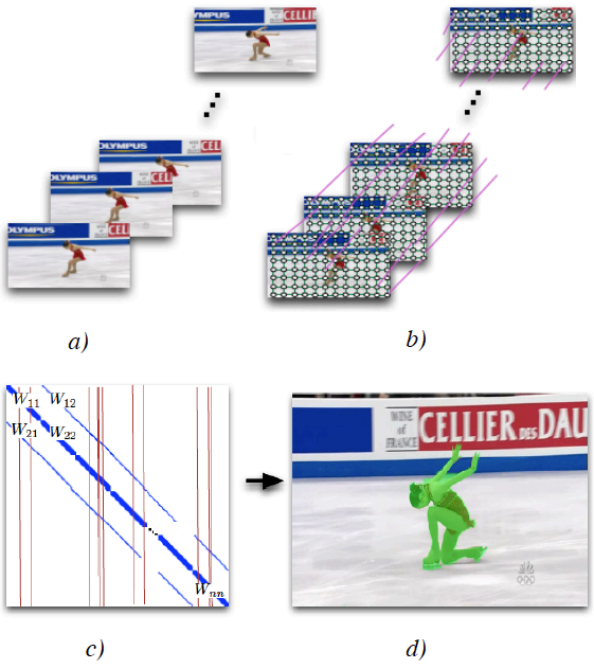

Figure 1: Approach overview: given a sequence of images a) we construct a graph b) connecting pixels in a neighborhood and also to their temporal correspondences. This is represented as a very large sparse matrix, from which we select a random subset of columns (which correspond to pixels) – only the randomly selected pixels are used for graph and matrix construction. d) we employ spectral clustering segmentation based on efficient and accurate low rank factorization based on the Nyström method to approximate the graph Laplacian.

Spatio-temporal cues are powerful sources of information for segmentation in videos. In this work we present an efficient and simple technique for spatio-temporal segmentation that is based on a low-rank spectral clustering algorithm. The complexity of graph-based spatio-temporal segmentation is dominated by the size of the graph, which is proportional to the number of pixels in a video sequence. In contrast to other works, we avoid oversegmenting the images into super-pixels and instead generalize a simple graph based image segmentation. Our graph construction encodes appearance and motion information with temporal links based on optical flow. For large scale data sets naïve graph construction is computationally and memory intensive, and has only been achieved previously using a high power compute cluster. We make feasible for the first time large scale graph-based spatio-temporal segmentation on a *single core* by exploiting the sparsity structure of the problem and a low rank factorization that has strong approximation guarantees.

The central contribution of this paper is to introduce a set of strategies that enable us to compute a dense graph based segmentation using a single processor and fitting in core memory. This is done primarily through two innovations: (i) we exploit the sparsity structure of the spatio-temporal graph, and (ii) we make use of an efficient and accurate low rank factorization based on the Nyström method to approximate the graph Laplacian in a spectral clustering approach. Our results show that not all of the pixels contain meaningful information about images, and just a subset of pixels can be a good representation of the entire scene.

Given an image I, we create a graph $G = (V, E, W)$, where the graph nodes V are the pixels in the image and are connected by edge E if they are within distance r from each other. W measures the similarity of pixels connected by an edge. We define W as the following: $W_{ij} = \exp \frac{-d^2(s_i,s_j)}{\sigma_i \sigma_j}$ where $W_{ij} = 0$ for $i =$j, and s_i denotes pixel color and $d(s_i, s_j)$ is the Euclidean distance. σ is a local scaling parameter [3] which takes into account the local statistics of the neighborhood around pixels i and j. Local scaling parameter is defined by: $\sigma_i = d(s_i, s_k)$ where s_k is the K'th neighbor of pixel i. In order to extend this to video, we make use of optical flow and add temporal motion information to the graph. We use optical flow to compute the motion vectors between frames. Then we connect pixel (x, y) in frame t to its 9 neighbors along the backward flow (u, v) in frame $t-1$, *e.g.* $(x+u(x,y)+\delta_x, y+v(x,y)+\delta_y)$ for $\delta_x, \delta_y \in \{-1, 0, 1\}$.

The similarity matrix for the video is a sparse symmetric block diagonal matrix of the size $n = number\ of\ frames \times number\ of\ pixels\ in\ one\ frame$.

Next, we use a time and space efficient spectral clustering via column sampling [1], that is similar to Nyström method, but with a further rank-k approximation of the normalized Laplacian using the sampled sub-matrix of the similarity matrix. This algorithm has shown promising results, since it reduces the time and space complexity of Nyström method and also it is able to recover all the degree information of the selected points. The time complexity of the algorithm is $O(nmk)$ and there is no need to store large similarity matrix W or its sampled columns in the memory. Also we are using the proposed inexpensive algorithm by [1] to orthogonalized estimated eigenvectors.

After performing spectral clustering on the similarity graph and obtaining the clusters, we first quantize each cluster into 256 bins (16 bins for each channel) and compute the *RGB* histogram. Then we merge adjacent clusters repeatedly if their similarity is more than a threshold τ to achieve the final segmentation.

We compared our method against other dense and sparse methods and achieved comparable performance while using just a subset of pixels (30%-50%) to label all of the pixels. In conclusion, we have demonstrated a novel method for spatio-temporal segmentation of dense pixel trajectories based on spectral clustering. We found that fully connecting pixels to their spatial neighbors within a given radius is an effective strategy for improving segmentation accuracy. Additionally, we use optical flow to more accurately compute temporal connectivity than a simple method based on an interpretation of the video sequence as a 3D volume. In contrast to previous work, we do not resort to super-pixel segmentation to achieve computational tractability and memory efficiency. Instead, we exploit the natural sparsity structure of the graph, and employ a low rank approximation of the Laplacian closely related to the Nyström method. We have found that sampling 30-50% of the pixels to index columns of the low rank approximation leads to comparable performance with a method that uses 100% of the columns. This strategy results in a spectral clustering method that can run on a single processor with the graph representation fitting in core memory. We have demonstrated the effectiveness of the approach on the Hopkins 155 data set, where we have achieved the best reported results for dense segmentation using an order of magnitude less computation than [2].

[1] Mu Li, Xiao-Chen Lian, James T Kwok, and Bao-Liang Lu. Time and space efficient spectral clustering via column sampling. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2297–2304. IEEE, 2011.

[2] Narayanan Sundaram and Kurt Keutzer. Long term video segmentation through pixel level spectral clustering on gpus. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 475–482. IEEE, 2011.

[3] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *NIPS*, volume 17, page 16, 2004.