A 3D Convolutional Approach to Spectral Object Segmentation in Space and Time

Elena Burceanu[1,2,3], Marius Leordeanu[1,2,4]

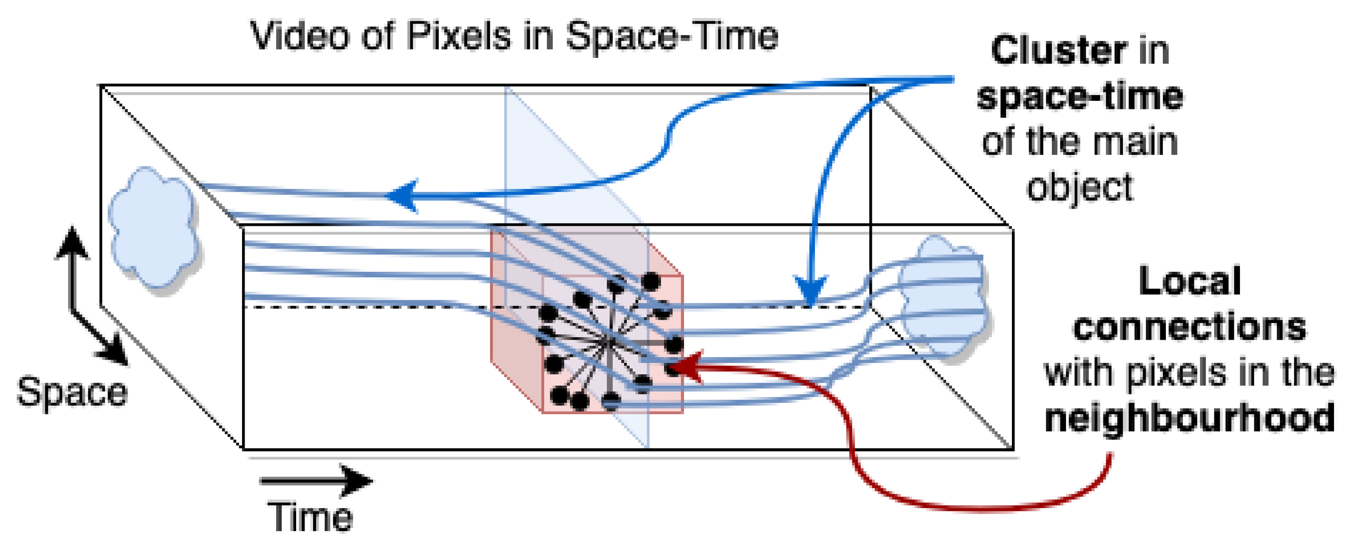
[1]Bitdefender, [2]Institute of Mathematics of the Romanian Academy, [3]University of Bucharest, [4]University Politehnica of Bucharest, Romania

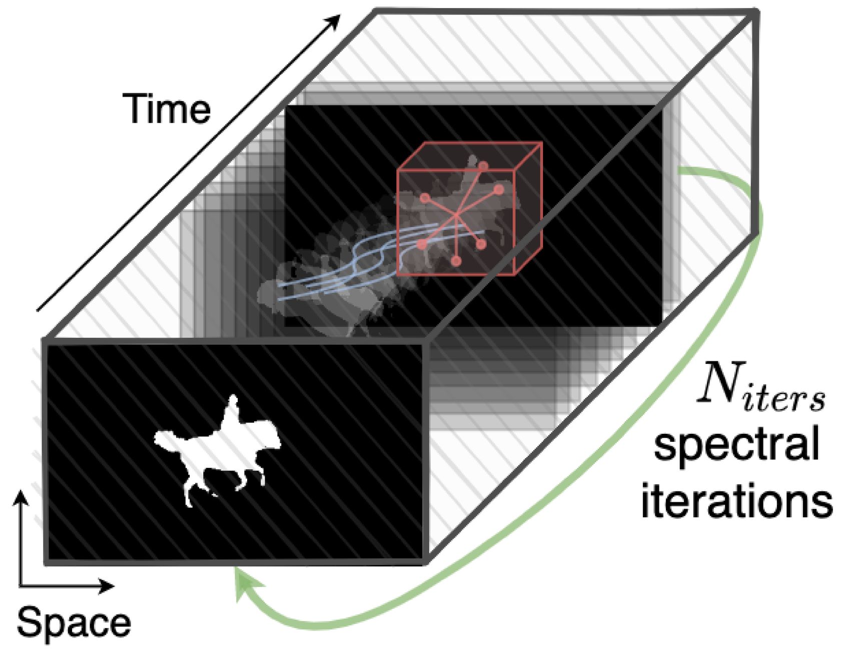
Contribution

1. **Formulating segmentation in video** as a problem of finding the **main space-time cluster**, represented by the leading eigenvector of the pixel-level adjacency matrix of the **video's graph in space-time**.
2. **Fast algorithm: SFSeg** is a **3D spectral filtering algorithm**, that computes the main eigenvector **without explicitly computing the graph's adjacency matrix**. This transforms the problem into a **tractable** one.
3. **State-of-the-art results** on **DAVIS-16** and **SegTrack2** datasets.
4. **Refinement**: **SFSeg** can be used as a powerful refinement method. It is also **faster and more accurate** then the well known space-time approach using CRF (denseCRF).

1. Formulation

- Improve the **instance segmentation** performance by tackling the problem of **better integrating the temporal aspect** when working with a video.





Instance segmentation in video as a spectral graph clustering problem in space and time, accurate and efficient at dense pixel-level.

2. SFSeg algorithm

- We consider **object segmentation** as a **graph partitioning** problem.
- **Nodes in the graph are pixels** from the **video space-time volume**, and **edges are relations** based on their similarities at the level of color or higher level features.
- **Segmentation solution** is the **leading eigenvector**.

Leading eigenvector of the **adjacency matrix M** → compute it with **power iteration**

$$\mathbf{M}_{i,j} = \mathbf{s}_i^p \mathbf{s}_j^p e^{-\alpha(\mathbf{f}_i-\mathbf{f}_j)^2-\beta \mathbf{dist}^2_{i,j}} = \mathbf{s}_i^p \mathbf{s}_j^p e^{-\alpha(\mathbf{f}_i-\mathbf{f}_j)^2}\mathbf{G}_{i,j}$$

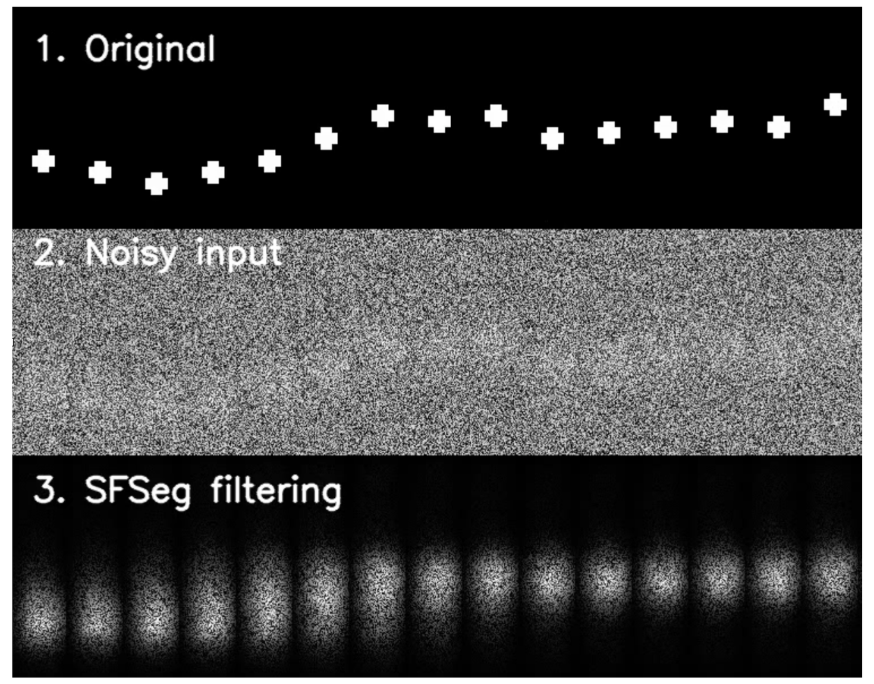
$$\approx^{Taylor} \underbrace{\mathbf{s}_i^p \mathbf{s}_j^p}_{\text{unary terms}} \underbrace{[1-\alpha(\mathbf{f}_i-\mathbf{f}_j)^2]\mathbf{G}_{i,j}}_{\text{pairwise terms}}$$

$$\mathbf{x}_i^{k+1} \leftarrow \sum_{j\in\mathcal{N}(i)} \mathbf{M}_{i,j}\mathbf{x}_j^k$$

Rewrite the classic power iteration into a special set of **fast 3D filtering operations**:

$$\mathbf{X}^{k+1} \leftarrow \mathbf{S}^p \cdot (\alpha^{-1}\mathbf{1} - \mathbf{F}^2) \cdot G_{3D} * (\mathbf{S}^p \cdot \mathbf{X}^k) - \mathbf{S}^p \cdot G_{3D} * (\mathbf{F}^2 \cdot \mathbf{S}^p \cdot \mathbf{X}^k) + 2\mathbf{S}^p \cdot \mathbf{F} \cdot G_{3D} * (\mathbf{F} \cdot \mathbf{S}^p \cdot \mathbf{X}^k)$$

3A. Space-time visualizations



- SFSeg **recovers most of the original segmentation**, even when it starts from a **very noisy input**.
- The clustering solution provides an **improved segmentation** of the main object.
- SFSeg **fluctuates less** compared with the input, while keeping track of the **detailed object shape**.

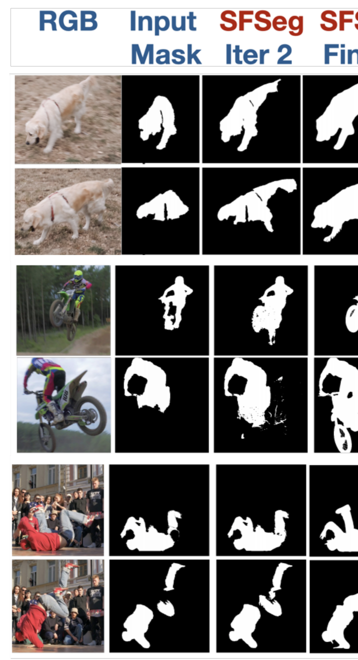
3B. Results

DAVIS-2016

	Input Method	Input Score (J)	**SFSeg** over Input (J)	Improved Videos (%)
Semi Supervised	OnAVOS	86.1	**86.3** (+0.2)	65
	OSVOS-S	85.6	**86.0** (+0.4)	90
	PReMVOS	84.9	**88.2** (+3.3)	90
	FAVOS	82.4	**83.0** (+0.6)	95
	OSMN	73.9	**75.9** (+2.0)	95
Un Supervised	COSNet	80.5	**80.9** (+0.4)	65
	MotAdapt	77.2	**77.5** (+0.3)	65
	PDB	77.2	**77.4** (+0.2)	60
	ARP	76.2	**77.7** (+1.5)	90
	LVO	75.9	**78.8** (+2.9)	90
	FSEG	70.7	**72.3** (+1.6)	95
	NLC	55.1	**55.6** (+0.5)	65
	Average Boost		+1.1%	80%

SegTrack v2

Method	Score (J)
LVO	57.3
FSEG	61.4
OSVOS	65.4
NLC	67.2
MaskTrack	70.3
BB + SFSeg + denseCRF (ours)	**72.7**



denseCRF

Method	DAVIS (J)	SegTrackv2 (J)
BB	67.2	72
BB + denseCRF	68.1	72
BB + SFSeg	68.7	72.1
BB + SFSeg + denseCRF	**69.2**	**72.7**

References

[1] Leordeanu and Hebert, ICCV 2005
[2] Shi and Malik, 2000
[3] Meila and Shi, AISTATS 2001