Spacetime Graph Optimization for Video Object Segmentation

Haller Emanuela
ehaller@bitdefender.com

29 June 2019

Emanuela Haller

- PhD student
  - Coordinators:
    - Marius Leordeanu
      (Institute of Mathematics of the Romanian Academy)
    - Adina Magda Florea
      (University Politehnica of Bucharest)
- Researcher
  - Bitdefender

Table of contents

Video object segmentation

- Video frames ⇒ Object segmentation masks
- Unsupervised task

- Object of interest
- Object / Group of strongly connected objects
- Most noticeable
- What is the sequence about

Perazzi et al. [10]

Video Object Segmentation

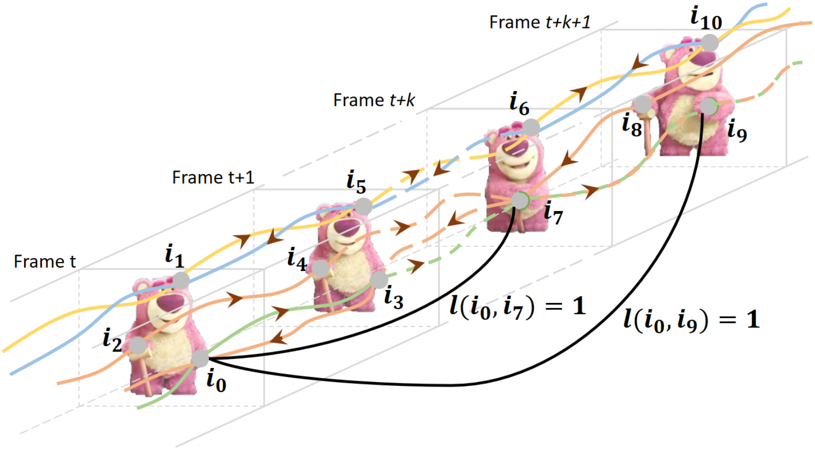
Perazzi et al. [10]

Motivation

- Move beyond traditional frame by frame approaches
- Exploit spacetime data
  - Use spacetime coherence as self-supervision signal
  - Accidental alignments are rare

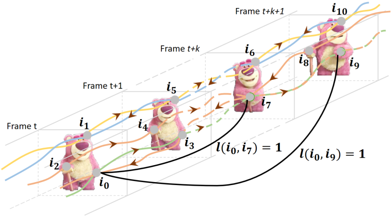
Spacetime graph

- Nodes connected through motion flows belong to the same object



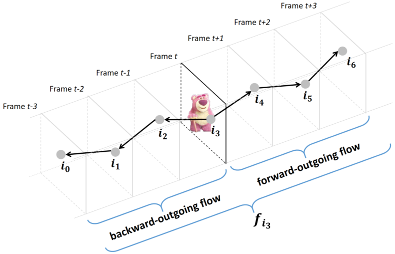
Spacetime graph

- $G = (V, E)$
  - Nodes correspond to video pixels
  - $|V| = n = m \cdot h \cdot w$
- Adjacency matrix
  - $\mathbf{M} \in \mathbb{R}^{n \times n}$
  - $\mathbf{M}_{i,j} = l(i,j) \cdot k(i,j)$
  - $l(i,j)$ - motion chains
  - $k(i,j)$ - $d_{temporal}(i,j)$



Space time graph

- Nodes features
  - $\mathbf{f}_i \in \mathbb{R}^{1\times d}$
  - Collected along outgoing motion flows
  - $\mathbf{F} \in \mathbb{R}^{n\times d}$
- Nodes labels
  - $x_i \in [0, 1]$
  - Soft segmentation labels
  - $\mathbf{x} \in \mathbb{R}^{n\times 1}$



Problem formulation

- Maximize graph clustering score
  - $\mathbf{S}_C = \sum_{i,j \in V} \mathbf{x}_i \mathbf{x}_j \mathbf{M}_{i,j} = \mathbf{x}^T \mathbf{M} \mathbf{x}$
  - Strong cluster in terms of motion flows
- Enforce feature-label consistency
  - $\|\mathbf{F}\mathbf{w} - \mathbf{x}\|_2$
  - Features should be able to predict node labels
- Subject to
  - $\|\mathbf{x}\|_2 = 1$
  - Interested in relative values of the labels

Problem formulation

- $(\mathbf{x}^*, \mathbf{w}^*) = \arg\max_{\mathbf{x},\mathbf{w}} S(\mathbf{x}, \mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{x}\|_2 = 1$

$$S(\mathbf{x}, \mathbf{w}) = \mathbf{x}^T \mathbf{M} \mathbf{x} - \alpha(\mathbf{F}\mathbf{w} - \mathbf{x})^T(\mathbf{F}\mathbf{w} - \mathbf{x}) - \beta \mathbf{w}^T \mathbf{w}$$

Algorithm

- Propagation:

$$\mathbf{x}^{(it+1)} \leftarrow \mathbf{M}\mathbf{x}^{(it)}$$

- Regression:

$$\mathbf{w}^{(it+1)} \leftarrow \left(\mathbf{F}^T\mathbf{F} - \beta\mathbf{I}_d\right)^{-1}\mathbf{F}^T\mathbf{x}^{(it+1)}$$

- Projection:

$$\mathbf{x}^{(it+1)} \leftarrow \mathbf{F}\mathbf{w}^{(it+1)}$$

Convergence

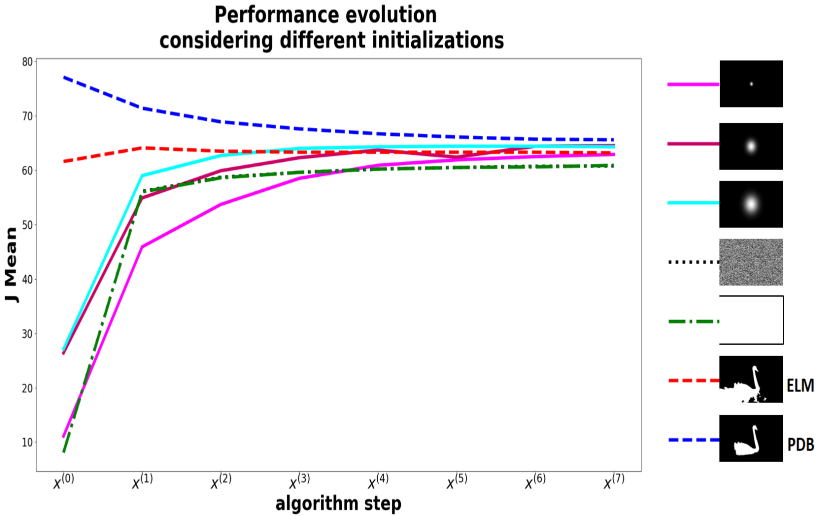
- Lead eigenvector of a specific matrix
- $\mathbf{x}^{(it+1)} = \frac{\mathbf{A}\mathbf{x}^{(it)}}{\|\mathbf{A}\mathbf{x}^{(it)}\|_2}$
- $\mathbf{A} = \mathbf{F}(\mathbf{F}^T\mathbf{F} - \beta\mathbf{I}_d)^{-1}\mathbf{F}^T\mathbf{M} = \mathbf{P}\mathbf{M}$
  - **P** - depends only on features
  - **M** - depends only on optical flow

Qualitative evolution over several iterations

random initialization

unsupervised features

Convergence - independence from initialization



The role of features

Quantitative Results - DAVIS dataset

Bitdefender

Task	Method		J Mean	F Mean	sec/frame
Unsupervised	Supervised features	PDB[12]	77.2	74.5	**0.05**
		ARP[7]	76.2	70.6	N/A
		LVO[14]	75.9	72.1	N/A
		FSEG[4]	70.7	65.3	N/A
		LMP[13]	70.0	65.9	N/As
		GO-VOS supervised + features of [12]	**79.9** (+2.7)	**78.1**	0.91
		GO-VOS supervised + features of [7]	78.7 (+2.5)	73.1	0.91
		GO-VOS supervised + features of [14]	77.0 (+1.1)	73.7	0.91
		GO-VOS supervised + features of [4]	74.1 (+3.5)	69.9	0.91
		GO-VOS supervised + features of [13]	73.7 (+3.7)	69.2	0.91
	Unsupervised	ELM[8]	61.8	**61.2**	20
		FST[9]	55.8	51.1	4
		CUT[6]	55.2	55.2	≈1.7
		NLC[2]	55.1	52.3	12
		GO-VOS unsupervised	**65.0**	61.1	**0.91**

Qualitative comparison

Quantitative Results - YouTube-Objects dataset

Bitdefender

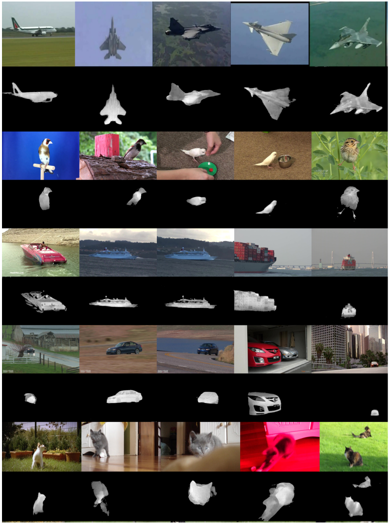
- YouTube-Objects v1.0

Method	aero	bird	boat	car	cat	cow	dog	horse	moto	train	avg	sec/frame
[11]	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5	N/A
[9]	65.4	67.3	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1	4
[15]	75.8	60.8	43.7	71.1	46.5	54.6	55.5	54.9	42.4	35.8	54.1	N/A
[5]	64.3	63.2	73.3	68.9	44.4	62.5	71.4	52.3	78.6	23.1	60.2	N/A
HPP[3]	76.3	71.4	65.0	58.9	68.0	55.9	70.6	33.3	69.7	42.4	61.1	0.35
[1]	77.0	67.5	**77.2**	68.4	54.5	**68.3**	72.0	**56.7**	44.1	34.9	62.1	0.04
GO-VOS unsupervised	**88.2**	**82.5**	62.7	**76.7**	**70.9**	50.0	**81.9**	51.8	**86.2**	**55.8**	**70.7**	0.91

- YouTube-Objects v2.2

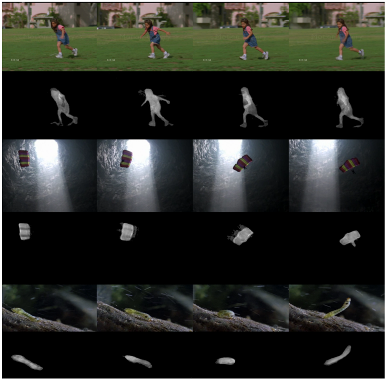
Method	aero	bird	boat	car	cat	cow	dog	horse	moto	train	avg	sec/frame
[1]	75.7	56.0	52.7	57.3	46.9	**57.0**	48.9	44.0	27.2	56.2	52.2	**0.02**
HPP[3]	76.3	68.5	**54.5**	50.4	**59.8**	42.4	53.5	30.0	**53.5**	**60.7**	54.9	0.35
GO-VOS unsupervised	**79.8**	**73.5**	38.9	**69.6**	54.9	53.6	**56.6**	**45.6**	52.2	56.2	**58.1**	0.91

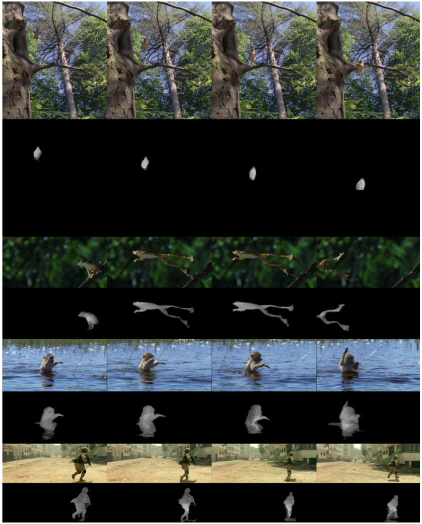
Qualitative Results - YouTube-Objects dataset



Quantitative & Qualitative Results
SegTrack dataset

Task	Method		IoU	sec/frame
Unsupervised	Supervised features	KEY [9]	57.3	>120
		FSEG [4]	**61.4**	N/A
		LVO [16]	57.3	N/A
		[10]	59.3	N/A
	Unsupervised	FST [11]	54.3	4
		CUT [6]	47.8	≈1.7
		HPP [3]	50.1	**0.35**
		GO-VOS unsupervised	**62.2**	0.91





Thank you!

References I

[1] I. Croitoru, S.-V. Bogolin, and M. Leordeanu. Unsupervised learning from video to detect foreground objects in single images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4335–4343, 2017.

[2] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014.

[3] E. Haller and M. Leordeanu. Unsupervised object segmentation in video by efficient selection of highly probable positive features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5085–5093, 2017.

[4] S. D. Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. *arXiv preprint arXiv:1701.05384*, 2(3):6, 2017.

References II

[5] Y. Jun Koh, W.-D. Jang, and C.-S. Kim. Pod: Discovering primary objects in videos based on evolutionary refinement of object recurrence, background, and primary object models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1068–1076, 2016.

[6] M. Keuper, B. Andres, and T. Brox. Motion trajectory segmentation via minimum cost multicuts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3271–3279, 2015.

[7] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 6, 2017.

References III

[8] D. Lao and G. Sundaramoorthi. Extending layered models to 3d motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[9] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784, 2013.

[10] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[11] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3282–3289. IEEE, 2012.

References IV

[12] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–731, 2018.

[13] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 531–539. IEEE, 2017.

[14] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. *arXiv preprint arXiv:1704.05737*, 2017.

[15] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia. Semantic object segmentation via detection in weakly labeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3641–3649, 2015.