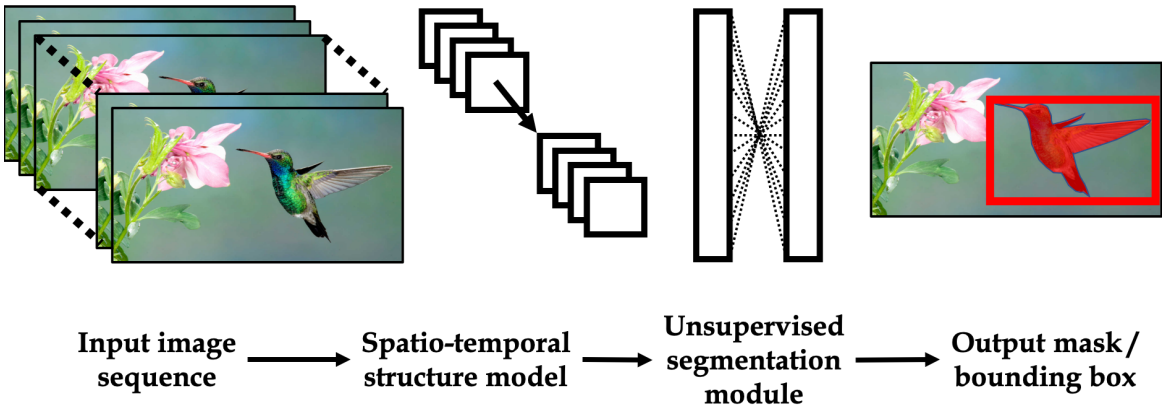
Unsupervised moving object segmentation in cluttered image sequences and videos

Michele Volpi, Nathanaël Perraudin and Fernando Pérez-Cruz

January 2021



1 Project description

Recent developments in deep learning have demonstrated unparalleled ability in accurately solving many computer vision tasks. Those usually involve estimation of attributes of interest from natural images, e.g. semantic object classes, bounding boxes, object masks. These achievements also translate to the processing of videos or, in general, image sequences. In this particular setting, one is usually interested to extract high-level visual attributes as done for natural still images. However, one does so by taking into account and exploiting temporal dependencies and structures across images (or frames) composing the input sequence. While supervised models learn to generalize from a given annotated training set (e.g. [1]), unsupervised models rely general on motion estimation [2], temporal cues [3]–[5], and spatio-temporal consistencies [5]–[8]. Unsupervised tasks can also be reformulated as self-supervised ones, in which a supervisory signal is extracted from the data itself without the need for external annotations. Notably, recent works deal with frame matching [9]–[11] or foreground-background mask propagation [12]. Interestingly, some of those approaches fill the gap between unsupervised and supervised learning strategies.

In this work, we aim at focusing on unsupervised video object segmentation models robust to motion blur, low sampling frame rate and occlusions. The reason is that this master thesis will be framed into a broader scope project, currently carried out at the Swiss Data Science Center, focusing on automated biodiversity monitoring using camera traps. In this scenario, complex and moving background (although camera pose is fixed) makes the detection of blurry and comparatively small moving objects extremely challenging. If localizing and segmenting moving

objects in biodiversity monitoring campaigns turns out to be possible with high accuracy, the whole biodiversity and wildlife monitoring field will undergo a revolution. Wildlife monitoring is currently bounded by the speed (and accuracy) of expert human annotators [13].

The project will develop along the following main axes: i) implementation of two or three baseline state-of-the-art methods (code is often available from paper authors); ii) assessment on challenging video benchmarks available and on hummingbird dataset; iii) development and adaptation of techniques to camera trap image sequences, with potential case studies involving detection of hummingbirds and moving objects; and iv) delivery of packaged, tested, executable and reproducible python code.

Datasets of interest

- DAVIS-2017 [link]
- YouTube-VOS: [link]
- Freiburg-Berkeley Motion Segmentation [link]
- Flying chairs [link]
- Hummingbirds [in house data]

  and potentially related:

- KITTI [link]
- MPI-Sintel [link]

2 Additional information

- **Difficulty of the project:** From moderate to very challenging.
- **What will you learn?** Applications of deep learning to a less common computer vision settings (image + time component), PyTorch, good scientific research practices.
- **Requirements:** Machine learning fundamentals, computer vision fundamentals, good python skills, good git understanding, motivation.
- **Supervisors:** Michele Volpi (michele.volpi@sdsc.ethz.ch), Nathanaël Perraudin (nathanael.perraudin@sdsc.ethz.ch) and Fernando Pérez-Cruz (fernando.perezcruz@sdsc.ethz.ch).

Bibliography

[1] F. Perazzi and et al., "Learning video object segmentation from static images," in *CVPR*, 2017.

[2] J. J. Yu, A. H. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *ECCV*, 2016.

[3] A. Dosovitstkiy and et al., "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015.

[4] J. Luitien, I. E. Zulfikar, and B. Leibe, "Unovost: Unsupervised offline video object segmentation and tracking," in *WACV*, 2020.

[5] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *AAAI*, 2018.

[6] Y.-T. Hu, J.-B. Huang, and A. G. Schwing, "Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation," in *ECCV*, 2018.

[7] D. Kang, J. Emmons, F. Abuzaid, P. Bailis, and M. Zaharia, "Noscope: Optimizing neural network queries over video at scale," in *arXiv:1703.02529*, 2017.

[8] S. Mahadevan and et. al., "Making a case for 3d convolutions for object segmentation in videos," in *BMVC*, 2020.

[9] Z. Lai and W. Xie, "Self-supervised learning for video correspondence flow," in *BMVC*, 2019.

[10] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," in *CVPR*, 2019.

[11] F. Zhu, L. Zhang, Y. Fu, G. Gui, and W. Xie, "Self-supervised video object segmentation," in *arXiv:2006.12480*, 2020.

[12] Y. e. a. Wang, "Occlusion aware unsupervised learning of optical flow," in *CVPR*, 2018.

[13] B. Weinstein, "Scene-specific convolutional neural networks for video-based biodiversity detection," *Methods in Ecology and Evolution*, vol. 9, no. 6, pp. 1435–1441, Jun. 2018.