Video Object Segmentation Using Adversarial Networks and Mathematical Morphology

Amin FEHRI
MINES ParisTech, PSL Research University, CMM - Center of Mathematical Morphology

Contact Information:
Email: amin.fehri@mines-paristech.fr

Problem

Video object segmentation is a two classes labelling problem desiring to separate foreground objects from the background region of a video.

- The choice of the object to follow across the video is highly dependant of the application.
- Thus this object is often specificied by the user on one or a few frames, making the video object segmentation a semi-supervised problem. It can also be seen as a mask propagation problem.

ID	Description
BC	*Background Clutter*. The back- and foreground regions around the object boundaries have similar colors (χ^2 over histograms).
DEF	*Deformation*. Object undergoes complex, non-rigid deformations.
MB	*Motion Blur*. Object has fuzzy boundaries due to fast motion.
FM	*Fast-Motion*. The average, per-frame object motion, computed as centroids Euclidean distance, is larger than $\tau_{fm} = 20$ pixels.
LR	*Low Resolution*. The ratio between the average object bounding-box area and the image area is smaller than $t_{lr} = 0.1$.
OCC	*Occlusion*. Object becomes partially or fully occluded.
OV	*Out-of-view*. Object is partially clipped by the image boundaries.
SV	*Scale-Variation*. The area ratio among any pair of bounding-boxes enclosing the target object is smaller than $\tau_{sv} = 0.5$.
AC	*Appearance Change*. Noticeable appearance variation, due to illumination changes and relative camera-object rotation.
EA	*Edge Ambiguity*. Unreliable edge detection. The average ground-truth edge probability (using [11]) is smaller than $\tau_e = 0.5$.
CS	*Camera-Shake*. Footage displays non-negligible vibrations.
HO	*Heterogeneus Object*. Object regions have distinct colors.
IO	*Interacting Objects*. The target object is an ensemble of multiple, spatially-connected objects (*e.g.* mother with stroller).
DB	*Dynamic Background*. Background regions move or deform.
SC	*Shape Complexity*. The object has complex boundaries such as thin parts and holes.

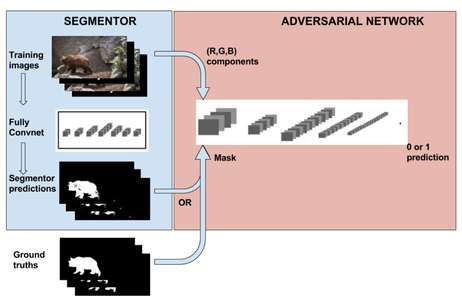
Difficulties met in video object segmentation. [4]

Contribution

- A video object segmenter trained using adversarial networks to learn a representation of the object, handling higher-order inconsistencies in the spirit of [2].
- An efficient morphological filters selection, which in practice complements and improves the performance of the convolutional network.

Representation Learning using Adversarial Networks

- Fully ConvNets approach as in SOTA [1, 3].
- Input: RGB + previous mask ; Output: probability map.
- Adversarial networks as a regularizer $\rightarrow$ label variables (foreground/background) are no more predicted independently from each other (as in [2]).

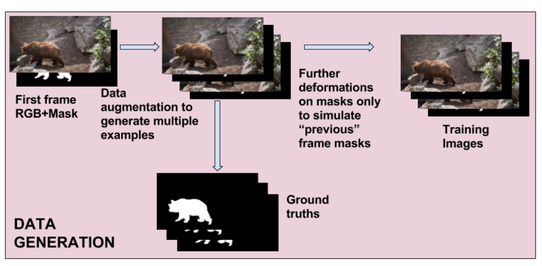


Fully ConvNets adversarial training.

- Loss function:

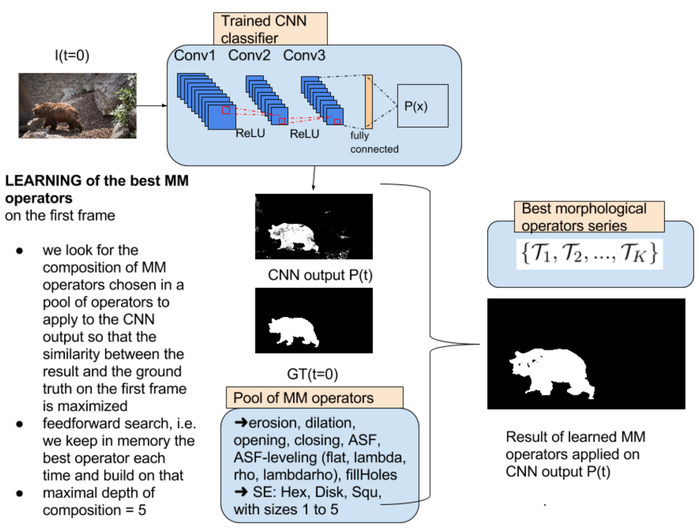
$$\begin{cases} l(\theta_s, \theta_a) = \sum_{n=1,\dots,N} l_{bce}(s(x_n), y_n) \\ -\lambda[l_{bce}(a(x_n, y_n), 1) + l_{bce}(a(x_n, s(x_n)), 0)] \\ \text{With: } l_{bce}(z, \hat{z}) = -[z \log(\hat{z}) + (1 - z) \log(1 - \hat{z})] \end{cases}$$

- Loss minimization w.r.t. θ_s the segmentor model parameters
- Loss maximization w.r.t. θ_a the adversarial model parameters



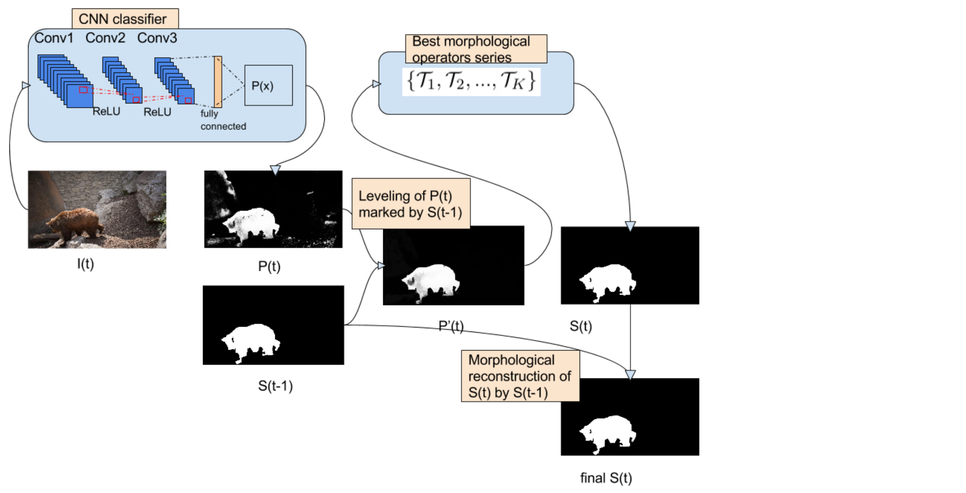
Data augmentation procedure

Learning a suited MM-based filtering of the CNN output



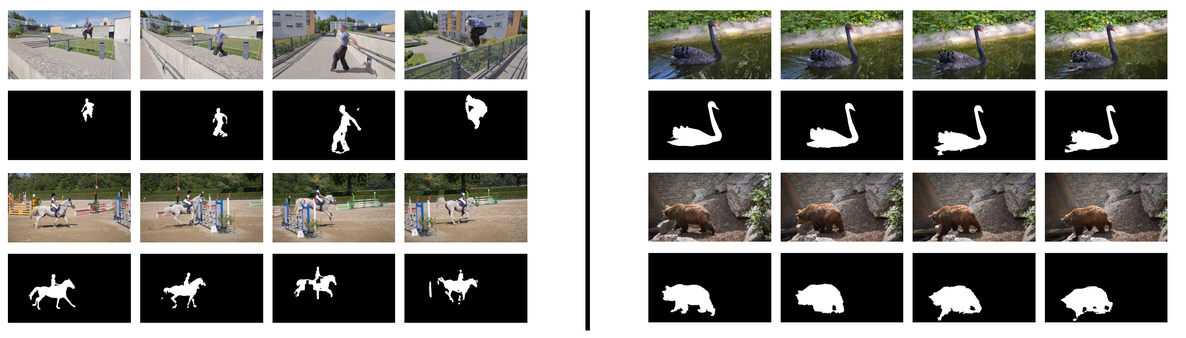
Following the object using MM

- Levelings: morphological filtering techniques allowing to filter an image without creating any new contours. We do a leveling of the filtered CNN output using the previous frame segmentation as a reference.
- Morphological reconstruction from one frame to the next one. Safeguard condition: the area of the object must not vary too much.



Ensuring temporal coherency using levelings and morphological reconstruction.

Preliminary results



Perspectives

- More thorough evaluation on reference databases (such as DAVIS [4]).
- Transfer learning to take advantage of powerful existing architectures.
- Learning the more suited morphological operators within the CNN (difficult because MM operators are non-differentiable for the most part).

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017.

[2] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic Segmentation using Adversarial Networks. In *NIPS Workshop on Adversarial Training*, 2016.

[3] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A.Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017.

[4] 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 *CVPR*, 2016.