Instance-level video segmentation from object tracks

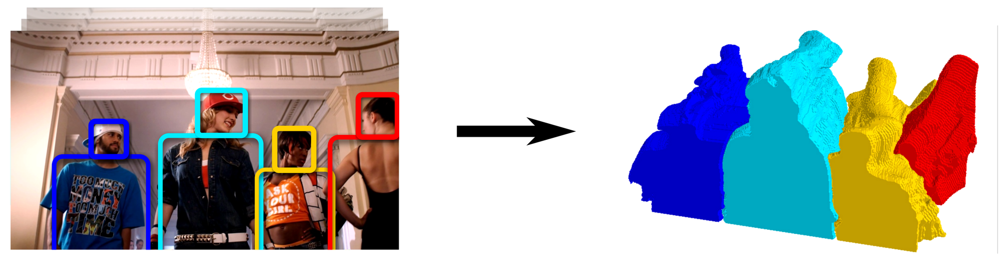
Guillaume Seguin* Piotr Bojanowski* Rémi Lajugie† Ivan Laptev*

*WILLOW Team / †SIERRA Team - Inria / École normale supérieure / CNRS - Paris, France

Partly supported by MSR-Inria and ERC grants Activia and VideoWorld.

Goal

- Object segmentation in video at instance level.
- Bounding box supervision only.



Motivation

- Manual pixel-wise annotation is expensive

- Segmenting each sheep in herd is difficult

+ Object bounding boxes can be used as form of weak supervision

+ Object detectors have reached maturity

+ Video provides redundant observations

Contributions

- Weakly-supervised model for object instance segmentation in video.
- Video dataset for instance-level person segmentation.

Overview

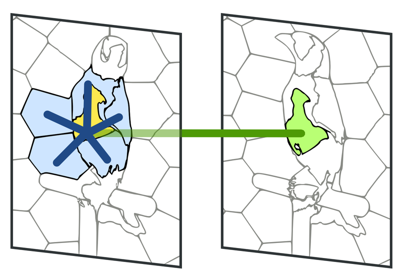
Input: video clip and object tracks.
Output: pixel-wise assignment of pixels to either background or object instance labels.

Pipeline:
- Segment video into superpixels using the TSP algorithm
- Use Optical Flow and appearance cues to measure similarity of neighboring superpixels
- Solve a graph labeling problem over superpixels, which uses:
  - a spatio-temporal grouping term to ensure local consistency of the solution,
  - a discriminative term to separate foreground from background, learning a long-term model of the target object class jointly from all frames, and
  - flexible linear constraints encoding priors derived from bounding boxes as to guide the segmentation.

RGB + Optical Flow + Object tracks + Superpixels

Video representation

N: number of superpixels.
K: number of labels (objects + background).
y : binary matrix in $\{0,1\}^{N\times K}$ which assigns superpixels to labels,
$y_{nk}=1$ iff superpixel n belongs to label k.



Model

- We formulate video segmentation as a graph-labeling problem.
- Our model is a quadratic cost function with linear constraints over the segmentation space:

$$E(y,\xi) = \boxed{E_G(y)} + \boxed{\alpha\ E_D(y)} + \boxed{\beta\|\xi\|^2}$$

Grouping term:

$$\boxed{E_G(y) = \frac{1}{N}\mathrm{Tr}(y^T L y).}$$

Laplacian matrix associated to the similarity matrix between adjacent superpixels

Penalizer for the slack variables of constraints

Foreground/background discriminative term:

$$\boxed{E_D(y) = \min_{\substack{w\in\mathbb{R}^{d\times 2}\\ b\in\mathbb{R}^2}} \frac{1}{N}\|yM - \Phi w - 1_N b^T\|_F^2 + \kappa\|w\|_F^2.}$$

mapping from K labels to foreground/background labels feature matrix model parameters regularization

with y s.t.

$$\boxed{\forall c, 0 \geq \sigma_c\left(\mathbb{1}_{R_c}^T\ y\ \mathbf{e}_{k_c} - \rho_c\right) - \xi_c.}$$

constraint id controls if it is an "at least" or "at most" constraint superpixels selector label selector threshold slack variable

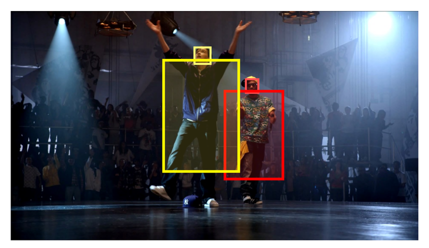
For instance, to enforce that the solution contains at least $\rho_c = 100$ pixels of region R_c assigned to object instance $k_c = 2$, we set $\sigma_c = -1$ and add the constraint $\mathbb{1}_R^T\ y\ \mathbf{e}_2 \geq 100 - \xi_c$.

Instance-level segmentation can be written as the **minimization of a quadratic cost under linear constraints:**

$$\min_{y\in\mathcal{Y},\ \xi\in\mathbb{R}_+^C} E(y,\xi) = \frac{1}{N}\Big(\mathrm{Tr}(y^T L y) + \alpha\mathrm{Tr}\big((yM)^T A(yM)\big)\Big) + \beta\|\xi\|^2.$$

Constraints

Linear constraints over y encode prior knowledge:



At least X_b % of pixels far from bounding boxes should belong to the background.

At least X_p% of pixels inside a bounding box must belong to the instance.

Pixels further than D_b pixels from a bounding box cannot belong to that instance.

Optimization

- Continuous relaxation on y, minimized with the Frank-Wolfe algorithm (only requires solving linear problems over $\mathcal{Y}$).
- Rounding to the closest integer point in terms of Frobenius norm.
- Non-convex refinement by adding $\mathrm{Tr}(y^T(1-y))$ to the cost and using the Frank-Wolfe algorithm again, to push the solution away from $1/K$ values, improving final segmentation quality.

Inria 3DMovie dataset v2

- New dataset for instance-level person segmentation in video from StreetDance 3D.
- Challenging poses and motions of people over 27 video clips, 2476 frames in total.
- Instance-level annotation of 632 people in 235 frames.

http://www.di.ens.fr/willow/research/instancelevel/

Results on multi-person segmentation

- Evaluated on the Inria 3DMovie dataset v2 using manually annotated ground truth object tracks, analyzing the influence of each component.

Method	F_1	Precision	Recall	Overlap
Ours	**78.3%**	**80.8%**	**77.3%**	**66.0%**
No temporal smoothness	76.4%	79.2%	75.4%	63.7%
Single frames	76.4%	77.9%	76.4%	63.7%
Grouping term only	77.6%	79.4%	77.2%	65.0%
Discriminative term only	66.9%	70.7%	64.7%	52.1%
No constraint	12.8%	10.4%	40.0%	09.0%
Convex only	75.6%	78.0%	74.1%	62.4%

- Comparison with baselines: purely unsupervised video segmentation [2,3], semantic segmentation [1,5] and video GrabCut [4] methods.

Method	F_1	Precision	Recall	Overlap
Ground truth tracks:				
Ours	78.3%	80.8%	77.3%	66.0%
Ours (+ semantic cue)	**80.1%**	81.9%	79.6%	**68.6%**
CRF as RNN [5]	78.5%	**83.2%**	77.7%	66.5%
Pose & segm. [4]	68.5%	68.3%	76.1%	55.0%
Multi-modal motion segm. [2]	27.4%	41.0%	30.4%	19.4%
FB/BG motion segm. [3]	52.2%	65.1%	49.8%	38.8%
SDS detections:				
Ours	72.5%	68.4%	**80.8%**	59.3%
SDS [1]	65.1%	73.5%	62.8%	52.6%

Qualitative results

Orig

[4]

Ours

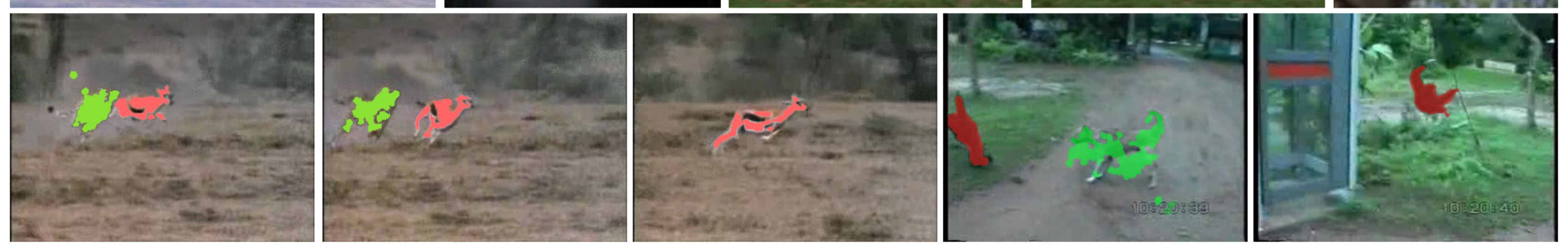
Results on SegTrack v1

- Our method is directly applicable to other object classes.
- Given readily-available object detectors/trackers no additional supervision is required.

Clip	No BB	BB tracks	GT BBs	[Jain '14]	[Fathi '11]
birdfall	221	**169**	**168**	**189**	342
cheetah	2196	1305	**724**	**1170**	**711**
girl	2733	**1606**	**1602**	2883	**1206**
monkeydog	2405	1021	**658**	**333**	**598**
parachute	305	**251**	278	**228**	**251**
penguin	**787**	848	**830**	**443**	1367

Legend: **Best** **2nd best** **3rd best**

No BB: no bounding box constraint, only simple constraints over the whole image.
BB tracks: automatic visual tracking (only takes first frame GT segmentation as input).
GT BBs: ground truth tracks.



Related work

[1] Hariharan et al., *Simultaneous Detection and Segmentation*, ECCV '14
[2] Ochs, Malik and Brox, *Segmentation of moving objects by long term video analysis*, PAMI '14
[3] Papazoglou and Ferrari, *Fast Object Segmentation in Unconstrained Video*, ICCV '13
[4] Seguin et al., *Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies*, PAMI '15
[5] Zheng et al., *Conditional random fields as recurrent neural networks*, ICCV '15
[6] Joulin, Bach and Ponce, *Discriminative Clustering for Image Co-segmentation*, CVPR '10
[7] Bojanowski et al., *Finding Actors and Actions in Movies*, ICCV '13