Unsupervised object segmentation in video by efficient selection of highly probable positive features

Emanuela Haller and Marius Leordeanu
ehaller@bitdefender.com, marius.leordeanu@imar.ro
Published in the IEEE International Conference on Computer Vision (ICCV), 2017

Introduction

- We introduce an efficient method for unsupervised foreground object segmentation in video.
- It has state of the art accuracy while being 10x faster than competition.

Key insights:

- Foreground and background are complementary and in contrast to each other, having different sizes, appearance and movements.
- Exploit foreground-background complementarity to select positive samples with high precision.
- Learn with highly probable positive (HPP) features.

Step 1 and 2 - VideoPCA [5]

Initial foreground regions are extracted using VideoPCA, where principal components analysis returns a subspace of the background in which the object is expected to be an outlier.

- $\mathbf{u}_i,\ i \in [0 \dots n_u]$ - principal components
- $\mathbf{f}_r \approx \mathbf{f}_0 + \sum_{i=1}^{n_u}((\mathbf{f} - \mathbf{f}_0)^\top \mathbf{u}_i)\mathbf{u}_i$ - frame $\mathbf{f}$ projected on the subspace
- $\mathbf{f}_{diff} = |\mathbf{f} - \mathbf{f}_r|$ - reconstruction errors
- High reconstruction errors => high foreground probabilities

Color segmentation:

- Automatically selected foreground pixels are used to estimate color distributions.
- $p(fg|c) = \frac{p(c|fg)}{p(c|fg)+p(c|bg)}$
- $p(c|fg) = \frac{n(c,fg)}{n(c)}$
- Discovered object masks are often very accurate.
- Computation is fast ($\approx$ 20 fps).

Step 3 - Object proposals refinement

Find consistencies in video soft-segmentations in order to reduce the noise.

- Compute PCA subspace associated with the soft-segmentations of previous step.
- Use projections on the computed subspace as new soft masks.

Step 6 - Motion

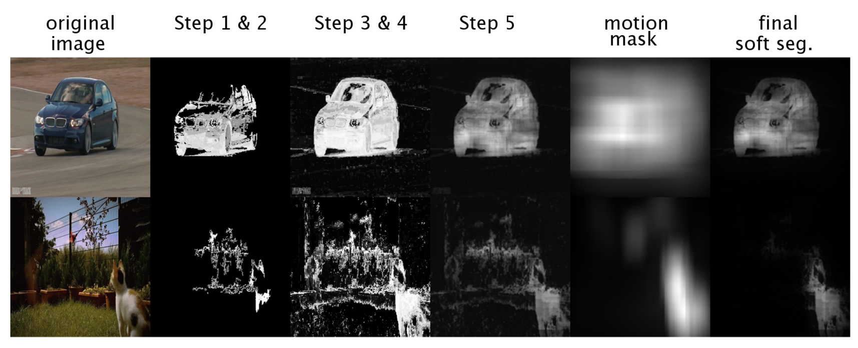
Objects usually have different motion patterns than the background.

- Learn affine background motion model and consider pixels with large deviations from this model as being more likely to belong to the object.
- For each background pixel consider $[\mathbf{I}_x, \mathbf{I}_y, x\mathbf{I}_x, x\mathbf{I}_y, y\mathbf{I}_x, y\mathbf{I}_y]$ and form the motion data matrix $\mathbf{D}_m$
- Estimate motion parameters: $\mathbf{w}_m = (\mathbf{D}_m{}^T\mathbf{D}_m)^{-1}\mathbf{D}_m{}^T\mathbf{I}_t$
- Compute model deviations as $|\mathbf{D}_m(p)\mathbf{w}_m - \mathbf{I}_t(p)|$

Approach

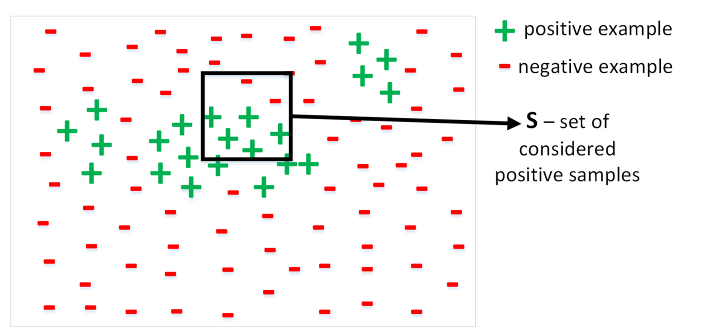
Step 1: select highly probable foreground pixels using PCA.
Step 2: estimate color distributions of foreground and background pixels based on the results of Step 1.
Step 3: refine the foreground masks by reducing the inconsistencies of the soft-segmentations computed at Step 2.
Step 4: re-estimate color distributions of foreground and background pixels based on the results of Step 3.
Step 5: train patch level discriminative classifier based on color co-occurences.
Step 6: combine appearance model with foreground motion cues.

■ pixel level ■ patch level ■ higher level



	Step 1&2	Step 3&4	Step 5	Step 6
F1 (SegTrack)	59.0	60.0	72.0	74.6
F1 (YTO)	53.6	54.5	58.8	63.4
sec/frame	0.05	0.03	0.25	0.02

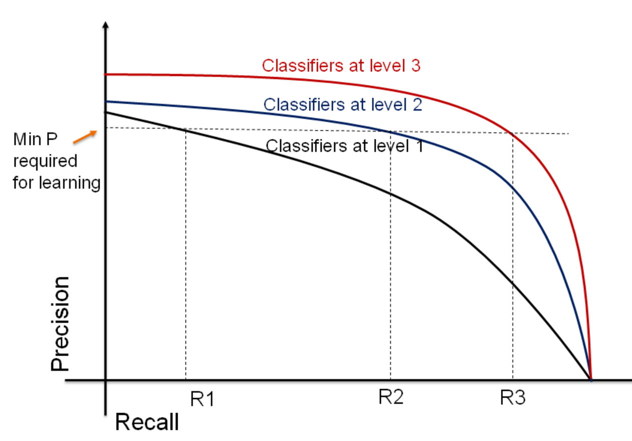
Learning with HPP features



Theoretical result:
$p(\mathbf{x}|S) > p(\mathbf{x}|\neg S) <=> p(\mathbf{x}|E_+) > p(\mathbf{x}|E_-)$
Assumptions:

- The object is smaller than the background.
- Features likelihood is independent of image location.
- S is selected with high precision.

We expect the classifier learned on S and $\neg S$ to have similar performances as one trained on true sets of positives and negatives (E_+ and E_-).



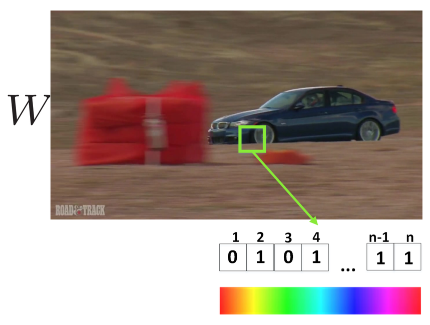
	Step 1&2	Step 3&4	Step 5
precision	66 → 70	62 → 60	64 → 74
recall	17 → 51	45 → 60	58 → 68
F-measure	27 → 59	53 → 60	61 → 72

Qualitative results

Step 5 - Color co-occurences

$\mathbf{d}_W$ descriptor of patch W

$$\mathbf{d}_W(c) = \begin{cases} 1 & \text{if color } c \text{ present in } W \\ 0 & \text{otherwise} \end{cases}$$



- Discretized HSV space => 1155 possible colors.
- It captures color co-occurences without detailed spatial constraints.

Unsupervised descriptor learning [2]:

$1155 \rightarrow k$ features, $k \ll 1155$
$\mathbf{w}^* = \mathrm{argmax}_{\mathbf{w}}\, \mathbf{w}^T\mathbf{C}\mathbf{w}$
s.t. $\sum_{i=1}^{n} w_i = 1, w_i \in [0, \frac{1}{k}]$

- The solution is guaranteed to be a binary mask which acts as a feature selector.
- Learn regularized regression model over the selected features.

Qualitative results

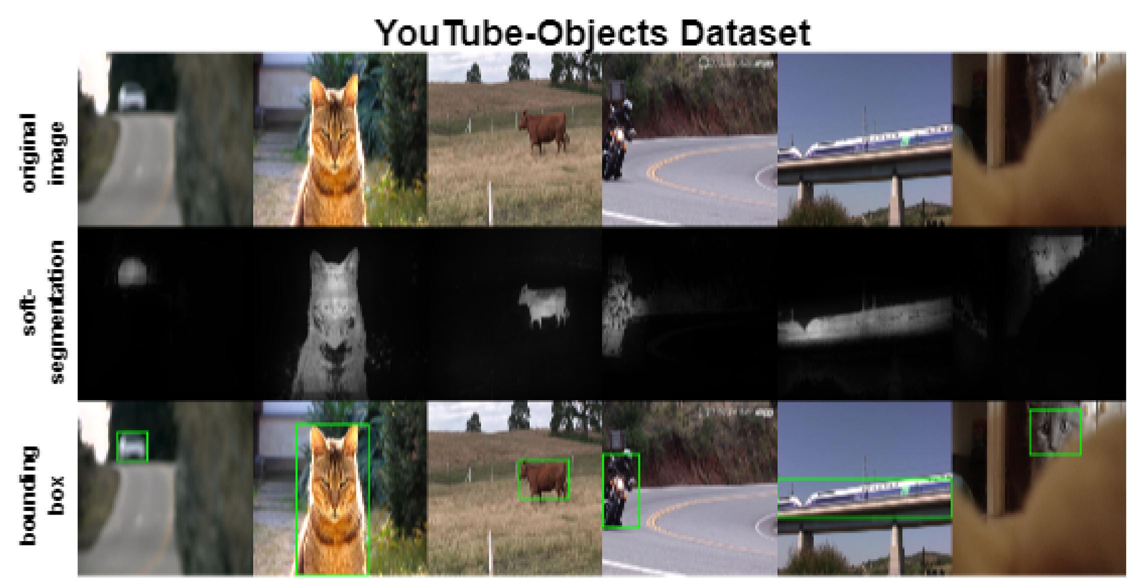
In the first example the appearance model of the car is sufficient for correct object segmentation. In the second example the motion model plays an important role in the discovery of the cat.

Quantitative results

Method / Supervised?	[1] Y	[6] Y	[4] N	[5] N	[3] N	Ours $v1.0$ N	Ours $v2.2$ N
aeroplane	64.3	75.8	51.7	38.3	65.4	**76.3**	76.3
bird	63.2	60.8	17.5	62.5	67.3	**71.4**	68.5
boat	73.3	43.7	34.4	51.1	38.9	**65.0**	54.5
car	68.9	71.1	34.7	54.9	**65.2**	58.9	50.4
cat	44.4	46.5	22.3	64.3	46.3	**68.0**	59.8
cow	62.5	54.6	17.9	52.9	40.2	**55.9**	42.4
dog	71.4	55.5	13.5	44.3	65.3	**70.6**	53.5
horse	52.3	54.9	**48.4**	43.8	**48.4**	33.3	30.0
motorbike	78.6	42.4	39.0	41.9	39.0	**69.7**	53.5
train	23.1	35.8	25.0	**45.8**	25.0	42.4	60.7
Avg	60.2	54.1	30.4	49.9	50.1	**61.1**	54.9
time sec/frame	N/A	N/A	N/A	6.9	4	**0.35**	

CorLoc scores - YouTube-Objects dataset.

Qualitative results



References:
[1] Koh et al. In: *ICCV*. 2016. [2] Leordeanu et al. In: *AAAI*. 2016. [3] Papazoglou et al. In: *ICCV*. 2013. [4] Prest et al. In: *CVPR*. 2012. [5] Stretcu et al. In: *BMVC*. 2015. [6] Zhang et al. In: *CVPR*. 2015.

Code and demo available online:
https://goo.gl/2aYt4s