Segmentation Free Object Discovery in Video

Giovanni Cuffaro, Federico Becattini, Claudio Baecchi, Lorenzo Seidenari, Alberto Del Bimbo
{giovanni.cuffaro,federico.becattini,claudio.baecchi,lorenzo.seidenari,alberto.delbimbo}@unifi.it

Introduction

- We propose to extend the Object Proposal paradigm from static images to video sequences.
- Temporal Proposal Tracks are generated by matching bounding boxes, relying on spatial correlations through time.
- No segmentation or visual features involved.
- Tracks can be used to discover meaningful content without any supervision.
- Novel entropy-based Object Proposal evaluation.
- Dataset-independent evaluation with no annotation required.

Method

- A set of Tracks is generated by matching proposal boxes (i.e. EdgeBoxes [1]) in consecutive frames of a video.
- A Track is a succession of bounding boxes b_i^k for which the Intersection over Union (IoU) between two boxes b_i^m (belonging to frame f_i) and b_{i+1}^n (belonging to frame f_{i+1}) is above a defined threshold θ_τ.
- Time To Live counter (TTL) τ indicating the number of frames before considering a track t_j terminated. Missing frames are linearly interpolated.

$$\tau_{i+1}(t_j) = \begin{cases} \tau_i(t_j) + 1, & \text{if } \exists\, n : \text{IoU}(b_i^m, b_{i+1}^n) > \theta_\tau \\ \tau_i(t_j) - 1, & \text{otherwise} \end{cases}$$

Motion Compensation

- To reduce unalignment, bounding boxes are registered with optical flow before computing IoU.
- A Track is composed of unregistered boxes.

Proposal Suppression

- Short tracks are removed to exclude occasional background matching.
- Still tracks are removed to exclude logos and writings impressed on the video.

Temporal NMS

- Standard Non-Maximal Suppression (NMS) is extended to consider time, computing IoU between volumes instead of areas.

$$\text{vIoU}(t_j, t_k) = \frac{v_j \cap v_k}{v_j \cup v_k}$$

Track Ranking

- Tracks are ranked according to EdgeBoxes scores E_t and IoU values between adjacent boxes in the track I_t.

$$S_t = \lambda E_t + (1 - \lambda) I_t$$

Entropy driven evaluation

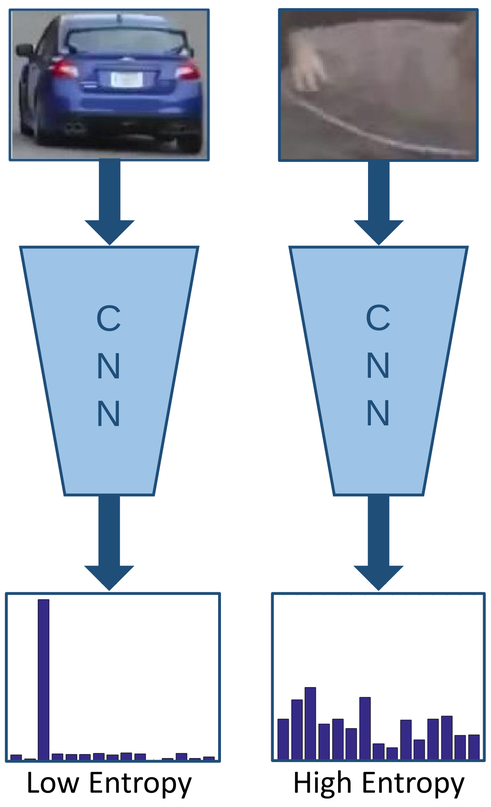
- Evaluation metrics for object proposals only consider how well annotated objects are convered and therefore depend on a dataset.
- To consider unannotated objects we propose an entropy based evaluation which indicates how the proposal is likely to be recognized as an object.
- Given a classifier capable of providing for an image a probability distribution $X = \{x_1, \ldots, x_N\}$ over N classes, we compute the Shannon entropy H for the probability vector X:

$$H(X) = -\sum_{i=1}^{N} x_i \log(x_i).$$

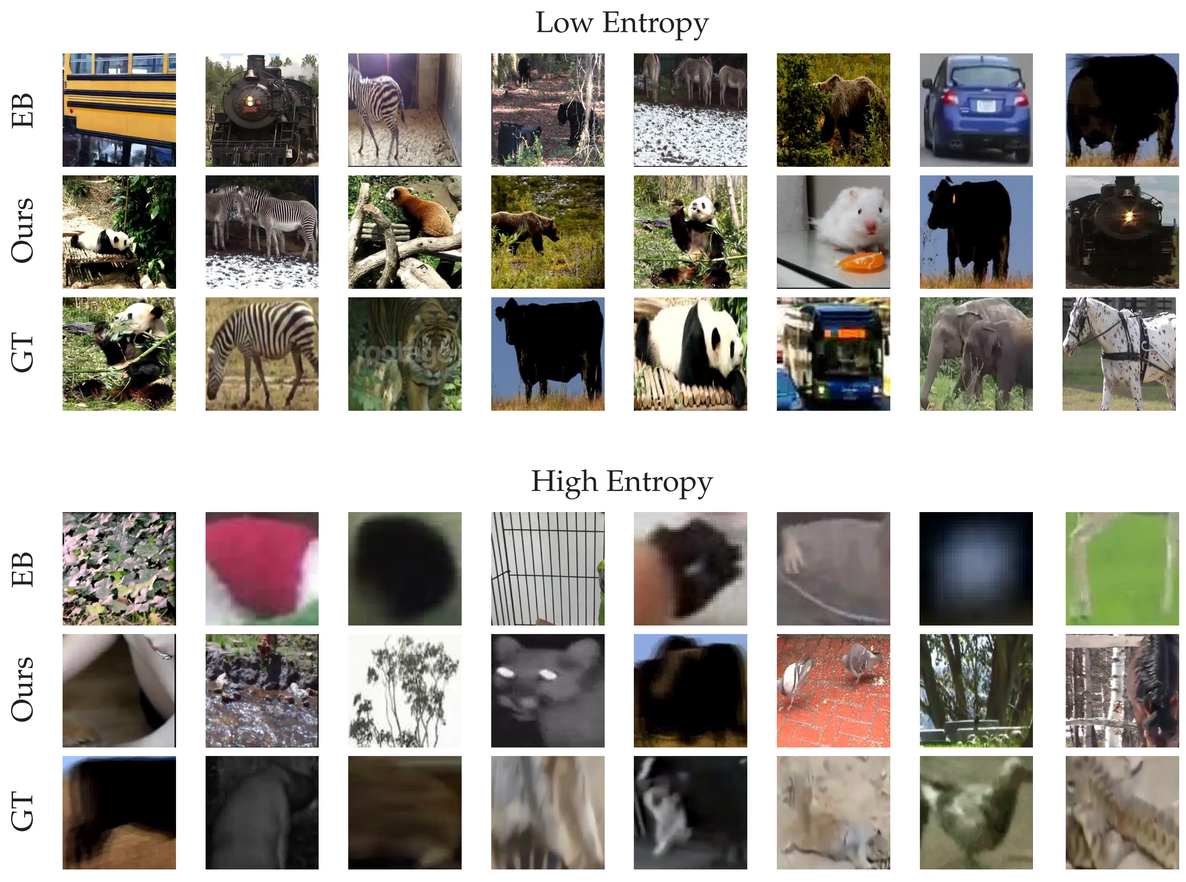
- If the classifier is able to cover effectively a sufficiently large number of classes, then the entropy can be interpreted as a measure of *objectness*.

Evaluation

- The entropy for the best 25 proposals is computed using both Tracks and the bounding box oracle (EdgeBoxes).
- For our method we select the highest scoring 25 tracks of each video and take the box with the best objectness score.

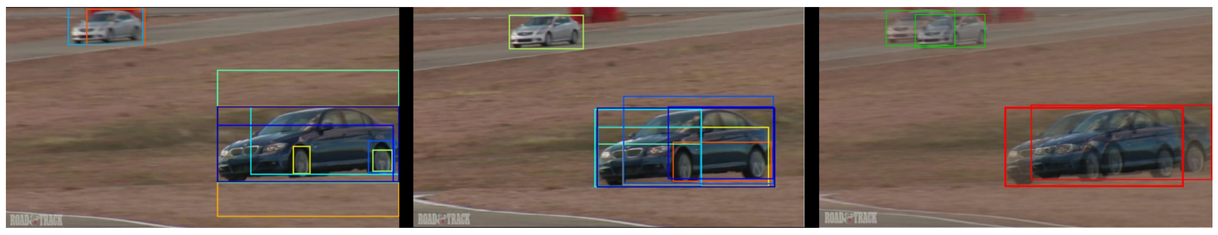


Low Entropy vs. High Entropy



- The proposals with the highest and the lowest entropy are reported for the VID dataset.
- Tracks (Ours) are compared to EdgeBoxes (EB) and the GroundTruth (GT)

Bounding box matching for Temporal Proposals



Experiments: Track Entropy

- Datasets
  - YouTube-Objects (YTO) [2]: 10 classes (subset of PASCAL VOC [3]), mostly one object per video.
  - ILSVRC2015-VID (VID) [4]: 30 classes, multiple objects per video.
- Etropy driven evaluation using VGG-16 [5] as classifier (1000-dimensional output).
- Results for YTO are shown in the Table. The same trend is observed on VID with an average Entropy of 4.73, 3.96 and 3.71 for EdgeBoxes, our method and the ground truth respectively.

Method											Mean
EB [1]	5.02	5.19	5.48	4.52	5.92	6.27	6.16	6.54	5.68	5.19	5.60
Ours	3.58	3.25	3.10	2.45	4.02	3.00	3.58	3.25	3.10	2.45	3.18
GT	0.58	1.33	1.03	1.83	2.31	2.41	2.28	2.58	2.57	2.41	1.93

Experiments: High Precision Proposals

- We evaluate the proposal method as an object detector on YTO, measuring mean Average Precision (mAP).
- Classification is performed with Fast-RCNN [6], restricted to the ten common classes.
- Tracks achieve a mAP 8.5 times higher than EdgeBoxes, proving that our proposal is much more precise.

Method											Mean
EB [1]	0.94	0.40	0.49	1.80	10.96	0.57	0.56	0.61	0.26	2.95	0.98
Ours	9.15	7.16	5.98	14.94	10.95	8.43	6.10	2.26	3.42	14.91	8.33

Qualitative Results

- The highest ranking proposals obtained for different videos with our method and with EdgeBoxes are shown.
- Our proposals are more focused on objects instead of background or object parts.

Discussion

- In this work we have shown how a generic image-based bounding box oracle can be extended to videos generating semantically meaningful Tracks.
- Completely unsupervised and no need of segmentation or visual descriptors.
- We introduce a novel proposal evaluation method based on entropies of classifier scores, which is dataset-independent and takes into account unannotated objects.
- Results show improvements against frame-wise proposals.
- Future work will focus on action proposals for tasks of action recognition and localization.

Bibliography

[1] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. of ECCV*, 2014.

[2] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, "Learning object class detectors from weakly annotated video," in *Proc. of CVPR*, IEEE, 2012.

[3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] R. Girshick, "Fast r-cnn," in *Proc. of ICCV*, 2015.