Segmentation Free Object Discovery in Video

Giovanni Cuffaro, Federico Becattini(✉), Claudio Baecchi,
Lorenzo Seidenari, and Alberto Del Bimbo

University of Florence, Florence, Italy
{giovanni.cuffaro,federico.becattini,claudio.baecchi,
lorenzo.seidenari,alberto.delbimbo}@unifi.it

Abstract. In this paper we present a simple yet effective approach to extend without supervision any object proposal from static images to videos. Unlike previous methods, these spatio-temporal proposals, to which we refer as "tracks", are generated relying on little or no visual content by only exploiting bounding boxes spatial correlations through time. The tracks that we obtain are likely to represent objects and are a general-purpose tool to represent meaningful video content for a wide variety of tasks. For unannotated videos, tracks can be used to discover content without any supervision. As further contribution we also propose a novel and dataset-independent method to evaluate a generic object proposal based on the entropy of a classifier output response. We experiment on two competitive datasets, namely YouTube Objects [6] and ILSVRC-2015 VID [7].

1 Introduction

Image and video analysis can be considered similar on many levels, but whereas new algorithms are continuously raising the bar for static image tasks, advancements on videos seem to be slower and hard going. What makes video comprehension more difficult is mainly the huge amount of data that has to be processed and the need to model an additional dimension: time.

We believe that focusing on relevant regions of videos, such as objects, will reduce the complexity of the problem and ease learning for models like Deep Networks. The same concept has been successfully applied to images using object proposals, which analyse low level properties, such as edges, to find regions that are likely to contain salient objects. Advantages are twofold, first the search space is considerably reduced, second, as a consequence, the number of false positives generated by classifiers is lowered.

In this work we propose a technique to include time into a generic object proposal, by exploiting the weak supervision provided by time itself to match spatial proposals between adjacent frames. This results in spatio-temporal tracks that represent salient objects in the video and can therefore be used instead of the whole sequence. To the best of our knowledge we are the first to adopt a fully unsupervised matching strategy that only relies on bounding box coordinates without any semantic content or visual descriptor apart from optical flow.

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part III, LNCS 9915, pp. 25–31, 2016.
DOI: 10.1007/978-3-319-49409-8_4

We also introduce a novel dataset-independent proposal evaluation method based on the entropy of classifier scores.

2 Related Work

Object proposals [3] provide a relatively small set of bounding boxes likely to contain salient regions in images, based on some *objectness* measure. Different proposals, such as EdgeBoxes [12], are commonly used in image related tasks to reduce the number of candidate regions to evaluate. Recently, there have been some attempts to adapt the paradigm of object proposals to videos to solve specific tasks, by generating consistent spatio-temporal volumes. In [6] motion segmentation is exploited to extract a single spatio-temporal tube for video, in order to perform video classification. The task of object discovery is tackled in [9] by generating a set of boxes using a foreground estimation method and matching them across frames using both geometric and appearance terms. Kwak *et al.* [4] combine a discovery step matching similar regions in different frames and a tracking step to obtain temporal proposals. In [5] a classifier is learnt to guide a super-voxel merging process for obtaining object proposals. Temporal proposals have been exploited to segment objects in videos in [10] by discovering easy instances and propagating the tube to adjacent frames. Other methods to generate salient tubes have been proposed for action localization in [11] using human and motion detection.

Differently from the above approaches we do not rely on segmentation, which is a time-consuming task especially for videos. Our method is simply based on the response of a frame-wise proposal method. The weak supervision obtained from the temporal consistency of the video is exploited to generate tracks. Our method aims at generating few, highly precise, tracks containing objects in the video.

3 Video Temporal Proposals

In this section we introduce the concept of "track", describing in details how these are generated from a set of bounding boxes extracted by an object proposal in the video.

Given a video V, for each frame f_i we extract a set B_i of bounding boxes b_i^k using an object proposal. We propose a method to match boxes that exhibit a temporal consistency in consecutive frames through the video, yielding to a set T of tracks t_j. A track is defined as 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.

Starting from the first frame, each time a match is found, the corresponding bounding box is added to the end of the track and becomes the reference box for the following frame. If no match is found the last box of the track is compared with the following frames until a good match is obtained. An example of matching is shown in Fig. 1.

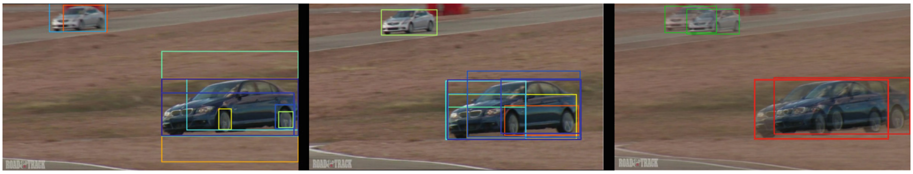

Fig. 1. Example of frame matching; matched boxes are inserted into their respective track. *(left)* reference frame where the top 10 proposals extracted with EdgeBoxes are shown; *(center)* following frame with top 10 EdgeBoxes proposals; *(right)* 2 matched proposals between the two frames; these will be part of two different tracks

When one or more consecutive matches are not found, tracks become fragmented, i.e. there are frames for which a track is active but there is no bounding box. This is usually due to a lack of good bounding boxes for that frame, occlusion or appearance changes of the object. It is thus necessary to avoid matching boxes in frames too far apart that therefore do not represent the same content, but at the same time we want to be able to tolerate some missing boxes without prematurely terminating the track.

To this end we introduce a Time to Live counter (TTL) τ for each track. We define $\tau_i(t_j)$ as the number of frames, at frame i, that the method can still wait before considering the track t_j terminated. TTL starts from an initial value γ; each time a box can not be matched in a consecutive frame the TTL is decremented, otherwise is incremented (up to γ). More formally, given a track t_j and its last bounding box b_i^m we increment or decrement its TTL as follows:

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

When the TTL for a track reaches 0, the track is considered terminated. Missing frames caused by track fragmentation are linearly interpolated using the positions of the previous and following bounding boxes in the track.

Proposal Motion Compensation. Proposals around objects in consecutive frames are usually unaligned due to movements of the object or the camera. This causes the IoU score to decrease even if the matching is good. We work around this problem by registering the boxes with optical flow before computing the IoU. The registration is performed on the last box of each track, by computing the mean offsets along the x and y axes inside the boxes. Shifted boxes are only used for matching and tracks consist only of unaltered boxes.

Temporal NMS. As in the spatial case, temporal proposals also suffer of high redundancy. To reduce this effect we extend spatial non-maximal suppression to time, defining a temporal NMS where instead of computing IoU on areas it is computed over volumes (vIoU). If α_j^k is the area of the k-th bounding box in track t_j, then the volume υ_j of the track is calculated as $\upsilon_j = \sum_{k\,=\,0}^{K} \alpha_j^k$ where K is the length of the track. Then, vIoU is defined as:

$$\mathrm{vIoU}(t_j, t_k) = \frac{v_j \ \cap \ v_k}{v_j \ \cup \ v_k} \tag{2}$$

Using vIoU we apply the standard NMS.

Proposal Suppression. Once all the tracks are computed for a given video, we apply a post-processing to remove the ones that are unlikely to represent an object. To this end we remove those tracks which have a length smaller than a value l. In this way we exclude very short tracks that are likely to be composed by background boxes that happen to have a high IoU.

Another problem is posed by logos and writings impressed on the video. In fact both of these are very well located by an object proposal but are usually of no interest. To prevent such objects to be considered as valid tracks, we take the mean optical flow magnitude in all the boxes of the track and we discard it if under a threshold s.

Track Ranking. It is important to compute a score for temporal proposals, in order to account for the likelihood of objects in such proposal. To this end, we propose to consider two factors: the object proposal score used to generate the bounding boxes at each frame and the values given by the IoUs between frames of the tracks. For the former we define E_t as the mean of the scores given by EdgeBoxes, for the latter we define I_t as the mean of all the IoUs of the frames in the track. Using these two figures we define a track score as:

$$S_t = \lambda E_t \ + \ (1 - \lambda) I_t \tag{3}$$

where $\lambda \in [0, 1]$ is a weighting factor used to balance the contributions of the two scores.

4 Method Evaluation

Object proposals are usually evaluated measuring how well objects are covered by the generated boxes. These kind of evaluation does not take into account unannotated objects, and therefore provide a benchmark not reflecting the real capabilities of the proposal method.

The method presented in this paper is a general framework for discovering salient spatio-temporal tracks in videos, which is built upon a generic bounding box oracle. To evaluate it, we introduce a novel method to establish the effectiveness of a generic video proposal, which is also dataset-independent since it does not rely on annotations. We evaluate whether a proposal effectively represents an instance of some object, since the goal of an object proposal is to locate good candidates and not to produce the candidate of a given class (i.e. the one of the ground truth). To this end 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 $X, H(X) = -\sum_{i=1}^{N} x_i \mathrm{log}(x_i)$.

The rationale behind this choice is that, given a good classifier, for a known object the output probability distribution will be high for the relative class and near zero for the others, thus producing a small entropy. On the contrary, for inputs that the classifier is unsure of, e.g. background patches, the output probability will be distributed non-uniformly among all the possible classes, resulting in a higher entropy. Therefore, 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* for the given proposal.

5 Experiments

We experiment on the YouTube-Objects (YTO) [6] and on the ILSVRC2015-VID (VID) datasets [7], which both provide a per-frame annotation of the objects. YTO is composed by 10 classes and most videos contain a single object per video. VID instead is a more challenging dataset with 30 classes with multiple objects per video.

Here we evaluate our method using the entropy measure introduced in Sect. 4. In all experiments we use EdgeBoxes [12] as object proposal to generate bounding boxes and as baseline. For the entropy-based proposal scoring we chose the VGG-16 [8] network, trained on the ImageNet [7] dataset as image classifier, yielding a 1000-dimensional output probability vector.

For each video we classify 25 proposals and compute the entropy score. For our method we select the best 25 tracks of each video, according to Eq. 3, and for each of them we classify, as representative, the box with the best EdgeBoxes score. We compare the entropy scores against the best 25 boxes given by EdgeBoxes for the whole video. As a lower-bound reference value we run the classifier on the dataset ground truth. This value is what can be expected to be obtained when proposing only meaningful objects.

Results for YTO are shown in detail in Table 1; it can be seen that our method yields a much lower entropy than EdgeBoxes, also it is close to the ground truth reference. The same trend can be observed on the VID dataset where we measured an average Entropy of 4.73, 3.96 and 3.71 for EdgeBoxes, Our method and the ground truth respectively.

High Precision Proposals. As a further evaluation, we treated our proposal as an object detector measuring the mean Average Precision (mAP) for the YTO dataset. This aims at measuring the precision of a proposal method. Since the

Table 1. Entropy comparison (lower is better) between the proposals provided by EdgeBoxes (EB) and our method (Ours) and Ground Truth boxes (GT).

Method	✈	🐦	⛵	🚗	🐈	🐄	🐕	🐎	🏍	🚂	Mean
EB [12]	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

Table 2. AP comparison (higher is better) for object detection between EdgeBoxes (EB) and Our method.

Method	✈	🐦	⛵	🚗	🐈	🐄	🐕	🐎	🏍	🚆	Mean
EB [12]	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

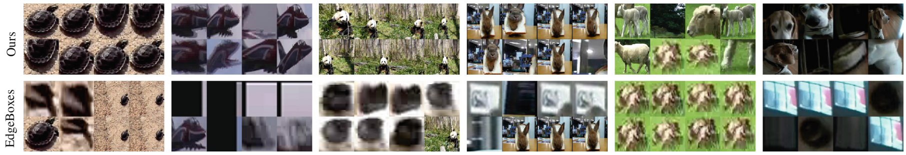

Fig. 2. Keyframes of the top 10 tracks in the VID dataset, compared with the top 10 EdgeBoxes proposals. Our method has less redundancy and frames objects more clearly.

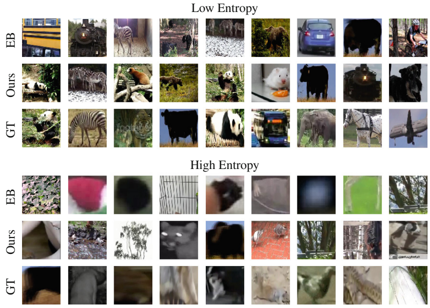

Fig. 3. Highest and lowest entropy proposals for our method (Ours), EdgeBoxes (EB) and ground truth boxes (GT).

class set of YTO is a subset of the one of Pascal VOC [1], for this evaluation we used Fast-RCNN [2], restricted to the ten common classes.

Table 2 shows a comparison between our proposed tracks and EdgeBoxes. In order to make the comparison fair, we evaluated the best 25 boxes proposed by both methods for each video, similarly to the entropy evaluation in Sect. 5. The mAP of our tracks is 8.5 times higher than EdgeBoxes, proving that our proposal is much more precise.

Qualitative Results. We report some qualitative results, showing a comparison of content extracted by our proposal with respect to EdgeBoxes. We compare the best boxes and tracks in a given video. In Fig. 2 it can be seen how our proposals are more diverse and frame an object correctly with respect to the top proposal chosen from EdgeBoxes.

In Fig. 3 we show an example of high and low entropy proposals. For our method and EdgeBoxes we report the 10 boxes with the lowest and highest entropies among the first best 25 proposals. As reference we also report high and low entropy boxes from the ground truth. It can be seen that our method is more focused on objects even in its highest entropy proposals.

6 Conclusions

We proposed a novel and unsupervised method to extract from videos, tracks containing meaningful objects. Our track proposal can build on any object bounding box proposal method. The matching process only relies on bounding box geometry and optical flow, resulting in a simple and effective method for high precision video object proposals. We also introduce a dataset independent method to evaluate the effectiveness of an object proposal, not relying on dataset annotations. The proposal has been evaluated on the YouTube Objects and ILSVRC-2015 VID datasets, showing a high precision and providing meaningful object proposals that can be used for any video analysis task without looking at the whole sequence.

References

1. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
2. Girshick, R.: Fast R-CNN. In: Proceedings of ICCV (2015)
3. Hosang, J., Benenson, R., Dollár, P., Schiele, B.: What makes for effective detection proposals? IEEE Trans. Pattern Anal. Mach. Intell. **38**(4), 814–830 (2016)
4. Kwak, S., Cho, M., Laptev, I., Ponce, J., Schmid, C.: Unsupervised object discovery and tracking in video collections. In: Proceedings of ICCV (2015)
5. Oneata, D., Revaud, J., Verbeek, J., Schmid, C.: Spatio-temporal object detection proposals. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 737–752. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_48
6. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: Proceedings of CVPR. IEEE (2012)
7. Russakovsky, O., Deng, J., Hao, S., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
8. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
9. Stretcu, O., Leordeanu, M.: Multiple frames matching for object discovery in video. In: Proceedings of BMVC (2015)
10. Xiao, F., Lee, Y.J.: Track and segment: an iterative unsupervised approach for video object proposals. In: Proceedings of CVPR (2016)
11. Yu, G., Yuan, J.: Fast action proposals for human action detection and search. In: Proceedings of CVPR (2015)
12. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_26