Video Detection by Learning with Deep Representation and Spatio-Temporal Context

Peng Han, Wenwu Yuan, Zhiwu Lu, and Ji-Rong Wen

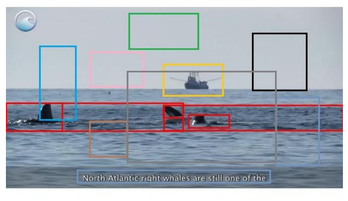
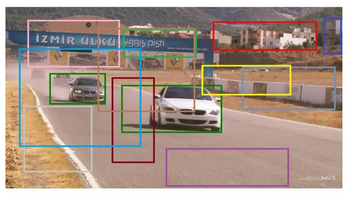
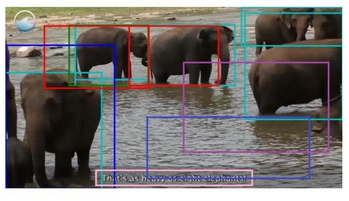
RUC_BDAI Team, Renmin University of China

Abstract - The main components of our algorithm are the R-CNN model [1] and the video segmentation algorithm [2]. Given a video clip, we use the well trained R-CNN model [1] to extract the potential bounding boxes and their object categories for each keyframe. Considering that R-CNN has ignored the temporal context across all the keyframes of the video clip, we further utilize the results (with temporal context) of the video segmentation algorithm [2] to refine the results of R-CNN. In addition, we also define several local refinement rules using the spatial and temporal context to obtain better object detection results.

1. Object Detection from Keyframes

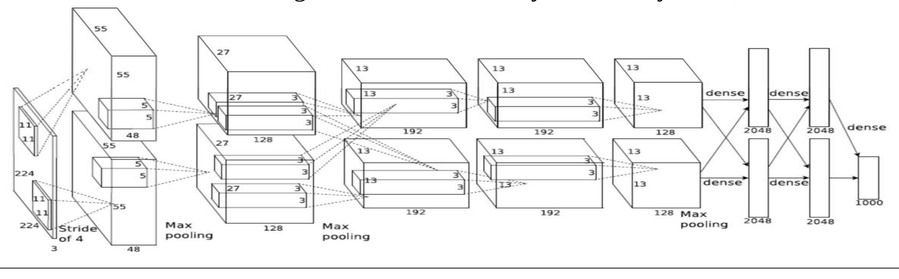
Proposal Selection

- Using selective search [3] to generate about 2000 region proposals for each keyframe of a video clip.

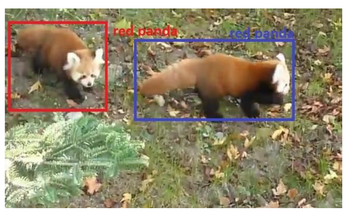
Feature Extraction

- For each region proposal, we extract a 4096-dimensional feature vector using the CNN described by Krizhevsky et al. [4].
- The architecture of this CNN is illustrated as :



Object Detection

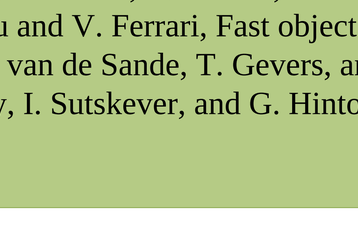
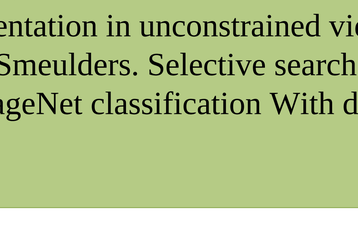
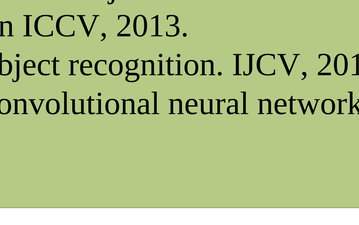
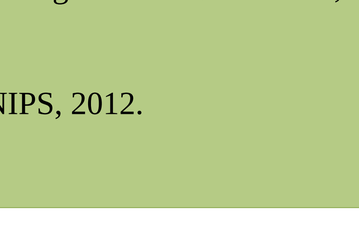
- Just as [1], we train one linear SVM per class. The positive examples are defined to be the ground-truth bounding boxes for each class, and we define the regions below an IoU overlap threshold as the negative examples.
- Then we can get the detection results from keyframes for the test dataset.



2. Object Segmentation in Video

Video Segmentation

- For each video clip, we generate the video segmentation results using the method proposed in [2].

3. Result Refinement

Result Refinement

- The results of object detection are refined using the temporal context of the results of video segmentation.
- For every k adjacent keyframes, the spatial and temporal context are used to further refine the results.

Reference:

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.
[2] A. Papazoglou and V. Ferrari, Fast object segmentation in unconstrained video. In ICCV, 2013.
[3] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. IJCV, 2013.
[4] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification With deep convolutional neural networks. In NIPS, 2012.