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


2. Object Segmentation in Video

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

3. 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: