Abstract

This thesis studies the problem of semantic segmentation in both images and videos. It targets at addressing two challenges of semantic segmentation: robustness and consistency. First, assigning pixels to the corresponding semantic categories may be unreliable especially under adverse imaging conditions. Second, existing semantic segmentation algorithms often produce inconsistent labeling at region boundaries or across neighboring frames. Visual context and consistency cues are highly useful to address the two challenges, hence the objective of this thesis is to develop techniques that exploit these cues to enhance the performance of semantic segmentation. The thesis consists of the following three technical works.

The first work is to exploit spatial layout context for semantic segmentation in images. When parsing images with regular spatial layout, the pixel location provides important prior for its semantic label and can be highly complementary to appearance cues. Therefore this thesis proposes a novel way to leverage both location and appearance information for pixel labeling. The proposed method utilizes the spatial layout by building a field of local pixel classifiers that are location-constrained, *i.e.*, trained with pixels from a local neighborhood region only. Our proposed local learning works well in challenging image parsing problems, such as pedestrian parsing, street-view scene parsing, and object segmentation, and outperforms existing methods that rely on one unified pixel classifier. To better understand the behavior of our local classifier, we perform theoretical analysis to explain why the local classifier is more discriminative and can handle misalignment.

The second work is to exploit spatio-temporal consistency for semantic segmentation in video streams. We propose an efficient online video smoothing method, called adaptive exponential smoothing (AES), to refine pixel classification maps in a video stream. We first trace each pixel in the past frames by finding an optimal spatio-temporal path; then temporal smoothing is performed over the found path with exponentially decreasing weights over time. Thanks to the pixel tracing, AES is adaptive and non-linear; thus it can better improve the "flickering" maps while avoiding over-smoothing. To enable real-time smoothing, a linear-complexity dynamic programming scheme is designed to trace all pixels simultaneously in the video stream. We apply the proposed smoothing method to improve both saliency detection maps and scene parsing maps. The comparisons with average and exponential filters, as well as more sophisticated approaches that rely on optical flow computations, validate that our AES can effectively refine the pixel classification maps in real-time.

The third work is to exploit region consistency for actor-action semantic segmentation, *i.e.*, joint labeling of actor and action categories for frames in videos. One major challenge is that different body parts provide different levels of action cues and may have inconsistent labeling when they are labeled independently. To address this issue, we utilize high-quality region masks from instance segmentation, and enforce pixels inside the region mask to take the same action label to achieve consistent labeling. Our approach uses a two-stream network which captures both appearance and motion information, followed by region-based actor-action segmentation networks which take the fused features from the two-stream network as the input. Our experiments on the A2D dataset demonstrate that both the region-based segmentation strategy and fused features from the two-stream network contribute to the performance improvements, which lead to significantly better results when compared with state-of-the-art methods.