Chapter 3
Semantic Object Segmentation

Xiaogang Wang

Abstract  Semantic object segmentation is to label each pixel in an image or a video sequence to one of the object classes with semantic meanings. It has drawn a lot of research interest because of its wide applications to image and video search, editing and compression. It is a very challenging problem because a large number of object classes need to be distinguished and there is a large visual variability within each object class. In order to successfully segment objects, local appearance of objects, local consistency between labels of neighboring pixels, and long-range contextual information in an image need to be integrated under a unified framework. Such integration can be achieved using conditional random fields. Conditional random fields are discriminative models. Although they can learn the models of object classes more accurately and efficiently, they require training examples labeled at pixel-level and the labeling cost is expensive. The models of object classes can be learned with different levels of supervision. In some applications, such as web-based image and video search, a large number of object classes need to be modeled and therefore unsupervised learning or semi-supervised learning is preferred. Therefore some generative models, such as topic models, are used in object segmentation because of their capability to learn the object classes without supervision or with weak supervision of less labeling work. We will overview different technologies used in each step of the semantic object segmentation pipeline and discuss major challenges for each step. We will focus on conditional random fields and topic models, which are two types of frameworks widely used in semantic object segmentation. In video segmentation, we summarize and compare the frameworks of Markov random fields and conditional random fields, which are the representative models of the generative and discriminative approaches respectively.
3.1 Introduction

The task of semantic object segmentation is to label each pixel in an image or a video sequence to one of the object classes with semantic meanings (see examples in Fig. 3.1). The object classes can be predefined or unsupervised learned from a collection of images or videos. It is different than unsupervised image and video segmentation, which is to group pixels into regions with homogeneous color or texture but without semantic meanings. It has important applications to image and video search, editing, and compression. For example, semantic regions with their 2D spatial arrangement sketched by users can be used as query to retrieve image. Segmented objects can be deleted from images or copied between images. Different regions of images can be enhanced in different ways based on their semantic meanings.

Semantic object segmentation is a very challenging problem, because there are a very large number of object classes to be distinguished, some object classes are visually similar, and each object class may have very large visual variability. These object classes can be structured, such as cars and airplanes, or unstructured, such as grass fields and water. Due to variations of viewpoints, poses, illuminations, and occlusions, objects of the same class have different appearance across images. In order to develop a successful semantic object segmentation algorithm, there are three important factors to be considered: local appearance, label consistency between

Fig. 3.1 Examples of images (first row) and manually segmented objects (second row) from PASCAL VOC 2009 [1] (a) and MSRC 21 [2] (b). Different colors represent object categories