Semantic Segmentation Survey

Semantic Segmentation is the task of clustering parts of images together which belong to the same object class. However, Object Detection has to distinguish different instances of the same object. Although using semantic segmentation has its own advantages, there are a couple of problems with that. First of all, neighboring pixels of the same class might belong to different object instances. Secondly, regions which are not connected may belong to the same object instances.

There are four different criteria by which Segmentation Algorithms can be classified: Allowed classes, Class affiliation of pixels, Input data, and Operation state. Semantic Segmentation is a classification algorithm, so there are most of the time fixed set of classes that we can choose our label, therefore allowed classes is one of the criteria that we can classify our algorithms with. Class affiliation of pixels means you can assign two different classes to a pixel or set of pixels for instance if there is a glass of water on a table you can assign to one of the pixels: water, glass, and table. All of them are correct classes. You can classify your algorithms by the type of input data that you are using. For instance, there are gray scale vs. colored image, excluding vs. including depth data, single image vs. stereo image vs. co-segmentation, 2D vs. 3D. Lastly, Operation state means that you are able to change in environment. As a matter of fact there two major operation state which are passive, and active. Passive operation means that the image cannot be influenced.

Unfortunately, there are no standard on quality measurements. However, there are a couple of way that we can measure the quality of our algorithms. Accuracy is one of the most common ways that let us to show the correctness of the segmentation hypothesis. One of the common ways of computing accuracy is using confusion matrix which overall means is there any misclassified segmentation. Per-pixel rate is taking pixel wise classification accuracy. However, there are two major drawbacks. First, tasks like segmenting images for cars have large regions of one class. Therefore accuracy with a little knowledge would be huge. Second, manually label image could have a coarser labeling. There are some accuracy method which do not suffer the first problems. Also there are another problem which is there could be no label assigned to a pixel.

There are some other quality measurements, such as: speed (a maximum upper bound on the execution time), Stability (robustness of algorithm over slight changes), and memory usage (peak memory matters when segmentation algorithm are used in devices with low computational power).

Luckily, the computer vision community produced a couple of different datasets which are publicly available. PASCAL VOC (a challenge, and a competition), MSRCV2 (Microsoft Research), Medical Databases.

Typically, Semantic Segmentation is done with a classifier which operates on fixed-size feature inputs and a sliding-window approach. However, in this manner we probably only get a subset of the real segmentation. Markov Random Fields and Conditional Random Fields are some of the alternatives which we can use in image segmentation and get reasonable results.

Traditional approaches in image segmentation algorithms make heavy use of domain knowledge, since they don't apply neural networks. They use a lot of old fashioned features such as: Pixel colors (RGB, which computer supports, and HSI, which is invariant to illumination), HOG (Histogram of Gradients), SIFT (Scale invariant feature transform), BOV (Bag of Visual words), Poselets (rely on

manually added extra keypoints), Textons (a minimal building block of vision, such as edge detectors, and things that CNN learns in the first filters), Dimensionality Reduction (There are a lot of pixels on an image and we can add a lot of label to each pixel, too much features!).

Unsupervised segmentation algorithms can be used in supervised segmentation as another source of information or to refine a segmentation. Since these algorithms are not semantic we call them non-semantic segmentations. Semantic Segmentation algorithms store information about the classes they were trained to segment while non-semantic segmentation algorithms try to detect consistent regions or region boundaries. One of the most common algorithm to run on the images are clustering algorithms which applies on the pixels and their features. Clustering done in two ways: It tries to cluster centers by centroids at mean coordinates or It tries to use find best match with k-means.

Another way to do non-semantic segmentation is Graph based image segmentation which means pixels are nodes and edges are a measure of dissimilarity. Edges are in either in four (four major directions) neighborhood or an eight neighborhood. Also there is Random Walk which is one of the subsets of graph based algorithms which means it uses Hog and by which it will calculate how we can approach to a seed point. It is an interactive method usually. However, it is possible to find the seed points by other algorithms. Another method to use is Active Contour Models which uses an energy function to smooth the edges and use those edges to draw a conclusion on a boundary and use that border and bound to segment the image. The last method that we speak of is Watershed Segmentation. In this method we use a grayscale image and interprets it as a height map. Low values are catchment basins and the higher values between two neighboring catchment basins is the watershed. The catchment basins contains what we are capturing. There are two flaws in watershed method which are over-segmentation due to local minima and thick watersheds due to plateaus.

Random Decision Forests such as ID3 and C4.5 apply methods so called ensemble learning to classify images. The random subspaces, bagging are two of the ways they do so. Training on random subspace of feature space, and training trees on random subset of the training set. There are two typical training modes: Central axis projection and perception training.

SVM (support vector machines) separate linearly separable data by a hyperplane. However, noises usually make this job as hard as it can. So we can define an error rate that we can tolerate. However, not every dataset is linearly separable so we can use kernel trick and map our space to another space that make the classes linearly separable. There are maybe more than one class which makes us to use one vs one strategies.

MRF (Markov Random Fields) are undirected probabilistic graphical models which are wide-spread model in computer vision. Let G =(V,E) be the associated undirected graph of an MRF and C be the set of all maximal cliques in that graph. Nodes represent random variables x, y and edges represent conditional dependencies. (Formula for MRF energy is sigma on c belongs to C of clique potentials which is weight times factorization and probability is normalized exp {-E(x, y)}).

CRF (Conditional Random Fields) are MRFs where all clique potentials are conditioned on input features. In this way since we don’t assume anything about x, it has advantage in comparison to MRFs. Also there are a lot less parameters as the distribution of x does not have to be estimated. (would be up instead of sigma would be pi).

Post Processing methods refine a found segmentation and remove obvious errors. Opening (Dilation followed by Erosion) and Closing are two operations we can use in this method.

There can be some problems in data that we collect because lens can flare and light get scattered in the picture, Corners can get very darks (Vignetting), Image can get blurry, an object can hide in the image (Camouflage of animals), Some objects are visible nevertheless you can see through them (Semi-transparent occlusion),View Points can change and it can affect our works, Partial Occlusions.