Scene Segmentation

Alex Konrad
akonrad@uci.edu

CS216 Final Project — June 13, 2020

Introduction

I reproduce a paper by Verbeek and Triggs [1] that performs scene segmentation with Conditional Random Fields. The approach to scene segmentation involves feature detection with SIFT descriptors, hue descriptors, and position descriptors. The responses from these feature descriptors are clustered separately and then assignments are chosen by nearest neighbor. I implement a Naive Bayes classifier based on the responses from these features.

1 Project Overview

Scene segmentation is the task of partitioning an image into regions corresponding to the objects in them and assigning the correct class labels to a given region. In this particular formulation of the task, the number of class labels is known in advance, and a training dataset of real-life photographs is accompanied by a ground-truth dataset of the images colored with one of k colors, each color corresponding to a class label. Hence the scene segmentation task requires that the classifier learn to color these images with the right colors.

Scene segmentation is a structured prediction problem, since the classifier needs to output multiple predictions for a single example. However, it is not necessarily desirable to predict a class label for every pixel in the example image. The ground truth datasets may be noisy or incomplete, so it may be desirable to model these on a coarser scale. There is also likely not enough information on such a small scale to make class-based inferences. So we instead model the image with 20x20 patches.

Instead of working directly with the color channels or image intensity values, the approach outlined by Verbeek and Triggs [1] projects the image patches into a lower-dimensional space. In this way, the feature detector responses on image patches function as a dimensionality reduction technique. By clustering the responses of the image patches together and then assigning them to centroids based on nearest neighbors, the predictor has a rough measure of similarity between patches of the image. The training set allows the classifier to build a visual dictionary of words, so that it scan a new image and 'look up' the visual words in a new image to make predictions.

The importance of using a Conditional Random Field to model the object class labels becomes apparent when considering how to model spatial regularities and context in the image. A CRF models the patches as random variables with edges to each of its patch-neighbors, and an edge going to the latent class label variable which is to be predicted. This conditional independence structure can help weight predictions based on the predictions of all the other neighbors, and so help keep objects connected together despite small textural differences. For example, different parts of a connected object might not have the same visual words, and so a naive classifier would classify them as different objects, but we realize that they are the same when considering how different they are from the background or another object in the image.

2 Data Sets

For this project I aimed to understand and reproduce a paper by Verbeek and Triggs from 2007 that applied Conditional Random Fields to the scene segmentation problem. Following Verbeeks and Triggs I used the Microsoft Research Cambridge (MSRC) 9-class label object recognition dataset. This dataset consists of 240 images and ground truth class labels, all images having the same aspect ratio and 320x200 resolution.

The contents of the image is mostly outdoor scenes, with many pictures of animals and vehicles. It is not entirely outdoors and there are a few pictures of humans, but for the most part the pictures are of one object and not more complicated scenes. While this dataset might seem small and relatively simple, I believe it is a good match for a Conditional Random Field, because it would be far too small for a more complex classifier like a Convolutional Neural Network. Still, a fairly accurate classifier could be learned on a small dataset like this: [1] were able to achieve almost 85% accuracy with further refinements of their approach.

I chose a 2/3 split for model validation purposes, so the training data set contains 160 training images and the test dataset 80 images. I cropped the bottom 13 pixels on the bottom of every image in order to convert the images into patches of size 20x20.

3 Algorithms

A conditional random field is a structured prediction classifier that can take context into account when making predictions by using a graphical model to add dependencies between regions. Conditional random fields are a distinct type of Markov random field because they directly model the conditional distribution, P(Y|X), rather than the joint distribution P(X,Y).

3.1 Features

To build features, I decomposed the image into 20x20 patches to build a feature vector. For each patch, I compute the 128-dimensional SIFT descriptor vector using OpenCV, a 36-dimensional color hue descriptor vector, and a position vector indicating which patch it belongs in. I encoded the position vector by overlaying a 16x10 grid on the image, so each patch would be in one of 160 tiles.

As mentioned above, the classifier does not work directly with the feature responses. For each of these features, I cluster the feature response from every patch in the entire training set together, and the actual features are indicator vectors which designate the centroid assignment for that patch. This allows us to classify a patch based on other most similar patches.

I used K-means clustering to cluster the training image patches and build a visual word dictionary. The feature detectors contain 1000 words, the hue detectors contain only 36 words, and the position vectors contain 170 possible words.

4 Results

Unfortunately my classifier is unable to make even very basic predictions. See the below figure for the predictions made by the classifier. It does not correctly segment the objects in the picture, and it does not in general predict the right class labels in the pictures either. This is the output of a Naive Bayes classifier working with the histograms of visual words created by the feature vectors.

I computed error on the test set and came up with 32.5% accuracy. This may be worse than random guessing for a k-class problem.

5 Assessment and Evaluation

I obviously made some mistake in coding my classifier which I haven't found out yet. I also was unfortunately unable to implement the Conditional Random Field aspect of the project. I was also hoping to read the tutorial paper about CRFs [2]. I picked the paper by Verbeek and Triggs because I was interested in learning more about CRFs, but the rough Naive Bayes classifier is as far as I was able to get.

I was hoping that the Naive Bayes classifier would be effective, since in the paper they were able to achieve about 67% accuracy with just the Naive Bayes approach, and then they added the CRF in to better model spatial dependencies.

I probably learned the most about feature descriptors in this project. I looked at the paper about SIFT descriptors, and while I ended up using some code to compute it I tried a few times to implement it myself. I was a little confused by the hue descriptors, since I have not seen many mentions of it in other papers or libraries. I ended up translating some code I found from MATLAB to compute the hue descriptor response. I think the visual words approach is really cool and elegant, and I like the way it

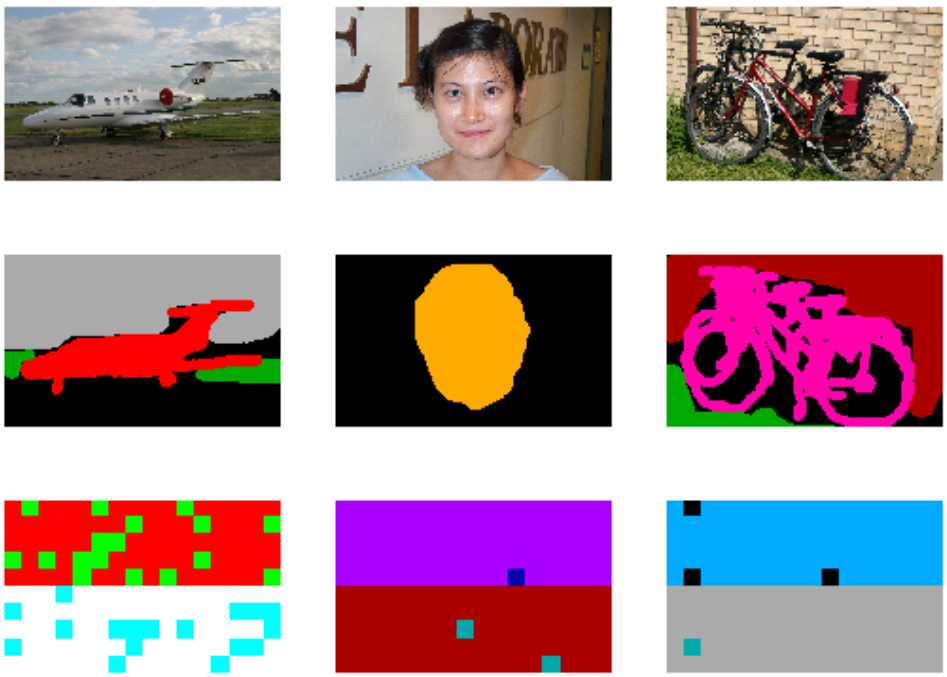

Figure 1: Three examples of the in-progress classifier shown. The image at the top is the original photo, the middle is the ground truth labeled segmentation, and at the bottom is the classifier prediction.

projects the data down to lower dimensions, encoding the relevant information for the classifier. I think that is one of the more interesting things I learned in general this quarter, that is, how images follow certain probabilistic regularities and contain information, but in computer vision we need to find ways to encode that information in a more structured way.

References

[1] Scene Segmentation with Conditional Random Fields Learned from Partially Labeled Images. Jakob Verbeek and Bill Triggs, 2007. http://ljk.imag.fr/membres/Bill.Triggs/pubs/Verbeek-nips07.pdf

[2] An Introduction to Conditional Random Fields for Relational Learning. Charles Sutton and Andrew McCallum, 2012. https://people.cs.umass.edu/ mccallum/papers/crf-tutorial.pdf

[3] Distinctive Image Features from Scale-Invariant Keypoints. David G. Lowe, 2004. https://www.cs.ubc.ca/ lowe/papers/ijcv04.pdf