Mangrove Monitoring - Semi Supervised Classification

Ashlesha Vaidya, Sidharth Suresh

Project Charter

Project Overview

Mangroves are trees and shrubs that have adapted to grow in the intertidal zone along subtropical coastlines. Mangroves protect shorelines from damaging storm and hurricane winds, waves, and floods. Mangroves also help prevent erosion by stabilizing sediments with their tangled root systems. They maintain water quality and clarity, filtering pollutants and trapping sediments originating from land. Mangroves play a major role in carbon sequestration. Therefore, they are a very important part of our ecosystem. In many areas, the prevalent mangrove species are threatened due to dumping and industrial development. Thus, the monitoring and tracking of these mangroves plays an important role in preserving them. In order to track these species we need to classify them as being mangroves or not.

Project Approach

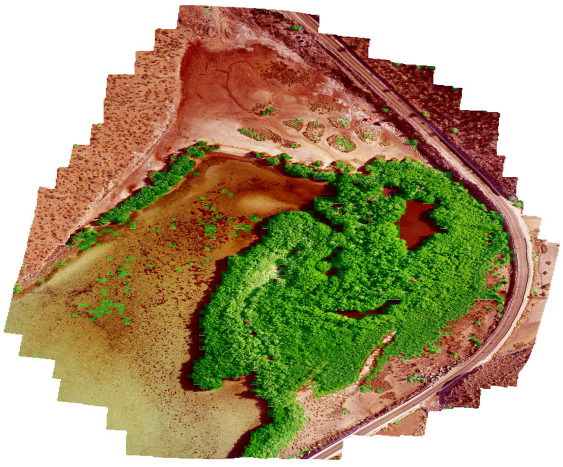

Fig. 1 : Drone captured Image

Motivation

The main idea behind the project is to identify mangroves from the drone images. Fig. 1 shows an example of one of images captured by a drone. These images are captured by high resolution capability drones. The region where the mangroves are present in these images need to be identified. Manually labelling them is time consuming due to the availability of a large amount of data.

Some models like support vector machines, convolutional neural networks have been developed in order to tackle this problem. Fig. 2 shows how the different existing models perform on classifying mangroves in a drone captured image. A Convolutional Neural Network Model is being used for classification of mangroves, it has an accuracy of about 97-99%. But Neural Networks are computationally expensive and they require a large amount of training data.

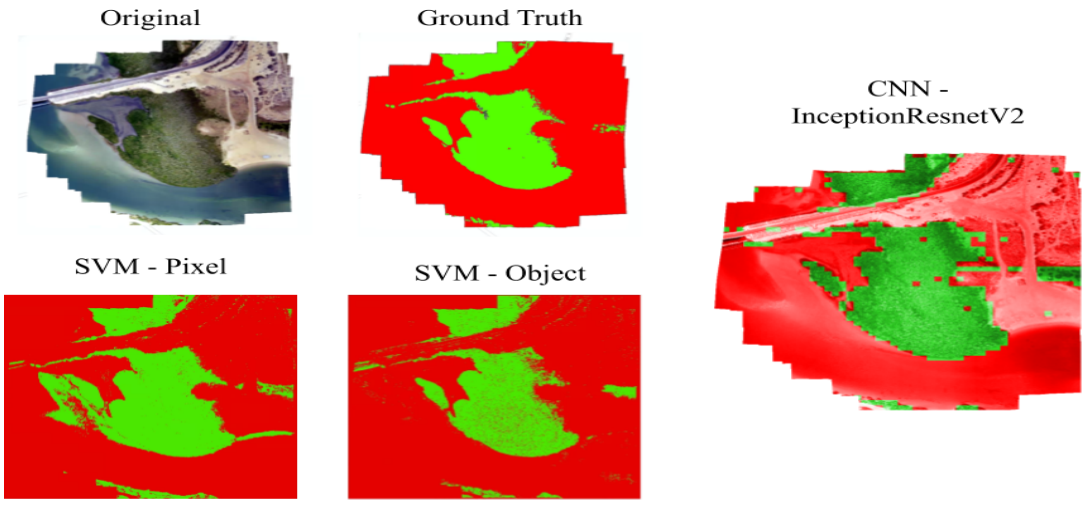

Fig. 2 Visualization of existing classification models

We plan on implementing a semi-supervised machine learning model to work with the limited labelled data and at a less computational cost compared to Neural Networks.

Our Approach

Semi-Supervised Learning is a machine learning approach that combines a small amount of labelled data and a large amount of unlabelled data for training. The mangroves are shown with a boundary in the image. As we know this process of labelling the large number of drone imagery is time consuming and hence this aids in overcoming the shortage of labelled data.

We will be using the semi-supervised clustering and classification approach to segment and classify images.We will be generating image tiles from a very large image of a mangrove site.An example of what this tile formation will look like is shown in fig. 3.

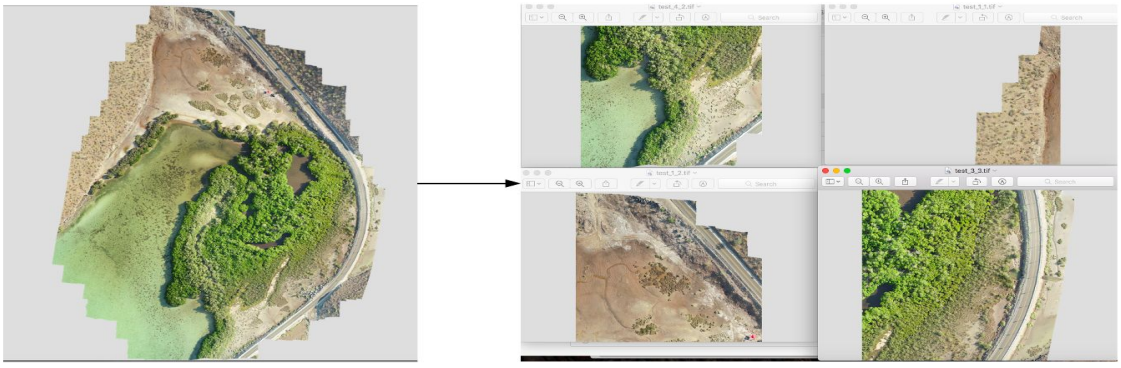

Fig. 3 Tile formation from images

A clustering algorithm is applied to these image tiles to clusters pixels that are of a mangrove together and the rest into another cluster. After clustering we assign labels to these image tiles. We intend to use the K-means clustering algorithm and Gaussian Mixture Models for clustering and then choose the one suitable for our application.

The next part consists of implementing a classification model for image segmentation. We plan to implement a model for classification using the available labelled image tiles and the ones that we obtain after clustering. Since we will be using the model to classify mangroves from image tiles where there is a clear distinction between the two labels, we will be using Support Vector Machines for classification due to its good performance on images and also because of its efficiency when there is a clear distinction between the classes/labels.

Project Objectives

The project objective is to come with an accurate, computationally less expensive Semi-Supervised Machine Learning Model to classify mangroves using the limited labelled data available and the large amount of unlabelled data available from the drone imagery.

Milestones

- Learn more about Semi-Supervised Learning and the approaches
- Divide the available site image into image tiles
- Apply clustering algorithms to cluster pixels of mangroves together and assign a label to each of the image tiles generated
- Implement a classification algorithm (SVM) using the available labelled image tiles
- Use the labelled image tiles and the ones obtained after clustering to train the model
- Visualization of the model's performance on test data, calculate the accuracy and compare the results with the already existing CNN model.
- Look for further scope of improvement to increase the accuracy of the model and test it out on a new dataset.

Major Deliverables

- Clustered Image tiles to be used for classification.
- A baseline semi supervised classification model with performance analysis.
- Optimized model to improve performance.
- Final project report to be delivered by both the team members.

Minimum Viable Product

Our Minimum Viable Product will be a machine learning model which takes in an image, divides it into tiles and then runs a semi supervised model on it to classify into mangrove and non mangrove images. The baseline model would be one which is able to classify the training images into mangrove and non mangrove. The baseline accuracy has not been decided yet and we can get an idea about this one we run a baseline model on some train data.

Potential long term goals include developing models with accuracy higher than neural networks.

Constraints, Risk, and Feasibility

We might mainly face computational constraints for this project because of the size and the type of data we are dealing with. Fast processing speeds are required in order to process such big images and run machine learning models on them.

Another major constraint we face is the dearth of labelled data. This makes it a harder problem but a feasible one as we can make use of some labelled and unlabelled data for our purposes of classification.

A potential risk involved is that the semi-supervised model might actually not be able to perform as well as the neural network model as the already achieved accuracy is around ~97-99%.

Group Management

Major Roles

Dillon Hicks - Technical Lead

Ashlesha Vaidya - Developer

Sidharth Suresh - Developer

Any decisions that need to be made will be made based on a consensus after discussing the appropriate matter and the technical lead will have a final say in case of disputes or when the developers may not be in agreement. Communication will mainly take place over slack and zoom will be used for any virtual meetings. We have planned weekly meetings over zoom with our technical lead to give weekly updates. Since we plan on having weekly meetings for the updates on the work we are doing, we will always be aware of our progress. This will help us realize if we are off schedule. Each member will be given specific tasks in the development progress but primarily we aim to work together as much as possible.

Project Development

In terms of the development roles, both the team members will be responsible for development tasks throughout the project .

Hardware/Software used

This project is mainly going to be software based. Although the dataset acquisition is done using drones, we will not be able to participate in that process. So we will be focusing on developing better machine learning models. The softwares we will make use of in our project are :

- QGis : for visualizing the images captured by the drones. This is an open source software that is used for viewing, editing and analysis of geospatial data.
- Anaconda : for running python scripts. We have set up our python environment with GDAL and other required libraries for processing the geospatial images of the mangroves.
- Microsoft Azure : we will be using virtual machines by azure in order to build our project. Since the data files are pretty big in size(each image is ~700MB),local machines may not have enough computational power to handle them.

Our team lead has already set up virtual machines for us on Azure and since we will not be working with hardware on this, we already have what we need.

In terms of testing we plan on testing our machine learning model on images captured by drones. The images captured by the drones vary a lot from site-to-site. Mangroves in one site may be very visibly clustered together while in other sites the mangroves may be spread across a wide area with very low density. We thus need to test the performance of our model in two parts-

- Testing the model on image data from the same site from where the training sample is taken. This will help us judge the performance of the model on similar data.

- Testing the model on image data from a different site than that of the training sample site. This will help us judge our model's performance on data which is different from the one on which the model is trained upon.

The testing metrics at this point include finding the accuracy of the model. Because we know the accuracy performance of existing models we can do a comparative analysis of our models with the existing ones.

We plan to document all our work on a shared google drive. This will include all our reports, codes and visualizations. Apart from this we also plan to maintain a github repository so that we can collaborate better.

Project Milestones and Schedule

P1 - High priority, P2 - Normal priority, P3- Low Priority

Timeline	Task	Priority	Handled by
Week 3, Week 4	Getting started ● Determining project approach ● Literature review - Semi Supervised Learning, GMMs ● Set up QGis for visualizing images ● Set up GDAL and Azure VMs	P1	Ashlesha, Sidharth
Week 5	Clustering of the tiles ● Dividing image into tiles ● Clustering of the tiles ● Experimenting with different clustering algorithms	P1	Ashlesha, Sidharth
Week 6	Classification ● Use labelled tiles available to implement SVM classification algorithm (Supervised Learning)	P1	Ashlesha, Sidharth
Week 7	Classification with the clusters ● Incorporate the clustered tiles into the dataset and train the classification model (Semi-Supervised Learning)	P1	Ashlesha, Sidharth
Week 8	Performance Evaluation ● Visualization of the model's performance on test data ● Evaluate the accuracy ● Compare the performance with already existing CNN model	P2	Ashlesha, Sidharth
Week 9	More testing ● Look for further scope of improvement to increase performance ● Test the model on new data (Different site)	P3	Ashlesha, Sidharth
Week 10	Making the video/final documentation	P1	Ashlesha, Sidharth

Fig. 4 Project Milestone table

Deliverables

We plan to achieve the following deliverables in the given timelines:

- Week 5 : github repo with the tiling code, tiles of image
- Week 7: clustering code with results, classification
- Week 9 : semi supervised model

Our MVP is creating a baseline classifier based on cluster then classify semi supervised approach. The performance of this model may be evaluated on the training set itself.

.