[[1]](#footnote-1)

Pixel Dawgs: Image Classifier

Manish R. Borse, Nick Rosetti, Veera Karri, Shreya Gupta  
Date: 12/9/2015

*Abstract*— The main objective behind this study is to make a tool that will classify the images based on the visual content. This study was performed using Python and IPython. Image classification will be preformed based on image segmentation in various different layers. Each layer holds a word value as its tag. Image processing tools like SLIC (simple linear iterative clustering), DBSCAN (density based spatial clustering of applications with noise), and Sobel filtering are used for segmenting the image in different layers. Layers that are in the same cluster are associated with same tag. Few thousand images are used as the training set. SVM (Support Vector Machine) is the most common machine learning algorithm used for classification and hence it is used to predict the tags for the test data set. Finally the predicted result will show five tags based on the visual content of the image. These five tags are short-listed based on the layer the covers the maximum are of the image. This tools hold a ????% accuracy.

*Index Terms* — Image Classification, Image processing, Machine Learning, Big Data, Data Science, Python, IPython, RDBMS.

T

# INTRODUCTION

HIS study is performed as a class project for the Big data and data science course. The goal of this project is use tools of big data and data science to design an image classifier. Classification of the images is done based on the visual content of the particular image. Data was collected form open source data that was provide by Yahoo – Flickr. Further the available data was process in a way that can be used. The open source data contains over one million images, but due to lack of computational power around 300,000 images were used in this study.

Previous studies suggest that the best way to perform image processing is to use SLIC for fragmenting the image in super pixels and then used DBSCAN for clustering together all the super pixels with same color. Later use Sobel-filtering for edge response and edge detection in the image.

Around two thousand images are used as training set for the machine-learning segment. The images in the training set were tagged manually since the initial tags were not associated with the segmented layers. To do this a Graphical User Interface was designed. Test data made up of 200 – 400 images is used to predict the results. Results show a decent accuracy of ????%.

# Data Acquisition and Data Engineering

Open source raw data is provided by Yahoo-Flickr. Data set includes two types of text files.

1) *Image URL (10 files, 5GB each)* – These set of files contains all the image ID’s and the associated URL to download the image, among other attributes.

2) *Auto-tags file (14GB - after extraction)* – This file contains all the image IDs and associated (auto-generated) tags, which were created by Flickr’s image processing tool.

Using the *Image URL* *file* ~300,000 images were downloaded. These images were later used for image processing. Due to computational limitation, subset of these images was used for segmentation. Further the tags for these images need to be associated with the image ID. This involved searching for the image ID through the *Auto-tags file.* Since the tags file was large (14GB) for traditional processing, it was divided into two different files. Hereby each image ID was associated with several different tags. For post processing and further using data in the correct format, index column was changed from image ID to tags. Now each tag is associated with different image IDs.

# Image processing tools

Three different image-processing tools were used for this project

## Simple Linear Iterative Clustering (SLIC)

SLIC is

## Density Based Spatial Clustering of Application Noise

## Sobel-Filtering:

# Machine Learning

## Graphical User Interface

## Support Vector Management

# Results

# Conclusion and Future Work

References

1. This study is a class project in Big Data and Data Science (CSE 4990/6990) course. Date: December 9, 2015. Team 5 – Pixel Dawgs, performed this study.

   *Manish Borse* with the Department of Mechanical Engineering, Mississippi State University, Mississippi State, MS, 39762 USA. (e-mail: mrb393@msstate.edu).

   *Nick Rosetti* with the Department of Computer Science, Mississippi State University, Mississippi State, MS, 39762 USA. (e-mail: ncr55@msstate.edu).

   *Veera Karri* with the Department of Computer Science, Mississippi State University, Mississippi State, MS, 39762 USA. (e-mail: svk38@msstate.edu).

   *Shreya Gupta* with the Department of Computer Science, Mississippi State University, Mississippi State, MS, 39762 USA. (e-mail: sg1312@msstate.edu). [↑](#footnote-ref-1)