ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 10, October 2018

Identification of Spam Images Using Image Processing and Machine Learning

[1] Mourya Gupta Vakacharla, [2] Bharath Kumar N.G, [3] Chilukuri Vishal Reddy
[1][2][3] Electronics and Communication dept, Christ Faculty of Engineering, Bengaluru, India

Abstract: - **The problem of space management of peripheral devices, especially like smart phones, is a pressing issues to everyday user who have to deal with the massive amount of media content that is received over the social media application like Whatsapp, Facebook to name a few. In such a scenario, the management of these images is very much necessary. In this paper, we are presenting an Image processing and Machine learning technique where the spam images are classified in a given set of images using Matlab software.**

Keywords: - Peripheral devices, Image processing, Machine Learning, Matlab.

I. INTRODUCTION

Billions and Billions of images are uploaded and downloaded from the Internet daily. Only Facebook and whatsapp alone handles 60 billion messages a day. Social media users grew by 121 million between Q2 2017 and Q3 2017. On an average, people have 5.54 social media accounts. With these statistics we can imagine the amount of data is being transferred. Out of all the data, Images are higher in number. Daily lots of images are being downloaded in the smart phones. Management or deletion of these images is a very tedious task. There is a requirement for classification of images. In this paper, based on user requirements a set of images is classified into spam or not.

II. TRAINING

A. Creating database

A database is needed to be created first by taking the images of users known persons like his friends, family. This database is required to classify whether the test image is a spam image based on the known faces. Larger the database, more the accuracy is acquired

B. HOG features extraction

After the database is created HOG (Histogram Oriented Gradients) features are extracted for every image in the database. A person label is given to each image based on the person in the image

C. Classification

After the HOG features are extracted for every image the features and training labels are sent to a classifier. This classifier is used while testing an image.

III. TESTING

A. loading a test set

Move all the test images into one folder. Load this test set to the Matlab using imageDatastore function. All the images in the test set are processed one by one for text and face detection.

B. Text Recognition

The image is loaded into a variable and the text in that image is recognized using Optical Character Recognition (OCR). When the image is processed through OCR it returns the text in that image and the probabilities that how much that recognized text is correct.

C. Face detection

After the text in the image is detected it passes through face detection. Face is detected using a function called CascadeObjectDetector. Cascade Object Detector detect objects using Viola-Jones algorithm . It has ability to detect many objects including face, nose, mouth, upperbody to name a few. A box is drawn around the face and if the box is empty it implies that there is no face detected in the image. The face that is detected in the box is cropped and saved in one variable. The image is resized and converted into grayscale image after which the HOG features of grayscale cropped image is extracted.

D. Prediction

After extraction of HOG features of the test face is done, These HOG features are compared to all of the HOG features extracted in the database using the classifier.

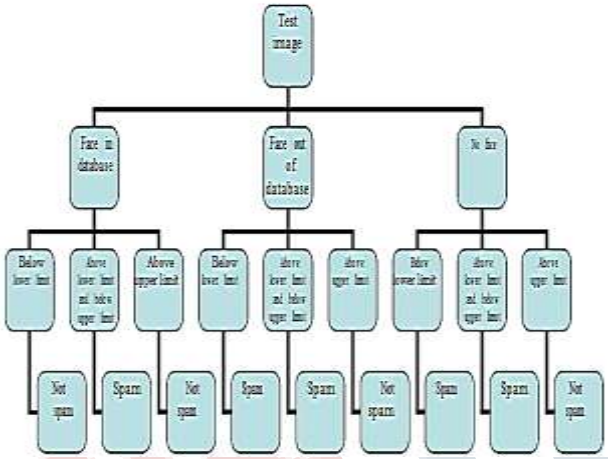
ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 10, October 2018

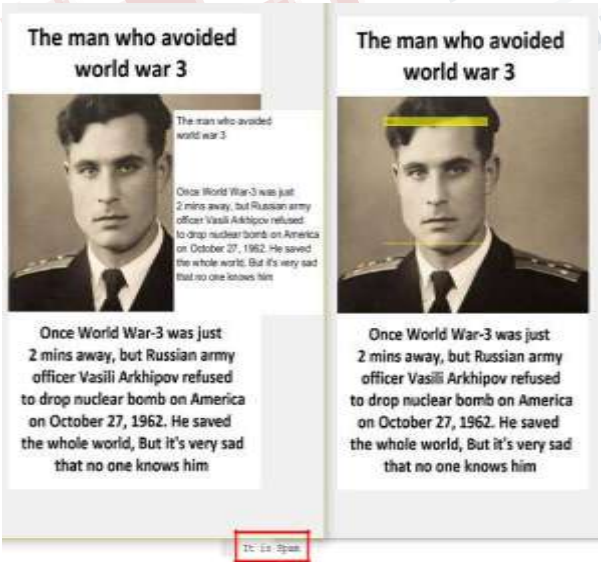
Predict function is used to compare the HOG features. A score is given as a measure of match between the test image and that of in the database. If the maximum score is above -0.4 then the test face is considered as a match in the database.

E. Classification of images

After Classification of images based on the face recognized and number of characters in the test image. There are two limits (lower and upper) are set. If number of characters are below lower limit the text is considered as less text, if they are between lower and upper limit the text is considered as medium text and if they cross upper limit the text is considered as high text.



IV. EXPERIMENTAL RESULTS



REFERENCES

[1] Arjit Sachdeva, Rishab kapoor, Amit sharma, and Akshit mishra "Categorical Classification and Deletion of Spam images on Smartphones using Image Processing and Machine Learning," 2017 International Conference on Machine learning and Data Scienc