**MAJOR PROJECT**

**SYNOPSIS**

**ON**

**Attribute Based Classification of Images for Identification of Spam**

**Submitted By**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Aayush Grover | Aayush Verma | | Komal Makkar | |  | |
| 500045032 | 500046853 | 500046239 | |  | |
|  |  |  | |  | |

***Under the guidance of***

**Dr. Monit Kapoor**

Assistant Professor

School of Computer Science



**School of Computer Science**

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**Dehradun-248007**

**2018**

****

**School of Computer Science**

**Universityof Petroleum & Energy Studies, Dehradun**

**Project Proposal Approval Form (2018-19)**

**** I

**Major**

**Project Title: ATTRIBUTE BASED CLASSIFICATION OF IMAGES**

**FOR IDENTIFICATION OF SPAM**

**ABSTRACT:**

The sharing and exchange of media, particularly images have increased exponentially with the increase in use of communication apps and emails. Spam is unwanted bulk content that is circulated on the internet nowadays. Spam images are such images which occupy a lot of space in devices but carry no specific use. This leads to problems like low storage, disorganization, manual deletion of images and so on. So, we present a solution to tackle these issues by automatically classifying every image in a folder into a set of categories, allowing user to delete spam images seamlessly making the management of images easy and less time consuming.

**Keywords:** image processing, classification, machine learning, face detection

**INTRODUCTION:**

Over the last decade, email and internet is flooded with spam content. A spam can be defined as unwanted content, distributed mostly via emails. Due to the effluence of spam emails over the Internet, a lot of techniques have surfaced which classify the spam from the valid content. Also, messaging apps have grown tremendously in recent years as the cost of mobile data has decreased, making it a much more cost-effective mode of communication. With sharing of important content, it has also promoted the exchange of rather less useful content. This has led to a typical phone gallery to be cluttered with various images that a user receives across messaging apps, which further leads to issues like high storage usage, filtering out relevant images, deletion of spam images, and classification/arrangement of images into well-defined categories.

In this project, we present an approach through which we could automatically scan and analyze every image on a device and classify it based on its type and relevancy to the user. Some categories would be Documents, Quotes, Facial images, and Spam. This makes segregation of useful images from spam images easy allowing user to delete spam images seamlessly. This project will be using Optical Character Reader for identifying texts in textual images to segregate text based spam images and Face Detection for identifying faces in non-textual images for proper classification.

**PROBLEM STATEMENT:**

While sharing media content on the internet is easiest now, it has also increased the sharing and exchange of spam images. This project aims at segregating such spam images from the useful images for better memory management and organization.

**OBJECTIVE:**

Perform classification of images into a set of predefined categories to identify spam by detection of image attributes using Image Processing and Machine Learning.

**LITERATURE REVIEW:**

Keeping a track of memory usage is important on any system. Organizing images into categories, on the basis of their attributes, gives user an insight on how much memory is being used on any type of image. This aspect of image classification can have various applications.

Billions of images are shared on various social media platforms every day. This results into collection of images that can be broadly specified as useful images and spam images. These spam images are stored with the useful images in the system which the user to delete them, one by one. With their certain attributes, they can be segregated as spam making it easier for the user to delete all similar type of images all at once. [1]

Various document images are shared over emails which are stored on the system with other images. Scanned images of documents can easily be found with the help of this classification. All document images will be collected at one place making it easier for the user to find the required document. [2]

**METHODOLOGY:**

To arrive at the classification of an image as spam(unimportant) or important, the system should be able to understand the contents of every image. After the analysis of various images collected of different categories, it is found that a majority of images in smartphones could be categorized in five sets, as follows:

1. Camera Images with faces

2. Screenshots or scanned Documents.

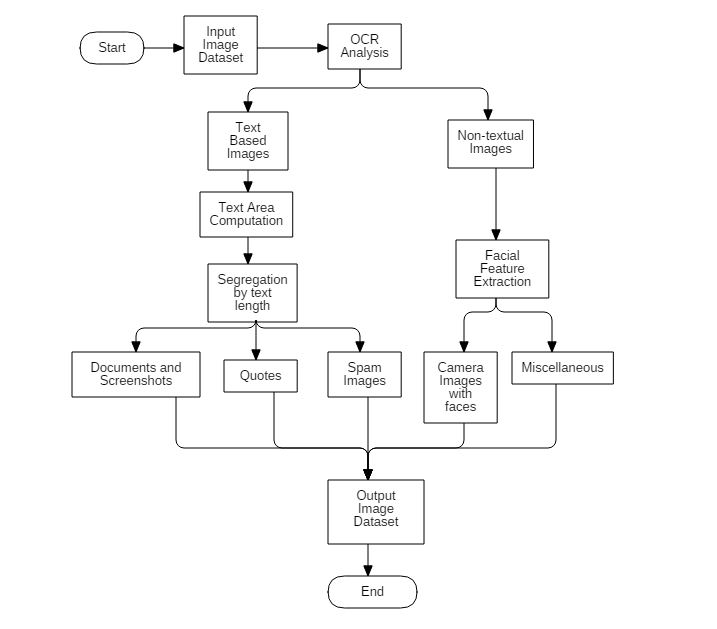
3. Quotes

4. Spam images/Wishes

5. Miscellaneous

Basically, our system is based on two modules. We first need to know what kind of images will be important for a user and what kind of images they wish to discard. The first module checks whether a photo contains text or not. Images that contain text can typically be a quote, scanned document or even a screenshot. We can detect this by using an OCR (Optical character Reader). The basic steps to classify images based on the OCR analysis are: Text Area Computation & Cropping Desired Area. The contoured area will be then passed to the OCR in order to detect the presence of text.The other images (which don’t contain text) will be passed for the detection of facial features. We segregate those images which contain pictures that have faces that could be of the user’s friends or family as these images would be relevant for him/her. This will differentiate the camera images with and without people. To find faces in images we use OpenCV face classifier (Haarcascade Frontal Face Classifier). In this way, the relevant images (camera pics, documents) are segregated from the spam images (wishes, quotes).

**FLOW CHART:**

****

**SYSTEM REQUIREMENTS:**

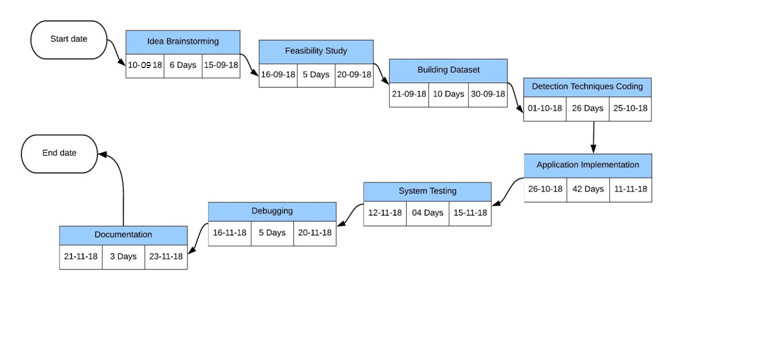
Hardware Requirements:

* 512MB of RAM or higher
* 32-bit Operating System or higher
* 50 GB of Hard Disk or Higher

Software Requirements for Window Based System:

* Python Compiler

**SCHEDULE: (PERT Chart)**

****

**REFERENCES:**

[1]G. Fumera, I. Pillai, F. Roli and B. Biggio, “Image spam filtering using textual and visual information,” in *MIT Spam Conference 2007*,Cambridge, 2007.

[2]R.Smith, “tesseract-ocr documentation,” [Online].

Available:https://github.com/tesseract-ocr/docs/blob/master/tesseracticdar2007.pdf.

[Accessed 10 August 2017].

**PROJECT PANEL**

|  |  |  |
| --- | --- | --- |
| **Faculty Name** | **Signature** | **Approve/Comments** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Approved By**

**(Name & Sign) (Name & Sign)**

**Project Guide Department Head**