**ATTRIBUTE BASED CLASSIFICATION OF IMAGES FOR IDENTIFICATION OF SPAM**

A Project File submitted in partial fulfilment of the

Requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

With IBM specialization in BANKING, FINANCIAL SERVICS AND INSURANCE

Submitted by

AAYUSH GROVER(R133215002)

AAYUSH VERMA(R133215003)

KOMAL MAKKAR(R133215016)

Under the guidance of

**Dr. MONIT KAPOOR**

Assistant Professor, CIT

Department of Computer Science & Engineering,



**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

UNIVERSITY OF PETROLEUM & ENERGY STUDIES

Dehradun – 248007.

**2018**

**DECLARATION**

I hereby declare that this submission is my own and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other Degree or Diploma of the University or other Institute of Higher learning, except where due acknowledgement has been made in the text.

AAYUSH GROVER (Enroll No. R133215002)

AAYUSH VERMA (Enroll. No. R133215003)

KOMAL MAKKAR (Enroll. No. R133215016)

**CERTIFICATE**

This is to certify that the project titled “ATTRIBUTE BASED CLASSIFICATION OF IMAGES FOR IDENTIFICATION OF SPAM” submitted by Aayush Grover (Enroll. No. R133215002), Aayush Verma (Enroll. No. R133215003), Komal Makkar (Enroll. No. R133215016) to the University of Petroleum & Energy Studies, for the award of the degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING is a bonafide record of project work carried out by them under my supervision and guidance. The content of the project, in full or parts have not been submitted to any other Institute or University for the award of any other degree or diploma.

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. Monit Kapoor Dr. T.P Singh

Assistant Professor, CIT Dept. of Analytics | SoCSE

**ACKNOWLEDGEMENT**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them. And their cooperation gives encouragement to us to work in this field.

We are highly indebted to Dr. Monit Kapoor for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

We would like to express my gratitude towards Mr. Ravi Singhal for their kind co-operation and encouragement which helps us in completion of this project.

We would like to express my special gratitude and thanks to industry persons for giving me such attention and time.

Our thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities

**ABSTRACT**

Image classification is a fundamental problem of computer vision and pattern recognition. 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. Image spam is unwanted content embedded inside the images. We regularly use communication apps like Facebook and WhatsApp on our smartphones, and the exchange of media, particularly images, has grown at an exponential rate**.** 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.

**TABLE OF CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| DECLARATION | | | ii |
| CERTIFICATE | |  | iii |
| ACKNOWLEDGEMENTS | | | iv |
| ABSTRACT | |  | v |
| LIST OF TABLES | | | vii |
| LIST OF FIGURES | | | viii |
| 1 | INTRODUCTION | | ix |
| 2. | PROBLEM STATEMENT | |  |
| 3. | LITERATURE REVIEW | | xiv |
| 4. | OBJECTIVE | |  |
| 5. | MODULES | |  |
|  | 5.1 DETECTION THROUGH CNN MODEL | |  |
|  | 5.2 BINARIZATION OF IMAGES | |  |
|  | 5.3 CONTOUR DRAW  5.4 BLOB DETECTION  5.5 BIRD EYE VIEW | |  |
| 5. | METHODOLOGY | | xix |
| 4 | RESULTS AND DISCUSSION | | xxi |
| 5 | CONCLUSIONS | | 5 |
| REFERENCES | |  | xxii |
| Annexure A: Screen shots | | | xxiii |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| 1.2 | Tax slabs in gst | xii |
| 2.1 | IMPACT OF GST IN VARIOUS SECTOR | xv |
| 2.2 | AGV vehicles reach cabinets | xvi |
| 2.2 | A PHARMACIST REACHING MEDICINE CABINET | xvii |
| 2.2 | Marg GUI | xviii |
| 3.1 | Methodology | xix |

.

CHAPTER 1

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. 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. With the rise in sharing of images, the problem of proper storage and arrangement of this media has turned into a major problem on mobile devices. 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 seaccgregation of useful images from spam images easy allowing user to delete spam images seamlessly. This project will be using Optical Character Recognition 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.

1.1 Optical Character Recognition

Optical Character Recognition (OCR) is a widespread technology to recognise text inside images, such as scanned documents and photos. OCR technology is used to convert virtually any kind of images containing written text (typed, handwritten or printed) into machine-readable text data.

1.2 PROBLEM STATEMENT

With the increase in the usage of social networking websites and media communicating apps, people share millions of texts comprising of images, videos and so on. This has led to accumulation of content on a person’s device. Some of the content might be useful but mostly the content is useless. Filtering of such less useful content is important to assure proper usage of the device. With our project, we provide a solution by separating useful images from the less useful one and putting them in different folders to make the task easier for the user with proper arrangement of images in well-defined categories.

1.3 OBJECTIVES

The core objective of this project is to perform classification of images into a set of predefined categories to identify spam by detection of image attributes using Image Processing. For this we have various sub-objectives:

* Extracting text-based images from the pool of available images
* Categorization of extracted images containing Latin characters on the basis of length of text
* Detection of facial features in the images
* Separating text based images and facial images in different categories for identifying spam

CHAPTER 2

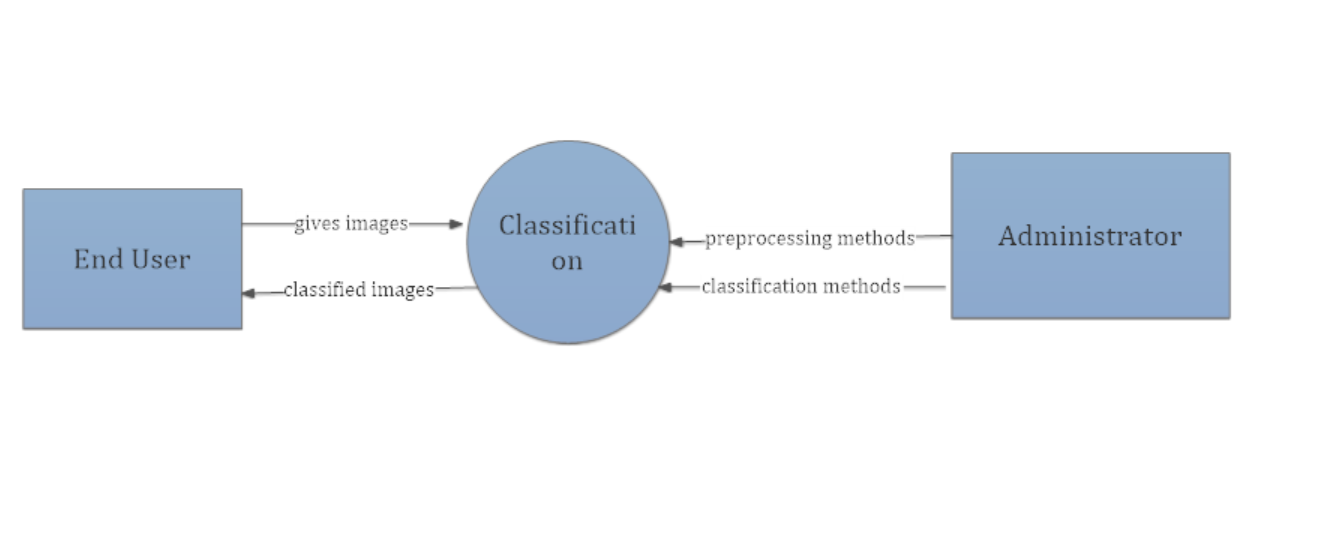
LITERATURE REVIEW

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.

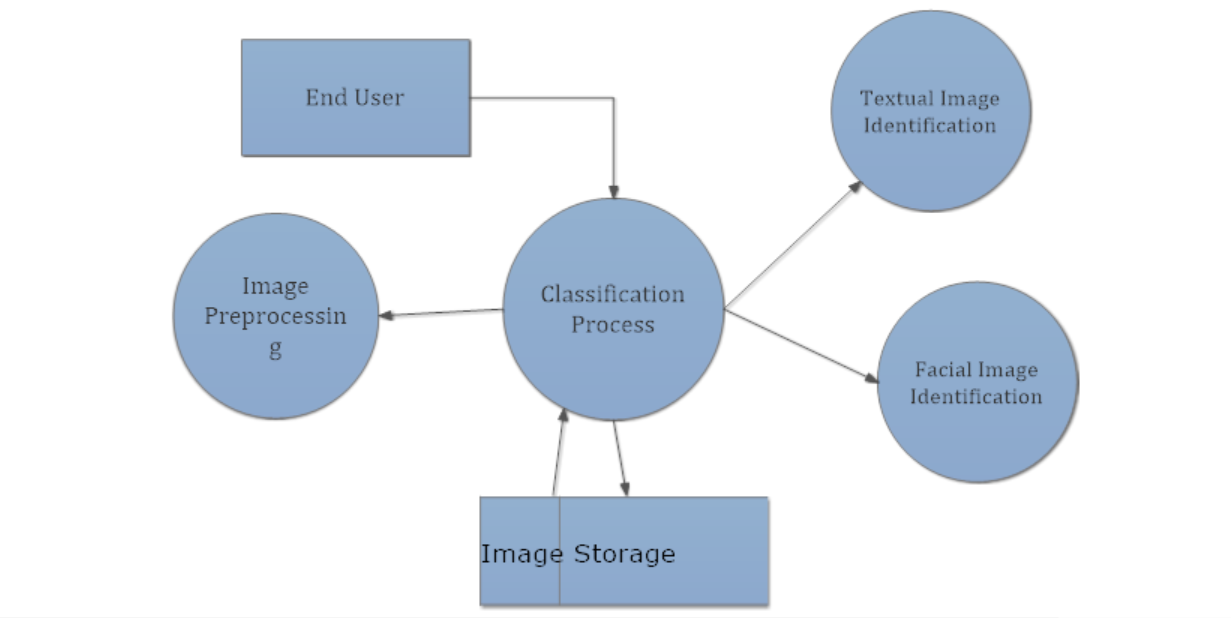
Billions of images are shared on various social media platforms every day. This results into collection of images that can be broadly specified as useful images and spam images. These spam images are stored with the useful images in the system which the user to delete them, one by one. With their certain 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]. For segregating text-based image first is necessary to identify the text in the image. The extraction of character from an image becomes easy if the boundaries of the character can be found through some algorithm. The boundaries of a character can be found in many ways depending on many factors like the edges of the character, intensity of the text present in the data [3]. Once, the text area is recognized, next step is to identify the text characters, which uses OCR. The extraction of character from an image becomes easy if the boundaries of the character can be found through some algorithm. The boundaries of a character can be found in many ways depending on many factors like the edges of the character, intensity of the text present in the data [4]. 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 a known face. After the text in the image is detected it passes through face detection [5]. Once faces are identified and text based images are identified too, they can be put in different folders and the remaining pictures can be declared as spam.

DESIGN

DFD LEVEL 0:



DFD LEVEL 1:



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.

[3] V.N.Santosh Kumar.Kuppili, Dhanaraj Cheelu, M.Rajasekhara Babu,P.Venkata Krishna, “Contour-Based Character Extraction from Text Regions of an Image”, *International Journal of Computer Science and Information Technologies*, Vol. 2 (3) , 2011

[4] Image Spam Filters Based on Optical Character Recognition (OCR) Techniques

[5] Mourya Gupta Vakacharla, Bharath Kumar N.G, Chilukuri Vishal Reddy, “Identification of Spam Images Using Image Processing and Machine Learning”, *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE),* Vol 5, Issue 10, October 2018