INFORMATION EXTRACTION FROM IMAGES USING PYTESSERACT AND NLTK

[1]Akash V Pavaskar, [2]Akshay S Accha, [3]Anoop R Desai, [4]Darshan K L
Computer Science and Engineering,
BMS College of Engineering, Bangalore, India

Abstract— Images are used in various fields such as advertisements, business purpose, and spreading awareness. Text data present in these images contain useful and helpful information like contact details, hyperlinks, QR codes. Extraction of this information involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image. However, variations of text due to differences in size, style, orientation, and alignment, as well as low image contrast and complex background make the problem of automatic text extraction extremely challenging in the computer vision research area. But the difficulty in implementation proves to be useful and fruitful. This project aims at using computer vision (Pytesseract) to extract useful information like text, contact details and hyperlinks from images. The android based app would allow user to upload a photo and enable user in storing the contact details, set a remainder, provide summary of the content of the image, opening of hyperlinks directly from the app without needing to type the URL inside the browser. Thus, making the images a more productive and making the job of the user more easy and convenient.

Index Terms— Text classification, Machine Learning, Android, Text extraction, Pytesseract, NLTK.

I. INTRODUCTION

The world is rapidly moving towards digitization. Multimedia sources like videos, images serve the majority of the content generated and spread for communication. However, it is seen that true value of images was not recognized. Images hold information in the form of text, numbers and encrypted codes. We are extracting useful content from images that are helpful to the user. We are using Pytesseract to extract text from images and then classifying it using NLTK (Natural Language Toolkit).

II. OBJECTIVES

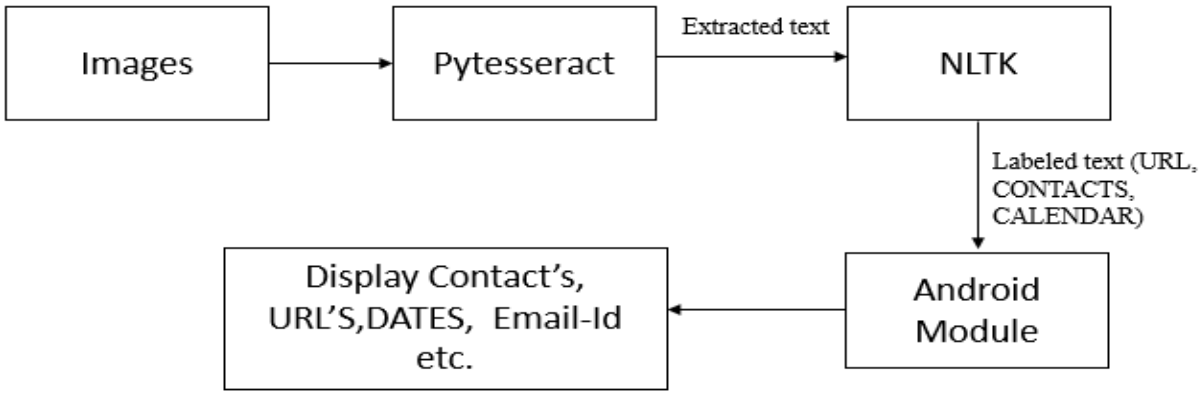
1) To extract textual data from images & automate the process of storing contact details and storing reminders.
2) Extraction of text and other form of data from images and using them for particular use. To extract URLs from the image and allow user to browse directly from the app using Android System Web View.

III. RELATED WORK

1) Google Goggles is an Image Detection System which identifies the content of an image and provides desired results to the user. It also uses Tesseract OCR to detect textual data in images and extracts the text into editable format. But, one of the limitations of Goggles is that it isn't able to classify the data present, and considers it to be in raw form.
2) Optical Character Recognition (OCR) is used in converting PDF files into editable DOC files. Another limitation is that the text present in images of PDF files aren't extracted into editable format.

IV.HIGH LEVEL DESIGN

Figure 1.1



ACKNOWLEDGMENT

We are very thankful to our guide Rajeshwari B **S (**Assistant Professor, Dept. of CSE BMSCE) for her constant guidance, support and motivation throughout the project.

CONCLUSION

The information extraction from image gives the opportunity to store certain details including contact information, URLs, Date/Day in the format user requires so as to be in sync with the fast-paced world. Thereby it integrates multiple functions under single app and reduces complexity of processing and time**.**

FUTURE WORK

1. To implement an API of our service that can be used by other apps.
2. Handling low resolution images.
3. Handling Images with handwritten text.

REFERENCES

[1] Information Extraction from Images by Muhammad Shahid Bhatti, Muhammad Usman Akram, Muhammad Ajmal, Ayesha Sadiq, Saif Ullah and Muhammad Shakil, Department of Computer Science, COMSATS Institute of Information Technology, Pakistan. World Applied Sciences Journal 29 (10): 1273-1276, 2014, ISSN 1818-4952 © IDOSI Publications, 2014, DOI: 10.5829/idosi.wasj.2014.29.10.38

[2] Text Information Extraction & Analysis from Images using Digital Image Processing Techniques by Partha Sarathi Giri, Department of E and C, M.E.M.S, Balasore, Odisha.

[3] Text Information Extraction in Images and Video: A Survey by Keechul Jung, Kwang In Kim, Anil K. Jain.

[4] Comparative Analysis of Machine learning algorithms in OCR by Vanita Jain, Arun Dubey, Amit Gupta, Sanchit Sharma.

[5] Cleaning Textual and Non-Textual Mixed Color Document Image with Uneven Shading by Xiaohua Zhang, Ning Xie, Masayuki Nakajima.

[6] Optical Character Recognition with Fast Training Neural Network by Huei-Yung Lin and Chin-Yu Hsu Department of Electrical Engineering Advanced Institute of Manufacturing with High-Tech Innovation National Chung Cheng University.

[7] Detecting Text Based Image with Optical Character Recognition for English Translation and Speech using Android by Sathiapriya Ramiah Tan Yu Liong Manoj Jayabalan School of Computing & Technology Asia Pacific University of Technology & Innovation, Technology Park Malaysia.

[8] Comparative Study of Distinctive Image Classification Techniques by Rajesh Sharma R, Beaula A, Marikkannu P, Akey Sungheetha, C.Sahana.

[9] A Survey: Text Extraction from Images and Video by Vivek Dhanapal Sapate, DKTE's Collage of Engineering, Ichalkaranji, India.

[10] TEXT EXTRACTION FROM IMAGES by Kumary R Soumya, Ancy.T.V, Athulya Chacko

[11] Text extraction in document images: highlight on using corner points by Vikas Yadav, Nicolas Ragot

[12] A Robust Algorithm for Text Extraction from Images by Najwa-Maria Chidiac, Pascal Damien, Charles Yaacoub.

[13] NLTK: The Natural Language Toolkit by Edward Loper and Steven Bird,Department of Computer and Information Science,University of Pennsylvania, Philadelphia, PA 19104-6389, USA

[14] Scikit-learn: Machine Learning in Python by Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort,Vincent Michel, Bertrand Thirion.

[15] An Efficient k-Means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D.Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE.

[16] A SURVEY ON TEXT EXTRACTION TECHNIQUES IN COMPLEX IMAGES AND VIDEOS Rosy K Philip, Gopu Darsan by Dept of computer Science Sree Buddha College of Engineering, Pattoor, Alappuzha, Kerala, (India).

[17] Natural Language Processing Based Instrument for Classification of Free Text Medical Records by Manana Khachidze, Magda Tsintsadze, and Maia Archuadze.

[18] An Efficient K-Means Clustering Algorithm by Khaled Alsabti, Sanjay Ranka, Vineet Singh.