International Research Journal of Engineering and Management Studies (IRJEMS)
Volume: 03 Issue: 04 | April -2019 ISSN: 1847-9790 || p-ISSN: 2395-0126

Extraction of PAN card data using OCR

Parmanand Ghodke, Saloni Kasture, Shrutika Shete, Saurabh Kawthekar

parmanand.ghodke15@vit.edu, saloni.kasture16@vit.edu, shrutika.shete16@vit.edu, saurabh.kawthekar15@vit.edu

---

Abstract - *Manual Data Collection from a PAN card is always a tedious job which exacts ample amount of time and effort. This paper has suggested a novel approach for developing an automatic, adaptive, fast and reliable system capable of recognizing Name of the card holder, Father's/Husband's name, Date of birth and corresponding PAN number from PAN card and storing it in the host computer in the form of excel file. In this system the input is PAN card which is being scanned through a scanner and hence captures image of the front page of each PAN card. This image is processed by using morphological operations and is then passed through Tesseract Optical Character Recognition (OCR) system which extracts characters. Accuracy of system depends on the sample space size of OCR system. In our experiment we have archived average 91 % accuracy in various light condition.*

***Key Words*: Tesseract Optical Character Recognition, OCR, PAN card, JSON, Morphing.**

1.INTRODUCTION

Optical character recognition (OCR) is a very active area of research and has become very successful in pattern recognition. OCR is mostly used in developing algorithms for reading text on the image taken by the camera, e.g. in reading registration plates, reading scanned books and documents, etc. It is based on algorithms for machine vision and artificial intelligence, i.e. neural networks, vector machines, fuzzy classifiers, etc.[1].

OCR is a widespread technology to recognize 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.

The advancements in image processing has accelerated recently due to the many emerging applications which are not only challenging, but also computationally more demanding, such evident in Optical Character Recognition (OCR), Document Classification, Computer Vision, Data Mining, Shape Recognition, and Biometric Authentication, for instance. The area of OCR is becoming an integral part of document scanners and is used in many applications such as postal processing, script recognition, banking, security (i.e. passport authentication) and language identification. The research in this area has been ongoing for over half a century and the outcomes have been astounding with successful recognition rates for printed characters exceeding 99%, with significant improvements in performance for handwritten cursive character recognition where recognition rates have exceeded the 90% mark [2].

Research has revealed that document-specific OCR systems, which are single-font OCR systems designed for a typeface, are far more accurate than omni-font systems [11][3]. For document-specific OCR systems, the training set should contain the representive character bitmaps from the page image we want to recognize (we call these character bitmaps prototypes). However, manually extracting prototypes from page images is very expensive. The lack of automatic prototype extraction methods has hindered the development of document-specific OCR systems.

Nowadays, many organizations are depending on OCR systems to eliminate the human interactions for better performance and efficiency. Optical Character Recognition also referred to as OCR is a system that provides a full alphanumeric recognition of printed or handwritten characters at electronic speed by simply scanning the document [4]. Documents are scanned using a scanner and are given to the OCR systems which recognizes the characters in the scanned documents and converts them into ASCII data.

Types of OCR:

- Optical character recognition (OCR) – targets typewritten text, one glyph or character at a time.
- Optical word recognition – targets typewritten text, one word at a time (for languages that use a space as a word divider). (Usually just called "OCR".)
- Intelligent character recognition (ICR) – also targets handwritten print script or cursive text one glyph or character at a time, usually involving machine learning.
- Intelligent word recognition (IWR) – also targets handwritten print script or cursive text, one word at a time. This is especially useful for languages where glyphs are not separated in cursive script.

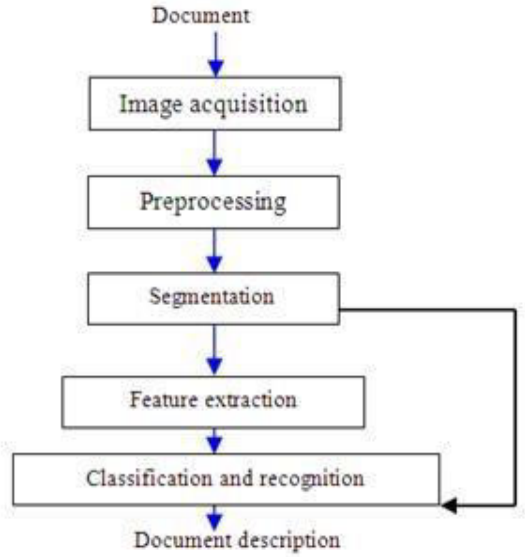

Fig. 1: Algorithm of OCR

2. RELATED WORK

In [5] presented OCR droid, a generic framework for developing OCR-based application on mobile phones. However, this discussion focused on using orientation sensor, embedded high-end camera and digital image processing technique to solve OCR issues related to camera-captured images. In [6] proposes a Text Extraction algorithm for the context of language translation of scene text images with mobile phones, which is fast and accurate at the same time. The author claims that the algorithm uses very efficient computations to calculate the Principal Color Components of a previously quantized image. The author also compares the algorithm with other algorithms using commercial OCR, achieving accuracy rates more than 12% higher, and performing two times faster, and the methodology is more robust against common degradations, such as uneven illumination, or blurring. However, no discussions have described the ability of OCR processing on multiple pages.

Alherbish et al. (1997) [7], introduces a parallel recognition system for Arabic characters. The objective of the system was to simultaneously attain high speed and full precision. The system uses distributed computing and parallel processing techniques to accomplish the goal. This multiprocessing system enhances Arabic character recognition systems of that time.

This paper proposes another area of application of OCR, namely in the extraction of text on the content of TV screens. The proposed text extraction system grabs the image representing the current TV screen content, prepares it for OCR and runs OCR to detect regions of text on the image and read the content. The proposed system is a part of the Black Box Testing (BBT) system [8][9] used for automated testing and functional verification of digital television sets. Text extraction is used to verify the functional operation of TV sets by, for example, reading the menu options presented on the screen in order to verify if the TV opened the correct menu when presented with a given set of remote-control commands.

Ohhira et al. (1995) [10], proposed a system using plural combination of Neural Networks and which could automatically recognize 6709 Chinese characters. The system consists of four parts: - rough classification part, fine classification part, recognition part, and auto judgement part. The system operates by classifying the input data by classifying by character density at the rough and fine classification parts. The multi-layered NN recognizes at the recognition part. The auto judgement part judges and output the values. The authors claim 100% recognition efficiency.

3. EXPERIMENTATION

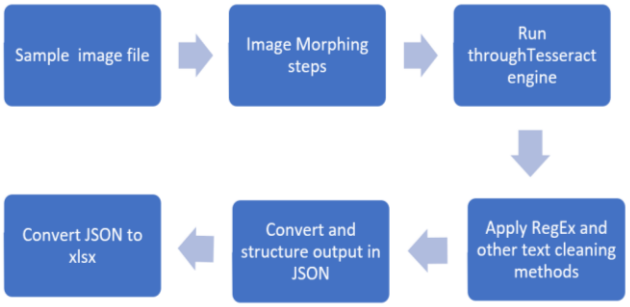

Fig. 2: Algorithm of project

The input given is sample PAN card image file. This image is scanned through scanner, scanned image is then processed through image morphing steps. Morphing is a special effect in motion pictures ad animations that changes (or morphs) one image or shape into another through a seamless transition. morphing means stretching or as part of a fantasy or surreal sequence. It is also used for the metamorphosis from one image to another. The idea is to get a sequence of intermediate images which when put together with the original images would represent the change from one image to the other. After this process image run through Tesseract engine. Tesseract is an OCR engine with support for Unicode and the ability to recognize more than 100 languages out of the box. It can be trained to recognize other languages. The OCR engine extracts the given string according to the characters specified in Allowed Characters. After this RegEx is applied. A regular expression is a special text string for describing a search pattern regular

expressions are wildcards on steroids. wildcard notations such as *.txt to find all text files in a file manager. The regex equivalent is .*.txt and some other text cleaning methods are applied on text which is extracted through OCR. Output of this step is converted and structured into a JSON file. JSON IS JavaScript Object Notation. It is a syntax for storing and exchanging data.it is text, written with JavaScript object notation. At the last JSON file is converted into CSV file. CSV is comma-separated values, it is tabular data that has been saved as plaintext data separated by commas.

4. RESULTS

	A	B	C	D
1	Name	Father Name	Date of Birth	PAN
2	PARMANAND VYANKATRAO GHODKE .	VYANKAT PANDHARINATH GHODKE	13-09-97	BVWPG3534P
3	SALONI SATISH KASTURE	SATISH BABURAO KASTURE	30-12-97	FWWPK6470F
4	SHRIYASH SATISH KASTURE	SATISH BABURAO KASTURE	24-12-99	GQMPK9401F
5				

Fig. 3: Result in form of excel sheet

Once all the processing is done on the scanned image of PAN card, running it through Tesseract OCR, denoising the image the output is obtained in from of JSON which is then converted to Excel file for the purpose of easiness in saving the data.

5. CONCLUSIONS

The Tesseract OCR has a good accuracy. From the proposed technique the details of PAN card are extracted which gives the accuracy of 91%.

REFERENCES

[1] Ivan Kastelan, Sandra Kukolj, Vukota Pekovic, Vladimir Marinkovic, Zoran Marceta, "Extraction of Text on TV Screen using Optical Character Recognition", SISY 2012 • 2012 IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics • September 20-22, 2012, Subotica, Serbia

[2] Yasser Alginahi, "Preprocessing Techniques in Character Recognition", esearchgate, publication,221909023

[3] Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar," Optical Character Recognition", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014

[4] "A Review on the Various Techniques used for Optical Character Recognition", Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi/International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012.

[5] A. Joshi, M. Zhang, R. Kadmawala, K. Dantu, S. Poduri and G. S. Sukhatme, *OCRdroid: a Framework to Digitize Text Using Mobile Phone*, University of Southern California, Los Angeles, CA 90089, USA.

[6] Canedo-Rodríguez, Adrián and Kim, Jung Hyoun, etc, Efficient Text Extraction Algorithm Using Color Clustering for Language Translation in Mobile Phone, *Journal of Signal and Information Processing, vol 3, pp. 228-237,* 2012.

[7] Alherbish J., Ammar R A, Abdalla M., "*Arabic character recognition in a multi-processing environment*", Proceedings of the second IEEE Symposium on Computers and Communications (1997), 286-291.

[8] D. Marijan, V. Zlokolica, N. Teslic, V. Pekovic, T. Tekcan: "Automatic functional TV set failure detection system", *IEEE Transactions on Consumer Electronics*, 2010, Vol. 56, pp. 125-133

[9] I. Kastelan, M. Katona, D. Marijan, J. Zloh: "Automated optical inspection system for digital TV sets ", *EURASIP Journal on Advances in Signal Processing*, 2011:140

[10] Ohhira T., Pecharanin N., Taguchi A., Iijima, N.; Akima, Y.; Sone, M., "Chinese *character recognition by the auto recognition system",* IEEE International Conference on Neural Networks, Volume: 5, (1995) 2222-2225.

[11] G. Nagy and G.L. Shelton, [a]Self-Corrective Character Recognition System, IEEE Trans. Information Theory, vol. 12, no. 2, pp. 215-222, Apr. 1966