An efficient extraction of information from Indian Government issued documents Aadhar and Pan Card.

Rachna Tewani[1], Arun Kumar Dubey[2], Achin Jain[3], Eshika Agarwal[4], Disha Mittal[5]

[1]Data Scientist, Great Learning
[2,3,4] Bharati Vidyapeeth's College of Engineering, INDIA
Emails: rachnatewani09@gmail.com; arudubey@gmail.com; achin.mails@gmail.com; eshika2812@gmail.com; dishamittal.it2@bvp.edu.in
* Correspondence: achin.mails@gmail.com

Abstract

In today's world, everything is getting digitized, and widespread use of data scanning tools and photography. When we have a lot of image data, it becomes important to accumulate data in a form that is useful for the company/organization. Doing it manually is a tedious task and takes an ample amount of time. Hence to simplify the job, we have developed a FLASK API that takes an image folder as an object and returns an excel sheet of relevant data from the image data. We have used optical character recognition and software like pytesseract to extract data from images. Further in the process, we have used natural language processing, and finally, we have found relevant data using the globe and regex module. This model is helpful in data collection from Registration certificates which helps us store data like chassis number, owner name, car number, etc., easily and can be applied to Aadhaar cards and pan cards.

Keywords: Optical character recognition; Aadhar; Pan Card; NLP

1. Introduction

In today's world, processing images such as invoices and handwritten bills has become an essential process in every sector, especially with extensive data scanning tools and photography. In a fast-moving world, it cannot be expected to spend much time typing the data into a particular form, leading to a waste of time. Although extracting text from images with 100% accuracy is quite a demanding task.[13] As deep learning expands and OCR technologies have progressed, semi or fully-automated solutions relating to document information extraction are seeing wider adoption. Modern OCR software is quick and precise and can manage common document processing constraints such as poorly or imperfectly formatted scans, handwritten documents, low-quality images/scans, and blemishes that would have traditionally required extended manual interventions.[14] Now organizations prefer automating document processing methods to become paperless and grasp cloud-based digital solutions. We have made use of optical character recognition and software like pytesseract for the extraction of data from images for the process. Further in the process, we have used natural language processing, and finally, we have found relevant data using the globe and regex module.

NLP(Natural Language Processing) is used to clean the data, remove all the irrelevant data from the textual content, and keep useful information.[3] One of the major steps involved in natural language processing is to remove the

Doi: https://doi.org/10.54216/FPA.040201
Received: April 12, 2021 Accepted: August 01, 2021

noise from the data to make it easier for the machine to detect patterns or, in our case, textual characters. The noise present in the data is in special characters such as hashtags, punctuations, and numbers. All of these are not important for the data and, therefore, should be excluded. Therefore we process the data to remove these elements. Similarly, there are stop words present in the text which introduce unnecessary noise and, therefore, should be removed. Tokenization and lemmatization are also done on the text to further clean the text to reduce their root words.[4] Part of speech tagging and chunking is also done for cleaning the data and keeping only the relevant information in the text.[5] All this is done using the natural language toolkit python library. We have divided the whole paper into six sections: introduction, Related work, Dataset, Opencv for OCR, Proposed methodology, Results and Comparison with existing methodology, and conclusion.

2. Related Work

The work of detecting and recognizing texts was previously done by recognizing the data from bills that are either handwritten or printed and updating them to the database automatically to reduce manual labor. Several deep learning methods are used for such processes as EAST algorithms or SVM.[10] using rnn for recognizing the text from images.[11] There have also been works where Hand character recognition is done using a histogram of the oriented gradient for the Recognition of text and then using a support vector machine. Pratik Madhukar Manwatkar and Dr. Kavita R. Singh [15] have reviewed different methods to extract characters from images in their paper. The basic architecture of the process for text recognition from images is reported in their work. They have also mentioned the order of image processing methods to extract textual content from the scanned image. D.Y. Turdakov [16] is based on a text extraction pipeline used to extract textual content from different quality images acquired from the internet. Their work largely pays attention to dividing the input images into various classes, and then preprocessing is done depending on the classes. This is further followed by text recognition using the OCR engine.

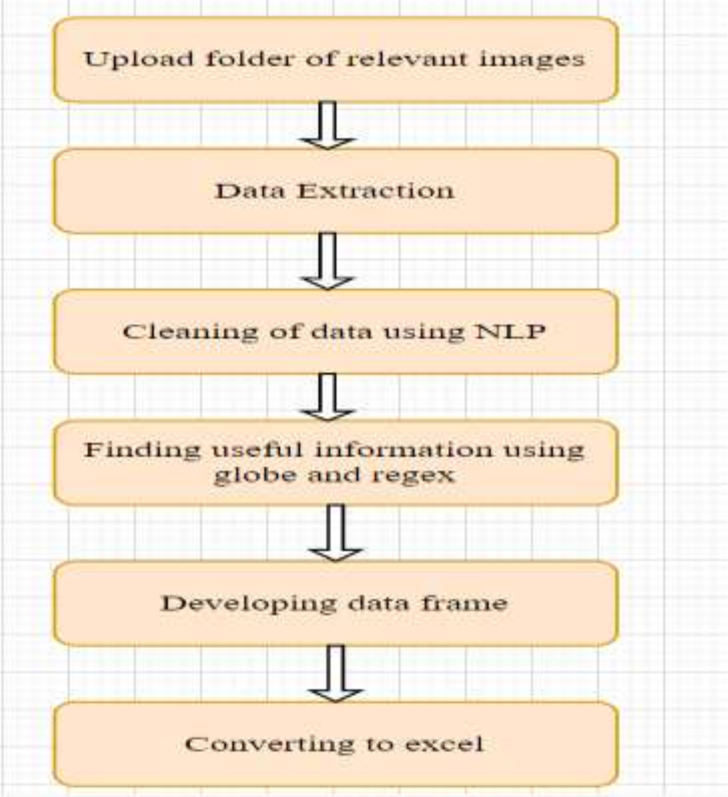

Upload folder of relevant images
Data Extraction
Cleaning of data using NLP
Finding useful information using globe and regex
Developing data frame
Converting to excel

Figure 1: Flowchart for our model

Doi: https://doi.org/10.54216/FPA.040201
Received: April 12, 2021 Accepted: August 01, 2021

3. Dataset

Various registration and identification certificates/documents like registration certificates, aadhaar cards, or pan cards comprise key information on the person such as name, gender, date of birth, chassis number, car number, etc. [12] Aadhaar is a 12-digit unique identity number that can be procured by one's own accord by residents or passport holders of India, based on their demographic and biometric data. A pan number consists of 10-digit numbers in alphabets and numbers administered by the income tax department to all the taxpayers and unique to each individual. We have used 46 images for our project and then made them run on pytesseract for text recognition, which is further filtered by using globe and regex modules. The collected data is filtered and relevant and is further stored in the form of an excel sheet.

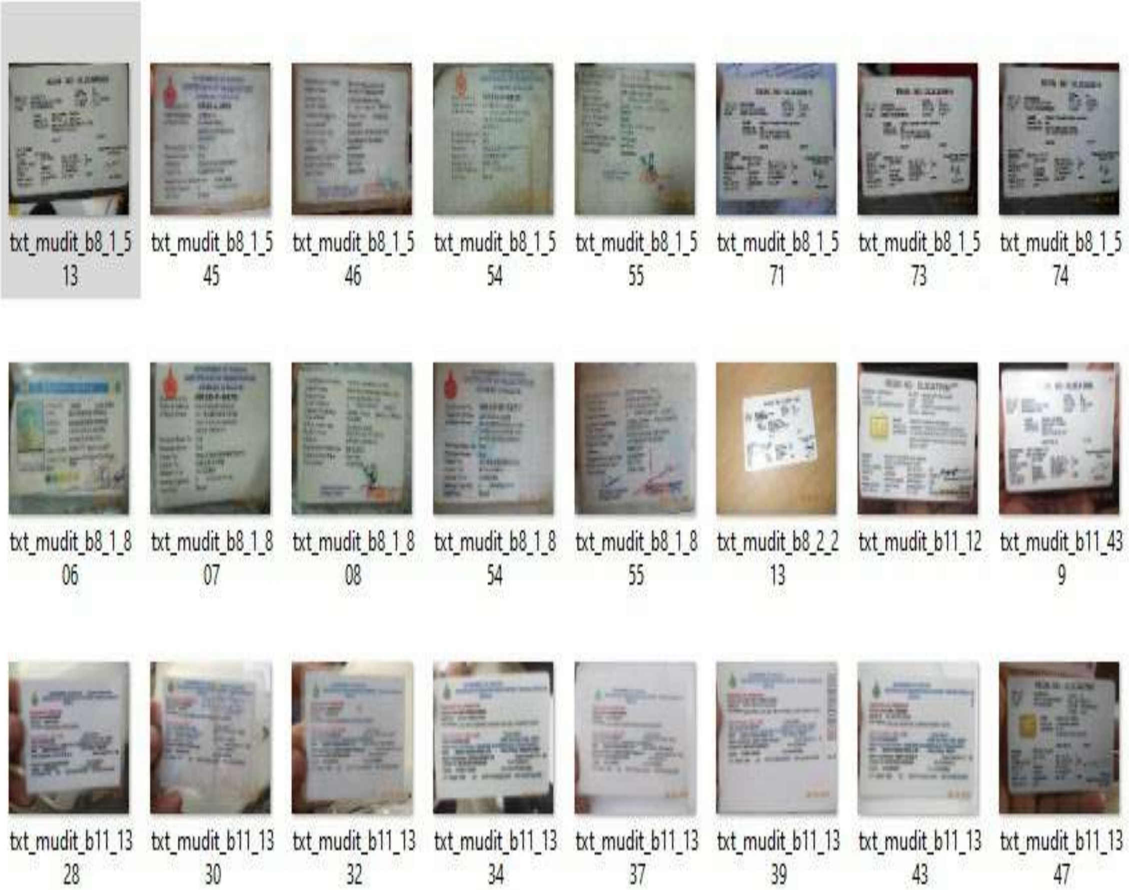


Figure 2: Sample of Aadhar and Pan Card dataset

4. OpenCV for OCR

OpenCV is a computer vision and deep learning software library. It has an ample number of optimized algorithms which provide us with important functions like tracking moving objects, recognizing faces, identifying objects, scanning real images, and processing and analyzing them.[77] Here we have used it for processing, to detect text from images.[8] Pytesseract requires a clean image to detect text; that is why OpenCV has been used.

OPTICAL CHARACTER RECOGNITION is a process of detection and Recognition of the text from the images. It is scanning, analyzing, and detecting the textual content within the images, determining them, extracting the text from the images, and finally translating the images to electronic or encoded text. [6]The accuracy of the OCR is

Doi: https://doi.org/10.54216/FPA.040201
Received: April 12, 2021 Accepted: August 01, 2021

mainly based on text processing and segmentation algorithms. Occasionally it is strenuous to recover textual data from the image due to various reasons such as differences in size, style, orientation, the compounded background of the images, etc. With the world being more digitized by the day, OCR has been immensely increased in various organizations to cut down traditional workloads. This makes it efficient to extract and store information like chassis number, owner name, car number, etc., easily and can be applied to aadhar cards and pan cards. We first processed the data of images using OpenCV libraries,[9] then we removed the unwanted data from the images by segmentation, tokenization, and lemmatization of the data. It all helps in recognizing text from images and extracting useful information.

5. Proposed Methodology

Python - Tesseract is an optical character recognition tool for python, which is used to scan the image and then read the text inserted in images. It works in a particular manner. The first step is Adaptive Thresholding [1], which is a method of converting the image into binary images—following that, the next step is connected component analysis[2], which is a method for extracting character outlines. These outlines then are converted into organized text lines, which are further analyzed for some fixed text size[2]. Text is divided into words using definite spaces and fuzzy spaces. Then the process of recognition starts, which includes recognizing each word from the text. Each word exceeded adequately is passed to an adaptive classifier as training data. The whole processing model of Tesseract follows a conventional step-by-step process, as shown in Figure 3:

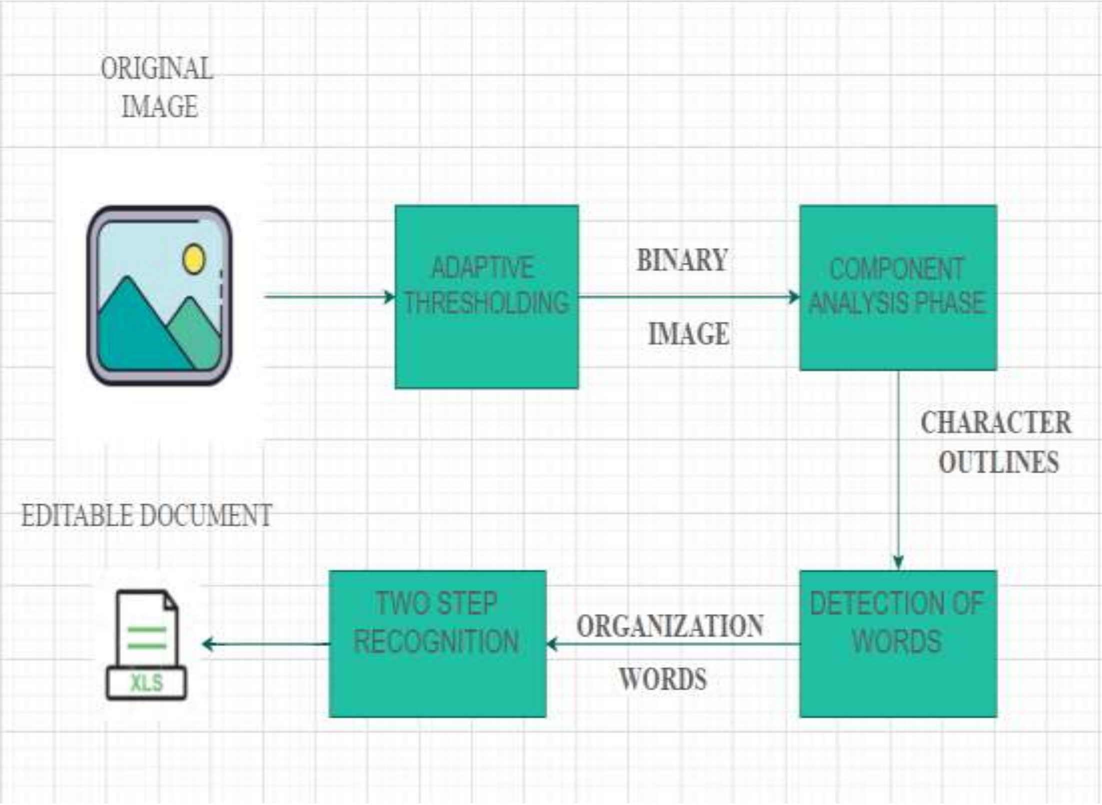


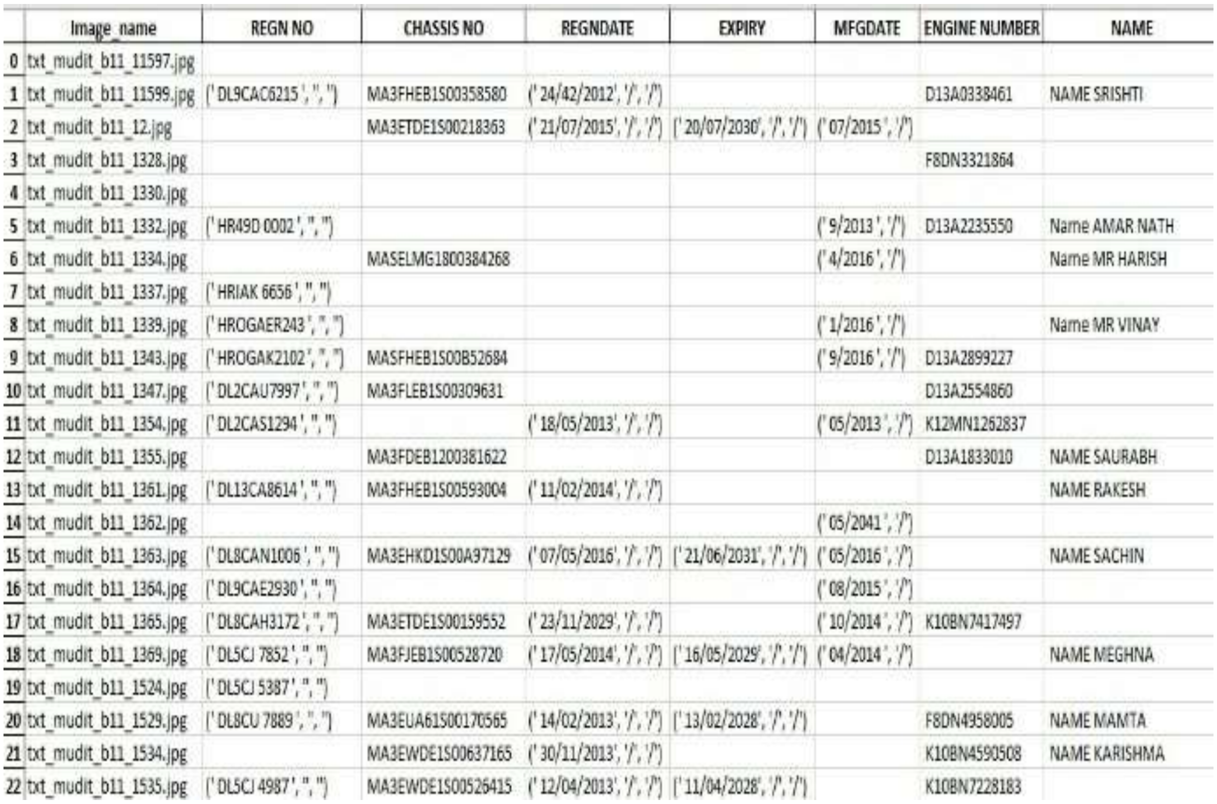
Figure 3: Document extraction model for Aadhar and Pan Card

6. Results and Comparison with existing methodology

Optical character recognition and software like pytesseract are used for the detection and extraction of data from images. To extract the printed text and the relevant data from the images and convert them into an excel sheet. The process starts by giving a folder of images as an object to FLASK API and using OCR for text recognition, and using globe and regex models to filter out the relevant data. Our project used 46 images as an input consisting of

Doi: https://doi.org/10.54216/FPA.040201
Received: April 12, 2021 Accepted: August 01, 2021

various aadhaar cards and pan cards. The end output is the combined data in the form of an excel sheet consisting of relevant information from those cards.



	Image_name	REGN NO	CHASSIS NO	REGNDATE	EXPIRY	MFGDATE	ENGINE NUMBER	NAME
0	txt_mudit_b11_11597.jpg							
1	txt_mudit_b11_11599.jpg	('DL9CAC6215 ', '', '')	MA3FHEB1S00358580	('24/42/2012', '/', '/')			D13A0338461	NAME SRISHTI
2	txt_mudit_b11_12.jpg		MA3ETDE1S00218363	('21/07/2015', '/', '/')	('20/07/2030', '/', '/')	('07/2015 ', '/')		
3	txt_mudit_b11_1328.jpg						F8DN3321864	
4	txt_mudit_b11_1330.jpg							
5	txt_mudit_b11_1332.jpg	('HR49D 0002 ', '', '')				('9/2013 ', '/')	D13A2235550	Name AMAR NATH
6	txt_mudit_b11_1334.jpg		MASELMG1800384268			('4/2016 ', '/')		Name MR HARISH
7	txt_mudit_b11_1337.jpg	('HRIAK 6656 ', '', '')						
8	txt_mudit_b11_1339.jpg	('HROGAER243 ', '', '')				('1/2016 ', '/')		Name MR VINAY
9	txt_mudit_b11_1343.jpg	('HROGAK2102 ', '', '')	MASFHEB1S00B52684			('9/2016 ', '/')	D13A2899227	
10	txt_mudit_b11_1347.jpg	('DL2CAU7997 ', '', '')	MA3FLEB1S00309631				D13A2554860	
11	txt_mudit_b11_1354.jpg	('DL2CAS1294 ', '', '')		('18/05/2013', '/', '/')		('05/2013 ', '/')	K12MN1262837	
12	txt_mudit_b11_1355.jpg		MA3FDEB1200381622				D13A1833010	NAME SAURABH
13	txt_mudit_b11_1361.jpg	('DL13CA8614 ', '', '')	MA3FHEB1S00593004	('11/02/2014', '/', '/')				NAME RAKESH
14	txt_mudit_b11_1362.jpg					('05/2041 ', '/')		
15	txt_mudit_b11_1363.jpg	('DL8CAN1006 ', '', '')	MA3EHKD1S00A97129	('07/05/2016', '/', '/')	('21/06/2031', '/', '/')	('05/2016 ', '/')		NAME SACHIN
16	txt_mudit_b11_1364.jpg	('DL9CAE2930 ', '', '')				('08/2015 ', '/')		
17	txt_mudit_b11_1365.jpg	('DL8CAH3172 ', '', '')	MA3ETDE1S00159552	('23/11/2029', '/', '/')		('10/2014 ', '/')	K10BN7417497	
18	txt_mudit_b11_1369.jpg	('DL5CJ 7852 ', '', '')	MA3FJEB1S00528720	('17/05/2014', '/', '/')	('16/05/2029', '/', '/')	('04/2014 ', '/')		NAME MEGHNA
19	txt_mudit_b11_1524.jpg	('DL5CJ 5387 ', '', '')						
20	txt_mudit_b11_1529.jpg	('DL8CU 7889 ', '', '')	MA3EUA61S00170565	('14/02/2013', '/', '/')	('13/02/2028', '/', '/')		F8DN4958005	NAME MAMTA
21	txt_mudit_b11_1534.jpg		MA3EWDE1S00637165	('30/11/2013', '/', '/')			K10BN4590508	NAME KARISHMA
22	txt_mudit_b11_1535.jpg	('DL5CJ 4987 ', '', '')	MA3EWDE1S00526415	('12/04/2013', '/', '/')	('11/04/2028', '/', '/')		K10BN7228183	

Figure 4: Extracted information from Aadhar and Pan Card

Table 1: Proposed model comparison with the existing model

Model	Accuracy (%)	Graphical component(%)	Text Component(%)
Xiaojing Liu et al. [17]	89	78	90
Seong Ah Chin et al. [18]	92	87	96
Proposed Model	95	91	95

In Table 1, we have compared our model with the latest existing model. It has been seen that our model is performing better than another model in graphical, but text component accuracy is slightly low in comparison to Seong Ah Chin et al. [18].

7. Conclusion

In conclusion, we have developed a flask API used to extract text from formatted cards using OpenCV through OCR. We have implemented this project using pytesseract. We have presented text recognition and extraction of relevant data from cards using various software. This project aims to inspect the method of classifying relevant text from the data into an excel format. This information is often extracted manually in multiple organizations; thus, in our project, automation is done, which could help reduce manual labor and thus fasten the process. This model's

Doi: https://doi.org/10.54216/FPA.040201
Received: April 12, 2021 Accepted: August 01, 2021

accuracy is better than the existing model by Xiaojing Liu et al. [17] and Seong Ah Chin et al. [18]. Overall accuracy is 95%.

References

[1] Shafait, F., Keysers, D., & Breuel, T. M. (2008, January). Efficient implementation of local adaptive thresholding techniques using integral images. In *Document recognition and retrieval XV* (Vol. 6815, p. 681510). International Society for Optics and Photonics.

[2] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and Recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE.

[3] Wen, Y., Lu, Y., Yan, J., Zhou, Z., von Deneen, K. M., & Shi, P. (2011). An algorithm for license plate recognition applied to intelligent transportation system. *IEEE Transactions on intelligent transportation systems*, *12*(3), 830-845..

[4] Fan, X., & Fan, G. (2008). Graphical models for joint segmentation and Recognition of license plate characters. *IEEE Signal Processing Letters*, *16*(1), 10-13..

[5] Wu, H., & Li, B. (2011, July). License plate recognition system. In *2011 International Conference on Multimedia Technology* (pp. 5425-5427). IEEE.

[6] Pan, Y. F., Hou, X., & Liu, C. L. (2008, September). A robust system to detect and localize texts in natural scene images. In *2008 The Eighth IAPR International Workshop on Document Analysis Systems* (pp. 35-42). IEEE..

[7] Liang, J., DeMenthon, D., & Doermann, D. (2008). Geometric rectification of camera-captured document images. *IEEE transactions on pattern analysis and machine intelligence*, *30*(4), 591-605..

[8] Wen, Y., Lu, Y., Yan, J., Zhou, Z., von Deneen, K. M., & Shi, P. (2011). An algorithm for license plate recognition applied to intelligent transportation system. *IEEE Transactions on intelligent transportation systems*, *12*(3), 830-845..

[9] Zheng, L., He, X., Samali, B., & Yang, L. T. (2013). An algorithm for accuracy enhancement of license plate recognition. *Journal of computer and system sciences*, *79*(2), 245-255..

[10] Deselaers, T., Gass, T., Heigold, G., & Ney, H. (2011). Latent log-linear models for handwritten digit classification. *IEEE transactions on pattern analysis and machine intelligence*, *34*(6), 1105-1117..

[11] Jiao, J., Ye, Q., & Huang, Q. (2009). A configurable method for multi-style license plate recognition. *Pattern Recognition*, *42*(3), 358-369..

[12] Kocer, H. E., & Cevik, K. K. (2011). Artificial neural networks based vehicle license plate recognition. *Procedia Computer Science*, *3*, 1033-1037..

[13] Desai, A. A. (2010). Gujarati handwritten numeral optical character reorganization through neural network. *Pattern recognition*, *43*(7), 2582-2589..

[14] Pal, U., Roy, P. P., Tripathy, N., & Lladós, J. (2010). Multi-oriented Bangla and Devnagari text recognition. *Pattern Recognition*, *43*(12), 4124-4136..

[15] Manwatkar, P. M., & Singh, K. R. (2015, January). A technical review on text recognition from images. In *2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO)* (pp. 1-5). IEEE..

[16] Akopyan, M. S., Belyaeva, O. V., Plechov, T. P., & Turdakov, D. Y. (2019, September). Text recognition on images from social media. In *2019 Ivannikov Memorial Workshop (IVMEM)* (pp. 3-6). IEEE.

[17] Xiaojing Liu, Feiyu Gao, Qiong Zhang and Huasha Zhao, "Graph convolution for multimodal information extraction from visually rich documents", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, no. Industry Papers, pp. 32-39, June 2019.

[18] Seong Ah Chin and Raashid Malik, "Extraction of Text in Images", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4459-4469, October-November 2018.

Doi: https://doi.org/10.54216/FPA.040201
Received: April 12, 2021 Accepted: August 01, 2021