JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Study of document verification system andvalidation against QR codes.

Utkarsh Pathak[a], Venu Sonavane[b], Mitali Gadiya[c], Suyash Musale[d]
,G.B.Sambare[e]Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India

Abstract*— As there is a need for products and services available is growing at an bizarre rate. And with this shift to adopting online platforms, it becomes important for businesses to identify their customers via online portals. The paper also includes studies of different document verification and validation techniques and with respect to it a model has been proposed that will verify the authenticity of the documents by applying machine learning algorithms and further validating it against QR/Barcodes

***Keywords*—Machine Learning, OCR, QR codes, K-means, SVM, Longest Common Subsequence, OCROpus, forgery.**

I. Introduction

Analysis has demonstrated that false documents usually spread more rapidly than a piece of genuine or valid news. This has led to a significant need for document verification which is fast, secure, trustworthy and reliable. Online identities and online documents are the norms nowadays. Documents need to be safe and unadulterated and businesses need assurance that they won't be tricked by these documents. To check if the document has been altered or tampered by anyone, Document Verification System is of utmost importance. Document verification is the method of checking the authenticity of a given file. These methods are mandatory for firms that daily come across the end-to-end transactions, such as account handling, financial organizations and cryptocurrency companies.

II. Literature Review

A. *Image Acquisition*

The images captured with the help of smartphones are inconsistent. Different phone models, lighting, angles, and distance from the document can result in different images with varying degrees of background included in the final image. This makes the data inconsistent and can reduce the overall efficiency. Thus, the image acquisition step is a crucial step of the proposed solution as it ensures a standard input for image verification. This stage mostly deals with separating the actual image from the background and the perspective alignment of the image. This can be achieved in various ways.

Identification of vertices in the wild was reported to have an accuracy of 68.57% [1]. Through the use of an SVM, Simon et al. [2] categorized 74 distinct ID kinds. They collected spatial information using a combination of HOG and Colour features, attaining a mean class-wise accuracy of 97.7%. The work by Rajiv Jain outperforms the traditional change detection techniques using OCR and LCS by using the Levenshtein edit distance and SIFT [3]. The solution proposed by Alejandra Castelblanco [4] uses semantic segmentation using the UNETS deep learning architecture to remove complex backgrounds. Konstantin Bulatov proposed the use of the Viola-Jones algorithm along with a classifier for text recognition in video clips [5]. The use of the GrabCut Algorithm for localizing number plates has shown an accuracy of about 99.8% with a processing time of 0.21 seconds on moderate hardware [6].

Comparison of algorithms used in Image Acquisition		
Author	***Algorithm/ Technique***	***Accuracy***
Attivissimo , F.	Iterative sampling of vertices	68.57%
Simon et al	HOG and Color features	97.7%
Castelblanco, A.	UNETS deep learning architecture with RF	97.7%
Ayodeji Olalekan Salau	Modified GrabCut	99.8%

Table 1: Comparison of algorithms and features used in Image Acquisition

B. *Image/Document Verification*

After the process of image acquisition, the next step is Image/document verification. Assurance of the originality of the document is essential for the verification step. To find the longest ordered subsets shared by two sequences that are to be compared, the longest common subsequence (LCS) algorithm is utilised. On the contrary, an approach consisting of a polar opposite idea to this algorithm which is to look out for exactly the shortest set of differences is required for change of detection. The modifications for standard text are sometimes referred to as an expansion of the LCS issue, when text differs between documents in LCS, it is referred to as a deletion if it is present in the original document and an addition if it appears exclusively in the new document. Line level changes are another significant specific to verify the authenticity of the documents. OCR contains some of the top open-source line segmentation techniques. OCROpus is a character recognition system which clears the noise present in the document and extracts certain zones that prioritize binarized images of each line. The segmentation used in OCROpus is heavily dependent on thresholds derived from the resolution of the image [2]. In this process, we also try to achieve the maximum accuracy for verification using the Machine learning models. The advantages of using machine learning for the verification process are improved forgery detection, scalability, and protection from fraud. Content Verification uses K-means clustering which deals with symbols and characters in a binary document are grouped into different classes. Grouping of noisy variants of the same character class[8]. Boundary-based and region-based characteristics are employed in K-means clustering. The perimeter and low-order Fourier descriptors calculated from a character's outer boundary are examples of boundary-based features. Compactness, second-order moments, orientation, top-to-bottom area ratio, left-to-right area ratio, and size of the largest segment along the 0, 45, 90, and 135-degree angles are some of the region-based characteristics.

Connected component labelling is used in CV to detect connecting regions in binary digital images [8]

Comparison of features and algorithms used in document verification		
Author	***Algorithm/Technique***	***Accuracy***
Rajapakshe[9]	1. SHA-256 & ECDSA (Elliptic curve digital	95%

Comparison of features and algorithms used in document verification		
Author	***Algorithm/Technique***	***Accuracy***
	signature algorithm) 2. Image processing and content extraction 3. Forgery Detection 4. Signature generation	
M. Jiang [10]	K-means clustering+ Connected component labelling	97.5%
Mthethwa [11]	AnyOCR.	96.0%
N. Ghanmi [12]	Supervised Learning: Image classification SSD (Sum of square difference), SAD (Sum of absolute difference)	96%
R. Jain [7]	OCR + LCS	49.85%
R. Jain [7]	OCR + LCS+edit distance	70.9%
R. Jain [7]	SIFT + LCS	91.1%

Table 2: Comparison of algorithms and features used in Document Verification

C. *Image/Document Validation*

After verifying different extracted features of the image, the next important step is to check whether the data matches with the original document. In this step, the text data received by OCR is cross-checked by using the QR codes on the document "Quick Response" code is a two-dimensional matrix code that considers mainly two things, i.e., it must shave the capability to store large amounts of 1D barcodes and must have high decoding speed. The Quick Response codes are used in almost every domain for validation and authentication purposes where there is a need to send text-based information There are around more than 30+ different quick response codes which are used in various domains.

Due to benefits like as high data holding capacity, instant scanning, error-correction, and convenience of use, QR codes have experienced significant adoption over the past few years[13]

OCR method takes any image which may be printed, typed or handwritten as input and generates machine-encoded text which is further used for the validation process[14]

OCR is the most advanced method in comparison to the traditional method where text was generated by training every character of the image and also was font specific.

The objective of optical character recognition, an active area of research, is to create a computer system that can automatically extract and interpret text from photographs [15]

Initially, the OCR of the image obtained is performed and text data is generated. On the order side, the QR code on the image is scanned and a URL, pdf or text file can be generated from it as per needed. Both the Text from OCR and the one received after scanning are cross-checked and the final results are estimated. Quick response Validation is one the most simple and efficient methods to validate the document.

III. Conclusion

After analyzing several methods and techniques for image acquisition, GrabCut algorithm is to be preferred over the rest for optimum results. Verification may constitute some significant features where extended LCS nay be used where the subset of the document to be verified is to be matched against an authentic one to find out the originality of the document. OCRopus is an open source used for line and page segmentation. Machine learning algorithm like K-means clustering is applied for grouping and detection of noisy variants to ensure compactness of the verification process.

REFERENC ES

[1] Attivissimo, F., Giaquinto, N., Scarpetta, M., Spadavecchia, M.: An automatic reader of identity documents. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, vol. 2019–10, pp. 3525–3530 (2019)

[2] Simon, M., Rodner, E., Denzler, J.: Fine-grained classification of identity document types with only one example. In: Proceedings of the 14th IAPR, MVA 2015, pp. 126–129 (2015)

[3] Jain, R., & Doermann, D. (2013). VisualDiff: Document Image Verification and Change Detection. 2013 12th International Conference on Document Analysis and Recognition. doi:10.1109/icdar.2013.17

[4] Castelblanco, A., Solano, J., Lopez, C., Rivera, E., Tengana, L. and Ochoa, M., 2020. Machine Learning Techniques for Identity Document Verification in Uncontrolled Environments: A Case Study. [online] Springer.com.

[5] Bulatov, K., Arlazarov, V. V., Chernov, T., Slavin, O., & Nikolaev, D. (2017). Smart IDReader: Document Recognition in Video Stream. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). doi:10.1109/icdar.2017.347

[6] Ayodeji Olalekan Salau, Thomas Kokumo Yesufu, Babatunde Sunday Ogundare, Vehicle plate number localization using a modified GrabCut algorithm, Journal of King Saud University - Computer and Information Sciences, Volume 33, Issue 4, 2021

[7] R. Jain and D. Doermann, "VisualDiff: Document Image Verification and Change Detection," 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 40-44, doi: 10.1109/ICDAR.2013.17.

[8] ROBUST DOCUMENT IMAGEAUTHENTICATION
Ming Jiang, Edward K. Wong∗, Nasir Memon
Dept. of Computer and Information Science Polytechnic
University
Brooklyn, New York 11201

[9] Rajapakshe, Madura & Adnan, Muammar & Dissanayaka, Ashen & Guneratne, Dasith & Yapa Abeywardena, Kavinga. (2020). Multi-Format Document Verification System. American Scientific Research Journal for Engineering, Technology, and Sciences. 74. 48-60.

[10] M. Jiang, E. K. Wong and N. Memon, "Robust Document Image Authentication," 2007 IEEE International Conference on Multimedia and Expo, 2007, pp. 1131-1134, doi: 10.1109/ICME.2007.4284854

[11] Mthethwa, S. and Dlamini, N.P. 2018. Verifying the integrity of hardcopy document using OCR. 2nd International Women in Science Without Borders (WiSWB)-Indaba, Johannesburg, South Africa, 21-23 March 2018

[12] N. Ghanmi and A. M. Awal, "A New Descriptor for Pattern Matching: Application to Identity Document Verification," 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), 2018, pp. 375-380, doi: 10.1109/DAS.2018.74.

[13] Sumit TiwariAn Introduction To QR Code Technology978-1-5090-3584-7/16 $31.002016 IEEEDept. of Technical Education SITS Educators Society Jabalpur, Madhya Pradesh, India

[14] JAMSHED MEMON 1 , MAIRA SAMI 2 , RIZWAN AHMED KHAN 3 , AND MUEEN UDDIN 4 Handwritten Optical Character Recognition1426423 rd September 2016 1School of Computing, Quest International UniversityPerak, Ipoh 30250, Malaysia.

[15] Karez Abdulwahhab Hamad * 1 , Mehmet Kaya 2 A Detailed Analysis of Optical Character Recognition TechnologyIJAMEC, 2016, 4(Special Issue), 244–249 3 rd September 2016.