Text recognition of images is beneficial in a wide range of computer vision purposes such as robot navigation, document analysis, and image search. The optical character recognition (OCR) technique presents a simple tool to combine text recognition functionality to many industrial and educational applications. Best OCR results can be acquired when the background of the text image is uniform and appears as a document picture. In contrast, the challenges to recognizing accurate texts occur when the image has a non-uniform background that require further preprocessing to obtain acceptable OCR result. This work discusses three scenarios. Initially, this work will test the OCR on a normal business card as an image with a uniform background. Next, discusses the text recognition of a keypad image including digits with a non-uniform background. Here, there are two preprocessing algorithms used to enhance the OCR function to overcome the negative effect of the non-uniform background of images and to detect text with high accuracy. Finally, the developed OCR method is tested on different scanned bills and discusses the variation of the obtained results. The two algorithms are the morphological reconstruction to eliminate artifacts and create cleaner images to be further processed by OCR and the Region of Interest ROI-based OCR to spot explicit regions in a tested image. Verification for the effectiveness of the Morphological-based OCR over the ROI-based method has been conducted on a dataset of scanned electricity bills images with an accuracy of 98.2 % for Morphological-based while it is only about 89.3 % for ROI-based OCR

Keywords: Morphological Reconstruction, Optical Character Recognition (OCR), document images, non-uniform illumination images

UDC 621
DOI: 10.15587/1729-4061.2022.252803

DEVELOPMENT OF TEXT EXTRACTION TECHNIQUE USING OPTICAL CHARACTER RECOGNITION AND MORPHOLOGICAL RECONSTRUCTION TO ELIMINATE ARTIFACTS OF IMAGE'S BACKGROUND

Wasan M Jwaid
Doctor Lecturer
Department of Banking and Finance
Administration and Economics
University of Thi-Qar
Nasiriyah, Iraq, 0096442
E-mail: Wasan.maktoof@utq.edu.iq

Received date 03.01.2022
Accepted date 03.02.2022
Published date 25.02.2022

***How to Cite**: Jwaid, W. M. (2022). Development of text extraction technique using optical character recognition and morphological reconstruction to eliminate artifacts of image's background. Eastern-European Journal of Enterprise Technologies, 1 (2 (115)), 50–57. doi: https://doi.org/10.15587/1729-4061.2022.252803*

1. Introduction

Optical Character Recognition (OCR) is a method of extracting text from photographs and translating it to an electronic format. Handwritten text, printed text such as documents, receipts, name cards, and so on, or even a snapshot of a natural setting could be included in these graphics. There are two parts to OCR. The first step is text detection, which determines the textual portion of the image. The second component of OCR, text recognition, is where the text is recovered from the image, and this localization of text inside the image is crucial. You can extract text from any image by combining these techniques. Pattern recognition in the second approach identifies the character as a whole. A line of text can be identified by looking for rows of white pixels separated by rows of black pixels. Similarly, we can determine where a certain character begins and stops. We convert the character's picture into a binary matrix of white pixels (zeros) and black pixels (ones). We can calculate the distance between the matrix's center and the farthest point. Then we make a circle with that radius and divide it into smaller portions. At this stage, the algorithm will compare each subpart and will be able to send a database of matrices representing characters in different fonts. To locate a character, look at whom it has the most in common with statistically. It's simple to move printed media into the digital world if you do this for every line in every character.

Optical character recognition is the electronic conversion of handwritten, typewritten, or printed text into machine-translated pictures (OCR). It's often used to recognize and search text in electronic documents, as well as to publish a text on the internet. Recognizing characters from text has been a prominent problem in the field of computer vision. OCR systems, which are widely used in a range of applications, allow human-machine interaction. Character recognition has been the subject of much research in some languages. A survey of OCR applications in many domains was offered in the paper [1]. The study [2] discussed the automatic text identification from natural photographs captured under uncontrolled lighting circumstances as a difficult task due to the existence of shadows. Due to the unique properties of the Arabic cursive script, establishing an OCR for printed Arabic text and Latin characters [3] remains a tough and open research subject. Filtering picture email spam is considered a difficult task because spammers constantly modify the graphics used in their campaigns using various obfuscation tactics [4]. Methods other than OCR-based text recognition were also used to perform this task like using a convolutional neural network [5], machine learning classifier [4]. These techniques provide high recognition accuracy but with complex software and consume a relatively long execution time.

Scanning printed papers into editions that can be updated with word processors similar to Google Docs or Microsoft Word is one example of how OCR may be used. The application of OCR also includes:

1) print material that is indexed for exploration engines;

2) data entry, processing, and extraction that are all automated;

3) documents that are deciphered into text that may be read loudly to blind people or visually impaired;

4) archiving historical material into searchable formats, such as phonebooks, magazines, or newspapers;

5) depositing checks electronically instead of going to a bank cashier;

6) signed legal and important documents that are entered into an electronic database;

7) using a camera or software to recognize text, such as license plates;

8) sorting letters in preparation for delivery;

9) words within an image that are translated into a specific language.

The key paybacks of OCR knowledge are reduced errors, reduced effort, and time savings. It also allows you to do things like compressing data into ZIP files, incorporating them into a website, highlighting keywords, and connecting them to an email that isn't possible with physical copies. From the aforementioned information, it is important to conduct research on developing text recognition using OCR as it is effective in many real-world applications such as digital archiving, relying heavily on OCR [6], fully automated OCR-based electricity billing [7], automatic number plate recognition [8–11], handling bank cheques [12], etc.

2. Literature review and problem statement

The paper [13] discussed the issue of solving the Javanese characters as a non-Latin alphabet text by the OCR technique texts. Although this work collected datasets of 5880 of these characters and trained them to be trained by Tesseract OCR methods, the paper only presented a classification of such types of characters. In the same context, the study [14] addressed the issue of understanding images including Bangla text. However, the limitation of this work is that it only involves extracting Bangla texts from corresponding images. The paper [15] also utilized the OCR model for Standard Yorùbá printed texts by creating image text lines to be trained by Recurrent Neural Network. Although its result shows a good recognition with Times New Roman font and Ariel font image dataset, it only solved these texts and it didn't discuss the background issues.

Image preprocessing method using local imaging entropy filter is presented by [16] with a dataset of 140 text images of different brightness to allow the expansion image threshold settings for text recognition applications. The results of this work were expressed as F-Measure values and Levenshtein distances to obtain the texts, but it was limited with black and white images. This issue was discussed by the study [17] as it presented OCR using a zoning-based technique and classification for 50 images with printing texts. However, this approach required a relatively high-quality input document to get high accuracy.

The paper [7] used the OCR to recognize seven-segment digits of a digital energy meter through using raw document images. Although this paper conducted OCR on images with day/night light surroundings, the approach was limited to classify seven-segment digits only. This application has been expanded by enhancing the OCR with a convolution-based pre-processing with specific kernels [18]. Therefore, detection capability to identify characters in more realistic has been improved. However, this will serve only the cleared documented images. This issue also was the disadvantage of the study [19], where a microcontroller-based OCR image scanning with the camera fixed on a fingertip was proposed, but it requires noise-free images.

The paper [20] compares different OCR tools with open-source data to extract texts (vehicle number) from an image. It discussed various OCR difficulties like style, digits orientation, different sizes, and the image background. However, the accuracy was very influenced by the surroundings.

According to the aforementioned literature, accurately recognizing texts, especially for images with a non-uniform background is promising as it is beneficial in a wide range of computer vision purposes such as robot navigation, and document analysis. Therefore, it is necessary to develop a text extraction technique combining an appropriate algorithm and the OCR to eliminate artifacts of the image's background to get accurate recognition results.

3. The aim and objectives of the study

The aim of the study is to develop a text extraction technique using optical character recognition to eliminate artifacts of the image's background to get accurate recognition results.

To achieve this aim, the following objectives are accomplished:

– to recognize a text from a document image including its location, and the recognition confidence of a non-text object in the original image;

– to analyze the performance of OCR-based text recognition using two different algorithms to overcome image background;

– to verify the effectiveness of the OCR and these algorithms on a dataset of scanned electricity bills images.

4. Materials and methods

Firstly, we examine the OCR to recognize particular digits on a normal business card as an example of an image with uniform background. Next, we test the effectiveness of using OCR to recognize digits of a keypad image with a non-uniform background. The negative effect of the non-uniform background is pre-processed using two algorithms:

1) morphological reconstruction algorithm, which is employed to eliminate artifacts and create cleaner images to be further processed by OCR;

2) ROI-based preprocessing method to spot explicit regions in a tested image, which the OCR must process.

Finally, these OCR techniques are verified on a dataset of different electricity scanned bills images. Therefore, the methodology will discuss these scenarios with three issues.

This work discusses three scenarios and the image processing MATLAB tools that have been used to perform all the tasks of this study based on the three scenarios and their corresponding algorithms.

4. 1. Uniform background

One sample for an image with a uniform background may be a business card, which was obtained from Google. This sample can be considered as a document image we want to extract text from, as well as its location, and the recognition confidence of a non-text object in this original image. The algorithm steps in such conditions can be demonstrated in Fig. 1.

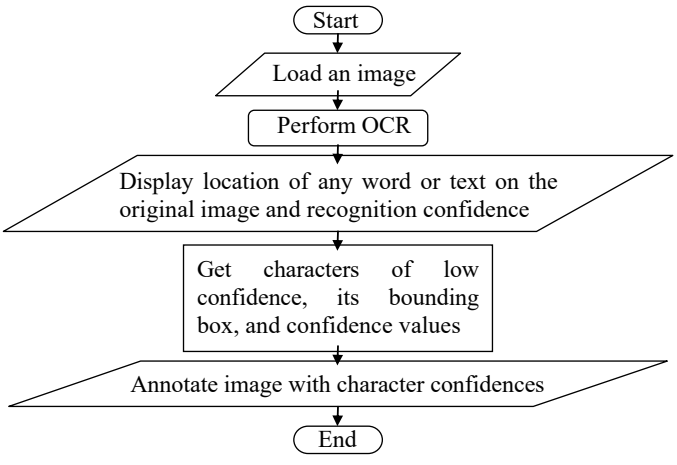


Fig. 1. Application of OCR on a business card

Since the business card, in this scenario, is usually clear, a direct application of OCR is sufficient to obtain a text of a document image, its location, and the recognition confidence of a non-text object in that image.

4. 2. Non-uniform background challenges

The difficulty of text extraction appears when the background becomes non-uniform after converting the image to a grayscale image. Pre-processing steps are required to filter images from noise and overcome such difficulty. One example of such a condition is an image of a keypad, which is initially examined by the OCR to see if it requires further processing. Despite it looks easy to handle, the text recognition of a keypad image is a challenge due to its non-uniform background and the existence of different sizes of texts in the original image. To overcome the unhelpful effect of the non-uniform surroundings, the first method is by using the morphological reconstruction algorithm, which is demonstrated in Fig. 2.

This extra processing to enhance the OCR method is employed to eliminate artifacts and create cleaner images. The second method for filtering images with noisy and non-uniform backgrounds is the ROI-based preprocessing method to spot explicit regions in a tested image, which the OCR must process. The flowchart showing this algorithm is depicted in Fig. 3.

These steps initially identify particular regions by ranging the areas and character sizes that are the digits in the example of the keypad photos.

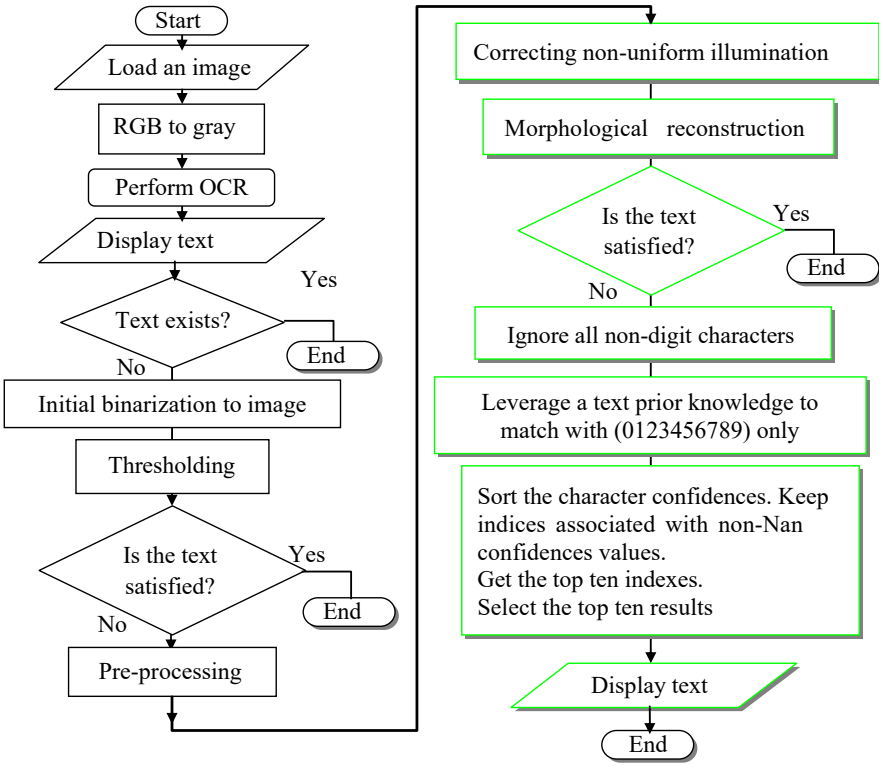


Fig. 2. Morphological reconstruction algorithm with the OCR method

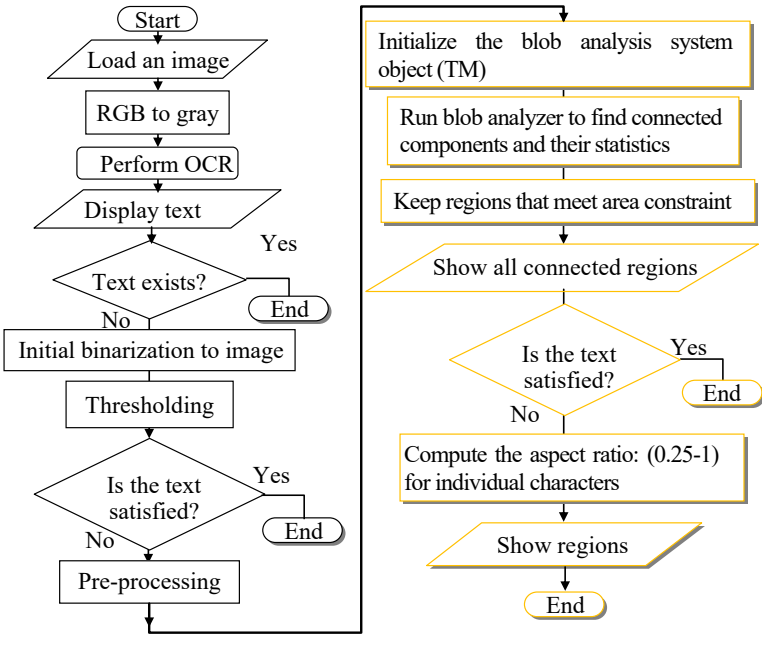


Fig. 3. ROI-based preprocessing to enhance the OCR method

The contact number of the business card owner was detected and its location on the image was identified as shown in Fig. 5, *b*. An arbitrary small object has been recognized with a confidence of 0.48.

5. 2. Results of applying OCR-based text recognition for images with non-uniform backgrounds

The results of the non-uniform background image when processed by Morphological reconstruction algorithm with OCR method are shown in Fig. 5.

The results of the non-uniform background image when processed by the ROI-based preprocessing method with the OCR method are shown in Fig. 6.

These results show that the Morphological reconstruction algorithm with the OCR method performs better than ROI-based preprocessing with OCR as the latter marks characters and things other than the numbers.

4. 3. Electricity bills verification

This work uses 195 scanned document images of electricity bills obtained from [21], which also includes a letter, form, email, handwritten, advertisement, scientific report, scientific publication, specification, file folder, news article, budget, invoice, presentation, questionnaire, resume, memo.

A quantitative comparison between the methods mentioned in Scenario 2 is conducted through a table, while the outcome of the best method would be presented by a graph showing the interesting numbers of these electricity bills.

5. Results of the developed text extraction technique

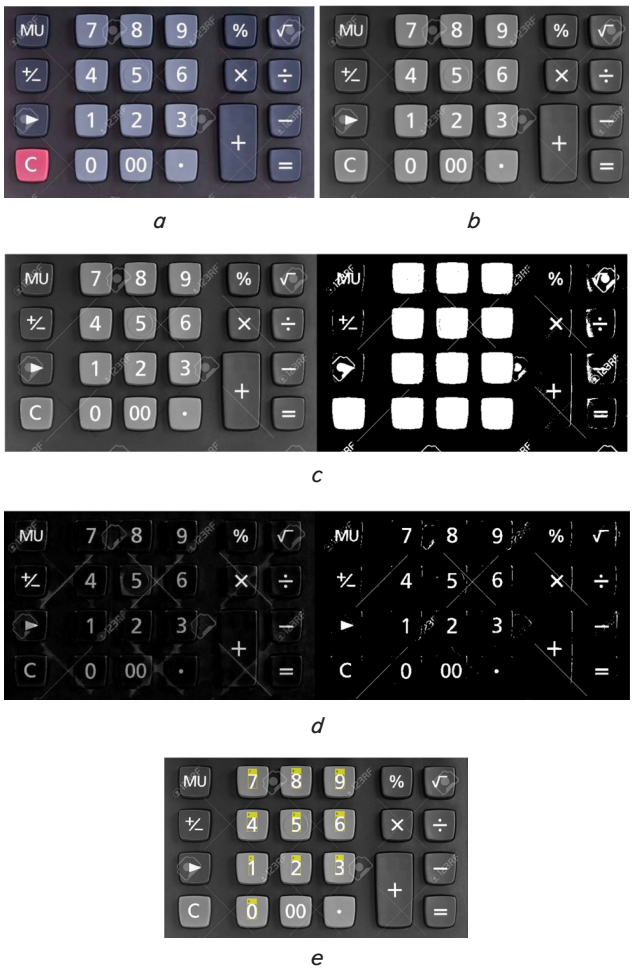
5. 1. Results of text recognition of a document image with uniform background

The results of the application of the OCR-based text recognition on an arbitrarily selected business card image with a uniform background are shown in Fig. 4.

a

b

Fig. 4. Results of applying OCR-based text recognition: *a* – original image highlighting the contact number; *b* – non-text confidence of an object on the image



a *b*

c

d

e

Fig. 5. The results when the image is processed by Morphological reconstruction algorithm with OCR method: *a* – original colored image of non-uniform background; *b* – applying OCR only; *c* – checking thresholding of OCR only; *d* – adjust thresholding result; *e* – applying Morphological reconstruction algorithm after the filter

a *b* *c*

Fig. 6. The results when the image is processed by ROI-based preprocessing with OCR method: *a* — applying blob analyzer to get all connected areas; *b* — show remaining blobs after area constraints, *c* — the outcome after area and aspect ratio constraints, removing all trailing characters

5. 3. Results of OCR verification on scanned electricity bills images

To verify the effectiveness of using the above algorithms, we applied these methods on scanned electricity bills images. One sample of these bills is shown in Fig. 7.

The red background represents texts with wrong characters, while the row of green background denotes the detected information of the image shown in Fig. 7.

Morphological reconstruction algorithm with the OCR method has been conducted as it proved that it performs better on the 195 scanned document images of electricity bills. The results of the first 20 images are listed in Table 1. All the results of the 195 bills are graphically expressed as shown in Fig. 8.

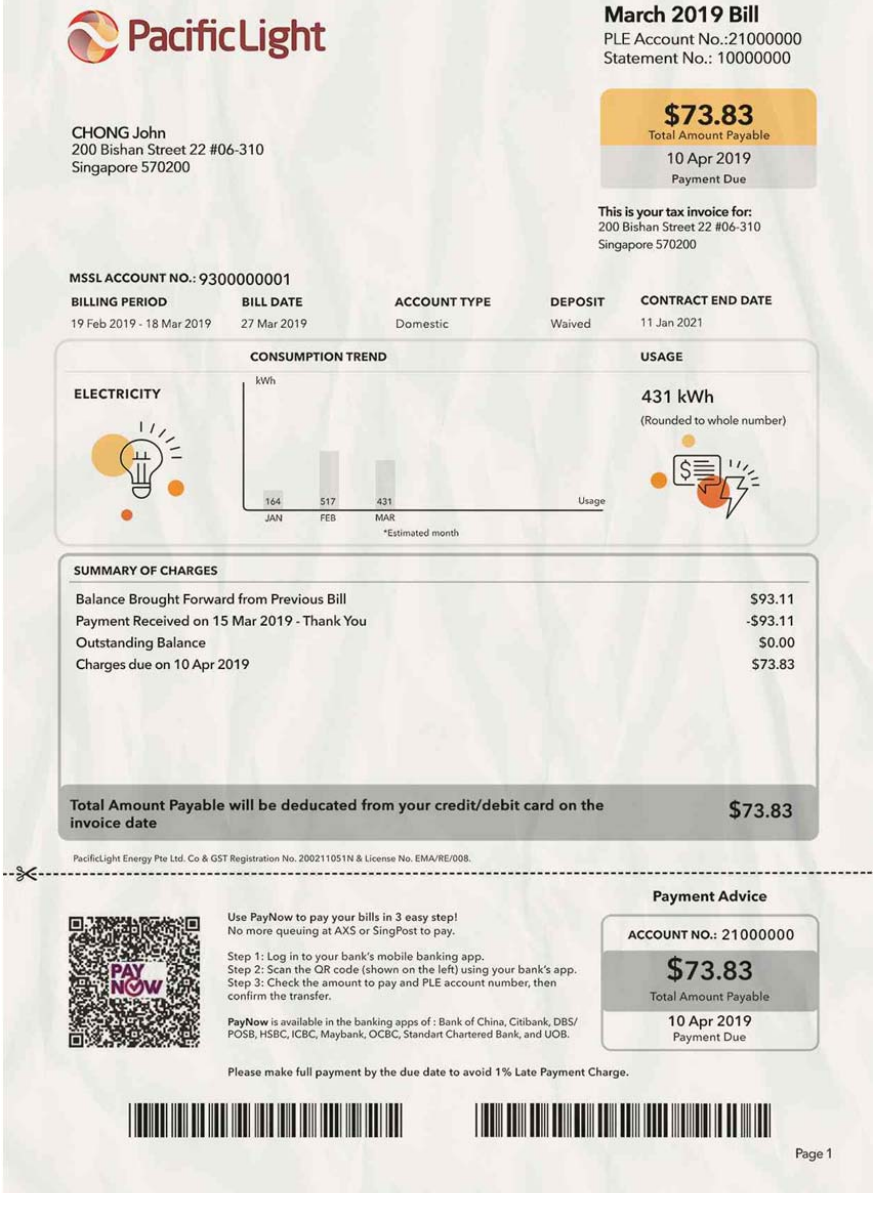
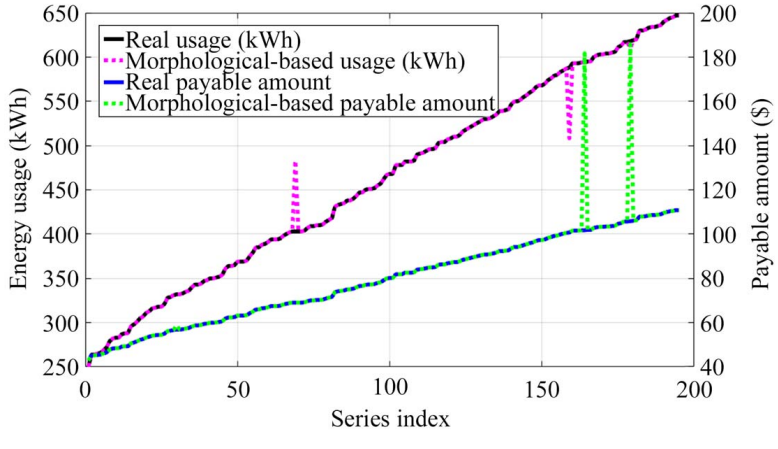


Fig. 7. One sample of scanned electricity bills images

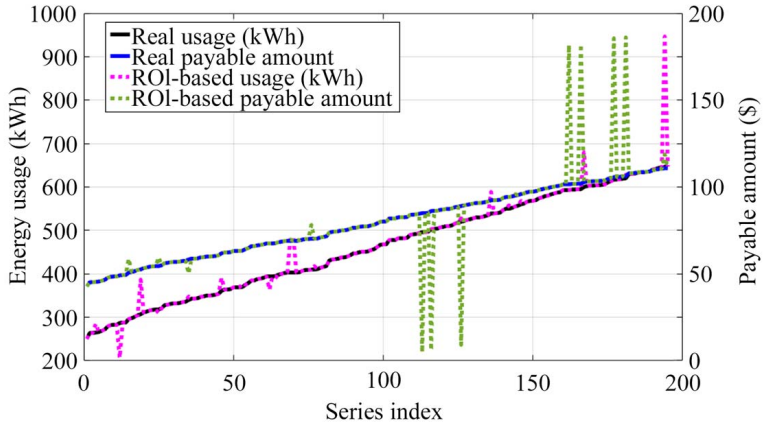
Table 1

The results of the first 20 images

Detected by Morphological-based				Detected by ROI-based OCR			
Date	Account No.	Usage (kWh)	Payable amount	Date	Account No.	Usage (kWh)	Payable amount
13-Apr-19	9300000160	402	68.86	13-Apr-19	9300000160	402	68.86
29-Apr-19	9300000030	403	69.03	29-Apr-19	9300000030	403	69.03
28-Apr-19	9300000043	483	69.03	20-Apr-19	9300000043	483	69.83
8-Apr-19	9300000146	403	69.03	8-Apr-19	9300000146	483	69.03
14-Apr-19	9300000182	403	69.03	14-Apr-19	9300000182	403	69.83
30-Apr-19	9300000082	404	69.2	30-Apr-19	9300000082	404	69.2
8-Apr-19	9300000016	407	69.72	8-Apr-19	9300000016	407	69.72
23-Apr-19	9300000095	409	70.06	23-Apr-19	9300000095	409	70.06
16-Apr-19	9300000136	409	70.06	16-Apr-19	9300000136	409	70.06
6-Apr-19	9300000020	410	70.23	6-Apr-19	9380080020	410	78.23
25-Apr-19	9300000048	410	70.23	25-Apr-19	9308080048	418	70.23
30-Apr-19	9300000058	411	70.4	30-Apr-19	9300000058	411	70.4
7-Apr-19	9300000037	414	70.92	7-Apr-19	9300000037	414	70.92
19-Apr-19	9300000164	416	71.26	19-Apr-19	9300000164	416	71.26
20-Apr-19	9300000036	418	71.6	20-Apr-19	9300000036	418	71.6
10-Apr-19	9300000001	431	73.83	10-Apr-19	9300000001	431	73.83
29-Apr-19	9300000132	433	74.17	29-Apr-19	9300000132	433	74.17
4-Apr-19	9300000061	434	74.34	4-Apr-19	9300000061	434	74.34
22-Apr-19	9300000184	435	74.52	22-Apr-19	9300000184	435	74.52
30-Apr-19	9300000085	438	75.03	30-Apr-19	9300000085	438	75.03



a



b

Fig. 8. The obtained texts of energy usage and payable amount of 195 bills are graphically expressed: *a* — comparison between the real and Morphological-based OCR; *b* — comparison between the real and ROI-based OCR

6. Discussion of the results of the three scenarios

Using OCR-based text recognition was sufficient to detect the contact number of a business card owner as well as its position on the image was identified as shown in Fig. 5. This can serve to build an automatically PC-based recording system for achieving a huge number of business cards. However, this method didn't perform well to extract texts of images with non-uniform backgrounds. Two algorithms have been discussed to overcome this difficulty: the Morphological reconstruction algorithm and the ROI-based processing. The results that have been attained from conducting these methods on document images with non-uniform backgrounds show that the Morphological reconstruction algorithm performs better than ROI-based preprocessing when enhancing the OCR to detect texts more accurately as shown in Fig. 5, 6 respectively. This result is verified by performing Morphological-based OCR on 195 scanned electricity bills images (a sample in Fig. 8) with an accuracy of 98.2 % for Morphological-based while it is only about 89.3 % for ROI-based OCR, which is satisfied visually (Table 1, Fig. 8). The developed technique is background noise-resistant and free of the artifacts that both global thresholding and fixed-value based local thresholding are known to cause.

Although the accuracy of text recognition from document images was acceptable for designing a computer-based electronic archiving, a complete automated system combining all the above steps is required, and visual satisfaction was a time-consuming method. Future work may consider these difficulties.

7. Conclusions

1. The results of the presented methodology show that the traditional OCR-based text recognition was able to extract text from a document image of uniform background such as a business card. The detection includes the text location and a non-text object with a particular confidence range in the original image.

2. It is found that the difficulty of text recognition for images with non-uniform background has been overcome throughout developing two different algorithms that are the Morphological reconstruction and the ROI-based algorithms to preprocess the input images and enhance the OCR function to detect texts more accurately. The analysis of the obtained results shows that the Morphological-based performs better than the ROI-based method.

3. The verification of the effectiveness of the developed approach using Morphological-based OCR over the ROI-based method has been conducted on a dataset of scanned electricity bills images with an accuracy of 98.2 % for Morphological-based while it is only about 89.3 % for ROI-based OCR.

References

1. Singh, A., Bacchuwar, K., Bhasin, A. (2012). A Survey of OCR Applications. International Journal of Machine Learning and Computing, 314–318. doi: https://doi.org/10.7763/ijmlc.2012.v2.137
2. Fang, Y., Yao, J. (2014). Multi-operator combination for character segmentation in complex background. 2014 International Conference on Audio, Language and Image Processing. doi: https://doi.org/10.1109/icalip.2014.7009896
3. Park, J., Lee, E., Kim, Y., Kang, I., Koo, H. I., Cho, N. I. (2020). Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter. IEEE Access, 8, 174437–174448. doi: https://doi.org/10.1109/access.2020.3025769
4. Al-Duwairi, B., Khater, I., Al-Jarrah, O. (2013). Detecting Image Spam Using Image Texture Features. International Journal for Information Security Research, 3 (4), 344–353. doi: https://doi.org/10.20533/ijisr.2042.4639.2013.0040
5. Qaroush, A., Awad, A., Modallal, M., Ziq, M. (2020). Segmentation-based, omnifont printed Arabic character recognition without font identification. Journal of King Saud University - Computer and Information Sciences. doi: https://doi.org/10.1016/j.jksuci.2020.10.001
6. Navitski, R. (2014). Reconsidering the Archive: Digitization and Latin American Film Historiography. Cinema Journal, 54 (1), 121–128. doi: https://doi.org/10.1353/cj.2014.0065
7. Kanagarathinam, K., Sekar, K. (2019). Text detection and recognition in raw image dataset of seven segment digital energy meter display. Energy Reports, 5, 842–852. doi: https://doi.org/10.1016/j.egyr.2019.07.004
8. Farhat, A., Hommos, O., Al-Zawqari, A., Al-Qahtani, A., Bensaali, F., Amira, A., Zhai, X. (2018). Optical character recognition on heterogeneous SoC for HD automatic number plate recognition system. EURASIP Journal on Image and Video Processing, 2018 (1). doi: https://doi.org/10.1186/s13640-018-0298-2
9. Arora, M., Jain, A., Rustagi, S., Yadav, T. (2019). Automatic Number Plate Recognition System Using Optical Character Recognition. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 986–992. doi: https://doi.org/10.32628/cseit1952280
10. Vaishnav, A., Mandot, M. (2019). Template Matching for Automatic Number Plate Recognition System with Optical Character Recognition. Advances in Intelligent Systems and Computing, 683–694. doi: https://doi.org/10.1007/978-981-13-7166-0_69
11. Akhtar, Z., & Ali, R. (2020). Automatic Number Plate Recognition Using Random Forest Classifier. SN Computer Science, 1 (3). doi: https://doi.org/10.1007/s42979-020-00145-8
12. Srivastava, S., Priyadarshini, J., Gopal, S., Gupta, S., Dayal, H. S. (2018). Optical Character Recognition on Bank Cheques Using 2D Convolution Neural Network. Applications of Artificial Intelligence Techniques in Engineering, 589–596. doi: https://doi.org/10.1007/978-981-13-1822-1_55

13. Robby, G. A., Tandra, A., Susanto, I., Harefa, J., Chowanda, A. (2019). Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application. Procedia Computer Science, 157, 499–505. doi: https://doi.org/10.1016/j.procs.2019.09.006
14. Rajbongshi, A., Ibadul, M., Amin, A., Mahbubur, M., Majumder, A., Ezharul, M. (2020). Bangla Optical Character Recognition and Text-to-Speech Conversion using Raspberry Pi. International Journal of Advanced Computer Science and Applications, 11 (6). doi: https://doi.org/10.14569/ijacsa.2020.0110636
15. Oni, O. J., Asahiah, F. O. (2020). Computational modelling of an optical character recognition system for Yorùbá printed text images. Scientific African, 9, e00415. doi: https://doi.org/10.1016/j.sciaf.2020.e00415
16. Michalak, H., Okarma, K. (2019). Improvement of Image Binarization Methods Using Image Preprocessing with Local Entropy Filtering for Alphanumerical Character Recognition Purposes. Entropy, 21 (6), 562. doi: https://doi.org/10.3390/e21060562
17. Barnouti, N. H., Abomaali, M., Al-Mayyahi, M. H. N. (2018). An efficient character recognition technique using K-nearest neighbor classifier. International Journal of Engineering & Technology, 7 (4), 3148–3153. doi: https://doi.org/10.14419/ijet.v7i4.18952
18. Sporici, D., Cuşnir, E., Boiangiu, C.-A. (2020). Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing. Symmetry, 12 (5), 715. doi: https://doi.org/10.3390/sym12050715
19. Sowmya, R., Jagtap, S. S., Kasthuri, G. (2020). Smart Reader for Visually Challenged Using Optical Character Recognition and Text-To-Speech. Innovations in Information and Communication Technology Series, 205–208. doi: https://doi.org/10.46532/978-81-950008-1-4_045
20. Majumdar, J., Gupta, R. (2019). An Accuracy Examination of OCR Tools. International Journal of Innovative Technology and Exploring Engineering, 8 (9S4), 5–9. doi: https://doi.org/10.35940/ijitee.i1102.0789s419
21. The RVL-CDIP Dataset. Available at: https://www.cs.cmu.edu/~aharley/rvl-cdip/