International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410111

Thinning Chinese, Korean, Japanese and Thai script for segmentation-free OCRs

[1]Abdul Majid, [1]Qinbo, [2]Dil Nawaz Hakro,[1] Saba Brahmani

[1]Department of Computer Science and Technology, Faculty of Information Science and Engineering, Ocean University of China

[2]Department of Software Engineering, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan

Corresponding Author : Abdul Majid, Abdul.majid827@yahoo.com

ARTICLEINFO

Article History:

Accepted: 02 Jan 2024

Published: 16 Jan 2024

Publication Issue

Volume 10, Issue 1

January-February-2024

Page Number

116-121

ABSTRACT

While searching on the internet, the OCR keyword will return a thousand research papers on optical character recognition. These papers are ranging from Latin language scripts, Cyrillic, Devanagari, Korean, Japanese, Chinese and Arabic scripts. Sindhi and many other languages extend the Arabic script in which base characters are same while the other characters are adopted in a same situation. Many of the languages possess OCRs for their languages but still there are some other languages which still require the OCRs for their language. The paper is organized in various sections such as introduction followed by Sindhi language characteristics. The OCR approaches and methods are explained. The last section describes the conclusion and future work. An OCR is a set of complex steps to convert image text to editable text. Skeletonization or shrining a word or character body is a method which helps to recognize text more easily. Multiple languages impose various challenges and are hard to recognize and skeletonization or thinning produces a new image which can be easy to recognize. The connected elements are found with this approach. A custom-built software has been developed to interface the generalized thinning algorithm so that the scripts of Chinese, Japanese, Korean and Thai be tested. The output of this algorithm is the final image to be used for the further processing of the OCR. Although the intention was to create algorithms for segmentation free OCRs, the study results and the software can also be used for segmentation-based algorithms. The generalized algorithm shows the accuracy of more than 95% for the experimented four scripts.

Keywords: OCR, Thai, Chinese, Japanese, Korean, languages, scripts, segmentation.

I. INTRODUCTION

The computer recognition of image text is called Optical Character Recognition (OCR) [1], The most technical job to recognize handwritten words and text is called Intelligent Character Recognition (ICR), sometimes referred as handwritten character recognition [2]. The technology that can read the image represented in x and y direction pixels to the editable text represented by the codes in form of ASCII or Unicode [3]. The process of recognizing text contains multiple stages and various algorithms have been proposed for these stages namely preprocessing, feature extraction, segmentation of lines, characters and words, and the classification or the recognition of the image text [4]. A text image must be prepared to make it more appropriate for the next stages of OCR process is call preprocessing [5]. During preprocessing multiple approaches are required as per need of the time such as skew detection is used to make an image text to align properly so that the recognition can be made easy. In the same way, thinning is one of the process or stage for the OCR preprocessing in which a word, character or text image is made into its skeleton form or the size of single pixel so that the recognition of the text image is more capable and increase the recognition accuracy [6]. As the name implies, the text image or word is thinned to its one dot skeleton and other dots are removed so that the text available in image can be found properly with more accuracy. For thinning, typically most of the researchers have used binary or two-tone images, hence the current study will utilize the same approach and the proposed algorithm will be applied to binary text images.

1. Comparison with existing studies

Many of the researchers have presented studies on thinning of the images and the possibilities of increased recognition rate including [7] and [8] where a generalized algorithm for thick images to thin images has been presented. The existing work has been mostly `used with Arabic script or its adopting languages but this study has applied the thinning algorithm for the images of Japanese, Chinese, Thai and Korean language text images. In other words, this can be considered as the extension of the exhaustive experiments to validate the algorithm. Various images of these four languages have been tested with interactive thinning algorithm and the results are presented here.

2. Custom built application

The earlier algorithm which has been tested with MATLAB 2020 version now has been refined and tested with MATLAB 2023b. The images of languages including Japanese and Chinese have been loaded to the custom-built application which has been created for thinning images [11]. The software is capable of loading images of word images, character images as well as sentences images. The software works as the integrative model for multiple algorithms where various algorithms are automatically called in sequence and performed with a single touch of the button. The text images upon loading can be edited and even the text image can also be edited during the process of skeletonization. The user has given authority and the independence to edit text image at any stage of the process and can stop and save the image at any stage of the thinning process. Figure 1 shows the design process of the custom-built software. Figure 2 illustrate the skeletonization process of Chinese [12] language script. The image dots can be added or removed during the process as the process can be stopped and repeated at any time during thinning. The skeletonizing process can be stopped at any time if the skeleton of the text image is found. The user is able to add or remove a particular dot to text image skeleton if it is broken or over dotted at some place. The software can edit, skeletonize, fine tune and save the images to a particular place of the computer.

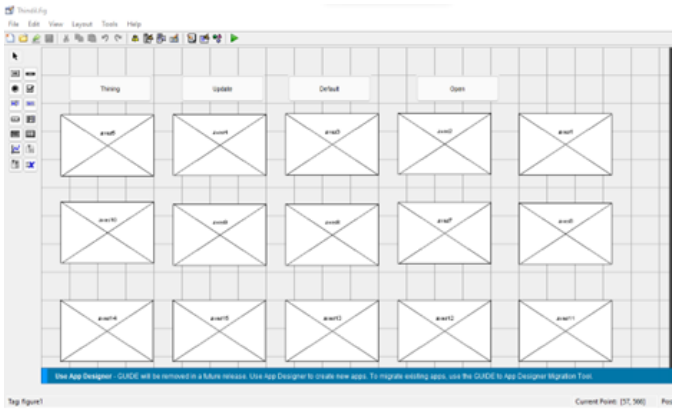


Figure 1: Design Process of the custom built Software

Figure 2: Chinese text thinning process

1. Types of OCR based on segmentation

The OCR process can be completed by segmentation of the words and sentences whereas the other approach is to recognize the words and sentences without segmenting. These approaches are called segmentation based [9] where the images are segmented and other approach is called segmentation free [10] which recognizes text without segmentation. The current study deals with both of the approaches where the interactive algorithm can be applied to segmentation based and segmentation free type of OCRs. The text images are skeletonized so that the recognition process will be easier and feature extraction process will be definite and clear to recognize. Thinning is also helpful in various steps of recognition process as well as the analysis of the images [4]. Various studies have reported that the thinned image is more helpful in recognition because it contains only strokes and lines of the characters or words in the image of the text.

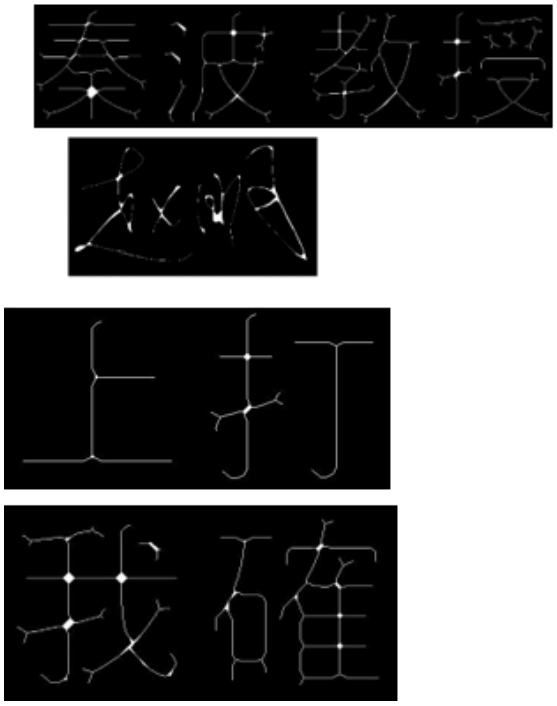
1. Interactive thinning algorithm

The original image is loaded for the thinning process in which a mask is used which gives output as the thinned image. The process contains a 2-pixel mask along with 3-pixel mask. The loaded text image is checked by the algorithm whether it is a black background with white text or white background with black text. The background is the indicator to identify the white or black background. To find out the overall blob of the text, area of the text and skeleton, sliding window has been used. The repetitive process is continued by using multiple factors and each and every time to check that the text image is remained with only one pixel. The iterative process is stopped while the skeleton is only one pixel thick.

1. RESULTS AND DISCUSSION

The interactive algorithm has been applied on Sindi images [14], Arabic language images and other languages adopting Arabic script including Persian, Pashto, Yugur and Urdu [15]. The interactive algorithm has much more to do with various scripts. This is the reason that the algorithm has been tested for further languages including Korean, Thai, Japanese and Chinese. The results of these languages are presented here. The current interactive algorithm is to be tested on multiple languages and scripts. The interactive algorithm will surely help in recognizing various scripts around the world and many of the researchers will benefit. For the sake of experiments, careful selection has been taken for the images and some of the images and thinned results are presented here. Practically any image can be included for testing of the algorithm. Some of the images from Japanese, Chinese, Thai and Korean languages are presented from Figure 3 to 6. The text has been taken from the Holy Quran translations available on internet.

(a) Original Image

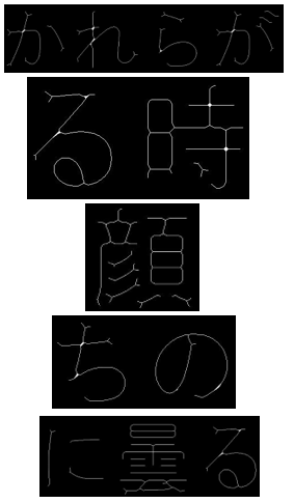


(b) Image after Thinning

Figure 3 : Thinning of Chinese Images: (a) Original Image (b) Image after Thinning

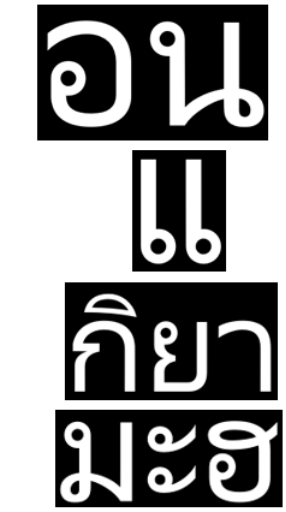
The custom-built software creates images called synthesized images. The images are part of the OCR study, and multiple billion-word images and character images were created [17]. Figure 3 illustrates the samples of the experiments done on the Chinese images. The results show that the many of the Chinese images are successfully thinned and the skeleton is found as shown in Figure 3. Japanese text images were also tested with the interactive algorithm and the Japanese characters were successfully skeletonized. The samples of Japanese text images are illustrated in Figure 4. The original and thinned images are illustrated.

(a) Original Image

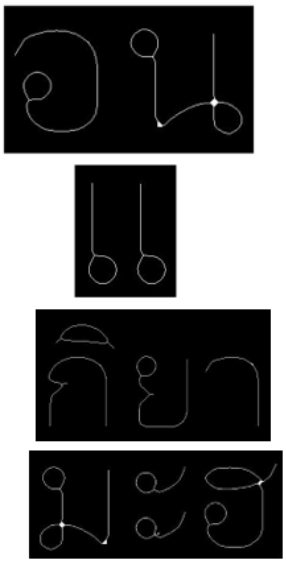


(b) Image after Thinning

Figure 4 : Thinning of Japanese Images: (a) Original Image (b) Image after Thinning



(a) Original Image



(b) Image after Thinning

Figure 5: Thai script Thinning: (a) Original Image (b) Image after Thinning

(a) Original Image

(b) Image after Thinning

Figure 6 : Thinning of Korean Images: (a) Original Image (b) Image after Thinning

II. CONCLUSION

OCR is a recognition of the various texts available in the images in the form of x and y coordinates of the pixels to the editable text. OCR is one of the integrated processes containing various algorithms and approaches to achieve text recognition. Most of the segmentation free OCRs are using the help of the thinning algorithm to increase the accuracy level. The interactive thinning algorithm has been tested on various language scripts and current study presents the testing results of Japanese, Chinese, Thai and Korean scripts. The algorithm has performed well on various text images of these language scripts. The algorithm has been tested on single word, double word and sentence text images. The custom-built application has been tested for interactivity which performed as per required results. The interactive algorithm can also be used with other multiple scripts and is expected to skeletonize the images with increased accuracy.

III. REFERENCES

[1]. Tote, A. S., Pardeshi, S. S., & Patange, A. D. (2023). Automatic number plate detection using TensorFlow in Indian scenario: An optical character recognition approach. Materials Today: Proceedings, 72, 1073-1078.

[2]. Berriche, L., Alqahtani, A., & RekikR, S. (2024). Hybrid Arabic handwritten character segmentation using CNN and graph theory algorithm. Journal of King Saud University-Computer and Information Sciences, 36(1), 101872.

[3]. Xue, S., Wang, S., Wu, T., Di, Z., Xu, N., Sun, Y., ... & Zhou, P. (2023). Hybrid neuromorphic hardware with sparing 2D synapse and CMOS neuron for character recognition. Science Bulletin, 68(20), 2336-2343.

[4]. Su, G., Zhao, S., Li, T., Liu, S., Li, Y., Zhao, G., & Li, Z. (2024). Image format pipeline and instrument diagram recognition method based on deep learning. Biomimetic Intelligence and Robotics, 4(1), 100142.

[5]. Li, J., Wang, Q. F., Huang, K., Yang, X., Zhang, R., & Goulermas, J. Y. (2023). Towards better long-tailed oracle character recognition with adversarial data augmentation. Pattern Recognition, 140, 109534.

[6]. Elaraby, N., Barakat, S., & Rezk, A. (2024). A generalized ensemble approach based on transfer learning for Braille character recognition. Information Processing & Management, 61(1), 103545.

[7]. Hakro (2015), ENHANCED SEGMENTATION AND FEATURE EXTRACTION FOR SINDHI OPTICAL CHARACTER RECOGNITION, PhD thesis, Submitted to University science Malaysia (USM), Malaysia.

[8]. Cowell J. and H. Fiaz (1992). “Thinning Arabic character feature extraction“, IEEE Transactions on Pattern Analysis Machine Intelligence, Vol. 14, No.11, 869-885,

[9]. Fan, X. and Verma, B. (2001). Segmentation vs. non segmentation based neural techniques for cursive word recognition: an experimental analysis, Computational Intelligence and Multimedia Applications, 2001. ICCIMA 2001. Proceedings. Fourth International Conference on, IEEE, Yokusika City, Japan, pp. 251–255.

[10]. Premaratne, H. and Bigun, J. (2004). A segmentation-free approach to recognise printed Sinhala script using linear symmetry, Pattern recognition 37(10): 2081–2089.

[11]. Zhang T. Y. and C. Y. Suen, (1984). “A fast Parallel Algorithms for Thinning Digital Patterns”, Research Contributions, Communications of the ACM. 27 (3): 236-239,

[12]. Gan, J., Chen, Y., Hu, B., Leng, J., Wang, W., & Gao, X. (2023). Characters as graphs: Interpretable handwritten Chinese character recognition via Pyramid Graph Transformer. Pattern Recognition, 137, 109317.

[13]. Abdalla, P. A., Qadir, A. M., Shakor, M. Y., Saeed, A. M., Jabar, A. T., Salam, A. A., & Amin, H. H. H. (2023). A vast dataset for Kurdish

handwritten digits and isolated characters recognition. Data in Brief, 47, 109014.

[14]. Hakro, D. N., Awan, S. A., Memon, M., AAMUR, A., & MOJAI, G. (2015). Interactive thinning for segmentation-based and segmentation-free Sindhi OCR. Sindh University Research Journal-SURJ (Science Series), 47(3).

[15]. Nabi, G., Shaikh, N. A., Rajper, R. A., & Shaikh, R. A. (2021). Thinning for segmentation-based and segmentation-free for Arabic script adopting languages. Sindh Univ. Res. J., 53(03), 271-274.

[16]. Wang, R., Cao, W., Wu, S., Jia, M., & Wang, X. (2023). Optical character correction of large-curvature annular sector text in polar coordinate system. Pattern Recognition Letters, 167, 157-163.

[17]. Hakro, D. N., & Talib, A. Z. (2016). Printed text image database for Sindhi OCR. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 15(4), 1-18.

Cite this article as :

Abdul Majid, Qinbo, Dil Nawaz Hakro, Saba Brahmani, "Thinning Chinese, Korean, Japanese and Thai script for segmentation-free OCRs", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 1, pp.116-121, January-February-2024. Available at doi : https://doi.org/10.32628/CSEIT2410111
Journal URL : https://ijsrcseit.com/CSEIT2410111