Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 53 (03) 271-274 (2021)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Thinning for Segmentation-based and Segmentation-free for Arabic script adopting languages

G. NABI, N. A. SHAIKH, R. A. RAJPER, R. A. SHAIKH

Department of Computer Science, Shah Abdul Latif University Khairpur

Received 17th May 2020 and Revised 26thMarch 2021

Abstract: Converting the text which is available in an image is not the same as the text in file as word processing is called Optical Character Recognition (OCR). The text image contains text which is not editable and recognizing by computer is called OCR process. Various approaches are used in preprocessing to prepare a text image for recognizing text available in an image. One of the approaches in preprocessing is thinning in which the available characters are thinned to one pixel of their strokes until the skeleton is found of the characters. The one-pixel skeleton is found of the characters of various languages so that the segmentation process of recognition is made easy. A thinning algorithm is presented in this paper for Arabic and its adopted languages such as Urdu, Persian, and Uyghur. The algorithm works in steps until the skeleton is found of a character. The thinning process always tries to keep all connected elements and the pattern of the character(s) intact so that the next phases of the OCR can be done easily. The custom-built application is an interactive process where the thinning process can be stopped and checked if it is the acceptable skeleton then the final image is produced. The interactive algorithm can be used with both types of OCR namely segmentation free and segmentation-based OCR. The thinning algorithm has been tested on various Arabic script and its adopting languages using multiple experiments. The other languages may be tested with our algorithms.

Keywords: OCR, Thining Segmentation

1. INTRODUCTION

Optical Character Recognition (OCR) is the process of understanding and recognizing text available in text images. Various phases of OCR such as preprocessing, segmentation, feature extraction and classification are followed (Hakro 2015). One of the processes is the thinning (Cowell and Fiaz 1992). in which the character or word imaged is thinned stepwise which means the extra dots are removed until a final one dot-based skeleton of the character or word image is found (Bag and Harit, 2011). Typically, the binary images are used for this process as the binary images are considered most suitable for OCR processing (Shang and Yi 2007).

2. Comparison with Hakro (2015)

The work in this paper is inspired from the work of (Hakro 2015) and the work has been extended in multiple aspects. Much of the existing work has been extended and modified to utilize and tested with Arabic and its adopting languages. The algorithm has been tested with single, multiple characters, words of various languages such as Urdu, Persian, Yugur and others. The application of the interactive algorithm has been tested with these languages and the results are presented.

3. MATERIAL AND METHODS

3.1 Custom built application

The algorithms have been implemented in MATLAB 2020 version to find out the step wise skeleton of the various script languages or the languages which adopt Arabic language. The algorithm or interactive program loads images of various Arabic and its adopted languages. These images are thinned or skeletonized so that the segmentation of the OCR of that language can be performed easily (Zang and Suen, 1984). The custom-built application has also been tested with words and full sentences. Various functions are called behind the scenes and these functions are integrated work as a single interactive application. When the image is loaded then the image can be edited at any stage, this allows user to fine tune any deficiencies which have been seen rare. The process of stepwise thinning of an Arabic script is shown in **(Fig. 1).** There are various step wise sections, and each section allows to edit the character or word to be edited if any missing structure or line is broken. The image can be finetuned and a pixel can be removed or added if needed. The iterative process can be stopped if the character skeleton is found.

++ Corresponding author email: rajperghulamnabi@gmail.com, noor.shaikh@salu.edu.pk, rahmatrajper@yahoo.com, riaz.shaikh@salu.edu.pk

The custom-built application is able to load, edit, thinning and saving of the image to the disk.

Fig. 1: Thinning steps For Arabic Language

Segmentation based and segmentation free OCR

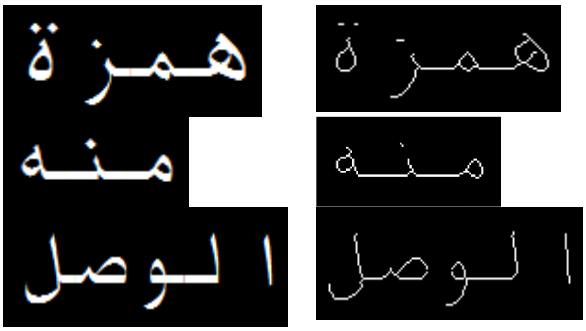
There are two types of OCR types of namely segmentation free and segmentation-based OCR (Fan and Verma, 2001). The segmentation based (Cavalin, 2006) is difficult and complex and the words of the script are broken into characters. The segmentation free (Premaratne and Bigun, 2004) OCR is typically do not segment words rather words are directly recognized. For this direct recognition, words are thinned to one pixel skeleton and then the features are extracted so that the words are recognized. The proposed algorithm can be applied for both types of the OCRs. According to [4], thinning can help in various types of analysis and the recognition process. A thinned image will help in recognition and other process steps as compared to the original as the skeleton has the only lines, strokes and necessary structure of the character or word. There are various benefits of thinning as specified by [].

4. Thinning algorithm:

The mask is created so that the original image of the character can be thinned. In this process the 2 pixel mask and 3 pixel mask has been created. The image is checked for white and black pixels. The lines, strokes and structure of the character is checked for connectivity and other features. The checking is done with the help of background. The sliding windows variant has been used to detected edges and background so that the skeleton can be found. The iteration process is repeated with various factors to check and then once the one pixel structure or stroke touching with the background is found then the pixel removing is stopped.

5. RESULTS AND DISCUSSION

The current study is the extension of the work [] in which the Sindhi language has been tested whereas the current algorithms with alterations and fine tunings are working with Arabic and all other Arabic script adopting languages such as Urdu, Persian, Pashto, Yugur. The algorithm will be extended to test on various languages of the world to test the suitability and versatility of the algorithm and approaches. The results of this study will help OCR research of various languages around the world. The tested images have been selected carefully in which isolated characters of these languages and ligatures (lehal and Rana, 2013) can be included in the testing images. **(Fig. 2).** to Figure 5 shows the various languages images in both forms without thinning and after thinning. The algorithm is working fine with Arabic language where the Arabic language which is cursive in nature has 28 characters and these character and word images are shown while the process of thinning. Fig. 2 shows the results of thinning process of Arabic language images.



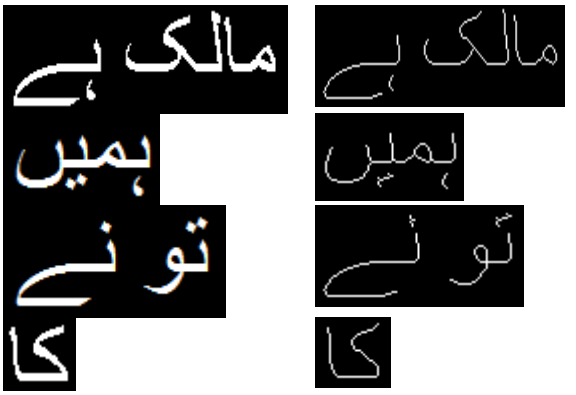
(a) Original Image (b) Image after Thinning

Fig. 2: Thinning of Arabic Images: (a) Original Image (b) Image after Thinning

These images were synthesized images, and the images were a part of another study on OCR and then the images were tested on the custom-built application []. **(Fig. 2)** presented the images of Arabic images, and it is obvious that the algorithm successfully thinned the Arabic script images. The results have been shown in **(Fig. 2)**. The algorithm also works on Persian language words and characters and successfully thinned some of the images which are shown in **(Fig. 3) (Fig. 4**a) shows the original images of the Urdu language and the thinned imaged of Urdu language are shown in **(Fig. 4b**)

(a) Original Image (b) Image after Thinning

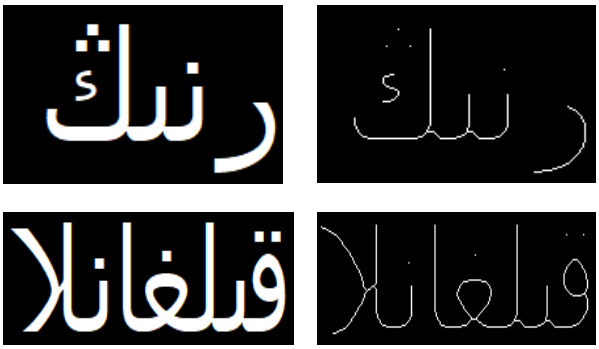
Fig.3: Thinning of Persian Images: (a) Original Image (b) Image after Thinning



(a) Original Image **(b.)** Image after Thinning

Fig. 4: Thinning of UrduImages: (a) Original Image (b) Image after Thinning

(Fig. 3) and **(Fig. 4)** illustrated the results of Persian and Urdu language respectively. The algorithm further applied to Uyghur scripts and the results are shown in Figure 5. The results can be observed that the algorithm worked fine for this language too. **(Fig. 5a)** shows the original Uyghur scripts words in original form whereas the **(Fig. 5b)** is the thinned images of Uyghur language.



(a) **Original Image** (b) **Image after Thinning**

Fig. 5: Thinning of UyghurImages: (a) Original Image (b) Image after Thinning

6. CONCLUSION

Thinning is one of the important process steps of OCR. The segmentation free OCRs are typically making use of thinning algorithms so that the feature extraction and the other steps. The interactive thinning algorithm has been tested with various languages and the algorithm worked fine and successfully thinned various images of multiple languages including Sindhi. The algorithm works fine for Arabic and its Adopting scripts namely Persian, Pashto, Urdu and Uyghur. The algorithm and its custom-built application can be used for thinning of words and sentences of Arabic and its adopting languages. The algorithm can be further tested on various languages of the world to test the accuracy and efficiency and versatility.

REFERENCES:

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

Bag, S. and G.Harit, (2011). An improved contour-based thinning method for character images. *Pattern Recognition Letters*, *32*(14), 1836-1842.

Cavalin, P. R., deSouza Jr., A. Britto, F. Bortolozzi, R. Sabourin, and L. E. S. Oliveira, (2006). An implicit segmentation-based method for recognition of handwritten strings ofcharacters, Proceedings of the 2006 ACM Symposium on Applied computing, SAC '06,ACM, Dijon, France, pp. 836–840.URL: http://doi.acm.org/10.1145/1141277.1141468

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

Fan, X. and B. Verma, (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, 251–255.

Hakro (2015), Enhanced Segmentation And Feature extraction For Sindhi Optical character Recognition, PhD thesis, Submitted to University science Malaysia (USM), Malaysia.

Lehal, G. S. and A. Rana, (2013). Recognition of Nastalique Urdu ligatures, Proceedings ofthe 4th International Workshop on Multilingual OCR, MOCR 13, ACM, Washington, DC,USA, 7:1–7:5.
URL: http://doi.acm.org/10.1145/2505377.2505379

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

Shang, L. and Z. Yi, (2007). "A class of binary images thinning using two PCNNs", Neurocomputing, Vol.: 70, 1096-1101,

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,

Hakro, D. N., S. A. Awan, M. Memon,, A. Aamur, and G. Mojai, (2015). Interactive thinning for segmentation-based and segmentation-free Sindhi OCR. *Sindh University Research Journal SURJ (Science Series)*, *47*(3).