Egyptian Informatics Journal (2016) 17, 189–197

Cairo University

Egyptian Informatics Journal

www.elsevier.com/locate/eij
www.sciencedirect.com

Text segmentation in degraded historical document images

A.S. Kavitha [a], P. Shivakumara [b,*], G.H. Kumar [a], Tong Lu [c]

[a] Department of Studies in Computer Science, University of Mysore, Karnataka, India
[b] Faculty of Computer Science and Information Technology, University Of Malaya, B-2-18, Malaysia
[c] National Key Lab for Novel Software Technology, Nanjing University, Nanjing, China

Received 2 March 2015; revised 1 October 2015; accepted 6 November 2015
Available online 2 January 2016

KEYWORDS

Text enhancement;
Sobel and Laplacian operations;
Indus document;
Clustering;
Text line segmentation

Abstract Text segmentation from degraded Historical Indus script images helps Optical Character Recognizer (OCR) to achieve good recognition rates for Hindus scripts; however, it is challenging due to complex background in such images. In this paper, we present a new method for segmenting text and non-text in Indus documents based on the fact that text components are less cursive compared to non-text ones. To achieve this, we propose a new combination of Sobel and Laplacian for enhancing degraded low contrast pixels. Then the proposed method generates skeletons for text components in enhanced images to reduce computational burdens, which in turn helps in studying component structures efficiently. We propose to study the cursiveness of components based on branch information to remove false text components. The proposed method introduces the nearest neighbor criterion for grouping components in the same line, which results in clusters. Furthermore, the proposed method classifies these clusters into text and non-text cluster based on characteristics of text components. We evaluate the proposed method on a large dataset containing varieties of images. The results are compared with the existing methods to show that the proposed method is effective in terms of recall and precision.

1. Introduction

India is a multilingual country, where all the states have provision to specify their own official language, which results in many official languages and various documents in different languages. Though work on segmentation of text is improved significantly, the recognition of old scripts like Indus is still difficult because of its complexity. Indus documents consist of symbols that look like ornamental in images [1]. Generally, these symbols are carved by hand on irregular surfaces such

* Corresponding author.
E-mail addresses: kavitha_sanjay_as@yahoo.co.in (A.S. Kavitha), hudempsk@yahoo.com (P. Shivakumara), ghk.2007@yahoo.com (G.H. Kumar), lutong@nju.edu.cn (T. Lu).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.eij.2015.11.003

Figure 1 Sample Indus document images.

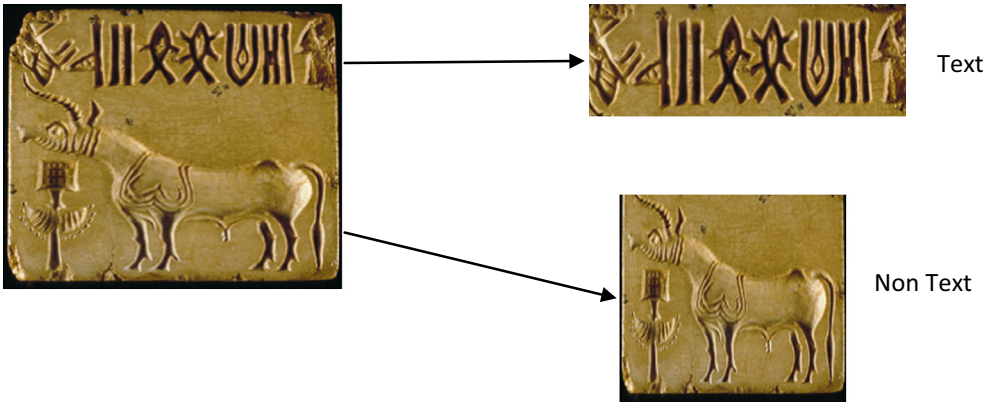

Figure 2 Illustrating text and non-text components in Indus documents.

as stones during the period of 3000 BC–1500 BC. As a result, Indus script is found in seal form that was used by people for the purpose of communication in the past. Fig. 1 shows a few images of these documents, where texts are associated with animal-like pictures in various forms such as a single horn and two horns. This complexity makes the segmentation problem more challenging and interesting. Due to the huge collection of such documents and the lack of scholars in the field of epigraphy, it is difficult to interpret all the scripts manually as it consumes a large amount of time. In order to reduce manual efforts, there is a need for the digitization of the scripts to preserve vital information for future study. Developing an automatic algorithm for converting raw script data to digital data involves four steps, namely text line segmentation, word segmentation, character segmentation, and character recognition. Text line segmentation is an important step as it facilitates other steps to achieve good recognition rates. In addition, text line segmentation is hard for the document like Indus due to the irregular structures of text components and unpredictable background variations [2,3]. Therefore, in this work, we focus on text line segmentation from Indus scripts. We can see some of the efforts toward text line segmentation [18–21] in the literature. Most of the methods are developed based on geometrical features such as aspect ratio and size for text line segmentation. Therefore, these methods may not be suitable for text line segmentation from Indus document images, where one cannot expect uniform size and structure due to complex background. Hence, we can conclude that there is an immense scope for developing a new method for segmenting text lines from Indus document images.

The paper is structured as follows. In Section 2, we give a brief survey of related work. Section 3 discusses the proposed work in detail. Finally, Section 4 discusses experimental results for the proposed method and the comparisons with the existing methods.

2. Previous work

There are several methods proposed for text extraction from scanned, handwritten, degraded and historical document images in the literature [4–10]. Most of the methods require plain and homogeneous background with a high contrast images for achieving good segmentation results. However, when we look at Indus documents as shown in Fig. 1, we cannot assume that such documents have plain backgrounds and structured text lines because these documents are handwritten with different tools on different surfaces. We consider Indus documents as a type of degraded historical document images, and text line segmentation from these documents still remains an unsolved problem. In this section, we will review the