Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 73 (2015) 115 – 121

Procedia Computer Science

The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)

Segmentation of Arabic handwritten text to lines

MOKHTARI Younes[a], Yousfi Abdellah[b]

[a]*Equipe TSE, ENSIAS, Université Mohammed V, Rabat- Maroc*
[b]*Equipe ERADIASS, FSJES, Université Mohammed V, Rabat- Maroc*

Abstract

Automatic recognition of writing is among the most important axes in the NLP (Natural language processing). Several entities of different areas demonstrated the need in recognition of handwritten Arabic characters; particularly banks check processing, post office for the automation of mail sorting, the insurance for the treatment of forms and many other industries. One of the most important operations in a handwriting recognition system is segmentation. Segmentation of handwritten text is a necessary step in the development of a system of automatic writing recognition. Its goal is to try to extract all areas of the lines of the text, and this operation is made difficult in the case of handwriting, by the presence of irregular gaps or overlap between lines and fluctuations of the guidance of scripture to the horizontal. In this paper, we have developed three approaches of handwritten Arabic text segmentation, then we compared between these three approaches.

Keywords—Image processing; Handwritten; Segmentation; histogram; Windows slippery.

1. Introduction

The handwriting recognition system is a technology that allows us to convert different types of documents (such as scanned paper documents, digital photos, etc.) into modifiable and exploitable formats.
In literature, there are two types of handwriting recognition:

- Static recognition also called offline recognition treats handwritten texts, in this context it is impossible to know how the different patterns were drawn.
- Dynamic recognition also called online recognition; it's kind of monitoring the writing, the lying and the raising of the electronic pen on a surface or a tab. 5

Peer-review under responsibility of organizing committee of the International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)
doi:10.1016/j.procs.2015.12.056

Handwriting recognition is generally done in five steps: pre-treatment, segmentation, learning, recognition and correcting recognition errors. In this paper, we interest of segmentation steps.

In literature, several approaches have been proposed in the field of handwriting segmentation. For example:

- BENSLIMANE and ZAROUNI used the partial differential equations (PDE) in order to detect the contours in the old documents. This method has been tested on more 100 manuscripts of different structures, mono, multi- oriented and multiscripts of different qualities .this approach can detect 2092 lines, with a detection rate of 83.2%. [3]
- ZAHOUR, TACONET and RAMDANE trimmed the Arabic manuscript text into 8 columns and horizontally projected each column. This method conducted over a dozen texts shows early encouraging results.
- BENNASRI and ZAHOUR proposed a method to extract the lines of Arabic manuscript text without any constraint for the writer. After detecting the starting points of all lines by a partial projection, he proceeds into monitoring the partial outline of each line, this method gives a rate of 98% 1. The drawback of these approaches is that the execution time is very long.

2. Segmentation

Segmentation is one of the most important steps in any handwriting recognition system; there exist two types of segmentations:

- Global segmentation: It allows us to isolate words of the documents and reference them by their position in the text. This type requires wide lexis so that the recognition could function properly. [5]
- Local segmentation, it exceeds the limits of the global approach, its goal is to make a segmentation into characters or graphemes, and this type gives better recognition rate, but the rate of implementation remains larger. [7]

In literature, there exist several segmentation techniques:

- segmentation by the contour, it rests only on filters and some thresholding techniques for the detection of objects and some homogenous regions, the weak point of this technique is the execution time that is very large.
- Segmentation from the histograms , it remains the most used approach in the recognition systems due to its high speed of retrieving results , but it is sensitive to the problem of overlapping between adjacent lines
- Sliding windows: wiping is the principle of this technique; it is based on the division of the image and the choice of the window size in order to get a good result.[4]

The problem of Arabic handwriting segmentation is still difficult, and this is due to the problem difficulty of overlapping between the lines, derived from the inclination of the writing, and erroneous positions of points and diacritical marks above and below the characters.[8]

In this article, we propose three approaches to carry out the segmentation of Arabic manuscripts to lines, in order to solve these types of problems and to reduce the segmentation execution time. These approaches are: