Chapter 23
A Metaheuristic Algorithm for OCR Baseline Detection of Arabic Languages

F. Daneshfar
University of Kurdistan, Iran

W. Fathy
University of Kurdistan, Iran

B. Alaqeband
University of Kurdistan, Iran

ABSTRACT

Preprocessing is a very important part of cursive languages Optical Character Recognition (OCR) systems. Thus, baseline detection, which is one of the main parts of the preprocessing operation, plays a basic role on OCR systems; improvement on baseline detection could be absolutely useful for decreasing errors in recognition words. In this chapter, a metaheuristic- and mathematical-based algorithm is recommended, which has improved the baseline detection process in relation to the well-known baseline detection algorithms. The most important advantages of the proposed method are simplicity, high speed processing, and reliability. To test this novel solution, IFN/ENIT database, which is a well-known and attending database, is utilized. However, the proposed solution is reliable to any standard database of cursive language's OCR.

INTRODUCTION

There isn't an exact definition for the baseline concept of handwritten texts, however in our mind the baseline is a supposed line which in cursive languages (here Arabic), passes into the most words on a line. The baseline may serve for several usages such as: elimination or normalization skews (Al-Shatnawi & Omar, 2008; Pechwitz & Maegner, 2003), segmentation scripts into words or letters (Al-Shatnawi & Omar, 2008; Amin, 1998; Arica & Yarman-Vural, 2002) and to extract dependent features (Al-Shatnawi

DOI: 10.4018/978-1-4666-7258-1.ch023

& Omar, 2008; El-Hajj et al., 2005). In Optical Character Recognition (OCR) systems, preprocessing is one of the most important parts of the system (Al-Rashaideh, 2006; Al-Shatnawi & Omar, 2008; Al-Shatnawi & Omar, 2009a; Farooq et al., 2005; Latfi et al., 2006) and the baseline detection, is a basic and necessary division of the preprocessing, too. Therefore, baseline detection is a very influential task for OCR systems, and if it does not work efficiently, it's impossible to get an acceptable result. In other word it has a straight effect on accuracy and credibility of character recognition (Al-Shatnawi & Omar, 2008).

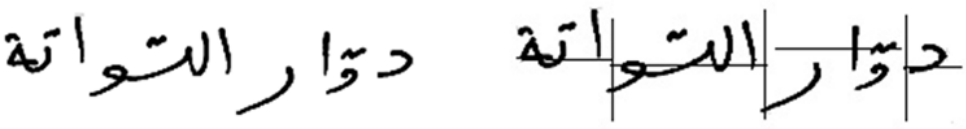
Generally, the aim of the current effort is to design an accurate, efficient and also simple baseline detection method for Arabic handwritten and typed texts, as by now there isn't any perfect and reliable baseline detection technique yet.

However there are many difficulties and problems to design an accurate baseline detection method for cursive languages' OCR systems. One of the most important related problems is that there are more letters with a non-geometric shape, so the places in handwritten texts are not totally evident toward the text baseline. Second or another problem is related to the sub-words. Each sub-word even in one word maybe have an own distinctive baseline. For example as it is shown in Figure 1, a given word with four sub-words could have four different baselines (AlKhateeb et al., 2011; Al-Shatnawi & Omar, 2008).

Until today several methods have been used for baseline detection in Arabic OCR like: methods based on Horizontal Projection (Parhami & Taraghi, 1981; Timsari & Fahimi, 1996; Olivier et al., 1996; Nawaz et al., 2003; Sarfraz et al., 2003; El-Hajj et al., 2005; Al-Rashaideh, 2006; AlKhateeb et al., 2008; AlKhateeb et al., 2011), The Word Skeleton (Pechwitz & Maergner, 2002), Word Contour Representation (Farooq et al., 2005), The Principal Components Analysis (Burrow, 2004), Sub-Words Treatment (Boukerma & Farrah, 2010) and Voronoi Diagram (Al-Shatnawi & Omar, 2009b). Most of the above techniques are based on mathematical (often statistical), geometric and visual or intuition methods; However in this chapter, the recommended solution is a metaheuristic and mathematical based algorithm. It uses the histogram diagram with only one pixel width horizontal lines which number of black pixels on the horizontal lines, are the primary materials of the proposed strategy (actually this is the main of the Horizontal Projection method).

Here the IFN/ENIT database for Arabic words has been chosen as the test bed, which is used by the most Arabic OCR projects yet (El-Hajj et al., 2005; Al-Rashaideh, 2006; AlKhateeb et al., 2008; Mozaffari et al. 2008; AlKhateeb et al., 2011), specifically the baseline detection projects. This database is a standard source including specified properties for each word (AlKhateeb et al., 2011). These properties are the most important constants in the current metaheuristic algorithm and will be added manually to the database as complementary information. Other sources are images of handwritten words that after enhancing and specifying essential properties will be recorded in the IFN/ENIT too. Then, each arbitrary word has the ability to add to the IFN/ENIT database and therefore can also examine with the current method.

Figure 1. (a) A sample word with four sub-words; (b) the word with four separated baselines



(a) (b)