Tifinaghe Document Converter

Mehdi Boutaounte, Informations Processing and Telecommunication Teams, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni-Mellal, Morocco

Driss Naji, Informations Processing and Telecommunication Teams, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni-Mellal, Morocco

M. Fakir, Processing and Telecommunication Teams, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni-Mellal, Morocco

B. Bouikhalene, Processing and Telecommunication Teams, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni-Mellal, Morocco

A. Merbouha, Processing and Telecommunication Teams, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni-Mellal, Morocco

ABSTRACT

Recognition of documents has become a basic necessity for two reasons: first to secure the existing data in paper because of the limited of their lives duration and the high rate of destruction insects, fire or humidity secondly to reduce space of archives. The aim of this work is to realize a converter that detects images and text within a document image taken by a scanner and applying a system for the recognition of characters (OCR) in order to obtain a web page (HTML extension) ready to be used in the same computer or on the web hosts to be accessible by everyone.

Keywords: *Converter, Document Decomposition, HTML, Neural Network, Optical Character Recognition (OCR)*

1. INTRODUCTION

The problem in the creation of a converter from image to document can be divided into two parts: first the Optical character recognition specially for Tifinagh characters in which we found some works using Neural networks (R.EL Ayachi et al., 2011) or other methods as Horizontal and Vertical Centerline of Character (Y.Es Saady et al., 2011)…etc. Second part the analyzing of document layout the physical structure (K. Hadjar et al., 2004) in the literature methods can be classed into two categories the top-down methods and bottom-up methods (S. N. Srihari et al., 1986)

Most work are reserved to the converter image files to doc or PDF extension which poses a large problem in conserving the original structure of the document (the positioning of images and text blocks inside) also the work reserve the transformation to an HTML page, are in the majority of these studies don't support the pictures. For those who do not ignores the images, they cannot recognize a Tifinagh letters.

The first converter transform a document image into an HTML page, but it acts in text

DOI: 10.4018/ijcvip.2013070104

blocks not in non-text blocks and Tifinagh characters did not taking into account. The second type of conversion software, whether they directly integrate the image in an HTML page or generate a sequence of character with different colors that looks like the original image and the last type of converter, these converters are not free of charge and give good results in terms of conservation and conservation of documents structure, but they also not support Tifinagh characters

To keep up with the evolution of technology in our lives and in order to create intelligent systems which spread our needs we try to describe in this paper a system, that convert a document image taken with a scanner into a HTML page ready to be used in a web site. Figure 1 illustrate the flow-chart of the converter Tifinagh document developed, that start by applying a preprocessing for the acquire, then segment the image and save the coordinates of each area. This coordinates will be used after stage of areas classification into text and non-text, and applying a OCR system on the text regions to create the structure of the page.

This paper is organized as follows: the first section describes the method used to analyze the physical structure of the document, in order to extract homogeneous components from the original image (text, title, image…etc) which will be used in the next section. In the second section we classify the components into text and non-text (images, graphic…etc), the text will be undergo into next processing, segmentation and recognition of characters using the neural network. The last section is reserved for the creation of the HTML page code.

2. PREPROCESSING

The acquired image is always accompanied by parasites: noise, tilt ... etc. Preprocessing applied in this study includes in this section is described as follows:

Binarisation is an operation that produces two classes of pixels represented by black pixels and white pixels. The method selected is the one adopted by "OTSU" (N. Thi Oanh et al., 2004) based on the calculation of an automatic threshold by calculating the histogram given by Equation (1).

$$h(i) = \frac{n_i}{\sum ni} \tag{1}$$

Where n_i represents the number of pixels of level i in the image.

Let q represent the estimate of class probabilities defined as:

$$q(k) = \sum_{i=1}^{k} P(k) \tag{2}$$

The separation takes place from the mean and variance given respectively by Equation (3) and (4).

$$mean = \sum_{i=1}^{k} i * h(i) \tag{3}$$

Equation 4 represents the individual class variances defined as

$$var = \sum_{i=1}^{k} (i - mean)^2 \frac{p(i)}{q(i)} \tag{4}$$

The problem of minimizing within class variance can be expressed as a maximization problem of the between class variance. It can be written as a difference of total variance and within class variance:

For each value of k=1….255, we calculate the square of S given by Equation (5).

$$s^2(k) = var_T{}^2 - var^2(k) = q(k) * (1 - q(k)) * (mean_T - mean(k))^2 \tag{5}$$

Classification of Multiple Interleaved Human Brain Tasks in Functional Magnetic Resonance Imaging

Manel Martínez-Ramón, Vladimir Koltchinskii, Gregory L. Heileman and Stefan Posse (2007). *Kernel Methods in Bioengineering, Signal and Image Processing (pp. 123-148).*

www.igi-global.com/chapter/classification-multiple-interleaved-human-brain/24821?camid=4v1a

Performance Analysis of Mobile Ad-Hoc Network Protocols Against Black Hole Attacks

Samy S. A. Ghoniemy (2013). *International Journal of Computer Vision and Image Processing (pp. 54-66).*

www.igi-global.com/article/performance-analysis-of-mobile-ad-hoc-network-protocols-against-black-hole-attacks/87251?camid=4v1a