Available online at www.sciencedirect.com
ScienceDirect
Applied Soft Computing 8 (2008) 118–126
Applied Soft Computing
www.elsevier.com/locate/asoc

Document page segmentation using neuro-fuzzy approach

Laura Caponetti [a], Ciro Castiello [a,*], Przemysław Górecki [b]

[a] *Università degli Studi di Bari, Dipartimento di Informatica, Via E. Orabona 4, 70126 Bari, Italy*
[b] *Wydział Matematyki i Informatyki, Uniwersytet Warmińsko-Mazurski ul. Oczapowskiego 2, 10-719 Olsztyn, Poland*

Received 16 December 2005; received in revised form 2 November 2006; accepted 22 November 2006
Available online 12 January 2007

Abstract

In this work, we propose a new document page segmentation method, capable of differentiating between text, graphics and background, using a neuro-fuzzy methodology. Our approach is based firstly on the analysis of a set of features extracted from the image, available at different resolution levels. An initial segmentation is obtained by classifying the pixels into coherent regions, which are successively refined by the analysis of their shape. The core of our approach relies on a neuro-fuzzy methodology, for performing the classification processes. The proposed strategy is capable of describing the physical structure of a page in an accurate way and proved to be robust against noise and page skew. Additionally, the knowledge-based neuro-fuzzy methodology allows us to understand the classification mechanisms better, contrary to what happens when other kinds of knowledge-free methods are applied.

Keywords: Document analysis; Page segmentation; Neuro-fuzzy classification; Multiresolution; Feature extraction

1. Introduction

Document analysis and understanding are relevant techniques for the automatic processing of paper documents [17]. Extraction of the structure or the layout from a document is referred to as document analysis and mapping the layout structure into a logical structure is referred to as document understanding. These techniques allow the recognition of document contents and the simplification of a number of complex tasks, such as re-editing, storage, maintenance, retrieval and transmission. Document image analysis plays an important role in the field of document processing and provides techniques for partitioning a document into a hierarchy of physical components (pages, columns, paragraphs, words, tables, figures, halftones, etc.) [15]. In this context, segmentation of document pages into coherent regions containing information such as text, graphics, pictures or background, can be seen as a preliminary phase in the construction of the physical and logical layout. Different document image segmentation methods have been proposed in the literature, both for printed pages [12,22] and for hand written pages [21]. In our work, we shall refer to the first category of approaches which can be categorised into top-down and bottom-up techniques. In the first case, a kind of knowledge-driven approach is applied, based on the expected nature of the original document. The page is divided into blocks that are successively split into sub-blocks, in an iterative fashion, to obtain the final text, graphics or image segments [1,2,15,16,25]. The bottom-up approach, instead, relies on a kind of data-driven technique that refines the data by layered grouping operations. In practice, single pixels are gathered on the basis of a low level analysis, to constitute blocks that can be merged into successively larger blocks [9,11,13,18,24].

A number of heuristics may be devised to produce region segmentation and a classical approach is to employ generic rules, based on thresholding [6]. Recently, the adoption of fuzzy logic has been proposed as a complementary approach to existing methods [10,20]. In fact, the concept of a fuzzy set allows a gradual transition from membership to non-membership, providing the management of a greater degree of vagueness, thus overcoming the threshold-based dimension of classical approaches. A number of works demonstrate that the introduction of fuzzy techniques can be very successful in the area of document image processing [5,10].

In this paper, we propose a methodology that aims at segmenting a document image into coherent and homogenous regions containing text, graphics and backgrounds. Following a

* Corresponding author.
E-mail addresses: laura@di.uniba.it (L. Caponetti), castiello@di.uniba.it (C. Castiello), pgorecki@matman.uwm.edu.pl (P. Górecki).

1568-4946/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2006.11.008

bottom-up approach, the overall methodology is based on the execution of consecutive steps. Initially, a multi-scale processing is applied to the document image to classify each pixel as a part of text, graphics or background. Successively, merging procedures group the identified pixels into coherent regions using a set of classical morphological operators. To refine the classification results, an additional region level analysis is performed, based on shape regularity and a region skew angle. At the end of this phase, a final accurate classification of the document regions is obtained.

The proposed methodology is based on a neuro-fuzzy learning approach, applied during the pixel level analysis and successively during the region refinement step. This approach, defined by a two-step learning of a neuro-fuzzy network, allows us to automatically derive a base of fuzzy rules, given a set of samples in the input feature space. This is in contrast to previous fuzzy approaches, where rules were defined only by an expert [5].

The experimental results reported show that the proposed methodology is able to produce highly satisfactory classification accuracy, and proves to be robust against noise and incorrect page alignment that may occur during scanning or photographing of a document page. For the sake of comparison, an alternative classification method has been tested on the same dataset of document images, based on the employment of a support vector machine. The results obtained highlight how the accuracy produced by the proposed neuro-fuzzy approach is definitely comparable with that achieved by the SVM techniques [7], with the additional benefit of generating a comprehensible knowledge base which is helpful in understanding better the mechanisms behind the automatic classification processes.

The paper is organised as follows. Section 2 describes the classification process, performed by the neuro-fuzzy learning methodology. The initial pixel level classification and the successive merging process are described in Section 3, followed by the detailed presentation of the text region refinement step, in Section 4. Section 5 presents the results of the experimental sessions and in Section 6 the results obtained are discussed and compared with other different approaches. Section 7 concludes the article with some final remarks.

2. Neuro-fuzzy learning methodology

Our strategy, developed for performing document image segmentation, is articulated in successive steps, where a neuro-fuzzy methodology is involved at different levels of application. In this section, we describe the neuro-fuzzy approach underlying both pixel level analysis (which constitutes the initial stage of the overall strategy) and region level analysis (which represents a refinement step, applied over text regions for improving the final results).

The adopted neuro-fuzzy methodology aims at deriving information in the form of fuzzy rules, based on features extracted from the document image dataset. The obtained fuzzy rule base encodes comprehensible knowledge, which can be employed to solve our classification tasks. The general

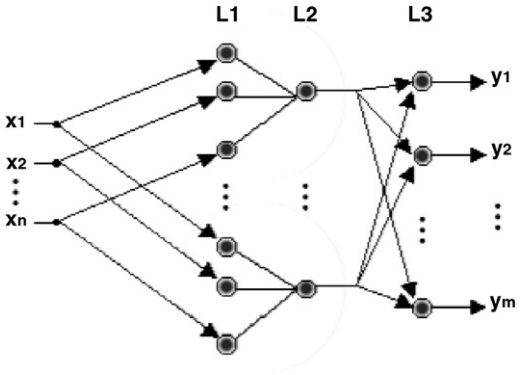

Fig. 1. The neuro-fuzzy network.

form of the kth rule in a knowledge base comprising K fuzzy rules is:

$$\begin{aligned} &\text{If } x_1 \in A_1^k \text{ And} \ldots \text{And } x_n \in A_n^k \\ &\text{Then } y \in C_h \text{ with degree } b_{hk}, \end{aligned} \tag{1}$$

for $h = 1, \ldots, m$, where $\{x_g\}_{g=1}^n$ are the input variables, A_g^k are fuzzy sets defined in terms of Gaussian membership functions, y represents the output variable, $\{C_h\}_{h=1}^m$ are the output classes and b_{hk} are fuzzy singletons indicating the degree to which the output belongs to a class. The fulfilment degree of each rule is given by the product operator (which stands for the particular T-norm for the AND connective). The output values are inferred as the weighted average of the activation strengths, with respect to the singletons.

The basic assumption of the neuro-fuzzy integration relies on the correspondence between the fuzzy inference engine (where the rules are involved) and a particular kind of three-layer feed-forward neural network, namely the neuro-fuzzy network (depicted in Fig. 1). Particularly, the connectionist component and the fuzzy inference system are associated in such a way as to share their respective free parameters. Moreover, the neuro-fuzzy network architecture corresponds to the inference mechanism of the fuzzy rule base. The first layer of the network evaluates the membership values for each input, according to:

$$\mu_{gk}(x_g) = \exp\left(-\frac{(x_g - c_{gk})^2}{\sigma_{gk}^2} \right),$$

where c_{gk}, σ_{gk} are the parameters (the centres and the widths) of the Gaussian functions. In the second layer, K units evaluate the fulfilment degree of each rule by means of the product operator, computing the activation strength of the kth rule by:

$$\mu_k(x_g) = \prod_{g=1}^{n} \mu_{gk}(x_g), \quad k = 1, \ldots, K.$$

The third layer supplies the final output of the system, according to the following average, weighted by the singleton parameters:

$$y_h = \frac{\sum_{k=1}^{K} \mu_k(x_g) b_{hk}}{\sum_{k=1}^{K} \mu_k(x_g)}, \quad h = 1, \ldots, m.$$