Expert Systems with Applications 42 (2015) 2593–2603

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Parameter tuning for document image binarization using a racing algorithm

Rafael G. Mesquita [a], Ricardo M.A. Silva [a], Carlos A.B. Mello [a,*], Péricles B.C. Miranda [a,b]

[a] *Centro de Informática, Universidade Federal de Pernambuco, Brazil*
[b] *Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco, Brazil*

ARTICLE INFO

Article history:
Available online 31 October 2014

Keywords:
Parameter tuning
Document image binarization
Racing algorithms

ABSTRACT

Binarization of images of old documents is considered a challenging task due to the wide diversity of degradation effects that can be found. To deal with this, many algorithms whose performance depends on an appropriate choice of their parameters have been proposed. In this work, it is investigated the application of a racing procedure based on a statistical approach, named I/F-Race, to suggest the parameters for two binarization algorithms reasoned (i) on the perception of objects by distance (POD) and (ii) on the POD combined with a Laplacian energy-based technique. Our experiments show that both algorithms had their performance statistically improved outperforming other recent binarization techniques. The second proposal presented herein ranked first in H-DIBCO (Handwritten Document Image Binarization Contest) 2014.

1. Introduction

Paper is still nowadays one of the most used medium to store and distribute information, even with the increase of the information technology that happened in the last century (Baird, 2003; de Mellao, de Oliveira, & dos Santos, 2012; Snellen & Harper, 2002). This is true due to some of its properties, like (i) the possibility to easily read and write simultaneously, (ii) the independence of a power source, (iii) its portability and (iv) its low cost (Baird, 2003; Snellen & Harper, 2002). It is possible to find a large amount of documents (in format of paper) of high cultural or historical value in libraries, museums or government archives. However, in some situations, the use of paper to store information is not recommended as (i) it can suffer degradation due to man handling or by aging and (ii) it is hard to perform a keyword search. Nowadays, for preservation purposes, most of the document are being digitized which does not solve none of the problems previously mentioned. Usually, after digitization of the document, its image is converted into black and white in a process called Binarization or Thresholding (Gonzalez & Woods, 2010).

Binarization is usually performed for two main reasons: the first one is that it can reduce the space needed to store an image, which is particularly important when dealing with large data sets. The second one is that Optical Character Recognition (OCR) algorithms usually require binary images to proceed with the recognition of the characters. Binarization is an important step in the document image analysis pipeline (that usually includes digitization, binarization Sezgin & Sankur, 2004, skew correction Mascaro, Cavalcanti, & A.B. Mello, 2010, text-line, word Sanchez, Mello, Suarez, & Lopes, 2011 and character segmentation Lacerda & Mello, 2013 followed by character recognition Cheriet, Kharma, Liu, & Suen, 2007; de Mellao et al., 2012) since its result affects further stages of the recognition. Thus, an unsuccessful binarization can also make it impossible to recognize characters, even for human beings (see Fig. 1(b)). Nevertheless, binarization of document images is considered a challenging task, especially in the case of old documents, because in this kind of images it is possible to find different issues, like uneven illumination, faded ink, smudges and smears, bleed-through interference and shadows (Mello, 2010b; Mesquita, Mello, & Almeida, 2014; Ntirogiannis, Gatos, & Pratikakis, 2013). Fig. 1 presents some unsatisfactory results obtained by three binarization algorithms.

Many algorithms that aim to solve computationally complex problems, including binarization, have a number of parameters that need to be properly configured in order to achieve satisfactory results (Birattari, Yuan, Balaprakash, & Stutzle, 2010). Those

* Corresponding author at: Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE), Av. Jornalista Anibal Fernandes, Cidade Universitária, 50740-560 Recife, PE, Brazil. Tel.: +55 8121268430; fax: +55 81 21268438.

E-mail addresses: rgm@cin.ufpe.br (R.G. Mesquita), rmas@cin.ufpe.br (R.M.A. Silva), cabm@cin.ufpe.br (C.A.B. Mello), pbcm@cin.ufpe.br (P.B.C. Miranda).

URL: http://www.cin.ufpe.br/~viisar (C.A.B. Mello).

http://dx.doi.org/10.1016/j.eswa.2014.10.039
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

parameters are typically configured based on some executions with different candidate configurations chosen based on some personal experience. Usually this is a tedious and time consuming procedure that may not lead to satisfactory results. In the specific case of binarization algorithms we can see the importance of selecting a good parametric configuration in Fig. 1(e)–(g), in which the results of the algorithm proposed in Mesquita et al. (2014) (refereed as POD_KO) with different parameter values are shown. We can also cite (Lin, Choy, Ho, & Ng, 2014) as another example, in which a model to design a green transportation scheme based on Genetic Algorithm (GA) is proposed. Two possible values are considered for each one of the four parameters of the GA and for each one of 24 different test scenarios the parameters of the GA are chosen based on at least five executions with different parametric configurations. Then, the best configuration is selected as the final solution for each scenario. Unfortunately, this scheme does not guarantee a good exploration of the search space (as only two possible values are considered for each parameter) and a considerable amount of time can be spent with executions using low performance settings. The process of selecting parameters is commonly treated as an optimization (Lin, 2010) problem in which a search technique is used to find the adequate ones for a given algorithm. In Zhang, Chen, and He (2010), an ant colony optimization-based algorithm used to optimize the parameters of a Support Vector Machine is presented. Furthermore, techniques as particle swarm optimization (PSO) (Kennedy & Eberhart, 1995), Tabu Search (Wang, Hao, Glover, & Lü, 2014) and racing algorithms are other approaches commonly used for parameter selection by many authors. Among the most widely used techniques for tuning algorithms, it is highlighted the racing algorithm (Birattari et al., 2010), that aims to find a good parametric configuration from a given finite set of alternatives through a sequence of steps. When sufficient evidence is gathered that some candidate has lower performance than at least another one, such candidate is discarded and the procedure is iterated over the surviving ones. The elimination of less promising candidates accelerates the execution and allows a more reliable evaluation focused on the promising ones. Besides, other positive point of racing algorithms is that unlike certain meta-heuristics, as GA or PSO, which need to be tuned adequately to perform well, they are simple to be designed. Racing algorithms have been used for meta-heuristics tuning as it can be seen in Pellegrini (2005), Becker, Gottlieb, and Stützle (2006) and Birattari, Stutzle, Paquete, and Varrentrapp (2002), and to select parameters of learning algorithms (Maron & Moore, 1997; Maron, 1994). These applications have shown that algorithms can be tuned efficiently by racing procedures.

In this paper, it is introduced a document image binarization method that achieved very good results on four datasets containing images of historical documents affected by several kinds of degradations. Furthermore, it is investigated the application of a racing algorithm to tune the parameters of the proposed technique. The rest of this paper is organized as follows. In Section 2, some classical and recent binarization algorithms are reviewed. Section 3 reviews I/F-Race approach for parameter tuning. Section 4 presents the proposal of application of a racing algorithm to tune the parameters of a binarization technique, while in Section 5 the experiments performed are explained. Finally, Section 6 concludes the paper.

2. Document image binarization techniques

A survey of classical image thresholding algorithms is presented in Sezgin and Sankur (2004). In that work, the algorithms are categorized as (i) histogram shape-based, (ii) clustering-based, (iii) entropy-based, (iv) object attribute-based, (v) spatial or (vi) local methods. Thresholding algorithms can also be classified as being parameter dependent or independent. For example, Otsu's (1979) clustering method, that defines an optimal threshold by minimizing the weighted sum of within-class variances and maximizing the between-class scatter, can be classified as parameter independent. On the other hand, as parameter dependent algorithms, there are classical local methods presented in Niblack (1986) and Sauvola and Pietikainen (2000), in which the image is divided into regions of size $b \times b$ and the threshold for each region is evaluated according to the local mean and standard deviation. In addition to the window size b, another parameter utilized by Niblack (1986) and Sauvola and Pietikainen (2000) is a bias k used in the calculation of the threshold.

Furthermore, it can be summarized more recent algorithms, like the ones proposed in Howe (2012), Su, Lu, and Tan (2013) and Moghaddam, Moghaddam, and Cheriet (2013). The method introduced in Su, Lu, and Tan (2010) works by performing stroke width estimation, high contrast pixel detection and thresholding. It has two parameters that need to be properly set: (i) the minimum number of high contrast pixels that must be counted within a neighborhood window so that a given pixel can be classified as text or ink and (ii) the size of the respective neighborhood window. In Su et al. (2013), the proposal in Su et al. (2010) is extended by (i) combining local image contrast with the local image gradient, (ii) detecting text stroke edge pixel by combining Canny's algorithm with the image contrast binarized by Otsu's method and (iii) adding some post processing steps to achieve final binarization. In Moghaddam et al. (2013), an unsupervised Ensemble of Experts (EoE) is introduced. Its main idea consists in selecting a set of appropriate binarization methods (experts) and combining their results. So, for each input, all methods are executed to generate a bi-level resultant image and a confidentness value for each classified pixel. Then, based on the confidentness maps, the set of experts that provides the better performance on the original image is identified and their binary results are combined into the final binary image. It is worth noting that EoE is also applicable to an ensemble of instances of the same technique with different parametric configurations (each parametric configuration for the given method is considered as a different expert); this is especially interesting for comparison reasons with the method proposed in this paper. In addition to the algorithms referred herein, in Ntirogiannis, Gatos, and Pratikakis (2014), seven other binarization techniques, submitted to the Handwritten Document Image Binarization Contest 2014 (H-DIBCO 2014), are presented. The methods are based on a variety of concepts as: (i) stroke width and slant; (ii) function minimization and Canny edge detection; (iii) Fuzzy C-Means; (iv) Location Cluster Model; (v) phase analysis; and (vi) filtering and local statistics.

In the rest of this section, the approach presented in Howe (2012) is outlined. The reason for reviewing Howe's algorithm with more detail than the other recent binarization approaches is that it is used as part of our proposal and also used in a statistical comparison.

2.1. A Laplacian energy based algorithm for document binarization

The work in Howe (2011) proposes an algorithm that uses a graph cut implementation (maximum flow) to find the minimum energy solution of an objective function that combines the Laplacian operator (Gonzalez & Woods, 2010) and Canny edge detection (Canny, 1986). The objective function to be minimized is defined as