Pattern Recognition Letters 34 (2013) 1299–1306

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

An optimization for binarization methods by removing binary artifacts

Marte A. Ramírez-Ortegón [a,c,*], Volker Märgner [a], Erik Cuevas [c], Raúl Rojas [b,1]

[a] Institut für Nachrichtentechnik, Technische Universität Braunschweig, Schleinitzstrae 22, 38106 Braunschweig, Germany
[b] Institut für Informatik, Freie Universität Berlin, Takustr. 7, 14195 Berlin, Germany
[c] Departamento de Ciencias Computacionales, Universidad de Guadalajara, Av. Revolución 1500, Guadalajara, Jalisco, Mexico

ARTICLE INFO

Article history:
Received 1 October 2012
Available online 23 April 2013

Communicated by M. Couprie

Keywords:
Historical documents
Threshold
Denoising
Binarization
Minimum error rate
Bayes theory

ABSTRACT

In this article, we introduce a novel technique to remove binary artifacts. Given a gray-intensity image and its corresponding binary image, our method detects and remove connected components that are more likely to be background pixels. With this aim, our method constructs an auxiliary image by the minimum-error-rate threshold and, then, computes the ratio of intersection between the connected components of the original binary image and the connected components of the auxiliary image. Connected components with high ratio are considered true connected components while the rest are removed from the output. We tested our method in binarization methods for historical documents (handwritten and printed). Our results are favorable and indicate that our method can improve the outputs from diverse binarization methods. In particular, a high improvement was observed for printed documents. Our method is easy to implement, has a moderate computational cost, and has two parameters whose model interpretation allows an easy empirical selection.

1. Introduction

In the context of document analysis, the process of distinguishing pixels that constitute ink strokes (foreground pixels) from the rest (background pixels) is known as *binarization*.

Binarization is a crucial task for document analysis methods since such methods rely on features extracted from foreground pixels. Hence, an inaccurate binarization tends to systematically propagate noise through the whole system. Some methods where binarization is crucial are methods for line detection (Louloudis et al., 2008), optical character recognition (Lázaro et al., 2010), text segmentation (Nikolaou et al., 2010), thinning (Bag and Harit, 2011), type of text classification (Peng et al., 2012), and writer identification (Brink et al., 2012).

During the past three decades, binarization methods for documents have been intensively researched; see surveys of binarization methods in Sahoo et al. (1988), Trier and Jain (1995), Sezgin and Sankur (2004) and Stathis et al. (2008). This is motivated in part because digital documents make easier the accessing and searching of document contents; and in part because the definition of binarization depends on the specific application. Actually, a great deal of research groups continue developing binarization methods specialized in historical documents (Gatos et al., 2011; Pratikakis et al., 2010, 2011) due to the importance of preserving cultural heritage and the complexity of historical documents.

For historical documents, binarization is a challenging task because such documents frequently have non-standard printing styles, like diverse fonts, ornamental strokes, background printing patterns, and irregular stroke widths. In addition, historical documents may have several types and degrees of degradation due to aging and mistreat, such as bleed-through, ink stains, smudged characters, and outlines of paper folds.

Because of the complexity of historical documents, some binarization methods specialized on historical documents like in Moghaddam and Cheriet (2012), Ben Messaoud et al. (2011), Lu et al. (2010), Ntirogiannis et al. (2009), Gupta et al. (2007) and Gatos et al., 2006 compute a preliminary binary image and, subsequently, the pixels are toggled from one class to the other in order to minimize the misclassification rate. For the purposes of this article, we refer as *binarization core* to the initial process of classifying the pixels, and as *binary restoration* to the process of toggling pixels.

Unlike the methods of binarization core that transform images from gray/color intensities to binary values, methods of binary restoration have a binary image as an input and as an output. Hence, any binarization method can be followed by any method of binary restoration. However, the election of a suitable binary restoration mainly depends on three factors: the type of noise in which the binarization core fails, the expected noise generated by the improper parameter selection for the binarization core, and the a priori

* Corresponding author at: Institut für Nachrichtentechnik, Technische Universität Braunschweig, Schleinitzstrae 22, 38106 Braunschweig, Germany. Tel.: +49 (0) 531 3912483; fax: +49 (0) 531 3918218.

E-mail addresses: mars.sasha@gmail.com (M.A. Ramírez-Ortegón), maergner@ifn.ing.tu-bs.de (V. Märgner), erik.cuevas@cucei.udg.mx (E. Cuevas), rojas@inf.fu-berlin.de (R. Rojas).

[1] Tel.: +49 (0) 30 83875130.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.04.007

knowledge of the objects of interest (ink strokes). While the former two factors are related with false binary strokes (binary artifacts) and diminished binary strokes, the latter factor is related to the data adequacy for further process.

We define a *binary artifact* as a connected component such that all of its pixels have been wrongly classified. Under this definition, we have two types of binary artifacts: foreground binary artifacts that are constituted by background pixels, and background binary artifacts that are constituted by foreground pixels.

Removing either fore- or background binary artifacts are dual problems. So, we detail our method only for foreground binary artifacts and, from now on, we refer to foreground binary artifacts as binary artifacts.

Methods to remove binary artifacts are important for binarization methods. In particular, all the binarization methods in Ramírez-Ortegón et al. (2010a,b), Ramírez-Ortegón and Rojas (2010), Moghaddam and Cheriet (2012), Ben Messaoud et al. (2011), Lu et al. (2010), Ntirogiannis et al. (2009), Gupta et al. (2007) and Gatos et al., 2006 compute one or more techniques to remove binary artifacts caused by splotches, paper cracks and folds, specks, bleed-through, and background printing patterns that frequently appear in historical documents.

Small binary artifacts tend to be caused by splotches and specks. To deal with this kind of noise, authors like Lu et al. (2010) and Gupta et al. (2007) remove connected components with n or less pixels. Lu et al. assume $n = 3$, but Gupta et al. compute n from an equation that involves the space (in pixels) between lines. Other approaches to eliminate small binary artifacts are binary templates (Ramírez-Ortegón et al., 2010a,b) and shrink filters (Gatos et al., 2006, 2008).

Counting only the number of pixels is not enough to distinguish between strokes and binary artifacts when the binary artifacts are large. Unfortunately, large binary artifacts are common in historical documents due to bleed-through images, paper cracks, outline folds, and bi-level background. To overcome these problems, Lu's method calculates the average of certain feature for each connected component (Lu et al., 2010), if such an average is higher than a threshold, then the evaluated connected component is considered as binary artifacts.

Instead computing only the average of some feature as in Lu's method, Moghaddam's method compute a vector of features for each connected component of both fore- and background (Moghaddam and Cheriet, 2012), then the vectors are clustered and classified to determine which connected components are binary artifacts.

Although both Lu's and Moghaddam's restoration methods have been applied with positive results. Both methods have the disadvantage of being complex in their implementation; both methods estimate a background surface that considerably increases the computational load and increase the number of parameters to be set. Furthermore, their parameter selection is not simple and, as a consequence, the incorporation of such methods in binarization techniques is difficult.

We propose a method to remove binary artifacts (Section 2) whose implementation is simple, whose computational cost is linear to the number of pixels, and whose number of parameters is two. Moreover, both parameters have model interpretation so that they can be empirically adapted for diverse applications.

Our method computes an auxiliary image (Section 2.2) from the initial binary input and, subsequently, it determines whether or not a connected component should be removed based on the intersection between the input binary image and the auxiliary image.

We evaluated the performance of our method based on DIBCO 2011 benchmark in order to standardize our evaluation.

2. Our method

Our method is a technique that strongly depends of the accuracy of binary input. Then, it may introduce more noise if the binary input is considerably inaccurate. However, we will show that our method has satisfactory results in general if the initial binarization is good.

Strictly speaking, our method has a single parameter α, but it calculates an auxiliary binary image. How to compute the auxiliary image is a parameter in itself and it may involves more parameters. To compute the auxiliary image, we suggest the minimum-error-rate threshold which only has a parameter r which controls the radius of the pixel neighborhood. In this manner, our implementation has a total of two parameters.

2.1. Notation

For the purpose of this paper, pixels are denoted in bold, and the gray intensity of a pixel $\boldsymbol{p}$ is denoted as $I(\boldsymbol{p}) \in \mathbb{N}$, where black is set to zero, and the color white is set to g. The image of gray intensities is denoted by I.

We denote the fore- and background sets by $\mathcal{F}$ and $\mathcal{B}$, respectively, such that $\mathcal{P} = \mathcal{F} \cup \mathcal{B}$. The neighborhood of a pixel $\boldsymbol{p}$, denoted by $\mathcal{P}_r(\boldsymbol{p})$, are those pixels within a square centered at the pixel $\boldsymbol{p}$ of sides with length $2r + 1$. Moreover, given a set of pixels $\mathcal{A}$, we will write $\mathcal{A}_r(\boldsymbol{p})$ as shorthand for $\mathcal{A} \cap \mathcal{P}_r(\boldsymbol{p})$. For instance, $\mathcal{F}_r(\boldsymbol{p}) = \mathcal{F} \cap \mathcal{P}_r(\boldsymbol{p})$ and $\mathcal{B}_r(\boldsymbol{p}) = \mathcal{B} \cap \mathcal{P}_r(\boldsymbol{p})$. In addition, the cardinality of a set $\mathcal{A}$ is denoted by $|\mathcal{A}|$.

In the following sections we adopt the thresholding approach as:

$$B(p) = \begin{cases} 1 \text{ (foreground)} & \text{if } I(\boldsymbol{p}) \leqslant T(\boldsymbol{p}), \\ 0 \text{ (background)} & \text{otherwise}, \end{cases} \tag{1}$$

where B denotes the binary image of I, and $T(\boldsymbol{p})$ is the threshold calculated for $\boldsymbol{p}$.

2.2. Auxiliary image

We compute an auxiliary image based on Bayes theory: the minimum-error-rate threshold. We elected this threshold because it is more robust to gray-intensity outliers than thresholds based on mean and variances of gray intensities.

2.2.1. Minimum-error-rate threshold

In the context of binarization for historical documents, we assume that ink strokes are darker than the background. This empirical idea is then exploited by assuming that the gray-intensity distribution of foreground pixels is left shifted from the gray-intensity distribution of background pixels and that the overlapping between these two distributions is small.

According to Bayesian decision theory, the probability of a pixel misclassification is minimized by Bayes decision rule: Classify $\boldsymbol{p}$ as foreground if

$$\Pr(\boldsymbol{p} \in \mathcal{F}_r(\boldsymbol{p}) | I(\boldsymbol{p}) = i) \geqslant \Pr(\boldsymbol{p} \in \mathcal{B}_r(\boldsymbol{p}) | I(\boldsymbol{p}) = i). \tag{2}$$

Otherwise, classify $\boldsymbol{p}$ as background.

Following Bayes criteria, let $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$ be the estimated fore- and background, respectively, of our input binary image. Then, the minimum-error-rate threshold for a pixel $\boldsymbol{p}$ within the neighborhood of radius r is defined by

$$T(\boldsymbol{p}) = \underset{t=0,1,\ldots,g}{\arg\min} \left\{ \sum_{i=0}^{t} h_b(i) + \sum_{i=t+1}^{g} h_f(i) \right\}, s \tag{3}$$

where $h_f(i) = |\{\boldsymbol{p} \in \hat{\mathcal{F}}_r(\boldsymbol{p}) \,|\, I(\boldsymbol{p}) = i\}|$ and $h_b(i) = |\{\boldsymbol{p} \in \hat{\mathcal{B}}_r(\boldsymbol{p}) \,|\, I(\boldsymbol{p}) = i\}|$. Note that no any particular distribution is assumed in this equation.