Pattern Recognition 44 (2010) 491–502

Contents lists available at ScienceDirect

Pattern Recognition

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

Unsupervised measures for parameter selection of binarization algorithms

Marte A. Ramírez-Ortegón [a,*], Edgar A. Duéñez-Guzmán [b], Raúl Rojas [a], Erik Cuevas [c]

[a] Institut für Informatik, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany
[b] Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford St., Cambridge, MA 02138, USA
[c] Department of Computer Science, University of Guadalajara, Av. Revolución 1500, Guadalajara, Jalisco, Mexico

ARTICLE INFO

Article history:
Received 23 March 2010
Received in revised form
24 September 2010
Accepted 29 September 2010

Keywords:
Binarization
Image pre-processing
Unsupervised evaluation method

ABSTRACT

In this paper, we propose a mechanism for systematic comparison of the efficacy of unsupervised evaluation methods for parameter selection of binarization algorithms in optical character recognition (OCR). We also analyze these measures statistically and ascertain whether a measure is suitable or not to assess a binarization method. The comparison process is streamlined in several steps. Given an unsupervised measure and a binarization algorithm we: (i) find the best parameter combination for the algorithm in terms of the measure, (ii) use the best binarization of an image on an OCR, and (iii) evaluate the accuracy of the characters detected. We also propose a new unsupervised measure and a statistical test to compare measures based on an intuitive triad of possible results: better, worse or comparable performance. The comparison method and statistical tests can be easily generalized for new measures, binarization algorithms and even other accuracy-driven tasks in image processing. Finally, we perform an extensive comparison of several well known measures, binarization algorithms and OCRs, and use it to show the strengths of the WV measure.

1. Introduction

Libraries such as *the National Archives of Egypt* and *the Library of Congress* (United States of America) have been digitalizing historical printed documents like ancient codices, maps, and books to preserve and spread their cultural heritage through digital libraries.

While digitization in itself is enough to preserve the contents of documents, a primordial benefit of digitization is the extraction of information from the digitalized images, and the access to this information through digital libraries.

The main challenge with the construction of digital libraries lies in the extraction of information from hundreds of thousands of ancient documents. This problem can be roughly divided into three parts: detection of objects of interest (binarization), text extraction (through an OCR), and text layout recognition (as a post-process). In this paper, we will focus on binarization and on how it can be used to maximize the accuracy of the OCR.

Conceptually, images often have a natural partition in foreground and background. Intuitively, *binarization* or *segmentation* (these terms will be used interchangeably throughout this article) consists of estimating such a partition, where we consider as foreground the set of pixels in an image containing the objects of interest and the background representing the rest of the image.

For the purposes of this paper, we will consider the *optical character recognition* (OCR) step of the process as a *black box* algorithm. That is, given values to its parameters it receives an input, and produces an output, but the details of implementation as well as any additional contextual information on the type of input are not considered. Although the OCR accuracy on an image depends on the OCR parameters and the quality of the input's segmentation, we assume that the main factor influencing recognition accuracy is not the OCR parameters, but the quality of the binarization; the better the binarization, the better the OCR recognition.

Historical documents usually present several challenges and varied forms of degradation, such as non-standard fonts, ink stains, weak ink strokes and wide variations in the background to mention some. Because of this, the parameters of binarization algorithms have to be tuned for each kind of degradation. For a large sets of images, however, the manual tuning of parameters is time-consuming and costly, and the use of *general parameters* may lead to a low binarization performance. Hence, the choice of binarization algorithms and their parameters play the most important role in the accuracy of recognition [1,2].

To address the problem of parameter selection in segmentation, unsupervised evaluation methods have been proposed to assess the quality of a segmentation [3,4]. Such methods allow for evaluation of many algorithms over large parameter spaces and on diverse images without the need for human intervention.

* Corresponding author.
E-mail addresses: mars.sasha@gmail.com (M.A. Ramírez-Ortegón), duenez@oeb.harvard.edu (E.A. Duéñez-Guzmán), rojas@inf.fu-berlin.de (R. Rojas), erik.cuevas@cucei.udg.mx (E. Cuevas).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.018

Consequently, they enable an objective comparison of both different segmentation methods and the different parameters of a single method. Moreover, such a comparison can be used to automatize the choice of parameter for a binarization algorithm.

Although unsupervised measures can be transformed into binarization algorithms, such a transformation may introduce bias and may not be suitable for neighborhoods which are completely contained either in the foreground or in the background (*uniform neighborhoods*); see Section 4.2. However, they can tune parameters of another binarization algorithms which are capable to deal with uniform neighborhoods or whose assumptions fit the images assumptions too. Hence, the importance to analyze which unsupervised measures are suitable to tune the parameters of a particular binarization algorithm.

Evaluation measures based on the variance of gray intensities have been used to assess binarization performance [5–7]. Especially in document images, both foreground and background are intuitively thought of as uniform and homogeneous regions. Unfortunately, few authors have analyzed the mathematical and experimental behavior of these measures [8,3], hence our interest to address the interaction between binarization methods and these evaluation measures. This interaction is analyzed under our model of *simple images*, which are images where the contrast of gray intensities between foreground and background pixels is bounded in small neighborhoods. Ideal images provide the mathematical basis to prove whether the optimal value of each evaluation measure leads to the estimation of an accurate foreground.

Section 2 introduces notation and preliminary concepts. We review and contrast many main stream binarization algorithms in Section 3. In Section 4 we introduce the unsupervised evaluation measures that will be used throughout to fine-tune segmentation algorithms and state the main theoretical results. We consider local variations of the binarization algorithms discussed with the aim of overcoming wide variations in gray intensities of both foreground and background. The design of our experiments is described in Section 5 where we also propose a novel statistical test, called *uncertainty test*, for performance comparison between pairs of unsupervised measures. Finally, we close in Section 6 with experimental results and conclusions.

2. Preliminaries

2.1. Notation

A continuous image can be represented by a finite partition $\mathcal{P}$ of $[0,w]\times[0,h]$. A pixel $\boldsymbol{p}$ is defined as an element of this partition.

An image function F can be defined as a function from a set of pixels $\mathcal{P}$ to a set representing colors. In this manner, a pixel $\boldsymbol{p}\in\mathcal{P}$ refers to a region in the image space, while the value (color) of that pixel is $F(\boldsymbol{p})$. Formally, F is such that $F:\mathcal{P}\to\mathbb{Z}$. In this article we restrict our study to gray-scale images; $\mathcal{P}$ is a rectangular partition of size $n\times m$, and pixel under F is interpreted as an intensity in the set $\mathbb{Z}_{g+1}$ with g+1 intensities of gray, where 0 represents black, and g represents white. While color images represent the most general family of images, a color image can be transformed to gray intensities by means of a mapping $\gamma:\mathbb{Z}\to\mathbb{Z}_{g+1}$. For simplicity, we define I as the gray image $\gamma\circ F$. Notice that while I depends both on the image F and the gray-intensity map γ, this dependency will always be clear from context and thus, will be left implicit.

A binarization of image I is given by a function $B:\mathcal{P}\to\{0,1\}$. The foreground, estimated by B, is given by $\widehat{\mathcal{F}}=\{\boldsymbol{p}\in\mathcal{P}|B(\boldsymbol{p})=1\}$, while the estimated background is $\widehat{\mathcal{B}}=\mathcal{P}\backslash\widehat{\mathcal{F}}$.

Throughout this paper, we will use upper-case "calligraphic" letters to denote sets of pixels ($\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, etc.). In addition, we will use the following notation: $\widehat{\ }$ to refer to an estimator; $|\mathcal{A}|$ to denote the cardinality of set $\mathcal{A}$; $\mathcal{N}_r(\boldsymbol{p})\subset\mathcal{P}$ as the neighborhood with radius r containing the 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 short-hand for $\mathcal{A}\cap\mathcal{N}_r(\boldsymbol{p})$. For instance, $\widehat{\mathcal{B}}_r(\boldsymbol{p})=\widehat{\mathcal{B}}\cap\mathcal{N}_r(\boldsymbol{p})$ denotes the pixels in the estimated background within a neighborhood of radius r around pixel $\boldsymbol{p}$; $\Pr(\cdot)$ denotes the probability of an event; and $E(x)$ denotes the expected value of a random variable x while $Var(x)$ denotes its variance.

We summarize the distribution of the gray intensities in a set of pixels $\mathcal{A}$ in the form of a histogram $H_{\mathcal{A}}(i)=|\{\boldsymbol{p}\in\mathcal{A}|I(\boldsymbol{p})=i\}|$, which gives the frequency of a gray intensity i in set $\mathcal{A}$.

2.2. Image model

Authors like Sahoo et al. [5] and Sezgin and Sakur [6] conjecture that both foreground and background should be uniform and homogeneous regions. However, that conjecture is false for images with composite foreground and/or background like the one shown in Fig. 1. Hence, we characterize the behavior of gray intensities locally.

Definition 1. Given r, an image follows Model 1 if the gray intensities of foreground in all neighborhoods of radius r can be modeled as random variables which are approximately independent and identically distributed (two different neighborhoods may follow different distributions). Gray intensities in the background are modeled in a similar manner.

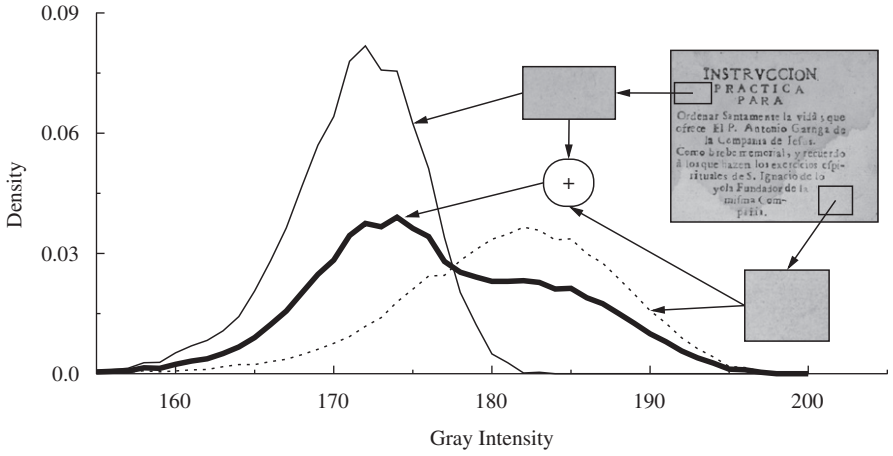

Fig. 1. Two different regions form the background.

From Model 1, the gray intensities of the fore- and background within a small neighborhood are considered as a sample of random variables that are independent and identically approximately (normally) distributed. In this model, the spatial position is irrelevant thought it may help to have a better model.

In our experience, historical documents fit Model 1 in a high percent of neighborhoods if their background has no patterns deliberately printed.

Kittler and Illingworth [9] pointed out that gray intensities of fore- and background are (approximately) normally distributed and proposed a threshold criterion based on it. The assumption of a priori distribution with Model 1 gives the mathematical basis to describe the behavior of unsupervised methods based on the variance of gray intensities. For our analysis, we assume that gray intensities of both foreground and background are approximately normally distributed. Therefore, the histogram of gray intensities is approximately given by

$$H_{\mathcal{P}_r(\boldsymbol{p})}(i) \approx |\mathcal{F}_r(\boldsymbol{p})| \cdot \phi(i; \mu_{\mathcal{F}_r(\boldsymbol{p})}, \sigma^2_{\mathcal{F}_r(\boldsymbol{p})}) + |\mathcal{B}_r(\boldsymbol{p})| \cdot \phi(i; \mu_{\mathcal{B}_r(\boldsymbol{p})}, \sigma^2_{\mathcal{B}_r(\boldsymbol{p})}) \quad (1)$$

where $\phi(x; \mu, \sigma^2)$ denotes the probability density function of the normal distribution with mean μ and variance σ^2; the notation $\mu_{\mathcal{A}}$ and $\sigma^2_{\mathcal{A}}$ refers to the mean and variance of gray intensities in $\mathcal{A}$, respectively.

In general, the probability that a pixel with a certain intensity belongs to the fore- or background depends on their distributions. This is clearly seen in (1), especially stressed by the pixels with intensities between $\mu_{\mathcal{F}_r(\boldsymbol{p})}$ and $\mu_{\mathcal{B}_r(\boldsymbol{p})}$. Thus, to minimize misclassification when using a threshold, it is better when these means are far apart and their variances are small, that is, when the contrast between fore- and background is large. This is illustrated in Fig. 2, where an image with good contrast is shown, and its histogram is compared with a hypothetical histogram that only differs in contrast (distance between the means). The shaded region in green represents the probability of misclassified pixels when using a threshold.

Given that contrast is crucial for an accurate segmentation, certain bounds are required for it. In our case, we formalize the requirement in the following definition, and will be used for the main results later on.

Definition 2. Assuming Model 1, an image is an *r-simple image* if all neighborhoods with radius r such that $|\mathcal{F}_r(\boldsymbol{p})| > 1$ and $|\mathcal{B}_r(\boldsymbol{p})| > 1$ satisfy the inequality:

$$\|\mu_{\mathcal{B}_r(\boldsymbol{p})} - \mu_{\mathcal{F}_r(\boldsymbol{p})}\| > \sqrt{2} \cdot \max(\sigma_{\mathcal{B}_r(\boldsymbol{p})}, \sigma_{\mathcal{F}_r(\boldsymbol{p})}) \quad (2)$$

where $\|\cdot\|$ denotes the absolute value.

In this paper, we consider that the gray intensities of the foreground are darker than those in the background. That is, $\mu_{\mathcal{B}_r(\boldsymbol{p})} > \mu_{\mathcal{F}_r(\boldsymbol{p})}$.

Research indicates that gray intensities around the boundary between the foreground and background are lognormally rather normally distributed [7,10]. Therefore, under Model 1, we can consider the distribution to be lognormal:

$$H_{\mathcal{P}_r(\boldsymbol{p})}(i) \approx |\mathcal{F}_r(\boldsymbol{p})| \cdot \lambda(i; \tilde{\mu}_{\mathcal{F}_r(\boldsymbol{p})}, \tilde{\sigma}^2_{\mathcal{F}_r(\boldsymbol{p})}) + |\mathcal{B}_r(\boldsymbol{p})| \cdot \lambda(i; \tilde{\mu}_{\mathcal{B}_r(\boldsymbol{p})}, \tilde{\sigma}^2_{\mathcal{B}_r(\boldsymbol{p})}) \quad (3)$$

where $\lambda(i; \tilde{\mu}, \tilde{\sigma}^2)$ denotes the probability density function of the lognormal distribution with parameters $\tilde{\mu}$ and $\tilde{\sigma}^2$; the notation $\tilde{\mu}_{\mathcal{A}}$ and $\tilde{\sigma}^2_{\mathcal{A}}$ refers to the mean and variance of the logarithm of gray intensities in $\mathcal{A}$, respectively.

3. Binarization algorithms

We can identify three categories of binarization algorithms [6]: *global algorithms* classifies a pixel using information from the whole image, *local algorithms* rely on information from the pixel neighborhood, and *hybrid algorithms* combine information from the whole image and pixel neighborhood.

Local binarization algorithms compute a threshold surface $T : \{\mathcal{P}_r(\boldsymbol{p}) | \boldsymbol{p} \in \mathcal{P}\} \to \mathbb{Z}_{g+1}$ over the whole image setting $B(\boldsymbol{p}) = 1$ if $I(\boldsymbol{p}) \leq T(\boldsymbol{p})$, where $T(\boldsymbol{p})$ is computed gathering information from $\mathcal{P}_r(\boldsymbol{p})$; otherwise $B(\boldsymbol{p}) = 0$. Note that global and hybrid algorithms can be implemented as local methods either restricting its global analysis to the pixel neighborhood (global methods) or computing the "global information" from a secondary local neighborhood (hybrid methods).

We especially study two kinds of binarization methods: histogram cluster methods and statistical methods. The former rely on information from the histogram of gray intensities, the latter rely on information from statistics of the gray intensities, like the mean, variance, third moment, maximum and minimum.

3.1. Histogram cluster methods

Histogram cluster methods assume that the foreground and background can be estimated by

$$\widehat{\mathcal{F}} = \{\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p}) | I(\boldsymbol{q}) \leq t_{opt}\} \quad \text{and} \quad \widehat{\mathcal{B}} = \{\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p}) | I(\boldsymbol{q}) > t_{opt}\} \quad (4)$$

respectively, where $t_{opt} \in [0,g]$ which satisfies the method's criterion optimally. Examples of methods to obtain t_{opt} include using entropy functions and mixture of two distributions, curvature analysis, and many more.

In images with composite background, t_{opt} may not exist such that $\widehat{\mathcal{F}}$ and $\widehat{\mathcal{B}}$ approximate $\mathcal{F}$ and $\mathcal{B}$ accurately. Therefore, its applicability may be restricted to neighborhoods where the method's assumptions are satisfied. However, computing $T(\boldsymbol{p})$ with the local version of a histogram cluster method will systematically produce false positives due to uniform neighborhoods. To solve this problem, several techniques have been proposed by binarization

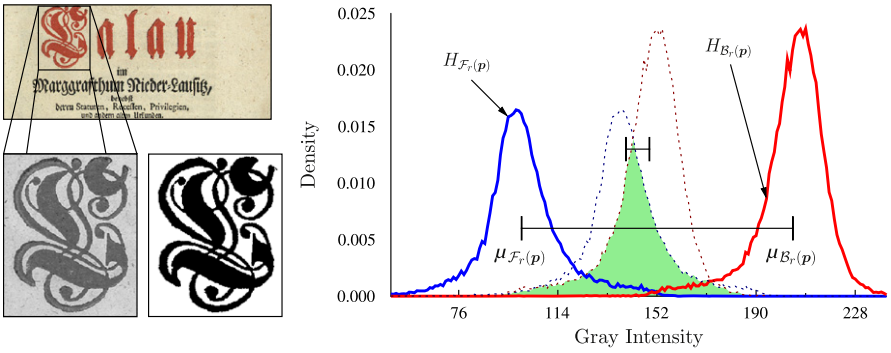

Fig. 2. Solid lines, example of "good" contrast in a neighborhood. Dashed lines, a hypothetical example of "bad" contrast.

researchers like in [11,1,12,13]. Although the analysis of these techniques is beyond the scope of this paper, we use a simple restriction that may help all these binarization methods without favoring a particular method:

$$T(\boldsymbol{p}) = \begin{cases} t_{opt} & \text{if } \widehat{\mu}_{\widehat{\mathcal{B}}} - \widehat{\mu}_{\widehat{\mathcal{F}}} < c \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where t_{opt} is the optimal method's threshold restricting the global analysis to $\mathcal{P}_r(\boldsymbol{p})$, and c depicts the minimum expected contrast between the foreground and background. We set $c=15$ in our experiments, since the human eye can approximately distinguish contrast between two gray intensities that differ in 15 or more levels in gray images with 256 levels; see [14, Chapter 2].

We simplify the frequency of gray intensities at level i in $\mathcal{P}_r(\boldsymbol{p})$ with h_i. Readers may be interested in an efficient implementation to compute the histogram of gray intensities described in [7,15].

- *Otsu's method* [16] is a global algorithm, which minimizes the variance of gray intensities of $\widehat{\mathcal{F}}$ and $\widehat{\mathcal{B}}$. It assumes that the gray intensities of foreground and background form two distinguishable clusters whose overlap is small. The local Otsu's threshold uses the criterion

$$t_{opt} = \arg\max_{t\in(0,g)} \{F_0(t)\cdot F_1(t)\cdot [\widehat{\mu}_1(t) - \widehat{\mu}_0(t)]^2\} \tag{6}$$

where

$$F_i(t) = \sum_{j=a}^{b} h_j \quad \text{and} \quad \widehat{\mu}_i(t) = \frac{1}{F_i(t)} \sum_{j=a}^{b} j\cdot h_j \tag{7}$$

and the lower limit a and upper limit b depend on the index $i=0$, 1. These limits are defined as: $a=0$ and $b=t$ if $i=0$; otherwise $a=t+1$ and $b=g$.

- *Johannsen and Bille's method* (Johannsen' threshold) [17] is a global algorithm, which minimizes the interdependence, in an information theoretic sense, between the gray intensities of the estimated foreground and background. The local Johannsen's method chooses t_{opt} from the relation

$$t_{opt} = \arg\min_{t\in(0,g)} \{C_0(t) + C_1(t)\} \tag{8}$$

where

$$C_0(t) = \ln\left(\sum_{j=0}^{t} p_j\right) - \frac{1}{\sum_{j=0}^{t} p_j}\left[p_t\cdot\ln(p_t) + \left[\sum_{j=0}^{t-1} p_j\right]\cdot\ln\left(\sum_{j=0}^{t-1} p_j\right)\right] \tag{9}$$

$$C_1(t) = \ln\left(\sum_{j=t}^{g} p_j\right) - \frac{1}{\sum_{j=t}^{g} p_j}\left[p_t\cdot\ln(p_t) + \left[\sum_{j=t+1}^{g} p_j\right]\cdot\ln\left(\sum_{j=t+1}^{g} p_j\right)\right] \tag{10}$$

and $p_j = h_j/|\mathcal{P}_r(\boldsymbol{p})|$ denotes the empirical probability of the gray intensity at level j in $\mathcal{P}_r(\boldsymbol{p})$.

- *The minimum error thresholding* (Kittler's threshold) [9] is a global algorithm, which minimizes a criterion related to the average classification error rate assuming that the gray intensities of both background and foreground are normally distributed with different mean and variance. The local Kittler's threshold is computed as

$$t_{opt} = \arg\min_{t\in(0,g)} \left\{\sum_{i=0}^{1} F_i(t)\cdot\ln\left(\frac{\widehat{\sigma}_i^2(t)}{[F_i(t)]^2}\right)\right\} \tag{11}$$

where

$$\widehat{\sigma}_i^2(t) = \frac{1}{F_i(t)}\left[\sum_{j=a}^{b} j^2\cdot h_j\right] - [\widehat{\mu}_i(t)]^2 \tag{12}$$

and $F_i(t)$, $\widehat{\mu}_i(t)$, a, and b are defined as in Otsu's threshold.

- *Kapur, Sahoo and Wong's method* (Kapur's threshold) [18] is a global algorithm, which maximizes the sum of the entropy of gray intensities in $\mathcal{C}_0$ and $\mathcal{C}_1$. The local optimal threshold is derived as

$$t_{opt} = \arg\min_{t\in(0,g)} \left\{-\sum_{i=0}^{1}\sum_{j=a}^{b}\left[\frac{h_j}{F_i(t)}\cdot\ln\left(\frac{h_j}{F_i(t)}\right)\right]\right\} \tag{13}$$

and $F_i(t)$, a, and b are defined as in Otsu's threshold.

- *Tsallis entropy's method* (Portes's threshold) [19] is a global algorithm proposed by Portes de Albuquerque, which maximizes the information measure between background and foreground. Locally, it derives the optimal threshold from *Tsallis entropy*[1] as

$$t_{opt} = \arg\max_{t\in(0,g)} \{C_0(t) + C_1(t) + (1-\alpha)\cdot C_0(t)\cdot C_1(t)\} \tag{14}$$

$$C_i(t) = \frac{1 - \sum_{j=a}^{b}\left[\frac{h_j}{F_i(t)}\right]^{\alpha}}{\alpha - 1} \tag{15}$$

where $F_i(t)$, a, and b are defined as in Otsu's threshold, and α is a parameter whose influence on the threshold was not determined in the original publication. Notice, however, that Tsallis entropy reduces to Boltzmann–Gibbs entropy if $\alpha\to 1$. That is,

$$\lim_{\alpha\to 1}\frac{1-\sum_i x_i^{\alpha}}{\alpha-1} = -\sum_i x_i\cdot\ln x_i \quad \text{where } \sum_i x_i = 1 \tag{16}$$

Therefore, Kapur's threshold is a particular case of Tsallis entropy's method for $\alpha = 1$.

Tsallis entropy is also used in [20,21]. They proposed a linear combination

$$t_{opt} = \arg\max_{t\in(0,g)} \left\{\sum_{i=0}^{1} w_i\cdot C_i(t)\right\} \tag{17}$$

where the weights w_i's were experimentally determined for each image type. However, the parameter space is enormous, considering the weights as parameters and the fact that the their ranges were not determined. Hence, this variant of Tsallis's entropy method was excluded from our experiments.

3.2. Statistical methods

Statistical methods rely on information from statistics of gray intensities. These methods usually compute the mean and variance of gray intensities in $\mathcal{P}_r(\boldsymbol{p})$. Both statistics can be computed in constant time [7] giving the advantage of speed over histogram cluster methods.

For the sake of brevity, let $\widehat{\mu} = \widehat{\mu}_{\mathcal{P}_r(\boldsymbol{p})}$ and $\widehat{\sigma}^2 = \widehat{\sigma}^2_{\mathcal{P}_r(\boldsymbol{p})}$.

- *Niblack's method* [22] is a local algorithm, which assumes that the gray intensities of the background form a dominant peak. The optimal threshold is computed as

$$T(\boldsymbol{p}) = \widehat{\mu} - \alpha\cdot\widehat{\sigma} \tag{18}$$

where α is a parameter which is usually greater than zero, the higher α, the lower $T(\boldsymbol{p})$. However, α could be negative if there is not a unique dominant peak or the dominant peak is mainly formed by foreground pixels in the histogram of gray intensities. Trier and Jain [15] suggested $\alpha = 0.2$.

[1] http://tsallis.cat.cbpf.br/biblio.htm

- *Sauvola and Pietikäinen's method* (Sauvola's threshold) [23] is a local algorithm, which computes a threshold similar to Niblack's threshold, but it incorporates a second parameter such that

$$T(\boldsymbol{p}) = \widehat{\mu} - \alpha \cdot \widehat{\mu} + \alpha \frac{\widehat{\sigma}}{\beta} \widehat{\mu} \tag{19}$$

where α behaves as in Niblack's threshold and $\beta > 0$. The influence of $\widehat{\sigma}$ on $T(\boldsymbol{p})$ is regulated by β so that $T(\boldsymbol{p}) \to \widehat{\mu} - \alpha \cdot \widehat{\mu}$ if $\widehat{\sigma} \to 0$ while $T(\boldsymbol{p}) \to \widehat{\mu}$ if $\widehat{\sigma} \to \beta$. Uniform neighborhoods may have a low $\widehat{\sigma}$ which implies that $T(\boldsymbol{p}) \approx \widehat{\mu} - \alpha \cdot \widehat{\mu}$ and, consequently, $I(\boldsymbol{p}) > T(\boldsymbol{p})$ with high probability. Sauvola and Pietikäinen suggest $\alpha = 0.5$ and $\beta = 128$ given $g = 255$.

- *Wolf and Jolion's method* (Wolf's threshold) [24] is a hybrid algorithm, which replaces the parameter β of Sauvola's threshold with the maximum standard deviation of gray intensities in neighborhoods with radius r so that the influence of $\widehat{\sigma}$ on $T(\boldsymbol{p})$ is normalized. It also replaces the mean of gray intensities in the last two terms of (19) with the difference between the mean and minimum of gray intensities in the neighborhood. Wolf and Jolian thus reflect the idea that the optimal threshold should lie between such an interval. Wolf's threshold is given by

$$\begin{aligned} T(\boldsymbol{p}) &= \widehat{\mu} - \alpha[\widehat{\mu} - m] + \alpha \frac{\widehat{\sigma}}{s}[\widehat{\mu} - m] \\ m &= \min_{\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p})} \{I(\boldsymbol{q})\}, \quad s = \max_{\boldsymbol{q} \in \mathcal{P}_{r^*}(\boldsymbol{p})} \{\widehat{\sigma}_{\mathcal{P}r(\boldsymbol{q})}\} \end{aligned} \tag{20}$$

where $\mathcal{P}_{r^*}(\boldsymbol{p})$ is the secondary neighborhood with radius $r^* \geq r$ and $\alpha \leq 1$. The higher α, the lower $T(\boldsymbol{p})$. Wolf and Jolion suggest the parameter $\alpha = 0.5$.

- *Iterative global thresholding* (Kavallieratou's threshold) is a hybrid and iterative method, which was originally proposed in [25] and subsequently improved in [26]. In each iteration i, the gray intensities are linearly transformed from $[m, \widehat{\mu}_i]$ to $[0, g]$, where m and $\widehat{\mu}^{(i)}$ are the minimum and mean of the gray intensities in $\mathcal{P}$ at iteration i, respectively, setting gray intensities greater than $\widehat{\mu}^{(i)}$ to g.
- *Kavallieratou's threshold* (new approach) We propose a variant of iterative global thresholding. Instead of $\widehat{\mu}^{(i)}$, our modified version computes the mean of gray intensities in the pixel neighborhood of interest. Thereby

$$T(\boldsymbol{p}) = \begin{cases} I(\boldsymbol{p}) & \text{if } \widehat{\mu}^{(\alpha)}(\boldsymbol{p}) > I^{(\alpha)}(\boldsymbol{p}) \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

where α is the number of iterations, $I^{(1)}(\boldsymbol{p}) = I(\boldsymbol{p}) - \min_{\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p})}\{I(\boldsymbol{q})\}$, and

$$\widehat{\mu}^{(i)}(\boldsymbol{p}) = \frac{1}{|\mathcal{P}_r(\boldsymbol{p})|} \sum_{\boldsymbol{q} \in \mathcal{P}_r(\boldsymbol{p})} I^{(i)}(\boldsymbol{q}) \quad \text{for } i = 1, \ldots, \alpha \tag{22}$$

$$I^{(i)}(\boldsymbol{p}) = \min\left(\widehat{\mu}^{(i-1)}(\boldsymbol{p}), g \cdot \frac{I^{(i-1)}(\boldsymbol{p})}{\widehat{\mu}^{(i-1)}(\boldsymbol{p})}\right) \quad \text{for } i = 2, \ldots, \alpha \tag{23}$$

4. Unsupervised binarization measures

To evaluate binarized images, Levine and Nazif [27] stated that the uniformity of a feature over a region is inversely proportional to the variance of the values of that feature evaluated at every pixel belonging to that region. Adjusting their original measure to binarization context, the *uniformity measure* is defined as

$$U = 1 - \frac{1}{w}[w_{\widehat{\mathcal{F}}_r(\boldsymbol{p})} \cdot S^2_{\widehat{\mathcal{F}}_r(\boldsymbol{p})} + w_{\widehat{\mathcal{B}}_r(\boldsymbol{p})} \cdot S^2_{\widehat{\mathcal{B}}_r(\boldsymbol{p})}] \tag{24}$$

where the notation $S^2_{\mathcal{A}}$ refers to *the biased sample variance of gray intensities* in $\mathcal{A}$, and w_f and w_b are the weights associated to $\widehat{\mathcal{F}}_r(\boldsymbol{p})$ and $\widehat{\mathcal{B}}_r(\boldsymbol{p})$, respectively, and w is a normalization factor designed to limit the maximum value of the measure to one

$$w = [w_f + w_b] \cdot \frac{[I_{max} - I_{min}]^2}{2} \tag{25}$$

where I_{max} and I_{min} are the maximum and minimum gray intensities in $\mathcal{P}$.

Sahoo et al. [5] used a particular case of U with $w_f = w_b = 1$ to evaluate binarization methods. We simplified this particular case of U with the *gray-intensity uniformity* (GU) measure

$$GU_r = S^2_{\widehat{\mathcal{F}}} + S^2_{\widehat{\mathcal{B}}} \tag{26}$$

which is linearly equivalent to Sahoo et al.'s evaluation measure.

Proposition 1. *Let $\mathcal{P}$ be an r-simple image. Then, the minimum of the expected value of $GU_r(\boldsymbol{p})$ is not necessarily reached for $\widehat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{F}_r(\boldsymbol{p})$ or $\widehat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})$; see proof in supplementary material.*

Proposition 1 indicates that GU_r does not lead to the best binarization for all r-simple images. What is more, if one wanted to minimize the expected value of $GU_r(\boldsymbol{p})$, then it could happen that the estimated background would *swallow* the foreground.

Another measure derived from U is the *region non-uniformity* (NU), which was proposed by Sezgin and Sankur [6] as

$$NU = \frac{|\widehat{\mathcal{F}}| \cdot S^2_{\widehat{\mathcal{F}}}}{|\mathcal{P}| \cdot S^2_{\mathcal{P}}} \tag{27}$$

NU can be transformed to the local measure $NU_r(\boldsymbol{p})$ by replacing $\mathcal{P}$, and $\widehat{\mathcal{F}}$, with $\mathcal{P}_r(\boldsymbol{p})$ and $\widehat{\mathcal{F}}_r(\boldsymbol{p})$, respectively. Unfortunately, $NU_r(\boldsymbol{p})$ lacks desirable properties: $NU_r(\boldsymbol{p})$ is zero if $\widehat{\mathcal{F}}_r(\boldsymbol{p}) = \emptyset$, which means that NU_r leads to *white images*.

Otsu [16] proposed several discriminant measures in order to evaluate the "*goodness*" of the threshold (at level t). One of these global measures is the *weighted variance* (WV), defined as

$$WV = \frac{1}{|\mathcal{P}|}[|\widehat{\mathcal{B}}| \cdot S^2_{\widehat{\mathcal{B}}} + |\widehat{\mathcal{F}}| \cdot S^2_{\widehat{\mathcal{F}}}] \tag{28}$$

Ng and Lee [28] proved that WV is equivalent to U if $w_f = |\widehat{\mathcal{F}}|$, $w_b = |\widehat{\mathcal{B}}|$, and $w = |\mathcal{P}|$.

Let WV_r be the measure which replaces $\widehat{\mathcal{F}}$ and $\widehat{\mathcal{B}}$ with $\widehat{\mathcal{F}}_r(\boldsymbol{p})$ and $\widehat{\mathcal{B}}_r(\boldsymbol{p})$ in WV. Then,

Proposition 2. *In an r-simple image, the minimum of the expected value of WV_r is not necessarily reached for $\widehat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{F}_r(\boldsymbol{p})$ or $\widehat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})$; see proof in supplementary material.*

Ramírez-Ortegón et al. [7] proposed the *uniform variance measure* (UV), which is defined with the local gray-intensity standard deviations as

$$UV_r(\boldsymbol{p}) = \frac{1}{|\mathcal{P}_r(\boldsymbol{p})|}[|\widehat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \widehat{\sigma}_{\widehat{\mathcal{B}}_r(\boldsymbol{p})} + |\widehat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \widehat{\sigma}_{\widehat{\mathcal{F}}_r(\boldsymbol{p})}] \tag{29}$$

In terms of a measure M_r, the binarization performance over a whole image is the accumulation of the binarization performances over all neighborhoods with radius r. We denote this evaluation by

$$Eval(M_r, \widehat{\mathcal{F}}) = \sum_{\boldsymbol{p} \in \mathcal{P}} M_r(\widehat{\mathcal{F}}_r(\boldsymbol{p})) \tag{30}$$

A measure is useful if the better the binarization obtained, the smaller the measure on to which the segmented image evaluates. In particular, we would desire the minimum of the measure to be attained only at the perfect segmentation $\widehat{\mathcal{F}} = \mathcal{F}$.

4.1. Unbiased weighted variance

To overcome the statistical bias of WV_r, we propose the *unbiased weighted variance measure*

$$\widehat{WV}_r(\boldsymbol{p}) = \frac{1}{|\mathcal{P}_r(\boldsymbol{p})|}[|\widehat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \widehat{\sigma}^2_{\widehat{\mathcal{B}}_r(\boldsymbol{p})} + |\widehat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \widehat{\sigma}^2_{\widehat{\mathcal{F}}_r(\boldsymbol{p})}] \tag{31}$$

if $|\widehat{\mathcal{B}}_r(\boldsymbol{p})| \geq 2$ and $|\widehat{\mathcal{F}}_r(\boldsymbol{p})| \geq 2$. Otherwise $\widehat{WV}_r(\boldsymbol{p}) = \widehat{\sigma}^2_{\mathcal{P}_r(\boldsymbol{p})}$. Similarly,we define the *unbiased uniform variance* which replaces $\widehat{\sigma}^2_{\widehat{\mathcal{F}}_r(\boldsymbol{p})}$ and $\widehat{\sigma}^2_{\widehat{\mathcal{B}}_r(\boldsymbol{p})}$ (variances) with $\widehat{\sigma}_{\widehat{\mathcal{F}}_r(\boldsymbol{p})}$ and $\widehat{\sigma}_{\widehat{\mathcal{B}}_r(\boldsymbol{p})}$ (standard deviations), respectively.

Theorem 1. *In an r-simple image,the expected value of the unbiased weighted variance measure is minimal if $\widehat{\mathcal{F}} = \mathcal{F}$ or $\widehat{\mathcal{F}} = \mathcal{B}$;see proof in supplementary material.*

Corollary 2. *In an r-simple image,if r is such that $|\widehat{\mathcal{B}_r(\boldsymbol{p})}|, |\mathcal{F}_r(\boldsymbol{p})| \geq 1$ and $\sigma^2_{\mathcal{B}_r(\boldsymbol{p})}, \sigma^2_{\mathcal{F}_r(\boldsymbol{p})} > 0$ for all $\boldsymbol{p} \in \mathcal{P}$, then*

$$Eval(\widehat{WV}_r, \mathcal{F}) < Eval(\widehat{WV}_r, \mathcal{A}) \tag{32}$$

for all $\mathcal{A} \neq \mathcal{B}$ and $\mathcal{A} \neq \mathcal{F}$; see proof in supplementary material.

Assuming that the gray intensities of both foreground and background are lognormally distributed, we derived the measures $\widetilde{WV}_r(\boldsymbol{p})$ and $\widetilde{UV}_r(\boldsymbol{p})$ from $\widehat{WV}_r(\boldsymbol{p})$ and $\widehat{UV}_r(\boldsymbol{p})$. These measures replace $\widehat{\sigma}^2_{\widehat{\mathcal{F}}_r(\boldsymbol{p})}$ with *the unbiased sample variance of gray-intensity logarithms*

$$\widetilde{\sigma}^2_{\widehat{\mathcal{F}}_r(\boldsymbol{p})} = \ln\left(1 + \frac{\widehat{\sigma}^2_{\widehat{\mathcal{F}}_r(\boldsymbol{p})}}{\widehat{\mu}^2_{\widehat{\mathcal{F}}_r(\boldsymbol{p})}}\right) \tag{33}$$

where $\ln(\cdot)$ denotes the natural logarithm. Similarly, $\widetilde{\sigma}_{\widehat{\mathcal{B}}_r(\boldsymbol{p})}$ replaces $\widehat{\sigma}_{\widehat{\mathcal{B}}_r(\boldsymbol{p})}$.

4.2. Bias of binarization algorithms derived from unsupervised measures

Unsupervised measures can be transformed into binarization algorithms like Otsu's algorithm corresponds to weighted variance measure, and Hou's algorithm [29] corresponds to gray-intensity uniformity measure. However, binarization methods based on measures usually introduce bias due to the foreground selection. For example, the means and variances in Otsu's threshold come from the distributions whose tails are truncated by the threshold value. The distribution on the left (foreground approximation) has its right tail truncated and the one on the right (background approximation) has its left tail gone. Then, the variances computed in Otsu's threshold do not correspond to such distributions and, in consequence, bias is introduced. Furthermore, the assumption of independence in the weighted variance measure now fails because $\widehat{\mathcal{F}}$ and $\widehat{\mathcal{B}}$ are strongly correlated given by any threshold candidate. Another bias factor is given by the space searching domain. While the optimal search domain for the weighted measure is $\Omega = \{\widehat{\mathcal{F}}_r(\boldsymbol{p}) | \widehat{\mathcal{F}}_r(\boldsymbol{p}) \subset \mathcal{P}_r(\boldsymbol{p})\}$, Otsu's threshold has a search domain Ω_{otsu}, where $\widehat{\mathcal{F}}_r(\boldsymbol{p}) \in \Omega_{otsu}$ if there exists t such that $I(\boldsymbol{p}) \leq t$ if $\boldsymbol{p} \in \widehat{\mathcal{F}}_r(\boldsymbol{p})$ and $I(\boldsymbol{p}) > t$ if $\boldsymbol{p} \in \widehat{\mathcal{B}}_r(\boldsymbol{p})$. Similar arguments are given for Hou's algorithm.

The larger the overlap between fore- and background distributions, the larger bias in the Otsu's and Hou's algorithms. Moreover, Otsu's and Hou's algorithms cannot deal with uniform neighborhoods; see alternatives proposed in [1,11–13]. Nevertheless, our experiments indicate that some unsupervised measures and some binarization algorithms in combination attain a high OCR accuracy while avoiding the manual parameter tuning.

5. Design of experiments

The purpose of the experiments presented here is to establish the relationship between unsupervised methods and OCR accuracy. Given an image, a binarization algorithm, and an unsupervised method, we expect that the best binarized image, obtained via the unsupervised measure, attains the highest OCR accuracy. That is, given an image, for instance, let e_i and a_i be the measurements of an unsupervised measure and OCR accuracy, respectively. Let e_x be the minimum of e_i's, which means that the binary image B_x is the best according the unsupervised method, and let a_y be the maximum of a_i's, which means that the binary image B_y has the highest OCR accuracy. Then, we expect that either $B_x = B_y$ (ideal case) or $a_x \approx a_y$ (good case).

The rest of this section is divided as follows: Section 5.1 describes the database of images used for the experiments; parameter datasets are described in Section 5.2; definitions of OCR accuracy and OCR measures are given in Section 5.3; finally, we propose the uncertainty test in Section 5.4 to assess which algorithms are better.

5.1. Test images

The binarization algorithms were tested with digitalized images of the historical atlas "*Theatrum orbis terrarum, sive, Atlas novus*" (Blaeu Atlas) [30] at 150 dpi resolution.

We report the results for $n = 86$ color images randomly extracted from 61 maps. These images are mainly composed of map headers, map comments and region labels without stylized handwriting characters; see Fig. 3. Each color image i is

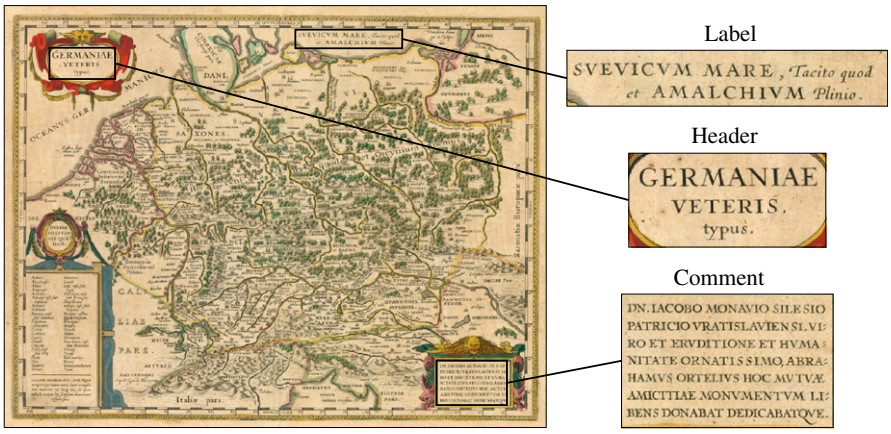

Fig. 3. Example of map which contains a header, label and comment.

Table 1
Each parameter is sampled according the increments of the third column between the range specified in the second column.

Algorithm	Parameter: from/to	Parameter: increment
Johannsen's, Kapur's, Kittler's and Otsu's	$r: 10/50$	$r: 5$
Kavallieratou's	$\alpha: 1/20, r: 10/50$	$\alpha: 1, r: 5$
Niblack's	$\alpha: 0/6, r: 10/50$	$\alpha: 0.1, r: 5$
Portes's	$\alpha: 0/5, r: 10/50$	$\alpha: 0.1, r: 5$
Sauvola's	$\alpha: 0/1, \beta: 32/196, r: 10/50$	$\alpha: 0.01, \beta: 32, r: 5$
Wolf's	$\alpha: 0/1, r: 10/50, r^*: 50$	$\alpha: 0.01, r: 5$

transformed to the gray image I_i by $I_i(\boldsymbol{p}) = 0.299I_{i1}(\boldsymbol{p}) + 0.587I_{i2}(\boldsymbol{p}) + 0.114I_{i3}(\boldsymbol{p})$, where $I_i(\boldsymbol{p}) \in [0,255]$, and $I_{i1}(\boldsymbol{p})$, $I_{i2}(\boldsymbol{p})$ and $I_{i3}(\boldsymbol{p})$ denote the red, green and blue intensities of $\boldsymbol{p}$ in the color image i [14].

5.2. Binarized images

Table 1 presents the range and increments of the parameter sampling for each algorithm. We denote $\Omega_{j,k}$ the parameter combination k of the binarization algorithm j, which is constructed combining the sampled parameters.

For instance, Sauvola's threshold algorithm has $5454 = 101 \times 6 \times 9$ parameter combinations since α, β and r are sampled with 101, 6, and 9 different values, respectively.

Let $\widehat{\mathcal{F}}_{i,j,k}$ denote the estimated foreground of the image I_i by the binarization algorithm j with parameters $\Omega_{j,k}$. We compute the best estimated foreground of a binarized image, for each measure u, image I_i and binarization algorithm j as

$$\widehat{\mathcal{F}}_{i,j}^{(u)} = \underset{\widehat{\mathcal{F}}_{i,j,k}}{\operatorname{argmin}}\{Eval(M_r^{(u)}, \widehat{\mathcal{F}}_{i,j,k})\} \tag{34}$$

where $M_r^{(u)}$ denotes the u-th measure of the list: GU_r, NU_r, $\widehat{UV}_r$, $\widehat{UV}_r$, $\widehat{WV}_r$, and $\widehat{WV}_r$.

In our experiments, the radius of all measures was set to $r = 50$ because it is approximately the minimum radius that entirely contains any character in the tested images.

5.3. OCR measure

Measures like Damerau–Levenshtein and Levenshtein distance compute the cost of transforming a string A into a string B by counting the number of edit operators: insertion, deletion, or substitution of a single character, or a transposition of two characters; see [31]. However, the probability of retrieving B from A depends strongly on the transformation string algorithm and on the strings' structure themselves. Because of that, we chose an OCR measure that is proportional to the number of correct recognized characters.

We define the *accuracy measure* (AC) of a binary image with foreground $\widehat{\mathcal{F}}$ as

$$AC(\widehat{\mathcal{F}}) = \frac{\#(\text{characters of } T_{\text{match}})}{\#(\text{characters of } T_{\text{in}})} \tag{35}$$

where T_{in} is the original text in the image, and T_{match} is a maximum matching string between T_{in} and the recognized text from the image. Since we would like to count the number of correct recognized characters from the binarized image, we define a maximum matching string between two strings as: Given two strings A and B, we say that A is a substring of B ($A \prec B$) if B can be transformed to A by removing characters from it; a maximum matching string C of A and B is a string of maximum length such that $C \prec A$ and $C \prec B$. Under this definition, T_{match} has at least the same number of correct recognized characters. Since Needleman–Wuntsh algorithm [32] fits our assumptions, we used it to compute T_{match}.

The AC measure is an important measure for OCRs, because the higher the AC measurement, the greater the possibility to extract, by further algorithms, relevant information from the recognized text. We used TopOCR to recognize the binarized images. A comparison of six OCRs can be found in the supplementary material.

We define the following values in order to evaluate OCR performance:

$$w_i^* = \max_{j,k} AC(\widehat{\mathcal{F}}_{i,j,k}) \quad (\textit{absolute potential } AC) \tag{36}$$

$$w_{i,j} = \max_{k} AC(\widehat{\mathcal{F}}_{i,j,k}) \quad (\textit{relative potential } AC) \tag{37}$$

In this manner, w_i^* approximates the maximum accuracy that can be computed for I_i with our OCR software in combination with any of the nine binarization methods; $w_{i,j}$ approximates the maximum accuracy with binarization algorithm j.

The absolute and relative potential AC may change if the number of sampled parameters or tested algorithms is incremented; nevertheless, we consider such values as the groundtruth.

Note that the value $w_{i,j}$ is not a good absolute measure for the ability of a binarization algorithm to maximize OCR performance. Given a fixed j, $w_{i,j}$ as indicator of binarization performance depends not only on w_i^*, but also on the rest of $w_{i,l}$ for other binarization algorithms (l's). For example, suppose that the best OCR accuracy is 0.5 ($w_i^* = 0.5$). If we have that $w_{i,j} = 0.45$ for some j, this could be interpreted either as a low OCR performance or as a low binarization method performance. However, the ratio $w_{i,j}/w_i^*$ is 0.90, which means that the binarization method j effectively maximizes the OCR accuracy despite the intrinsic low OCR performance in I_i. Hence, our observations are mainly based on pairwise tables and statistics of the ratios

$$x_{i,j} = \frac{w_{i,j}}{w_i^*} \quad (\textit{potential } AC \textit{ efficacy}) \tag{38}$$

and

$$y_{i,j}^{(u)} = \frac{AC(\widehat{\mathcal{F}}_{i,j}^{(u)})}{w_i^*} \quad (AC \textit{ efficacy}) \tag{39}$$

The ratio $x_{i,j}$ approximates the efficacy of the binarization algorithm j to maximize the accuracy in I_i. The ratio $y_{i,j}^{(u)}$ approximates the efficacy of measure u to tune the parameters of algorithm j in order to maximize the accuracy in I_i.

5.4. Uncertainty test for pairwise data from one sample

Given an image, suppose that we are able to compare the performance of two methods x and y based on some criterion. In this context, performance means how well the method performs its task. Also suppose that there are only three possible outcomes of the method's comparison in a single image: *Method x is better than*, *worse than*, *and as good as method y* (E_1,E_2, and E_3, respectively). Therefore, we ascertain that method x is better than method y in an image population if E_1 occurs more frequently than E_2. More formally, let $p_i = \Pr(E_i)$ for $i = 1,2,3$ be the probability of occurrence of E_i in an image which was randomly drawn from an image population. Then, our assessment is based on the numerical relation between p_1 and p_2.

Let the random variable N_i indicates the number of occurrences of E_i in a sample of n images which were independently and randomly drawn from a large population of images. Then, the triad (N_1,N_2,N_3) is approximately trinomially distributed.

Technically speaking, this is sampling without replacement, so the correct distribution is the multivariate hypergeometric distribution, but the distributions converge as the population size grows.

Assume that (n_1,n_2,n_3) is an observed vector of (N_1,N_2,N_3); the probability of observing (n_1,n_2,n_3) is given by

$$\psi(n_1,n_2,n_3;n,p_1,p_2,p_3) = \Pr(N_1=n_1,N_2=n_2,N_3=n_3) = \frac{n!}{n_1!\cdot n_2!\cdot n_3!}p_1^{n_1}\cdot p_2^{n_2}\cdot p_3^{n_3} \quad (40)$$

where $\cdot!$ denotes the factorial function. Therefore, p_i can be estimated by $\widehat{p}_i = n_i/n$. Unfortunately, large samples to ensure convergence may be unavailable, and the probability of observing $\widehat{p}_1 < \widehat{p}_2$ may be significant even if $p_1 > p_2$. Consequently, we focus on estimating how (un)likely it is that $\alpha\cdot\widehat{p}_1 \geq \widehat{p}_2$ for $\alpha < 1$, given that $p_1 \leq p_2$.

Observe that the upper bound of $\Pr(\alpha\cdot\widehat{p}_1 \geq \widehat{p}_2 | p_1 \leq p_2)$ is the maximum probability of observing $\alpha\cdot\widehat{p}_1 \geq \widehat{p}_2$ subject to $p_1 \leq p_2$. This upper bound represents our maximum risk of judging x better

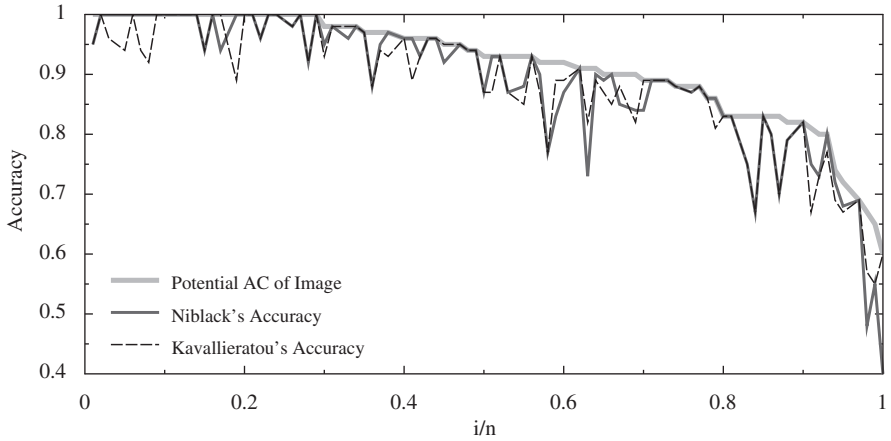

Fig. 4. Graph of the absolute potential *AC*.

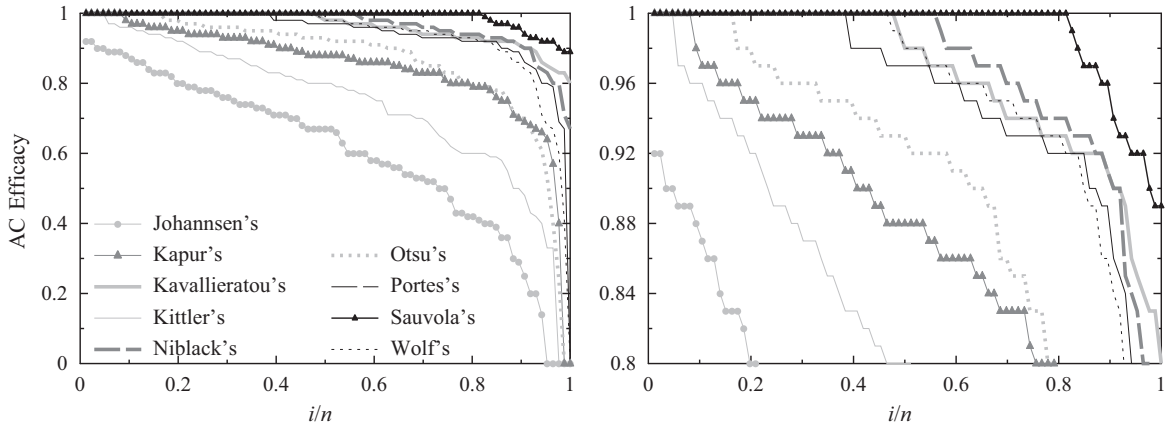

Fig. 5. Ordered graphs of the potential *AC* efficacy.

Table 2
Pairwise comparison of absolute efficacy.

	Rank	Joh.		Kap.		Kav.		Kit.		Nib.		Otsu		Por.		Sau.		Wolf	
		n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}	n_{yx}	p_{yx}
Johannsen	9	–	–	2	0.03	0	0.00	13	0.18	0	0.00	2	0.03	0	0.00	0	0.00	0	0.00
Kapur	7	78	0.98	–	–	1	0.01	57	0.89	0	0.00	18	0.31	0	0.00	0	0.00	2	0.03
Kavallieratou	3	85	1.00	73	0.99	–	–	80	0.99	10	0.32	57	0.92	30	0.70	4	0.11	28	0.60
Kittler	8	59	0.82	7	0.11	1	0.01	–	–	0	0.00	5	0.08	0	0.00	0	0.00	0	0.00
Niblack	2	85	1.00	73	1.00	21	0.68	80	1.00	–	–	62	0.98	33	0.77	7	0.17	31	0.65
Otsu	6	77	0.97	40	0.69	5	0.08	61	0.92	1	0.02	–	–	4	0.07	1	0.01	5	0.08
Portes	4(*)	84	1.00	71	1.00	13	0.30	79	1.00	10	0.23	52	0.93	–	–	4	0.08	19	0.48
Sauvola	1	85	1.00	77	1.00	34	0.89	81	1.00	34	0.83	69	0.99	46	0.92	–	–	40	0.83
Wolf	4(*)	85	1.00	70	0.97	19	0.40	76	1.00	17	0.35	57	0.92	21	0.53	8	0.17	–	–

Both Wolf's and Portes's methods marked with (*) are ranked fourth because their p_{yx}'s values differ from each other slightly.

than method y when indeed y is better than x. This is analogous to a confidence interval prevalent in other statistical methods.

We call this probability α-*uncertainty*, which can be estimated by

$$UN(n,\alpha)=\max_{(y_1,y_2)\in\mathcal{Y}}\sum_{(x_1,x_2,x_3)\in\mathcal{X}}\psi(x_1,x_2,x_3;n,y_1,y_2,y_3) \tag{41}$$

where $\mathcal{Y}=\{(y_1,y_2)\in\mathbb{R}^2|0\le y_1\le y_2\le 1 \text{ and } y_1+y_2\le 1\}$, $y_3=1-y_1-y_2$, $\mathcal{X}=\{(x_1,x_2,x_3)\in\mathbb{N}^3|\alpha\cdot x_1\ge x_2 \text{ and } x_1+x_2+x_3=n\}$.

$\alpha-$uncertainty for different values of n and α can be found in the supplementary material.

6. Results and conclusions

The absolute potential *AC* is greater than 0.60 in all our test images; see Fig. 4. Indeed, in 93% of them are at least 0.80, which indicates that our OCR is capable of recognizing most of the characters in our test images. In the same figure, the corresponding relative potential *AC* measurements of Niblack's and Kavallieratou's algorithms fluctuate irregularly. A visual comparison between Niblack's and Kavallieratou's graphs is consequently difficult. Hence, all the following graphs are in decreasing order to make the visual inspection easier.

The results of our experiment are summarized in Fig. 5 (graphs of potential efficacy), Table 3 (mean and variances of *AC* efficacy), and Table 2 (pairwise tables of potential *AC* efficacy). Each cell (y-row,x-column) of Table 2 contains two values n_{yx} and p_{yx}. In terms of absolute efficacy measure, the number n_{yx} represents the times that the algorithm y has a higher score than the algorithm x; $p_{yx}=n_{yx}/(n_{yx}+n_{xy})$ represents the conditional probability of y's score being higher than x's score. Since $UN(86,0.75)<0.09$ ($\alpha-$uncertainty), we ascertain that algorithm y is better than algorithm x if $n_{xy}\le 0.75n_{yx}$, which is equivalent to $p_{yx}\ge 0.57$.

Fig. 6 shows the ranking of all six evaluation measures for each binarization method. This ranking is given by pairwise tables of *AC* efficacy (supplementary material) with an $\alpha-$uncertainty less than 0.9.

A visual inspection of the binarized images suggests that Johannsen, Kapur's, Kittler's, and Otsu's threshold wrongly classify pixels whose neighborhoods are completely contained in the background. In contrast, the rest of the algorithms, which have one or two parameters more besides the radius, can successfully binarize this kind of neighborhoods by tuning their parameters. Our conclusions are also supported for the means and standard deviations of the relative potential *AC* presented in Table 3.

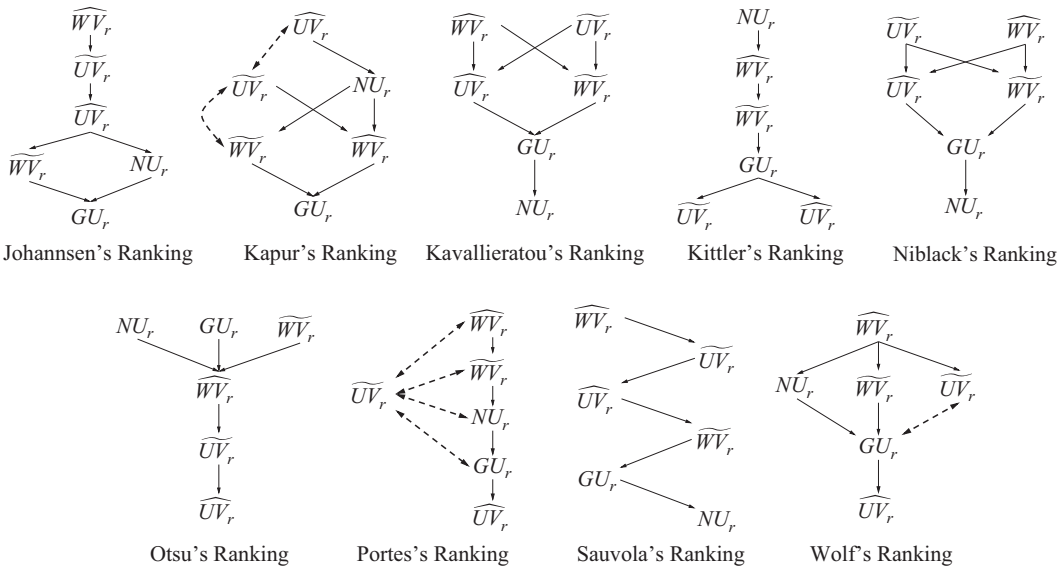

Fig. 6. Measure ranking for each binarization algorithm. The ranking is in decreasing order from top to bottom. Two algorithms with the same ranking either lie on the same level or are linked with a dash line with double arrow.

Table 3
Mean (μ) and standard deviation (σ) of the *AC* efficacy for each binarization algorithm and unsupervised evaluation method.

	Potential		$y_{ij}^{(u)}$											
	μ	σ	GU_r		NU_r		$\widehat{UV}_r$		$\widetilde{UV}_r$		$\widehat{WV}_r$		$\widetilde{WV}_r$	
			μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Johannsen	0.600	0.239	0.483	0.253	0.487	0.258	**0.496**	**0.250**	0.493	0.256	0.486	0.257	0.496	0.252
Kapur	0.845	0.168	0.750	0.201	**0.756**	**0.197**	0.756	0.199	0.751	0.198	0.751	0.200	0.750	0.200
Kavallieratou	0.963	0.048	0.601	0.220	0.517	0.227	**0.763**	**0.224**	0.728	0.222	0.715	0.195	0.763	0.227
Kittler	0.741	0.215	0.640	0.244	**0.658**	**0.243**	0.631	0.250	0.629	0.252	0.646	0.239	0.651	0.238
Niblack	0.964	0.063	0.538	0.233	0.007	0.046	0.767	0.230	0.716	0.241	0.711	0.207	**0.773**	**0.227**
Otsu	0.864	0.189	0.795	0.217	0.796	0.219	0.789	0.217	0.787	0.217	**0.797**	**0.217**	0.794	0.216
Portes	0.941	0.122	0.777	0.184	0.777	0.184	0.770	0.220	0.753	0.216	0.778	0.185	**0.785**	**0.209**
Sauvola	**0.989**	**0.027**	0.531	0.229	0.058	0.117	0.761	0.247	0.724	0.244	0.712	0.206	**0.798**	**0.210**
Wolf	0.936	0.141	0.801	0.204	0.804	0.191	0.769	0.235	0.740	0.249	0.806	0.193	**0.812**	**0.220**

For each algorithm, the best values of $y_{ij}^{(u)}$ are shown in bold.

In our test images, the radius used to compute the best binarized images, in terms of relative potential *AC*, are random and independent of the binarization used; they range randomly from 10 to 50. However, a value of $r=50$ was found optimal for most evaluation measures, regardless of the binarization method, with an exception of Wolf's method in which usually $r=10$. This behavior led Otsu's threshold to have almost the same mean and variance whenever evaluation measures adjusted the Otsu's radius; see Table 3 and Fig. 7 (right). For example, the histogram in Fig. 7 (left) shows in light gray bars the radius's probability of being selected by $\widehat{WV}_r$. The probability of $r=50$ is close to 0.90, even though the radius's probability of being optimal in terms of the relative potential *AC* of Otsu's method is around 0.3. This unwanted effect also appears in Johannsen's, Kapur's and Kittler's methods, pointing out that all six measures are ineffective to adjust the neighborhood radius. We conjectured that all four binarization methods estimate the foreground in such a manner that, for a given binarization method, all six measures reach their minimum with the same $\widehat{\mathcal{F}}$ (the same radius). Unfortunately, our mathematical analysis is unable to explain this pattern.

We observed that the OCR accuracy in an image depends mostly on how well binarized the image is. The OCR accuracy of two binarized images mainly differs due to broken characters, large false positive spots, and overestimated foreground boundaries. Moreover, the presented ranking is based on pairwise tables and not in the measurement magnitude. Therefore, the *AC* efficacy ranking for our dataset may be similar in other OCRs, but not so the accuracy measurements.

6.1. GU and NU discussion

In Section 4 we showed that NU_r does not penalize false negatives and that GU_r estimates the background in such a manner that it tends to contain $\mathcal{F}_r(\boldsymbol{p})$ if $|\mathcal{F}_r(\boldsymbol{p})|$ is small. Therefore, NU_r and GU_r are unsuitable for binarization methods whose parameters allow the generation of *white images* or *images with degraded text*. Particularly, the threshold of Kavallieratou's, Niblack's, and Sauvola's methods can be interpreted as *the acceptable deviation from the expected gray intensity* such that the higher the parameter α is, the more pixels are classified as background. NU_r led Niblack's and Sauvola's methods to generate *white images* and led to Kavallieratou's method to generate images with degraded characters. Likewise, Kavallieratou's, Niblack's and Sauvola's methods yielded images with degraded characters when their parameters were tuned by GU_r; see Fig. 8. Table 3 summarizes the low performance of NU_r and GU_r in these binarization methods.

6.2. WV and UV discussion

After visually inspecting the binarized images, we concluded that $\widetilde{UV}_r$ outperforms $\widehat{UV}_r$ in all binarization algorithms (Table 3) because $\widehat{UV}_r$ leads to generate more false positive spots (connected components with four or more pixels) which are scattered all around the background. In addition to this noise, binarization algorithms which are evaluated with $\widehat{UV}_r$ overestimate the foreground contours occasionally; see Fig. 9. In our tests, measures based on the lognormal distribution yielded sharper foreground boundaries than those based on the normal distribution. This supports the previous observations that gray intensities at foreground boundaries are lognormally distributed [7,10].

In our experiments, $\widehat{WV}_r$ and $\widetilde{UV}_r$ were found the best for parameter selection of binarization methods with potential *AC* efficacy over 0.9 (Kavallieratou's, Niblack's, Porte's, Sauvola's and Wolf's methods); see Table 3 and Fig. 6. Particularly, $\widehat{WV}_r$ is better than $\widetilde{UV}_r$ for Sauvola's and Wolf's methods despite observing sharper

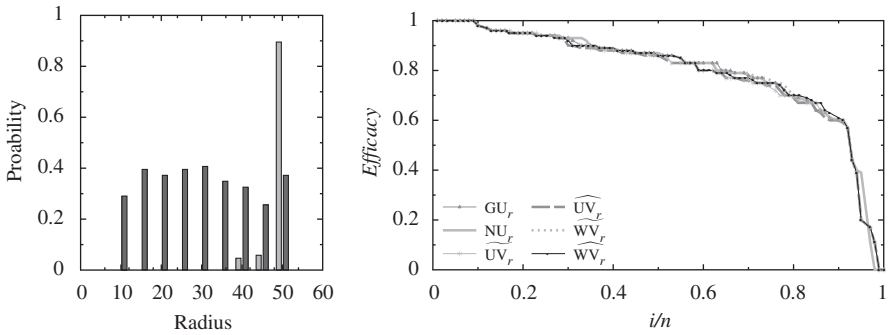

Fig. 7. On the left, light gray bars represent the radius's probability of being optimal in terms of $\widehat{WV}_r$ ($r=50$), in dark gray bars, the radius's probability of being optimal in terms of the relative potential *AC* of Otsu's method. (Right) The efficacy graphs of Otsu's method, one for each measure.

Fig. 8. Original image on the left. Center and right images were binarized by Kavallieratou's threshold after being tuned with GU_r and NU_r, respectively.

Fig. 9. Original image on the left, center and right images were binarized by Portes's threshold after being tuned with $\widetilde{UV}_r$ and $\widehat{UV}_r$, respectively.

Fig. 10. Left and right images were binarized by Wolf's threshold after being tuned with $\widetilde{UV}r$ and $\widehat{WV}_r$, respectively.

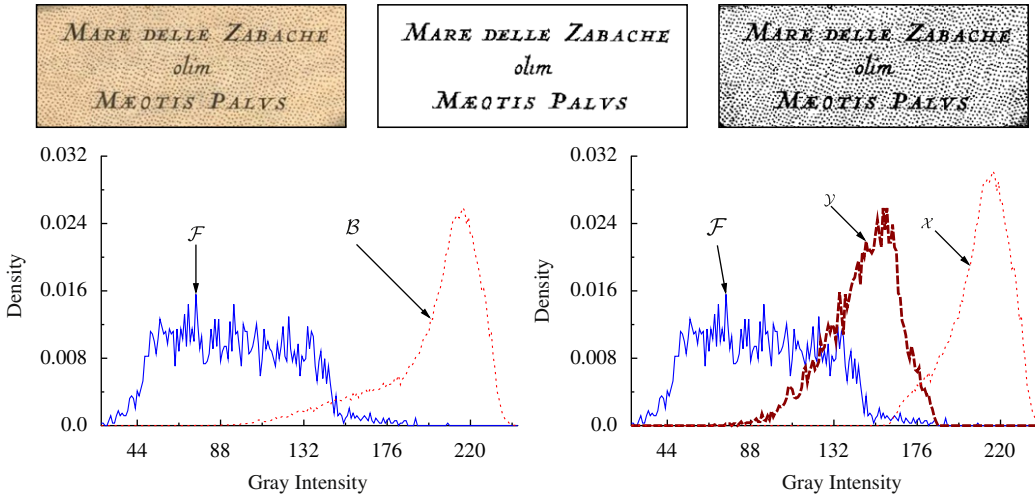

Fig. 11. At the top: an example of a non-ideal image (left), its corresponding groundtruth (center), and Wolf's binarization tuned with $\widehat{WV}_r$ (right). On the bottom-left, the density function of gray intensities of $\mathcal{F}$ and $\mathcal{B}$. On the bottom-right, the density function of gray intensities of $\mathcal{F}$, $\mathcal{X}$, and $\mathcal{Y}$, where $\mathcal{X}$ represents the set of false positive from the top-right image, and $\mathcal{Y}=\mathcal{F}\backslash\mathcal{X}$.

foreground contours with $\widetilde{UV}_r$. We suppose that $\widehat{WV}_r$ surpasses $\widetilde{UV}_r$ because it leads to well-conserved foreground contours and, at the same time, generated less noise than $\widetilde{UV}_r$; see Fig. 10. Another reason for this superiority can be attributed to TopOCR since it occasionally misclassifies a character with sharp contours.

In the practice, images fulfill the conditions of r-simple images only partially. In an image, the performance of $\widehat{WV}_r$ and $\widetilde{WV}_r$ is directly related with the number of neighborhoods with radius r which satisfy both Model 1 and (2). Fig. 11, for instance, shows an image where the percent of neighborhoods ($r \geq 10$) that satisfy (2) is close to 1, but the gray intensities in its background are not identically and independently distributed. The gray intensity of false positive pixels from Wolf's binarization, denoted by $\mathcal{Y}$, follows a different distribution to those in $\mathcal{X}=\mathcal{B}\backslash\mathcal{Y}$. As a consequence, $\widehat{WV}_r$ leads Wolf's method to generate $\widehat{\mathcal{F}}=\mathcal{F}\cup\mathcal{Y}$ since $\widehat{\mu}_{\mathcal{Y}}-\widehat{\mu}_{\mathcal{F}}<\sqrt{2}\cdot\max(\widehat{\sigma}_{\mathcal{Y}},\widehat{\sigma}_{\mathcal{F}})$ and $\widehat{\mu}_{\mathcal{X}}-\widehat{\mu}_{\mathcal{Y}}>\sqrt{2}\cdot\max(\widehat{\sigma}_{\mathcal{X}},\widehat{\sigma}_{\mathcal{Y}})$.

Acknowledgments

We thank the anonymous referees for helpful and relevant comments. This research was partially supported by The National Council on Science and Technology (CONACYT) of Mexico (Grant number: 218253) and by a Howard Hughes Medical Institute Collaborative Innovation Award.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.patcog.2010.09.018.

References

[1] M.R. Gupta, N.P. Jacobson, E.K. Garcia, OCR binarization and image pre-processing for searching historical documents, Pattern Recognition 40 (2007) 389–397.
[2] J. Lázaro, J.L. Martín, J. Arias, A. Astarloa, C. Cuadrado, Neuro semantic thresholding using OCR software for high precision OCR applications, Image and Vision Computing 28 (2010) 571–578.
[3] H. Zhang, J.E. Fritts, S.A. Goldman, Image segmentation evaluation: a survey of unsupervised methods, Computer Vision and Image Understanding 110 (2008) 260–280.
[4] Y.J. Zhang, A survey on evaluation methods for image segmentation, Pattern Recognition 29 (8) (1996) 1335–1346.
[5] P.K. Sahoo, S. Soltani, A.K. Wong, Y.C. Chen, A survey of thresholding techniques, Computer Vision, Graphics, and Image Processing 41 (2) (1988) 233–260.
[6] M. Sezgin, B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging 13 (1) (2004) 146–168.
[7] M.A. Ramírez-Ortegón, E. Tapia, L.L. Ramírez-Ramírez, R. Rojas, E. Cuevas, Transition pixel: a concept for binarization based on edge detection and gray-intensity histograms, Pattern Recognition 43 (2010) 1233–1243.
[8] S. Chabrier, B. Emile, C. Rosenberger, H. Laurent, Unsupervised performance evaluation of image segmentation, Journal on Applied Signal Processing 2006 (2006) 1–12.
[9] J. Kittler, J. Illingworth, Minimum error thresholding, Pattern Recognition 19 (1) (1985) 41–47.
[10] M.A. Ramírez-Ortegón, E. Tapia, R. Rojas, E. Cuevas, Transition thresholds and transition operators for binarization and edge detection, Pattern Recognition 43 (10) (2010) 3243–3254.
[11] Q. Huang, W. Gao, W. Cai, Thresholding technique with adaptive window selection for uneven lighting image, Pattern Recognition Letters 26 (2005) 801–808.
[12] C.-H. Chou, W.-H. Lin, F. Chang, A binarization method with learning-built rules for document images produced by cameras, Pattern Recognition 43 (4) (2010) 1518–1530.
[13] R.F. Moghaddam, M. Cheriet, A multi-scale framework for adaptive binarization of degraded document images, Pattern Recognition 43 (6) (2010) 2186–2198.

[14] R.C. Gonzalez, R.E. Woods, Digital Image Processing, third ed., Prentice Hall, 2007.
[15] Ø.D. Trier, A.K. Jain, Goal-directed evaluation of binarization methods, Transactions on Pattern Analysis and Machine Intelligence 17 (12) (1995) 1191–1201.
[16] N. Otsu, A threshold selection method from grey-level histograms IEEE Transactions on Systems, Man, and Cybernetics 9 (1) (1979) 62–66.
[17] G. Johannsen, J. Bille, A threshold selection method using information measures, in: Proceedings of the Sixth International Conference on Pattern Recognition, 1982, pp. 140–143.
[18] J.N. Kapur, P.K. Sahoo, A.K.C. Wong, A new method for gray-level picture thresholding using the entropy of the histogram, Computer Vision, Graphics and Image Processing 29 (1985) 273–285.
[19] M. Portes de Albuquerque, I. Esquef, A.G. Mello, M. Portes de Albuquerque, Image thresholding using Tsallis entropy, Pattern Recognition Letters 25 (2004) 1059–1065.
[20] C.A. Mello, L.A. Schuler, Thresholding images of historical documents using a Tsallis-entropy based algorithm, Journal of Software 3 (6) (2008) 29–36.
[21] C. Mello, A. Sanchez, A. Oliveira, A. Lopes, An efficient gray-level thresholding algorithm for historic document images, Journal of Cultural Heritage 9 (2) (2008) 109–116.
[22] W. Niblack, An Introduction to Digital Image Processing, Prentice Hall, Birkeroed, Denmark, 1985.
[23] J. Sauvola, M. Pietikäinen, Adaptive document image binarization, Pattern Recognition 33 (2) (2000) 225–236.
[24] C. Wolf, J.-M. Jolion, Extraction and recognition of artificial text in multimedia documents, Pattern Analysis and Applications 3 (2003) 309–326.
[25] E. Kavallieratou, A binarization algorithm specialized on document images and photos, in: ICDAR '05: Proceedings of the Eighth International Conference on Document Analysis and Recognition, IEEE Computer Society, Washington, DC, USA2005, pp. 463–467.
[26] E. Kavallieratou, S. Stathis, Adaptive binarization of historical document images, in: ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition, IEEE Computer Society, Washington, DC, USA2006, pp. 742–745.
[27] M.D. Levine, A.M. Nazif, Dynamic measurement of computer generated image segmentations, IEEE Transactions on Pattern Analysis and Machine Intelligence 7 (2) (1985) 155–164.
[28] W. Ng, C. Lee, Comment on using the uniformity measure for performance measure in image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (1996) 933–934.
[29] Z. Hou, Q. Hu, W.L. Nowinski, On minimum variance thresholding, Pattern Recognition Letters 27 (2006) 1732–1743.
[30] W. Janszoon, J. Blaeu, Theatrum Orbis Terrarum, Sive, Atlas Novus, Blaeu Atlas, 1645. URL: 〈http://www.library.ucla.edu/yrl/reference/maps/blaeu〉.
[31] G. Navarro, A guided tour to approximate string matching, ACM Computing Surveys 33 (2001) 1–58.
[32] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology 48 (3) (1970) 443–453.

Marte A. Ramírez-Ortegón is a Ph.D. student at the Department of Mathematics and Computer Science of the Freie Universität Berlin. In 2002 he received a B.S. degree (Computer Science) from the University of Guanajuato (UG)/the Centre for Mathematical Research (CIMAT), Guanajuato, Mexico (1998–2002). His current research fields include pattern recognition, edge detection, binarization, multithresholding and text recognition.

Edgar A. Duéñez-Guzmán is currently a Post-doctoral Fellow at Harvard University, Department of Organismic and Evolutionary Biology. He has a B.Sc. in Mathematics from the University of Guanajuato, a M.Sc. in Computing and Industrial Mathematics from the Mathematics Research Center (CIMAT) and a Ph.D. in Computer Science from the University of Tennessee. His research interest range from theoretical evolutionary biology and evolutionary computation to optimization and image processing.

Raúl Rojas received the B.S. and M.S. degrees in Mathematics from the National Polytechnic Institute (IPN), Mexico City, Mexico. He obtained the Ph.D. degree at the Freie Universität Berlin, Berlin, Germany. Currently, Dr. Rojas is a full professor in Computer Science at the Freie Universität Berlin and leader of the Work Group on Artificial Intelligence at that University. His research interests include artificial intelligence, computer vision, and robotics.

Erik Cuevas received the B.S. and M.S. degrees in Computer Engineering from the University of Guadalajara, Guadalajara, Mexico, in 1996 and 2000, respectively. He received the Ph.D. degree in Computer Engineering at the Freie Universität Berlin, Berlin, Germany, in 2006. Currently, he is a full professor in the Computer Science Department at the University of Guadalajara. His research interests include computer vision, learning systems, and soft computing.