Unsupervised Evaluation Methods Based on Local Gray-Intensity Variances for Binarization of Historical Documents

Marte A. Ramírez-Ortegón* and Raúl Rojas
Institut für Informatik,Freie Universität Berlin,
Takustr. 9, 14195 Berlin, Germany
* mars.sasha@gmail.com, rojas@inf.fu-berlin.de

***Abstract*—We attempt to evaluate the efficacy of six unsupervised evaluation method to tune Sauvola's threshold in optical character recognition (OCR) applications. We propose local implementations of well-known measures based on gray-intensity variances. Additionally, we derive four new measures from them using the unbiased variance estimator and gray-intensity logarithms. In our experiment, we selected the well binarized images, according each measure, and computed the accuracy of the recognized text of each. The results show that the *weighted* and *uniform variance* (using logarithms) are suitable measures for OCR applications.** [1]

***Index Terms*—binarization; unsupervised; evaluation;**

I. 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 the cultural heritage through digital libraries.

The main problem in the construction of digital libraries lies in the extraction of information from hundreds of thousands ancient documents. The digitization of bibliographic records is the only feasible solution to that problem.

This problem can be roughly divided in three parts: detection of object of interest (binarization), text extraction, and text recognition. Here, we ignore the text extraction problem and assume that the text recognition is performed by an *optical character recognition* (OCR) application, which works as a *black box* algorithm. This is, the OCR performance mainly depend on the input image while the OCR parameters has a low influence in the output. Therefore, the evaluation of the binarization algorithm and its parameters play the most important roll in the system. Then, the natural question is: Which parameters may be set in the binarization algorithm to maximize the OCR performance?

Manual tuning of the binarization parameters by human experts is inadequate because it implies time-consuming operations and high expenses; then, the binarization performance may be assess with unsupervised evaluation methods which analyze the segmentation quality by properties and principles of the segmentation. These methods do not need neither human intervention, nor groundtruth. Consequently, they can be used on a large scale. Furthermore, they enable the objective comparison of both different segmentation methods and different parameters of a single method. They also enable the self-tuning of algorithms based on evaluation results.

Measures based on gray-intensity variance are popular for evaluating binarized images [1], [2], [3] because, intuitively, both foreground and background should be uniform and homogeneous regions. Unfortunately, few authors have analyzed the mathematical and experimental behavior of these measures [4], [5]. This is why, we study the efficacy of them for tuning binarization methods in order to maximize the accuracy of OCR applications. In our test, we analyzed Sauvola's method [6] (binarization method) and TopOCR [7] (OCR software) but the same methodology can be applied to more binarization methods and OCR software.

We propose local implementations of classic and recent measures to overcome images with composite background (two or more sub-regions). Afterward, we propose modeling the distribution of gray intensities of both foreground and background as lognormally distributed.

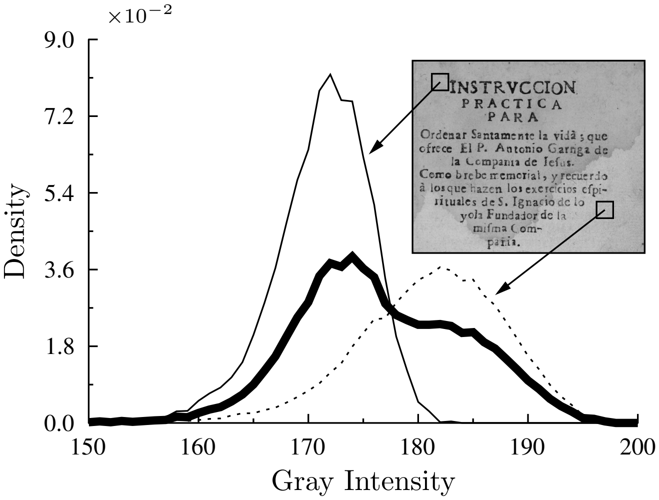
The rest of this paper is organized as follows. Section II introduces the examined unsupervised evaluation methods. The comparison study is described in Section III. Results of the experiment and conclusions are presented in Section IV.

II. Evaluation methods for binarization

Binarization is the process of dividing the set of pixels $\mathcal{P}$ into $\hat{\mathcal{F}}$ and $\hat{\mathcal{B}}$ with the aim of estimating the foreground $\mathcal{F}$ and background $\mathcal{B}$, respectively. In binarization context, $\mathcal{F}$ represents the set of pixels containing the objects of interest and $\mathcal{B}$ is the complement of $\mathcal{F}$ in $\mathcal{P}$.

All binarization algorithms reported on [3], [8], [9] assume that foreground pixels can be distinguished by extracting diverse features based on their gray intensities. Under this assumption, authors like [10], [2], [3] conjecture that the variance of gray intensities of both foreground and background in well binarized images are smaller than the corresponding

20th International Conference on Pattern Recognition, IEEE Computer Society, 2010, 2029 - 2032 August 23-26, 2010. Istambul, Turkey



INSTRVCCION
PRACTICA
PARA
Ordenar Santamente la vidâ ; que
ofrece El P. Antonio Garriga de
la Compania de Ieſus.
Como brebe memorial, y recuerdo
à los que hazen los exercicios eſpi-
rituales de S. Ignacio de lo
yola Fundador de la
miſma Com-
pañia.

Fig. 1. Two different regions form the background. Although the gray intensities of each background region are approximately normally distributed, the gray intensities of the entire background are not.

variances in wrong binarized images. However, this conjecture is false for images with composite foreground and/or background like Fig. 1. As a result, evaluation measures, based on gray-intensity variances, could be misleading. To overcome this difficulty we analyze local implementation of these measures.

A. Notation

We define the neighborhood $\mathcal{P}_r(\boldsymbol{p})$ as the set of pixels within a square centered at the pixel $\boldsymbol{p}$ of sides with length $2r+1$. We abbreviate the intersection between $\mathcal{F}$ and $\mathcal{P}_r(\boldsymbol{p})$ as $\mathcal{F}_r(\boldsymbol{p}) = \mathcal{F} \cap \mathcal{P}_r(\boldsymbol{p})$. Similarly we define $\hat{\mathcal{F}}_r(\boldsymbol{p})$, $\mathcal{B}_r(\boldsymbol{p})$, and $\hat{\mathcal{B}}_r(\boldsymbol{p})$. The cardinality of a set A is denoted as $|\mathcal{A}|$.

Given a set $\mathcal{A}$, we denote the following statistics and estimators of gray intensities: The expected value with $\mu_{\mathcal{A}}$; the variance with $\sigma_{\mathcal{A}}$; the mean with $\hat{\mu}_{\mathcal{A}}$ (an estimator of $\mu_{\mathcal{A}}$); the unbiased sample variance with $\hat{\sigma}^2_{\mathcal{A}}$ (an unbiased estimator of $\sigma_{\mathcal{A}}$), $\hat{\sigma}^2_{\mathcal{A}} = 0$ if $|\mathcal{A}| < 2$; the biased sample variance of gray intensities with $S^2_{\mathcal{A}}$ (an unbiased estimator of $\sigma_{\mathcal{A}}$), $S^2_{\mathcal{A}} = 0$ if $|\mathcal{A}| < 1$; *the unbiased sample variance of gray-intensity logarithms*

$$\tilde{\sigma}^2_{\mathcal{A}} = \ln\left(1 + \frac{\hat{\sigma}^2_{\mathcal{A}}}{[\hat{\mu}_{\mathcal{A}}]^2}\right). \tag{1}$$

B. Unsupervised evaluation methods

To evaluate binarized images, Levine and Nazif [11] proposed the *gray-intensity uniformity* (GU) measure. With the same aim, Sezgin and Sankur [3] derived the *region nonuniformity* (NU) measure from *GU*. These measures are defined as

$$GU = S^2_{\hat{\mathcal{B}}} + S^2_{\hat{\mathcal{F}}} \quad \text{and} \quad NU = \frac{|\hat{\mathcal{F}}| \cdot S^2_{\hat{\mathcal{F}}}}{|\mathcal{P}| \cdot S^2_{\mathcal{P}}}. \tag{2}$$

Otsu [12] proposed the *weighted variance* (WV) defined as

$$WV = \frac{|\hat{\mathcal{B}}| \cdot S^2_{\hat{\mathcal{B}}} + |\hat{\mathcal{F}}| \cdot S^2_{\hat{\mathcal{F}}}}{|\mathcal{P}|} \tag{3}$$

Ramírez-Ortegón et al. [1] proposed the *uniform variance measure* (UV) that is defined with the local gray-intensity standard deviations as

$$UV_r(\boldsymbol{p}) = \frac{|\hat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \hat{\sigma}_{\hat{\mathcal{B}}_r(\boldsymbol{p})} + |\hat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \hat{\sigma}_{\hat{\mathcal{F}}_r(\boldsymbol{p})}}{|\mathcal{P}_r(\boldsymbol{p})|}, \tag{4}$$

All four measures expect that the better the binarization, the lower the evaluation measurement.

GU, NU and WV can be transformed easily in the local measures GU_r, NU_r and WV_r by replacing $\mathcal{P}$, $\hat{\mathcal{F}}$, and $\hat{\mathcal{B}}$, with $\mathcal{P}_r(\boldsymbol{p})$, $\hat{\mathcal{F}}_r(\boldsymbol{p})$, and $\hat{\mathcal{B}}_r(\boldsymbol{p})$, respectively. However their local implementations lack desirable properties: NU_r measure is zero if $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \emptyset$, and the expected values of both WV_r and GU_r are not minimum if $\hat{\mathcal{B}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})$. For example, assume that all pixels are background and $\hat{\mathcal{B}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p})$ then

$$E(GU_r) = E(S^2_{\mathcal{B}_r(\boldsymbol{p})}) = \frac{|\mathcal{B}_r(\boldsymbol{p})| - 1}{\mathcal{B}_r(\boldsymbol{p})} \sigma_{\mathcal{B}_r(\boldsymbol{p})} \tag{5}$$

where $E(\cdot)$ denotes the expected value. However, if $\hat{\mathcal{B}}_r(\boldsymbol{p}) = \mathcal{B}_r(\boldsymbol{p}) \backslash \{\boldsymbol{p}\}$ and $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \{\boldsymbol{p}\}$ then

$$E(GU_r) = E(S^2_{\mathcal{B}_r(\boldsymbol{p}) \backslash \{\boldsymbol{p}\}}) = \frac{|\mathcal{B}_r(\boldsymbol{p})| - 2}{|\mathcal{B}_r(\boldsymbol{p})| - 1} \sigma_{\mathcal{B}_r(\boldsymbol{p})} \tag{6}$$

which is smaller than (5).

We propose *r-local weighted variance measure* whose expected value is minimum if $\hat{\mathcal{F}}_r(\boldsymbol{p}) = \mathcal{F}_r(\boldsymbol{p})$. [2]

$$\widehat{WV}_r(\boldsymbol{p}) = \begin{cases} \frac{|\hat{\mathcal{B}}_r(\boldsymbol{p})| \cdot \hat{\sigma}^2_{\hat{\mathcal{B}}_r(\boldsymbol{p})} + |\hat{\mathcal{F}}_r(\boldsymbol{p})| \cdot \hat{\sigma}^2_{\hat{\mathcal{F}}_r(\boldsymbol{p})}}{|\mathcal{P}_r(\boldsymbol{p})|} & (*) \\ \hat{\sigma}^2_{\mathcal{P}_r(\boldsymbol{p})} \text{ otherwise.} \end{cases} \tag{7}$$

$$(*) \text{ if } |\hat{\mathcal{B}}_r(\boldsymbol{p})| \geq 2 \text{ and } |\hat{\mathcal{F}}_r(\boldsymbol{p})| \geq 2.$$

Similarly, we define $\widehat{UV}_r(\boldsymbol{p})$.

Experiments in [1] suggested that both foreground and background gray intensities locally behave as lognormally rather than normally distributed. Hence, we derived $\widetilde{WV}_r(\boldsymbol{p})$ and $\widetilde{UV}_r(\boldsymbol{p})$ from $\widehat{WV}_r(\boldsymbol{p})$ and $\widehat{UV}_r(\boldsymbol{p})$. These measures replace $\hat{\sigma}_{\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and $\hat{\sigma}_{\hat{\mathcal{B}}_r(\boldsymbol{p})}$ with $\tilde{\sigma}_{\hat{\mathcal{F}}_r(\boldsymbol{p})}$ and $\tilde{\sigma}_{\hat{\mathcal{B}}_r(\boldsymbol{p})}$, respectively, see (1).

The binarization performance, in term of r-local weighted variance measure, is evaluated as

$$\widehat{WV}_r(B) = \frac{1}{|\mathcal{P}|} \sum_{\boldsymbol{p} \in \mathcal{P}} \widehat{WV}_r(\boldsymbol{p}), \tag{8}$$

where B represents the binarized image. Likewise, we define the rest of the measures.

[2] We have constructed a formal treatment of this argument, using some probability assumptions of gray-intensity differences. This work has been submitted for publication.

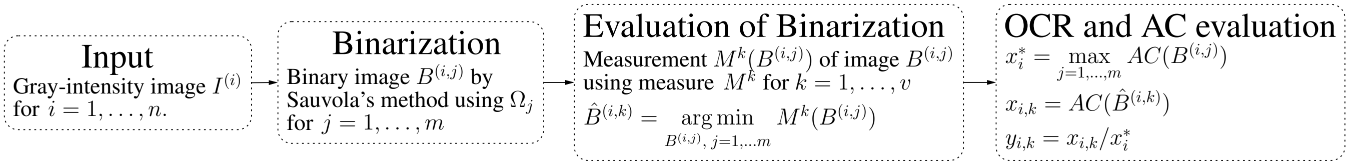

Input
Gray-intensity image $I^{(i)}$ for $i = 1, \ldots, n.$
Binarization
Binary image $B^{(i,j)}$ by Sauvola's method using Ω_j for $j = 1, \ldots, m$
Evaluation of Binarization
Measurement $M^k(B^{(i,j)})$ of image $B^{(i,j)}$ using measure M^k for $k = 1, \ldots, v$
$\hat{B}^{(i,k)} = \underset{B^{(i,j)},\, j=1,\ldots m}{\arg\min} \; M^k(B^{(i,j)})$
OCR and AC evaluation
$x_i^* = \max_{j=1,\ldots,m} AC(B^{(i,j)})$
$x_{i,k} = AC(\hat{B}^{(i,k)})$
$y_{i,k} = x_{i,k}/x_i^*$

Fig. 2. Overview of the evaluation process.

III. Comparative study

We evaluated the efficacy of $v = 6$ measures in relation to OCR performance. Figure 2 shows the evaluation processing flow. By simplicity, we denote $M^{(k)}$, for $k = 1, \ldots, v$, the k-measure of the list $M = \{GU_r, NU_r, \widehat{WV}_r, \widetilde{WV}_r, \widehat{UV}_r, \widetilde{UV}_r\}$.

Our test database is composed by $n = 86$ gray-intensity images $I^{(i)}$ for $i = 1, \ldots, n.$ that contain degraded text (ink stains and weak strokes for mention some kind of degradation). These images were extracted from 61 maps of the historical atlas *"Theatrum orbis terrarum, sive, Atlas novus"* (Blaeu Atlas) [13] with 150 dpi resolution.

We chosen Sauvola's method [6] to perform the binarization because it was top-ranked by [3], [8]. Sauvola's threshold is defined as

$$T(\boldsymbol{p}) = \hat{\mu}_{\mathcal{P}_{r'}(\boldsymbol{p})} \cdot \left[1 + \alpha \cdot \left(\frac{\hat{\sigma}_{\mathcal{P}_{r'}(\boldsymbol{p})}}{\beta} - 1\right)\right], \tag{9}$$

where r', α and β are parameters. The pixel $\boldsymbol{p}$ is classified as foreground if its gray intensity is lower than $T(\boldsymbol{p})$. Table I presents the range of each Sauvola's parameter that we used in our experiment. Varying the parameters of Sauvola's method, we computed $m = 5,454$ binary images $B^{(i,j)}$ for each image $I^{(i)}$. Later on, we computed $M^k(B^{(i,j)})$ which represents the measurement of $B^{(i,j)}$ with $M^{(k)}$. Then,

$$\hat{B}^{(i,k)} = \underset{B^{(i,j)}, j=1,\ldots,m}{\arg\min} \; M^{(k)}(B^{(i,j)}). \tag{10}$$

denotes the best-binarized image among $B^{(i,j)}$ in terms of measure $M^{(k)}$.

We used *TopOCR* [7] to recognize the text from the binarized images using four parameter sets. We measure the accuracy of the recognized text as

$$AC(B^{(i,j)}) = \frac{\#(\text{characters of } T_{\text{match}}^{(i,j)})}{\#(\text{characters of } T_{\text{in}}^{(i)})}, \tag{11}$$

where $T_{\text{in}}^{(i)}$ is the original text in $I^{(i)}$ and $T_{\text{match}}^{(i,j)}$ denotes the maximum matching text between $T_{\text{in}}^{(i)}$ and the OCR output. $T_{\text{match}}^{(i,j)}$ is computed using Needleman-Wuntsh algorithm [14]. The AC measure is an important measure for OCR applications, because the high AC measurement, the greater the possibility to extract, by further algorithms, relevant information from the recognized text.

In our evaluation, x_i^* represents the maximum AC among all the binarized images of $I^{(i)}$, and $x_{i,k}$ represents the OCR accuracy of the best-binarized image of $I^{(i)}$ in terms of measure $M^{(k)}$. Hence, our statistics and observations are mainly based on

$$y_{i,k} = \frac{x_{i,k}}{x_i^*} \text{ (AC efficacy)} \tag{12}$$

which represents the efficacy of $M^{(k)}$ for tuning Sauvola's method in order to maximize the accuracy. Observe that $x_{i,k}$ highly depends on x_i^* and, consequently, we cannot infer from it how efficient is $M^{(k)}$ to assess the binarization method. For example, suppose that however the parameters of Sauvola's method is, the OCR accuracy is lower or equal to 0.5 ($x_i^* = 0.5$); If $x_{i,k} = 0.45$, for instance, this could be interpreted either as low OCR performance, or as low binarization method performance, but the ratio of x_i^* to $x_{i,k}$ is $y_{i,k} = .90$, which means that $M^{(k)}$ is highly efficient to maximize the OCR accuracy despite of the intrinsic low OCR (binarization method) performance in $I^{(i)}$.

TABLE I
Range of Sauvola's parameters. Sweeping the parameters r', α and β, we generated $m = 5,454$ different parameter combinations $\Omega_j = \{r'_j, \alpha_j, \beta_j\}$.

Parameter	From/To	Increment
r'	10/50	5
α	0.0/1.0	0.01
β	32/196	32

IV. Results and Conclusion

In our experiment, we set $r = 50$ for all measures. Table II and Fig. 3 present statistics of points $(i, y_{z_{i,k},k})$ where $z_{i,k}$ are indexes such that $y_{z_{1,k},k} \geq \ldots \geq y_{z_{n,k},k}$ for $k = 1, \ldots, v$ Table II also present the pairwise comparison between values $y_{i,k}$.

The measure $\widehat{WV}_r$ is the best in overall performance (mean and variance). However, $\widetilde{UV}_r$ performed better in the first quartile of measurements $y_{z_{i,k},k}$. $\widehat{UV}_r$ and $\widetilde{WV}_r$ have an acceptable performance in a lesser degree.

Results in Table II indicate that GU_r and NU_r are ineffective to tune Sauvola's parameters, see Fig. 3. Notice that Sauvola's threshold can be interpreted as *the acceptable deviation from the expected gray intensity*. While incrementing the parameter α, this *tolerance* increases and, consequently, more and more pixels are classified as background up to all pixels are in the estimated background. Therefore, high α's are chosen for those evaluation measures which do not or lightly

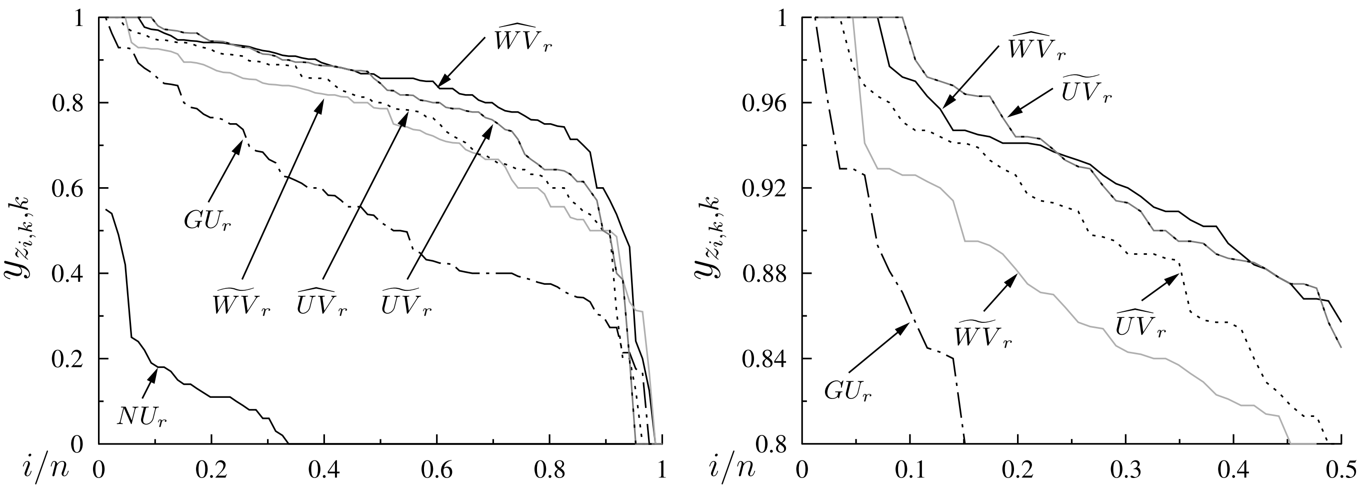


Fig. 3. Each of the AC efficacy graph are in decreasing order to make the visual inspection easier.

TABLE II
AC EVALUATION OVERVIEW.

Method	x_i^*	GU_r	NU_r	$\widehat{WV}_r$	$\widetilde{WV}_r$	$\widehat{UV}_r$	$\widetilde{UV}_r$
Mean	0.907	0.536	0.059	0.805	0.719	0.731	0.769
Std. Dev.	0.094	0.229	0.118	0.209	0.207	0.244	0.247
$y_{z_{i,k},k} \geq$				Quantiles i/n			
1.00	0.25	0.01	0.00	0.07	0.05	0.03	0.09
0.95	0.39	0.02	0.00	0.13	0.05	0.12	0.19
0.90	0.61	0.06	0.00	0.38	0.14	0.28	0.34
0.85	0.77	0.10	0.00	0.59	0.29	0.41	0.49
0.80	0.90	0.15	0.00	0.70	0.48	0.51	0.60
0.75	0.90	0.24	0.00	0.81	0.53	0.60	0.71
0.60	1.00	0.40	0.00	0.90	0.78	0.83	0.86
0.50	1.00	0.53	0.02	0.93	0.91	0.91	0.90

Pairwise comparison $P(y_{i,a} > y_{i,b})$ (a row, b column)

	GU_r	NU_r	$\widehat{WV}_r$	$\widetilde{WV}_r$	$\widehat{UV}_r$	$\widetilde{UV}_r$
GU_r	0.00	0.97	0.07	0.06	0.20	0.10
NU_r	0.00	0.00	0.00	0.00	0.00	0.00
$\widehat{WV}_r$	0.85	0.98	0.00	0.66	0.55	0.47
$\widetilde{WV}_r$	0.79	0.98	0.22	0.00	0.33	0.28
$\widehat{UV}_r$	0.76	0.95	0.21	0.52	0.00	0.23
$\widetilde{UV}_r$	0.77	0.93	0.30	0.62	0.51	0.00

penalize the estimated background. Particularly, NU_r yields *white images* while GU_r renders degraded characters.

After inspecting the binarized images visually, we concluded that $\widetilde{UV}_r$ outperforms $\widehat{UV}_r$ (Table II) because $\widehat{UV}_r$ generates more false positive spots (connected components with four or more pixels) which are scattered all around the background. In addition to this noise, binarized images which are evaluated with $\widehat{UV}_r$ overestimate the foreground contours occasionally,

We also concluded that measures based on the lognormal distribution yield sharper foreground boundaries than those based on the normal distribution. However, we suppose that $\widehat{WV}_r$ surpasses both $\widetilde{UV}_r$ and $\widetilde{WV}_r$ because $\widehat{WV}_r$ conserves the foreground contours fairly well and, at the same time, generates few noise in comparison with $\widetilde{UV}_r$ and $\widetilde{WV}_r$.

ACKNOWLEDGMENT

This research was partially supported by The National Council on Science and Technology (CONACYT)of Mexico (Grant number:218253).

REFERENCES

[1] M. A. Ramírez-Ortegón, E. Tapia, L. L. Ramírez-Ramírez, R. Rojas, and E. Cuevas, "Transition pixel: A concept for binarization based on edge detection and gray-intensity histograms," *Pattern Recognition*, vol. 43, pp. 1233 – 1243, 2010.

[2] P. K. Sahoo, S. Soltani, A. K. Wong, and Y. C. Chen, "A survey of thresholding techniques," *Computer Vision, Graphics. and Image Processing*, vol. 41, no. 2, pp. 233–260, 1988.

[3] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–168, January 2004.

[4] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent, "Unsupervised performance evaluation of image segmentation," *Journal on Applied Signal Processing*, vol. 2006, pp. 1–12, 2006.

[5] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, pp. 260 –280, September 2008.

[6] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[7] T. Soft, *Top OCR*. Top Soft, 2008. [Online]. Available: http://www.topocr.com/

[8] P. Stathis, E. Kavallieratou, and N. Papamarkos, "An evaluation technique for binarization algorithms," *Journal of Universal Computer Science*, vol. 14, no. 18, pp. 3011–3030, Octuber 2008.

[9] Ø. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1191–1201, 1995.

[10] M. D. Levine and A. M. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 7, no. 2, pp. 155–164, 1985.

[11] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

[12] N. Otsu, "A threshold selection method from grey-level histograms," *IEEE Transaction on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.

[13] W. Janszoon and J. Blaeu, *Theatrum Orbis Terrarum, Sive, Atlas Novus*. Blaeu Atlas, 1645. [Online]. Available: http://www.library.ucla.edu/yrl/reference/maps/blaeu

[14] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, March 1970.