Document downloaded from:

http://hdl.handle.net/10251/64376

This paper must be cited as:

Mauricio Villegas; Romero Gómez, V.; Sánchez Peiró, JA. (2015). On the modification of binarization algorithms to retain grayscale information for handwritten text recognition. En Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015, Proceedings. Springer International Publishing. 208-215. doi:10.1007/978-3-319-19390-8_24.

The final publication is available at

http://link.springer.com/chapter/10.1007/978-3-319-19390-8_24

Copyright Springer International Publishing

Additional Information

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-19390-8_24

On the Modification of Binarization Algorithms to Retain Grayscale Information for Handwritten Text Recognition

Mauricio Villegas, Verónica Romero, Joan Andreu Sánchez

PRHLT, Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{mauvilsa,vromero,jandreu}@prhlt.upv.es

Abstract. The amount of digitized legacy documents has been rising over the last years due mainly to the increasing number of on-line digital libraries publishing this kind of documents. The vast majority of them remain waiting to be transcribed to provide historians and other researchers new ways of indexing, consulting and querying them. However, the performance accuracy of state-of-the-art Handwritten Text Recognition techniques decreases dramatically when they are applied to these historical documents. This is mainly due to the typical paper degradation problems. Therefore, robust pre-processing techniques is an important step for helping further recognition steps. This paper proposes to take existing binarization techniques, in order to retain their advantages, and modify them in such a way that some of the original grayscale information is preserved and be considered by the subsequent recognizer. Results are reported with the publicly available ESPOSALLES database.

1 Introduction

In the last years, large amounts of handwritten historical documents residing in libraries, museums, archives and other institutions have been digitized both to preserve them and to make them available to the general public. The automatic transcription of these historical documents is a challenging problem related to Document Image Analysis and Natural Language Processing. The advances in these fields have made possible in recent years to start exploring this problem.

Available OCR technologies are not applicable to historical documents, since characters can not be isolated automatically in these images. Therefore, holistic, segmentation-free text recognition techniques are required. This technology is generally referred to as *"off-line Handwritten Text Recognition"* (HTR) [6]. Several approaches have been proposed in the literature for HTR based on hidden Markov models (HMM) [6,12], recurrent neural networks [2], or hybrid systems using HMM and neural networks [9]. These systems have proven to be useful in a restricted setting for simple tasks. However, in the context of historical documents, their performance decreases dramatically, mainly due to paper degradation problems encountered in this kind of documents, such as presence of smear, significant background variation, uneven illumination, and dark spots, require specialized image-cleaning and enhancement algorithms. In addition, show-through and bleed-through problems can render the distinction between background and foreground difficult [1]. The combination of these and other problems make the pre-processing of these documents a difficult task.

In this paper we present a new technique to extract handwritten text from noisy backgrounds and improve the quality of the image, making it more legible. It is based on the well known Sauvola binarization method [10], with the main difference that it produces a grayscale image, a feature that we show that is more adequate for an HMM-based HTR system. The following section gives a brief overview of the complete HTR system considered in this paper. Then in Section 3, the proposed pre-processing method is described in detail. The corpus and the experimental results are presented in Section 4 and 5. Finally, conclusions are drawn in Section 6.

2 Handwritten Text Recognition

The HTR system used in this paper followed the classical architecture composed of three main modules: document image pre-processing, line image feature extraction and HMM and language model training/decoding [12].

In the pre-processing module, the pages are first divided into line images as explained in [8]. Given that it is quite common for handwritten documents to suffer from degradation problems it is necessary to first apply appropriate filtering methods to remove the background, noise, improve the quality of the image and to make the documents more legible. This part of the pre-processing is the main focus of this paper. Three different methods have been tested. The first one is the background estimation and subtraction method used in [8], the second one is the classical Sauvola thresholding [10], and the final one is the proposed method described in Section 3. In Section 5 we can see the different results obtained with the three methods. Continuing with the other pre-processing steps, after filtering, the skew and slant of each line are corrected. Finally the size is normalized separately for each line. A more detailed description of this process is found in [8]. Then, each pre-processed line image is represented as a sequence of feature vectors representing gray levels and gradients [12].

Given a handwritten sentence image represented by a feature vector sequence ($\mathbf{x}$), the HTR problem can now be formulated as the problem of finding a most likely word sequence, $\mathbf{w}$, i.e., $\mathbf{w} = \arg\max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x})$. Using the Bayes' rule we can decompose this probability into two probabilities:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} P(\mathbf{w} \mid \mathbf{x}) = \arg\max_{\mathbf{w}} P(\mathbf{x} \mid \mathbf{w}) P(\mathbf{w}) \tag{1}$$

$P(\mathbf{x} \mid \mathbf{w})$ is typically approximated by concatenated character models, usually HMMs [3], while $P(\mathbf{w})$ is approximated by a word language model, usually n-grams [3].

In this work, the characters were modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. These models were estimated from training text images represented as feature vector sequences using the Baum-Welch algorithm. On the other hand, each lexical word was modeled by a stochastic finite-state automaton and the concatenation of words into text line sentences was modeled by bi-grams models, with Kneser-Ney back-off smoothing [5].

All these finite-state (character, word and sentence) models were *integrated* into a single *global* model on which a search process is performed for decoding the feature vectors sequence $\mathbf{x}$ into the words sequence $\mathbf{w}$. This search is optimally carried out by using the Viterbi algorithm [3].

3 Proposed Approach

In the context of ancient texts, it can be observed that the writing strokes vary greatly in width and darkness, not only between different pages or lines, but also within single words. Because of this, when applying thresholding techniques, there is always a risk that parts of the strokes are lost due to the hard black or white decision. Since the thresholding methods cannot be expected to work perfectly, this problem could be addressed somehow in order to improve recognition results.

To this end, instead of thresholding, i.e., assigning pixels to be either black or white depending on whether a pixel value is below or above a certain threshold, we relax the method such that values in a band near the threshold are mapped to a transition between black and white. The intuition behind this is that instead of a thresholded image, our aim is to use the probability that each pixel is above or below the threshold. In general this probability can't be estimated explicitly since many popular thresholding techniques are not derived from probabilistic principles. Nevertheless, pixels for which the thresholding technique is very certain should be assigned values very close to or equal to black or white, thus retaining the strength of the technique in question, and only for the least certain pixels the resulting image will be assigned a shade of gray. Little can be assumed about the distribution of the pixels in these transitions, so we opt for a simple linear mapping function, symmetric with respect to the threshold value. The only parameter to decide is how wide is the transition from black to white, or in other words the slope of the line. For this, we can note that ink strokes tend to have similar transitions from ink color to background throughout a document which means that the amount of pixels in the transitions should be more or less constant for any region of text.

Due to the good performance of local based thresholding methods, in this work we have taken as reference Sauvola's method [10], and we have modified it following the previously mentioned idea. In Sauvola's, given an input image, for every pixel a threshold is computed based on the pixel's neighborhood. Like the original Niblack's method [7], Sauvola's method computes the threshold based on the neighborhood's mean and standard deviation values. The threshold $T(x, y)$

for a pixel position (x, y) is given as follows

$$T(x,y) = \mu(x,y)\left[1 + k\left(\frac{\sigma(x,y)}{R} - 1\right)\right] \ , \tag{2}$$

where $\mu(x, y)$ and $\sigma(x, y)$ are the neighborhood's mean and standard deviation, respectively, R is the dynamic range of the standard deviation, and k is a parameter that needs to be adjusted. The neighborhood is a square region of W pixels wide, centered at the corresponding pixel. This algorithm is very fast, since it is well known that the mean and standard deviation of any sub-window of an image can be efficiently computed by using integral images [11].

Now to obtain the function for the linear transition, the mapping should be symmetric with respect to the threshold value, which means that for an input value equal to the threshold T the output should be $\frac{G}{2}$ being G the maximum grayscale value. From this we can deduce that the y-axis intercept is $\frac{G}{2} - mT$, where m is the slope of the line. As mentioned before, the slope of the line m should be set such that the expected number of pixels that fall in the transition is constant. To achieve this approximately without requiring more computations, the same standard deviation σ of the Sauvola sub-window can be used by setting the transition to be a fixed factor s of σ, in which case the slope becomes $m = \frac{G}{2 \cdot s \cdot \sigma}$ (the 2 is just to simplify a bit the final equation). After a few manipulations it can be shown that the pixel values of the resulting output image $O(x, y)$ for a given input $I(x, y)$ should be assigned as follows:

$$O(x,y) = \mathrm{limit\_gray}\left(\frac{G}{2}\left[\frac{I(x,y) - T(x,y)}{s \cdot \sigma} + 1\right]\right) \ . \tag{3}$$

The factor s must be adjusted empirically, although in our experimentation it was observed that a good value is $s = 1$. In Eq. 3, limit_gray has been introduced so that the function is defined for all input range, not just for the transition, and it is defined as $\mathrm{limit\_gray}(v) = \{\ 0 \text{ if } v < 0,\ G \text{ if } v > G,\ v \text{ otherwise }\}$.

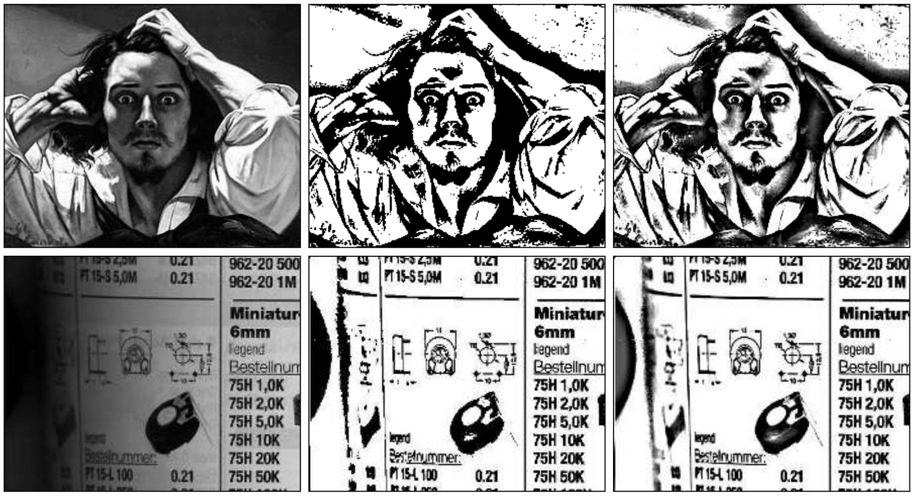
Figure 1 compares the results obtained for a couple of examples. Note that the proposed technique produces a very similar result to the binarized, important so that it is useful for text recognition, but can be observed how grayscale information is preserved.

4 The ESPOSALLES Corpus

In this paper the experiments have been carried out on the publicly available ESPOSALLES[1] database [8]. Most specifically, we used the LICENSES part, which was compiled from a handwritten marriage license book conserved at the Archives of the Cathedral of Barcelona.

The book was written by only one person between the years 1617 and 1619 in old Catalan. It has 173 pages in total. Figure 2 presents an example page. These

[1] The corpus is publicly available at: http://www.cvc.uab.es/5cofm/groundtruth



(a) (b) (c)

Fig. 1. Example images that illustrate the difference obtained by the Sauvola thresholding and the proposed technique. Columns: (a) original image, (b) binarized and (c) proposed. Top row: Gustave Courbet's portrait. Bottom row: example from [11].

pages contain 5,447 lines grouped in 1,747 licenses. The whole manuscript was transcribed line by line by an expert palaeographer. The complete annotation of LICENSES contains around 60,000 running words from a lexicon of around 3,500 different words. More information can be found at [8].

5 Experimental Results

The objective of the experimentation was to compare the different methods for the image filtering pre-processing step (as described in Section 2), while keeping the rest of the HTR system fixed. The parameters for both the Sauvola thresholding and the proposed algorithm were chosen by manually analyzing the resultant images for a few examples. With this visual inspection it was verified that for the selected parameters that when the thresholding discarded parts of the strokes, the proposed algorithm was capable of keeping part of that lost information. The final parameters were $W = 30$ pixels, $R = 128$, and $k = 0.2$, which are common to both algorithms, and for the proposed algorithm the slope parameter of the linear transition was set to $s = 1$. With these parameters the performance of the complete handwriting recognition system was estimated by a 7-fold cross-validation procedure using the partitions proposed in [8]. The rest of the processing steps and training of the models was exactly the same as the ones from [8].

The quality of the transcription is given by the well known *Word Error Rate* (WER). It is defined as the minimum number of words that need to be

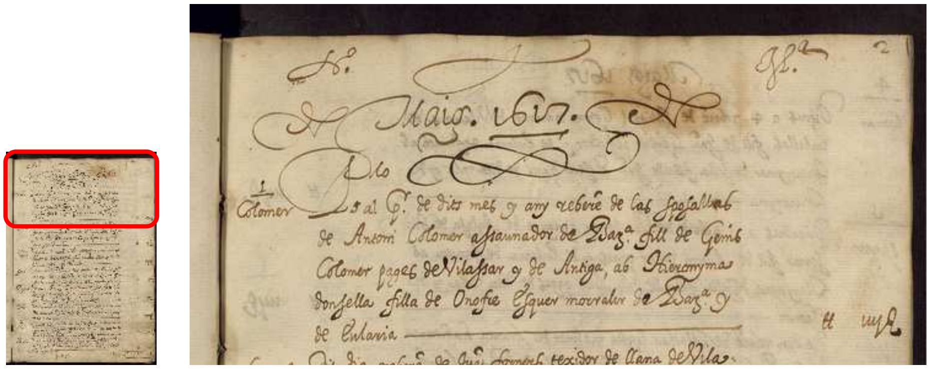


Fig. 2. Example of an ESPOSALLES marriage license page.

Table 1. Word error rate results for the LICENSES of the ESPOSALLES corpus comparing the proposed technique with the baseline and previously published results.

Approach	**WER (%)**	**95% conf. int.**[*]
Previously published [8]	16.1	15.8–16.4
with Sauvola thresholding	13.4[**]	13.1–13.7
with proposed technique	13.1[**]	12.8–13.4

[*] Wilson interval estimation.

[**] Proposed method better than Sauvola for a confidence level of 90% using a two-proportion z-test.

substituted, deleted or inserted to convert the sentences recognized by the system into the reference transcriptions, divided by the total number of words in these transcriptions. Table 1 shows the WERs obtained for the three different methods studied. First the result from [8], then replacing the background removal by a Sauvola thresholding, and then again replacing the background removal by the proposed technique. The first thing to note in the table is that for this corpus, the previously used background removal technique performs considerably worse than local thresholding, obtaining a relative improvement of about 16.8%. Then comparing the results of the proposed technique and the thresholding, there is a smaller, although still significant improvement. From this we can observe that the hard decision that a thresholding technique imposes, definitely discards useful information that benefits a recognizer based on HMMs and Gaussian mixtures.

In Figure 3 a few example images are presented comparing the original, and the result after applying the thresholding or the proposed technique. In these images it can be observed that strokes that are locally lighter than other neighboring strokes, with thresholding there is a risk that these are discarded. On the other hand, with the proposed technique there is a much lower chance that locally light strokes be completely removed, instead these are mapped to

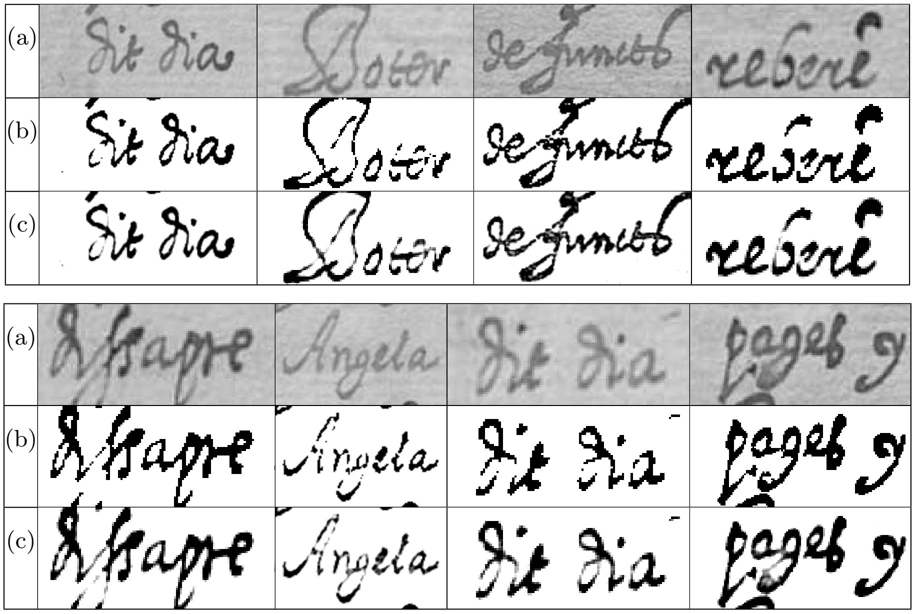


Fig. 3. Example images from the ESPOSALLES corpus comparing (a) the original image with (b) the thresholded image and (c) the proposed approach.

a shade of gray. Furthermore, the handwritten text looks smoother and more natural since all of the boundaries between background and strokes have a softer transition.

6 Conclusions

In this paper we have presented a new image pre-processing technique designed for extracting handwritten text from noisy backgrounds and improving the quality of the image making it more legible. The method is based on the well known Sauvola thresholding technique [10], although with the key difference that the objective is to obtain a grayscale image instead of simply a bitmap. This diminishes the risk that a relatively light writing stroke be completely removed due to a hard black or white decision of thresholding. In the experiments performed, the first finding was that a local thresholding technique performs much better than previously used methods based on background estimation and subtraction on the ESPOSALLES corpus. The second finding was that as expected, avoiding the conversion of the images to bitmaps, more information from the original image is preserved, which does result in an improvement in recognition performance for an HMM Gaussian mixture-based HTR system. Presented here is only a small experiment, thus in future works a more extensive experimentation should be carried out trying out several corpora and analyzing the effect of the parameters of the algorithm. Also similar modifications could be proposed for other thresholding methods that have been more recently proposed in the literature such as the ones mentioned in [4].

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 600707 - tranScriptorium and the Spanish MEC under the STraDA project (TIN2012-37475-C02-01).

References

1. Drida, F.: Towards restoring historic documents degraded over time. In: Proc. of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL 2006), pp. 350–357. Lyon, France (2006)
2. Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(5), 855–868 (2009)
3. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1998)
4. Khurshid, K., Siddiqi, I., Faure, C., Vincent, N.: Comparison of niblack inspired binarization methods for ancient documents. In: Berkner, K., Likforman-Sulem, L. (eds.) 16th Document Recognition and Retrieval Conference, DRR 2009. SPIE Proceedings, vol. 7247, pp. 1–10. SPIE, San Jose, CA, USA (January 18-22 2009), doi:10.1117/12.805827
5. Kneser, R., Ney, H.: Improved backing-off for m-gram language modeling. vol. 1, pp. 181–184. Detroit, USA (1995)
6. Marti, U., Bunke, H.: Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. IJPRAI 15(1), 65–90 (2001)
7. Niblack, W.: An Introduction to Digital Image Processing, pp. 115–116. Prentice-Hall, Englewood Cliffs, NJ (1986)
8. Romero, V., Fornés, A., Serrano, N., Sánchez, J., Toselli, A., Frinken, V., Vidal, E., Lladós, J.: The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. Pattern Recognition 46, 1658–1669 (2013), doi:j.patcog.2012.11.024
9. S. España-Boquera and M.J. Castro-Bleda and J. Gorbe-Moya and F. Zamora-Martínez: Improving offline handwriting text recognition with hybrid hmm/ann models. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(4), 767–779 (2011)
10. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern Recognition 33(2), 225–236 (2000), doi:10.1016/S0031-3203(99)00055-2
11. Shafait, F., Keysers, D., Breuel, T.M.: Efficient implementation of local adaptive thresholding techniques using integral images. In: Proc. SPIE 6815, Document Recognition and Retrieval XV, 681510. pp. 1–6 (Jan 2008), doi:10.1117/12.767755
12. Toselli, A.H., Juan, A., Keysers, D., González, J., Salvador, I., Ney, H., Vidal, E., Casacuberta, F.: Integrated Handwriting Recognition and Interpretation using Finite-State Models. Int. Journal of Pattern Recognition and Artificial Intelligence 18(4), 519–539 (June 2004), doi:10.1142/S0218001404003344