Image Thresholding of Historical Documents Based on Genetic Algorithms

Carmelo Bastos Filho, Carlos Alexandre Mello, Júlio Andrade,
Marília Lima, Wellington dos Santos, Adriano Oliveira and Davi Falcão
Department of Computing and Systems, University of Pernambuco
Brazil

1. Introduction

Digital Libraries have been developed nowadays as a way to dispose digital information through the Internet. This is particularly very useful when the information comes from historical documents. This research takes place in the PROHIST Project [Mello et al., 2008] which aims the creation of a digital library with methods to preserve and broadcast images of historical documents. In general, the access to original documents has to be done carefully as, because of its age, the paper is more susceptible to the wear and tear over time. In order to make the documents more easily accessible, digitization comes as the most efficient solution. In a digital media, as digital images, the documents can be visualized and copied. This also helps the preservation of the documents as they are digitized in high resolution and in true color format. It is common to use JPEG file format (Sayood, 1996) to store these images ensuring a good space storage/quality ratio. However, even in this format, to access an archive of thousands of high quality true color images is not an easy task even with the extended use of broad band Internet.

The storage space of the images can be reduced with its conversion to black-and-white images. In this bi-level format and stored using GIF file format, the size of the file can be five times lower than the original true color JPEG image. Binarization or thresholding (Parker, 1997) is the process that converts an image into black-and-white: a threshold value is defined and the colors above that value are converted into white, while the colors below it are converted into black. This is a very simple process in digital image processing when one has a document with black ink written on a white paper. Historical documents, however, have several types of noises. The degradation yellows the sheet of paper and creates some noise that is perceptible to the digitizing process. Even more, in some cases, the ink has faded. This is particularly important when the document is written on both sides of the paper. In some cases, the ink of one side interferes in the other creating an effect called "ink bleeding". Because of these problems, it is very difficult to find the best threshold value that separates the colors that belong to the paper from the colors that belong to the ink. An example of such a document is presented in Figure 1-left.

In this paper, we present a new thresholding algorithm for color quantization based on genetic algorithms and image fidelity metrics. These metrics are used to define the convergence point of the genetic algorithm. The quantized image is then binarized based on the number of classes defined in the quantization phase. The algorithm is adjusted to work on images of historical documents. Figure1-centre presents the final bi-level image of Figure 1-left. This is the final kind of image that we are looking for in our method. Fig. 1-right presents the results of an incorrect choice of a threshold value. The comprehension of the foreground text is severely damaged.

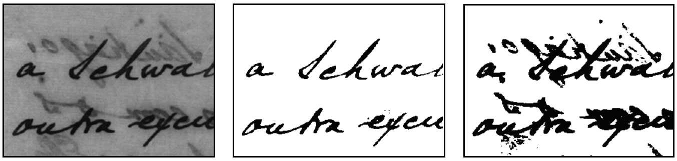


Fig. 1. (left) Zooming into part of an historical document written on both sides of the paper, presenting the ink bleeding effect, (centre) the ideal bi-level image and (right) an incorrect threshold value can create a highly noisy image.

In the next Section, we briefly review some important aspects of color quantization, genetic algorithms and image fidelity assessment. Section 3 describes our method while Section 4 presents and analyzes the results. Section 5 concludes the Chapter.

2. Fundamentals

2.1 Color quantization

Color quantization (Parker, 1997) is the process of selecting a set of significant colors to represent an image. This must be necessary to show images in devices which have limited color support or broadcast capacity as hand held devices as PDA's (Personal Digital Assistants), mobile phones, etc. After a color quantization, an image has its color resolution decreased to a specific quantity. This process can reduce the quality of the image if the process were not applied with high precision and specific algorithms. The images produced by the color quantization must be as similar as possible as the original ones which can be evaluated with the use of image fidelity indexes.

Color quantization algorithms can be classified into two classes: splitting algorithms and clustering algorithms. Splitting algorithms divide the color space of an image interactively into disjoint cells according to some criteria until the number of desired cells is reached. Some of the splitting algorithms are: popularity algorithm (Heckbert, 1982), median-cut (Heckbert, 1982), error diffusion (Kang, 1999), Floyd-Steinberg (Kang, 1999), and Stucki (Stucki, 1981). Clustering-based algorithms perform a clustering of the color space into K-desired clusters. The methods involve an initial selection of color map followed by repeatedly updating cluster representatives. C-Means (Parker, 1997) is the most common algorithm of this class.

Algorithms for dithering also work with the reduction of the color space. A review about dithering algorithms is shown in (Alasseur et al, 2003). Dithering is not the best solution for some applications as image processing of historical documents where the background (the paper) must be removed for a character recognition process. A dithering algorithm based on genetic algorithms (GA) and K-Means is proposed in (Freisleben & Schrader, 1997).

The use of computational intelligence for quantization is not new: a technique based on Competitive Hopfield Neural Networks is presented in (Wu et al. 2001). However, this algorithm converges rapidly but it easily finds a local minimum as the solution.

Another problem associated with color quantization is the analysis of the performance of color quantization algorithms. The authors in (Tremeau et al., 1994) define two metrics: LSE (Local Squared Error) and SCAP (Spatial Correlation Among Pixels). Nowadays, however, there are most appropriated metrics as the use of the concepts of image fidelity indexes (Janssen, 2001). We use herein image fidelity metrics in order to evaluate the similarity of the quantized image and the original one.

Thresholding or binarization is a specific type of color quantization that reduces the color palette to just two colors; in general, black and white tones. Some well-known algorithms are: Pun, Kapur, Renyi, Brink, Otsu, Kittler, Percentage of Black and C-Means, for example. Details on these and other algorithms can be found in (Sezgin & Sankur, 2004).

2.2 Image fidelity assessment

Image quality can be defined (Janssen, 2001) “in terms of the satisfaction of two requirements: usefulness (i.e. discriminality of image content) and naturalness (identifiability of image content)”. When one has two images and wants to compare them, it is a fidelity value that is searched. This is the main problem of fidelity metrics: the requirement of two images (a reference image and a target one) to make this comparison.

The definition of image fidelity metrics is subject of several studies that come from subjective measures as Mean Opinion Score (MOS) to objective ones as Peak Signal-to-Noise Ratio (PSNR) and Mean Square Error (MSE). Our interest is in objective measures as our work involves sets of thousand of images. PSNR (in dB) is evaluated by:

$$PSNR = 10\log_{10}(\frac{C^2}{MSE}) \tag{1}$$

where C represents the maximum color value (for images).

A fidelity index, *Q*, is defined in (Wang & Bovik, 2001) in terms of the linear correlation coefficient and the similarities between the mean and variance of two images. This index is defined as:

$$Q = \frac{4.\mu_x.\mu_y.\sigma_{xy}}{(\mu_x{}^2 + \mu_y{}^2).(\sigma_x^2 + \sigma_y^2)} \tag{2}$$

where *x* and *y* are the original and tested images respectively, μ_x and μ_y are their means, σ_x and σ_y are their variances and σ_{xy} is the correlation. As defined in (Wang & Bovik, 2001), the range of *Q* is [-1, 1]. The value of 1 happens when the images are the same (or $y_i = x_i$, for every *i*). The lowest value, *i.e.* $Q = -1$, occurs when $y_i = 2.\mu_x - x_i$, for every *i*.

2.3 Genetic algorithms

Genetic algorithms are very useful to solve search problems (Mitchell, 1998), especially for complex, multivariable and non-analytical problems. Therefore, it can be used to solve problems such as identify grayscale levels and the limits between them in a quantization process. This intelligent computing technique is important for the thresholding process proposed herein.

The Genetic Algorithm used in our method follows the flow chart presented in Fig. 2. First, an initial population with *P* individuals is created. In our method, each individual represents a set of grayscale levels involved in the process codified in a bit string. For each individual, a simulation tool is run and the fitness function value is returned. Therefore, individuals with the best performance are the stronger ones. For each generation, new individuals are created through crossover processes to compose the next generation. The mutation operator helps to avoid local minimum. The selection operation finds the *P* individuals with higher fitness function and deletes the weakest individuals. At the end of the selection operation, the algorithm checks if the predefined number of generations was reached. The algorithm keeps running until it reaches the limit of generations or it finds a predefined condition.

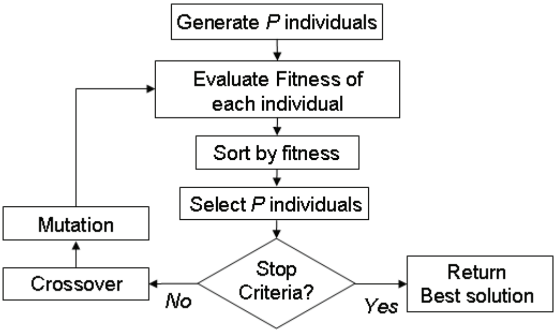


Fig. 2. Flow chart of the genetic algorithm used in our simulations.

The crossover operation is applied to two individuals and it generates two new individuals mixing information of the parents bit strings. Only new individuals are added to the population, clones are discarded. Two individuals in the population have a probability P_c of performing crossover in each generation. In this work we used $P_c = 50\%$.

The mutation operation is applied in a single individual. It consists of complementing bits in the individual string of bits. An individual in the population have a probability P_m of suffering mutation in each generation and, considering mutation in an individual, each bit has a probability of suffering mutation P_{mb}. We used $P_m = 5\%$ and $P_{mb} = 10\%$ in our simulations.

3. Proposed thresholding algorithm

The main objective of this method is to generate a bi-level image from a historical document. This image shall represent the text as black and background or back-to-front interference as white. So forth, we introduced an intelligent algorithm using genetic algorithm to quantize the original image. The target is to reduce the palette so that the remaining colors (or gray levels) represent classes such as text, background or back-to-front interference. Notice that those classes can be represented by more than one color.

Furthermore, supposing that the text is composed by the darkest grayscale levels, the process is followed by a threshold to classify the pixels as text or non-text, excluding the background and the back-to-front interference. The thresholding divides the classes as text (darkest remaining colors) and non-text.

We used historical documents images stored in 256 gray levels. The information about each gray level that represents a class was codified into a binary bit string. We also codified the limits that define the threshold in the quantization process. We used 8 bits to represent each color and each limit. Therefore, the individual consists of a bit string with *2z-1* segments, where *z* is the number of gray levels in the novel palette. We used Wang and Bovik's fidelity index (*Q*) as the fitness function. The fitness for an individual is obtained by performing the *Q* factor comparison between the image after the quantization process and the original image. We observed that the *Q* values obtained in our simulation were always in the interval [0, 1]; so this range is considered instead of [-1, 1] as defined by Wang and Bovik. After crossover or mutation, the individual segments shall be sorted according their values.

To illustrate the method we present an example: consider the Figure 1-left as the document to be treated. Figure 1-center shows the best result achieved from binarization and manual exclusion of non desired information. Therefore the target is to define the method to achieve automatically images as showed in Figure 1-center without *a priori* knowledge.

The first step is to execute the quantization based on genetic algorithm, maximizing the *Q* factor.

The second step is to define which of the new gray levels from the quantization should be classified as text. If one considers the image presented in Fig. 1-Left quantized to 4 gray levels, there are three different possibilities to threshold that image: the first one is to classify the darkest grayscale level as text and the others as non text (up ahead classify just the darkest grayscale level will be called *limit1*); the second possibility is to classify the two darkest grayscale levels as text and the others as non text (up ahead classify the two darkest grayscale levels will be called *limit2*). The last possibility is to classify the three darkest grayscale levels as text and the other as non text (from now on this classification will be called *limit3*). Figure 2 shows the resultant images after binarization considering these three limits. The threshold limits were defined as 53, 106 and 137, for limit1, limit2 and limit3, respectively. Visually, the best image was achieved for limit1.

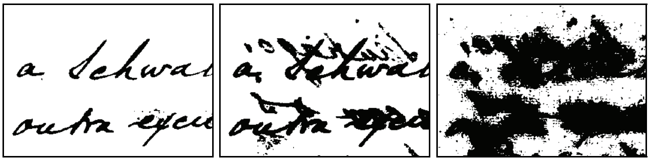


Fig. 3. Thresholding images of the example for binarization with threshold (left) 53, (center) 106 and (right) 137, after the intelligent quantization.

4. Results

It is necessary to determine how many gray levels should be used in the quantization to achieve the best results. For the first step (quantization), the *Q* factor compared with the original grayscale image increases as the number of the gray levels increases. We achieved *Q* = 0.949594, *Q* = 0.967749, *Q* = 0.978475, *Q* = 0.980648, *Q* = 0.985494, *Q* = 0.987862, *Q* = 0.987115, *Q* = 0.99008, *Q* = 0.991768 and *Q* = 0.992203 for 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 gray levels, respectively. But it must be observed that as our final purpose is binarization as the number of gray levels increases more difficult is to find the better threshold value.

A visual approach is not the best way to quantify the quality of an image. To provide an efficient way to validate the method and to find the optimum point, we compared *Q* factor, PSNR (Peak Signal-to-Noise Ratio) and ROC (Receiver Operating Characteristics) curves of the generated images to the perfect binary image of the example. Figure 3 shows the *Q* factor and PSNR results of the binarized images as a function of the number of grayscale levels in the GA quantization and the number of grayscale levels classified as text (the darkest ones). The best results were obtained for 3 gray levels in the quantization process and limit1 (*Q* = 0.577892 and PSNR = 15.18) and 10 gray levels in the quantization process and limit2 (*Q* = 0.5533 and PSNR = 15.86). The second option presents the best *Q* factor and PSNR. In spite of this, all the three options are quite similar and the first option seems to be more adequate to generalize the data for a group of documents.

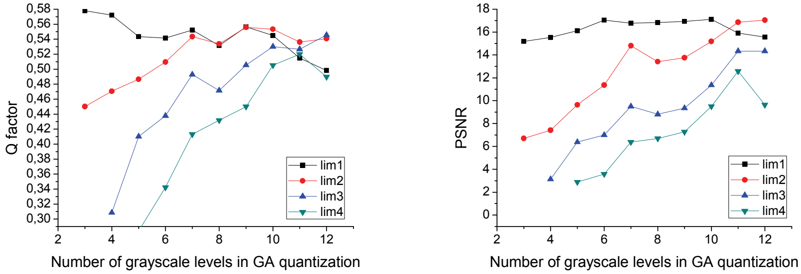


Fig. 3. (left) Q factor and (right) PSNR results as a function of the number of grayscale levels in the GA quantization for each limit case defined.

Another way to evaluate the method is using some measures from Signal Detection Theory (McMillan & Creelman, 2005): precision, recall, accuracy and specificity. Figure 4 presents the plotting of these measures for limit1 and limit2 cases.

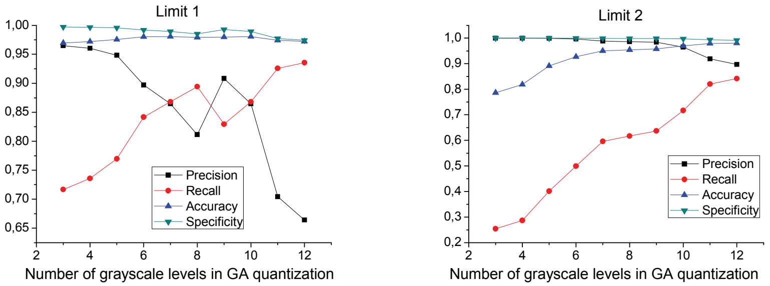


Fig. 4. Precision, recall, accuracy and specificity for (left) Limit 1 and (right) Limit 2 cases as a function of the number of grayscale levels in the GA quantization for each limit case defined.

In order to evaluate the performance of the algorithm using these metrics, a “clean” image was produced for each image in a set of 140 documents. This “clean” image is a bi-level document with only the pixels of the ink. These images were generated manually by visual inspection. With these clean images, we can evaluate the values precision, recall, accuracy and specificity. An efficient algorithm must have these four measures tending to 1.

Table 1 presents the average result for these measures in a comparison between a set of 140 documents binarized by classical algorithms (Sezgin & Sankur, 2004) and the best response achieved by the new proposed algorithm (images with the higher *Q* value) with their "clean" images. The new algorithm (labeled as GA) achieved high values for all four measures. Table 1 also presents the average values of PSNR and MSE for this set.

Algorithm	Precision	Recall	Accuracy	Specificity	PSNR	MSE
GA	**0.8290**	**0.8273**	**0.9851**	**0.9694**	**21.5167**	**0.0306**
Bernsen	0.839	0.8135	0.9636	0.9847	21.4086	0.0364
Brink	0.8972	0.7176	0.9378	0.9898	20.5486	0.0622
C-Means	0.9656	0.4752	0.7276	0.9919	15.5124	0.2724
daSilva-Lins-Rocha	0.8882	0.7277	0.9459	0.9908	20.3931	0.0541
Fisher	0.9085	0.7274	0.9329	0.9903	20.5406	0.0671
Huang	0.9174	0.6946	0.9146	0.9904	19.8345	0.0854
Johannsen	0.9398	0.6548	0.9317	0.9942	19.3038	0.0683
Kapur	0.0148	0.4073	0.9021	0.9013	16.8976	0.0979
Kittler	0.8587	0.7982	0.9429	0.9847	21.1809	0.0571
Li-Lee	0.872	0.7313	0.9447	0.9899	20.3123	0.0553
Mean Grey Level	0.9116	0.7518	0.9654	0.9891	21.11	0.0346
Niblack	0.862	0.3911	0.8531	0.985	14.6877	0.1469
Otsu	0.8431	0.7128	0.9422	0.9683	19.1462	0.0578
Percentage of Black	0.9902	0.2156	0.6165	0.9985	10.3754	0.3835
Pun	0.941	0.6513	0.9395	0.9957	19.3791	0.0605
Renyi	0.9894	0.1716	0.3683	0.9011	8.6996	0.6317
Sauvola	0.8925	0.1888	0.6045	0.9755	10.0915	0.3955
Iterative Selection	0.3368	0.3356	0.9418	0.9514	18.9542	0.0582
TwoPeaks	0.9402	0.6288	0.9236	0.9947	18.7636	0.0764
White	0.3855	0.757	0.7348	0.9524	15.6284	0.2652
Wu-Lu	0.7792	0.7711	0.9411	0.9595	19.7098	0.0589
Yager	0.7985	0.2502	0.7475	0.9609	12.0418	0.2525
Yen	0.5276	0.8058	0.9358	0.944	19.0102	0.0642

Table 1. Average value of precision (P), recall (R), accuracy (A), specificity (S), PSNR and MSE evaluated by a comparison of bi-level images without noise and the images generated by classical algorithms and the new proposal (labeled GA).

7. Conclusion

It is presented in this paper a technique for automatic thresholding images of historical documents, in special, documents written on both sides of the paper, presenting back-to-front interference. The new method uses genetic algorithms to achieve a quantized image and proceed with a binarization. The resulting images were analyzed using a fidelity index, PSNR and measures from signal detection theory. The method can also be extended to optimize quantization processes for other types of images.

The method proved to be more efficient than several other classical thresholding algorithms in an evaluation using precision, recall, accuracy and specificity.

Currently the bi-level images are being used to improve several steps on an optical character recognition process of these documents such as text segmentation and the recognition *per se*.

8. References

Mello, C.A.B.; Oliveira, A.L.I.; Sanchez, A. Historical Document Image Binarization, *Proceedings of the International Conference on Computer Vision Theory and Applications*, pp. 108-113, ISBN 9789898111210, Funchal, January 2008, INSTICC, Portugal.

Sayood, K. (1996). *Introduction to data Compression*, Morgan Kauffman, ISBN 1558603468, San Francisco.

Parker, J.R. (1997). *Algorithms for Image Processing and Computer Vision*, John Wiley and Sons, ISBN 0471140562, New York.

Heckbert, P. (1982). Color image quantization for frame buffer display. *ACM SIGGRAPH Computer Graphics*, Vol. 16, No. 3, (July 1982), pp. 297-307, ISSN 00978930.

Stucki, P. (1981). MECCA - A multiple error correcting computation algorithm for bilevel image hardcopy reproduction, *Research Report RZ1060*, IBM Research Laboratory, Zurich, Switzerland.

Alasseur, C. ; Constantinides, A.G. ; Husson, L. (2003). Colour Quantisation Through Dithering Techniques, *Proceedings of International Conference on Image Processing*, pp. 469-472, ISBN 0780377516, Barcelona, September 2003, IEEE Press, New Jersey.

Freisleben, B.; Schrader, A. (1997). Color Quantization with a Hybrid Genetic Algorithm, *Proceedings of the International Conference on Image Processing and its Applications*, pp. 86-90, ISBN 085296692X, Ireland, July 1997, IEEE Press, New Jersey.

Wu, Y.; C.Yang; T.Wang. (2001). A New Approach of Color Quantization of Image Based on Neural Networks, *Poceedings of International Joint Conference on Neural Networks*, pp. 973-977, ISBN 0780370465, Washington, USA, July 2001, IEEE Press, New Jersey.

Tremeau, A.; Calonnier, M. ; Laget, B. (1994). Color Quantization Error in Terms of Perceived Image Quality, *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, pp. V93-V96, ISBN 078031775094, Adelaide, Australia, June 1994, IEEE Press, New Jersey.

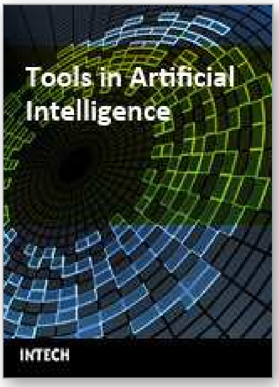
Janssen, T.J.W.M. (2001). Understanding image quality, *Proceedings of the International Conference on Image Processing*, pp. 7, ISBN 0780367251, Thessaloniki, Greece, October 2001, IEEE Press, New Jersey.

Sezgin, M.; Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation, *Journal of Electronic Imaging*, Vol. 1, No.13, (January 2004), pp. 146-165, ISSN 10179909.

Wang, Z.; Bovik, A.C. (2002). A Universal Image Quality Index. *IEEE Signal Processing Letters*, Vol. 9, No. 3, (March 2002), pp. 81-84, ISSN 10709908.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms*, MIT Press, ISBN 0262631857, Cambridge.

McMillan, N.A.; Creelman, C.D. (2005). *Detection Theory*. LEA Publishing, ISBN 0805842306, New Jersey.



Tools in Artificial Intelligence
Edited by Paula Fritzsche

ISBN 978-953-7619-03-9
Hard cover, 488 pages
Publisher InTech
Published online 01, August, 2008
Published in print edition August, 2008

This book offers in 27 chapters a collection of all the technical aspects of specifying, developing, and evaluating the theoretical underpinnings and applied mechanisms of AI tools. Topics covered include neural networks, fuzzy controls, decision trees, rule-based systems, data mining, genetic algorithm and agent systems, among many others. The goal of this book is to show some potential applications and give a partial picture of the current state-of-the-art of AI. Also, it is useful to inspire some future research ideas by identifying potential research directions. It is dedicated to students, researchers and practitioners in this area or in related fields.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821