Quantile Linear Algorithm for Robust Binarization of Digitalized Letters

Marte Ramírez, Ernesto Tapia, Marco Block and Raúl Rojas
Freie Universität Berlin
Institut für Informatik
Takustr. 9, 14195 Berlin, Germany
{marte,tapia,block,rojas}@inf.fu-berlin.de

Abstract

We describe a threshold-based local algorithm for image binarization. The main idea is to compute a transition energy using pixel value differences taken from a neighborhood around the pixel of interest. By filtering the pixels with low positive and negative energy, we keep two subsets in the neighborhood, corresponding to higher positive and negative energy values. The binarization threshold is calculated using a statistical model of the high energy pixels. Experiments show that this new approach is faster and better than current state-of-the-art algorithms.

1. Introduction

Binarization is one of several steps used in most document image analysis systems. It consists on labeling each pixel in an image as foreground and background. The former is a pixel subset $\mathcal{F}$ that represents a region of interest containing forms and objects used for further document analysis and recognition, while the latter $\mathcal{B}$ is the complement of $\mathcal{F}$.

Most binarization algorithms for gray-scale images label pixels as foreground, if their gray values are above some threshold. Global algorithms extract the information from the whole image and use only one threshold for all pixels, while local binarization algorithms compute a threshold by extracting information from a neighborhood of the pixel [6].

In other words, local algorithms compute a *threshold surface* over the whole image: a pixel (i, j) belongs to the foreground, if its intensity $I(i, j)$ is higher than the threshold value $T(i, j)$. A classical local algorithm is based on Niblack's method [5]. He computes the threshold value as

$$T(i, j) = \mu(i, j) + \gamma\sigma(i, j), \quad (1)$$

where $\mu(i, j)$ and $\sigma(i, j)$ are the mean and standard deviation of image intensities, which are taken from a neighborhood of the pixel (i, j). An ideal neighborhood size maintains a trade-off between local-detail preservation (small size) and noise elimination (large size). Kavallieratou [2] developed an iterative algorithm that uses the mean gray value $\mu(i, j)$. In each iteration, the pixels with an intensity value above $\mu(i, j)$ are removed. The remaining pixels are considered as the current foreground and used to recompute $\mu(i, j)$. This process is repeated until a fixed number of iterations is reached.

Li et al. [3] developed a method based on the discrete Laplacian

$$\begin{aligned}\nabla^2 I(i, j) \quad = \quad & I(i + 1, j) + I(i - 1, j) + \\ & I(i, j + 1) + I(i, j - 1) - 4I(i, j). \quad (2)\end{aligned}$$

They use the sign change in the Laplacian at different image scales as a criterion to select edge pixels from the image. The remaining edge points are used to construct a *co-occurrence matrix*, which is the base to compute the binarization threshold.

The key ideas in the methods are, apart from theoretical and implementation details, the following. Niblack and Kavallieratou *use statistics on a neighborhood* of the pixel of interest, while Li et al. *define a criterion to select pixels* that have high information content.

The method we present in this paper combines the above mentioned ideas. We define a criterion to select pixels in the image using an *energy function* whose computation depends on an small neighborhood of the pixel. We select the pixels that reach extreme positive and negative energy values, and compute some statistics that are used to compute the threshold used for the binarization.

2. Transition pixels and energy function

The ideal situation for a binarization algorithm occurs when there is a large difference between pixel values of foreground and background, and the rest of the pixel values do not differ significantly from those values. Actually, it is expected that the background and foreground correspond to the pixels that reach the maximal and minimal pixel intensity. Under these assumptions, binarization consists only on locating a region in the image, where there is a large difference between image values, computing the maximum and minimum values in the region, and finally computing some threshold that lays between them [1].

Large differences are observed at the boundary of our region of interest, because a neighborhood of these pixels is a *transition* from foreground to background. They contain more information than other pixel neighborhoods where the image values are practically constant.

The previous reasoning let us to define the *transition energy* for a pixel (i, j) as

$$E(i,j) = I_{\max}(i,j) + I_{\min}(i,j) - 2I(i,j) \qquad (3)$$

where $I_{\max}(i,j)$ and $I_{\min}(i,j)$ are the maximal and minimal pixel values in a small neighborhood $S(i,j)$ of the pixel (i,j). Observe the similarities between the energy function and (2).

Considering the transition energy and a positive integer e, the pixels fall into three sets:

- $\mathcal{E}_+(e) = \{(i,j)|E(i,j) > e\}$
- $\mathcal{E}_-(e) = \{(i,j)|E(i,j) < e\}$
- $\mathcal{E}_0(e) = \{(i,j)|E(i,j) = e\}$

It is not difficult to note that the transition energy *characterizes* the foreground and background: pixels with *high positive energy* are foreground with high probability, and the pixels with *high negative energy* are background pixels with high probability.

2.1. Global transitions and energy thresholds

To keep track of the pixels in relation to their transition energy, we compute two histograms $H(\mathcal{E}_+(0))$ and $H(\mathcal{E}_-(0))$, where $H(\mathcal{P})$ computes the gray-level vs frequency histogram over the pixel set $\mathcal{P}$. Similarly, the value $H(\mathcal{P}, i)$ is the frequency of the pixels that have a i-level energy in $\mathcal{P}$.

Let us denote with Φ the *transition pixels*, which are the pixels containing background and foreground pixels in some neighborhood. In an ideal case, there is a energy value $e_+^* > 0$ such that

$$\Phi \cap \mathcal{F} = \mathcal{E}_+(e_+^*), \qquad (4)$$

with

$$\frac{1}{|\Phi \cap \mathcal{F}|}\sum_{i=0}^{l-1} H(\Phi \cap \mathcal{F}, i) \cdot i \approx \frac{1}{|\mathcal{F}|}\sum_{i=0}^{l-1} H(\mathcal{F}, i) \cdot i, \qquad (5)$$

where l is the number of gray-level in the image I. Equation (4) is interpreted as follows. If a transition pixels belongs to the foreground, they must belong to the set of pixels with highest energy. The intersection $\Phi \cap \mathcal{F}$ is a scattering of $\mathcal{F}$, consequently average intensity values on the intersection *is an estimator* of average intensities on the foreground. Linking equation (4) and equation (5) we approach the average of $\mathcal{F}$ using the suitable $\mathcal{E}_+(e_+^*)$ set.

$$\frac{1}{|\mathcal{E}_+(e_+^*)|}\sum_{i=0}^{l-1} H(\mathcal{E}_+(e_+^*), i) \cdot i \approx \frac{1}{|\mathcal{F}|}\sum_{i=0}^{l-1} H(\mathcal{F}, i) \cdot i \qquad (6)$$

In the same way, there is a value $e_-^* < 0$ such that

$$\Phi \cap \mathcal{B} = \mathcal{E}_-(e_-^*) \qquad (7)$$

with

$$\frac{1}{|\Phi \cap \mathcal{B}|}\sum_{i=0}^{l-1} H(\Phi \cap \mathcal{B}, i) \cdot i \approx \frac{1}{|\mathcal{B}|}\sum_{i=0}^{l-1} H(\mathcal{B}, i) \cdot i \qquad (8)$$

and

$$\frac{1}{|\mathcal{E}_-(e_-^*)|}\sum_{i=0}^{l-1} H(\mathcal{E}_-(e_-^*), i) \cdot i \approx \frac{1}{|\mathcal{B}|}\sum_{i=0}^{l-1} H(\mathcal{B}, i) \cdot i. \qquad (9)$$

In the case of real images, equation (4) is not fulfilled because there are noise since some elements of $\Phi \cap \mathcal{F}$ have low positive energy or there are pixels in $\mathcal{B}$ with hight positive energy. However, in a real image, equation (4) can be rewritten as

$$|\mathcal{X}_+(e) \cup \mathcal{Y}_+(e)| = \mathcal{E}_+(e) \qquad (10)$$

where $\mathcal{X}_+(e) = \{(i,j) \in \mathcal{F}, E(i,j) > e\}$ and $\mathcal{X}_+(e) = \{(i,j) \in \mathcal{B}, E(i,j) > e\}$. Thus, the optimal parameter $e_+^* \in [0, l)$ is the one that minimizes the error

$$Error(e_+) = \sum_{i=0}^{l-1}\left(\frac{H(\Phi \cap \mathcal{F}, i)}{|\Phi \cap \mathcal{F}|} - \frac{H(\mathcal{E}_+(e), i)}{|\mathcal{E}_+(e)|}\right)^2. \qquad (11)$$

Since the set $\Phi \cap \mathcal{F}$ is unknown, the function $Error(e_+)$ cannot be computed. We have found in our experiments that the set $\Phi \cap \mathcal{F}$ fulfills:

$$\frac{|\Phi \cap \mathcal{F}|}{|\mathcal{E}_+(0)|} \leq 0.2. \qquad (12)$$

Following the reasonings expressed above, we reduced the binarization problem to find the value *energy threshold* e_+^* that fulfills the inequality

$$\frac{|\mathcal{E}_+(e_+^*)|}{|\mathcal{E}_+(0)|} \leq \alpha, \qquad (13)$$

where α is our *tunning parameter*. The main idea is now to discard pixels with an energy value below e_+, while keeping pixels an energy above e_+ for further computations. The parameter α is the most important in our algorithm, because its variation influences dramatically the binarization results. See Fig. 1.

Figure 1. Good values for α are $.82$ (left image) and $.8$ (right image).

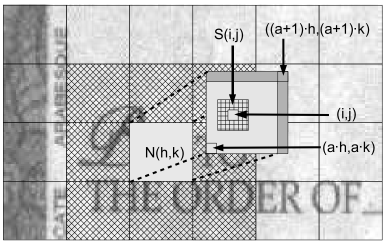


Figure 2. Neighborhood $N(h,k)$ is in gray, the neighborhood $M(h,k)$ is marked with cross pattern and the $S(i,j)$ is marked with grid pattern.

2.2. Local averaging and binarization threshold

In practice, images are far away from the ideal case, because local variations of the image intensity, such as shadows and noise, generates many potentially values for background and foreground. We deal with such situations by only considering determinated region in images, in particular, image partitions.

The partition we consider is constituted of rectangle areas, which are denoted with $N(h,k)$ and are limited by its lower left corner and upper right corner

$$(a \cdot h, a \cdot k) \text{ and } ((a+1) \cdot h, (a+1) \cdot k). \tag{14}$$

By convention, the lower and left edges are considered inside the rectangle and the upper and right edges are considered outside the rectangle, see Fig. 2.

In our computation, we consider the pixels with high energy within the rectangle $N(h,k)$ defined by

$$\Lambda_+(h,k) = \{(i,j) \in \mathcal{E}_+(e)\}. \tag{15}$$

If $\Lambda_+(h,k) \neq \emptyset$ compute:

$$\mu_+(h,k) = \frac{1}{|\Lambda_+(h,k)|} \sum_{(i,j)\in\Lambda_+(h,k)} I(i,j) \tag{16}$$

In the other case, we define $\mu_+(h,k) = 0$.

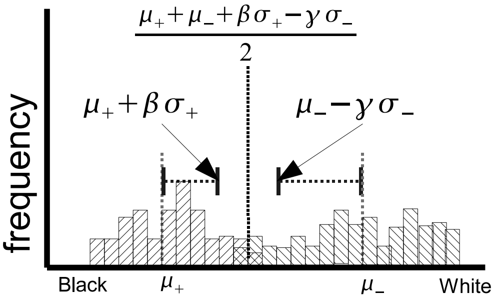


Figure 3. Quantile linear threshold.

2.3. Computing the binarization threshold

Given a rectangle $N(h,k)$, we define its positive neighborhood as

$$M_+(h,k) = \{N(i,j) | r \geq |i-h|, |j-k| \leq r, |\Lambda_+(i,j)| \neq \emptyset\}, \tag{17}$$

where r is usually one.

Now, it is possible to compute the average $\hat{\mu}_+(h,k)$ of the means $\mu_+(*)$ in the neighborhood $M_+(h,k)$ when it is not the empty set. Formally:

$$\hat{\mu}_+(h,k) = \frac{1}{|M_+(h,k)|} \sum_{N(i,j)\in M_+(h,k)} \mu_+(i,j) \tag{18}$$

In the same way:

$$\hat{\sigma}_+(h,k) = \frac{1}{|M_+(h,k)|} \sum_{N(i,j)\in M_+(h,k)} (\mu_+(i,j) - \hat{\mu}_+(h,k))^2 \tag{19}$$

Analogously, we can compute $\hat{\mu}_-(h,k)$ and $\hat{\sigma}_-(h,k)$, except the set $\Lambda_-(h,k)$ that must be

$$\Lambda_-(h,k) = \{(i,j) \in \mathcal{E}_-(-e)\}. \tag{20}$$

Finally, if the pixel (i,j) is inside the rectangle N(h,k):

$$p(i,j) = \begin{cases} 0 & : \quad I(i,j) \leq T(h,k) \\ 1 & : \quad I(i,j) > T(h,k) \end{cases} \tag{21}$$

with

$$T(h,k) = \frac{\hat{\mu}_+(h,k) + \hat{\mu}_-(h,k) + \beta \cdot \hat{\sigma}_+(h,k) - \gamma \cdot \hat{\sigma}_-(h,k)}{2}, \tag{22}$$

where β and γ are a free parameters that correspond to the parameter γ in (1) for Niblack's method, see figure 3.

2.4. Description and complexity

Our binarization algorithm consits on the following steps:

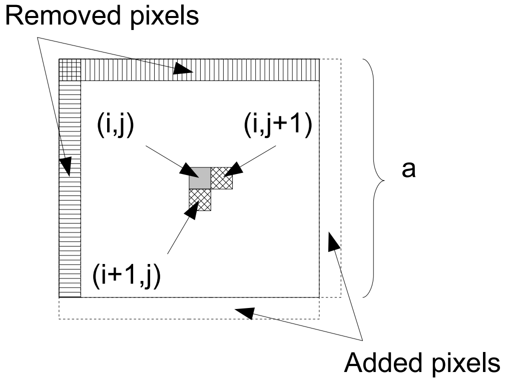


Figure 4. Using the values of histogram $H(N(i,j))$, we can compute the histograms $H(N(i \pm 1, j))$ or $H(N(i, j \pm 1))$ with effort $4 \cdot a$.

1. Compute transition energy for all the pixels.
2. Compute global transition energy thresholds (positive and negative).
3. Compute the means of gray level by neighborhoods using only pixels with high energy.
4. Compute the threshold by huge neighborhoods.
5. Cluster the pixels according his neighborhood.

We consider $|I| = n$ and $|S(i,j)| = m$. Then, computing the transition energy has a complexity $O(m \cdot n)$.

Given $S(i,j)$, computing $\mu_+(i,j)$ has a complexity a^2. Thus, to compute the mean in all the rectangles, we need an order of $O(a^2 \cdot \frac{n}{a^2}) = O(n)$ operations.

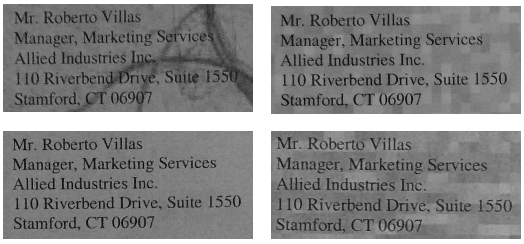
If $|M_+| = w$, then computing $\hat{\mu}_+(i,j)$ and $\hat{\sigma}_+(i,j)$ has a complexity $O(w \cdot \frac{n}{a^2} + w) = O(w \cdot \frac{n}{a^2})$. Reasonable parameter values are $m = 9$, $w = 9$ and $a = 10$. Then the final complexity is $O(9 \cdot n + n + \frac{n}{10}) = O(10.1n) = O(n)$.

3. Experimental results

Our implementations of Kavallieratou's and Niblack's algoritms use their local version, i.e. we computed a threshold $T(i,j)$ for each pixel (i,j) using a neighborhood $N(i,j)$ with size $a = 31$. Both algorithms compute the histogram of the intensity values on $N(i,j)$. Our computation is optimal, because we use the property that square neighborhoods of consecutive pixels differ from only one column or row, see Fig. 4. The Quantile Linear algorithm uses the parameters $a = 10$, $\beta = 1$ and $\gamma = 1$. Li's method was removed from our tests because it showed a very poor performance in previous experiments.

Table 1. Average Runtime.

Size	Kavallieratou	Niblack	Quantile
0.25	1123	974	32
0.5	2268	1952	63
1	3119	4384	125
2	6624	8964	246
3	9020	11827	319
4	11321	15495	418



Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Figure 5. Some examples of background.

We implemented the algorithms in C++ and ran our tests on a computer with a 3.2 GHz Pentium IV processor and 2 GB in RAM. Table 1 shows the runtime of the algorithms for different image sizes. The image are square and have a size measured in mega pixels. The runtime is expressed in milliseconds.

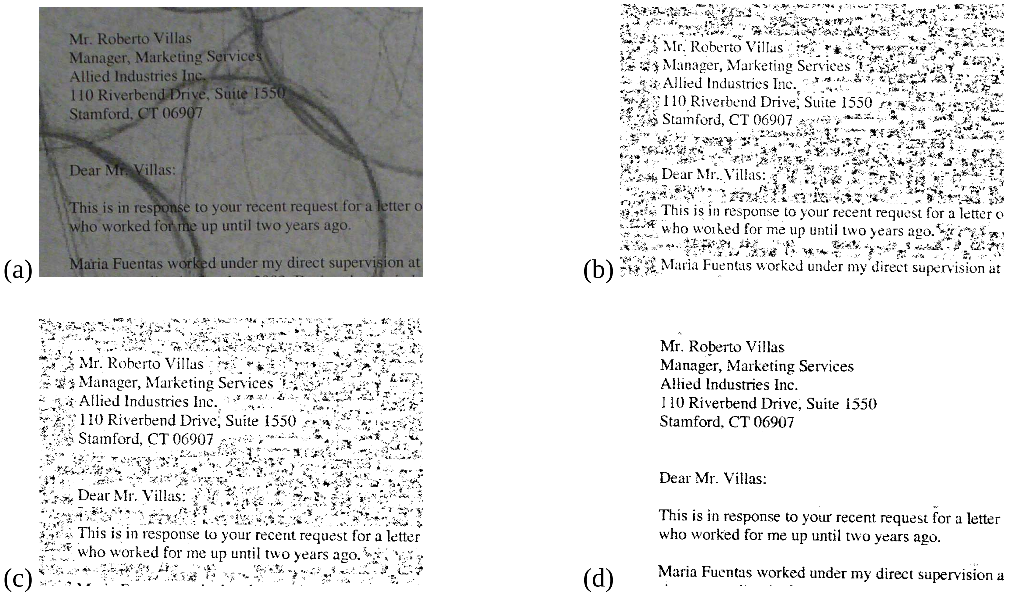
We also measure the quality of the binarization algorithms in a second experiment, using a benchmark database. The database is formed with two image sets. The first set consists of ten gray-scale photos of the same document, using different backgrounds, See Fig. 5. The second set are the binarized versions of the first set, generated by varying the parameters of the binarization algorithms, giving a total of 800 pictures, see Table 2. These images were recognized by the commercial OCR system ABBYY-FineReader version 8.0. In the last step we compare the recognized text against the original text using Needleman-Wunsch [4]. The result are summarized in Table 5.

4. Comments and further work

We presented a new binarization algorithm that combines global and local information. The global information establish criterion that selects pixels with high energy, using a global energy threshold. Given the pixel (i,j) contained

Table 2. Parameter sets.

Algorithm	Start value	End value	Increment
Kavallieratou	1	20	1
Niblack	-2.15	-0.7	0.05
Quantile Linear	0.7	0.99	0.1



Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Dear Mr. Villas:

This is in response to your recent request for a letter o
who worked for me up until two years ago.

Maria Fuentas worked under my direct supervision at

Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Dear Mr. Villas:

This is in response to your recent request for a letter o
who worked for me up until two years ago.

Maria Fuentas worked under my direct supervision at

Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Dear Mr. Villas:

This is in response to your recent request for a letter
who worked for me up until two years ago.

Mr. Roberto Villas
Manager, Marketing Services
Allied Industries Inc.
110 Riverbend Drive, Suite 1550
Stamford, CT 06907

Dear Mr. Villas:

This is in response to your recent request for a letter
who worked for me up until two years ago.

Maria Fuentas worked under my direct supervision a

Figure 6. (a) The Original image and its binarized version using (b) Kavallieratou, (c) Niblack and (d) Quantile Linear binarization. The images are only a detail of a whole letter.

Table 3. Recognition rates.

Text	Kavallieratou	Niblack	Quantile
Letter A	76.63	86.20	92.77
Letter B	86.70	94.30	96.82
Letter C	78.55	90.25	97.72
Letter D	94.13	94.09	97.39
Letter E	95.21	96.57	98.45
Letter F	88.10	91.70	97.20
Letter G	75.26	83.85	91.61
Letter H	86.41	90.66	97.15
Letter I	92.07	93.47	97.52
Letter J	91.37	93.60	97.95
Average	86.44	91.47	96.46

in $N(h,k)$, the local information uses the high energy pixels in a neighborhood $M(h,k)$ to compute the binarization threshold $T(h,k)$. Our experiments shown that our method has a high runtime performance, around 30 times faster than the other algorithms. Our algorithm also keep the highest recognition rates, when used as preprocessing step for an OCR system. Figure 4 also shows that our method also removes background artifacts that the other algorithms can not.

It could be interesting to experiment with two values of the parameter α_+ and α_-, corresponding for the positive and negative energy values. One can use another method to reduce the set of transition pixels, using for example the pixel extracted with some edge detection operator such as the Sobel filter. Other interesting research subject could be the integration of some statistical model for α and the image noise, in order to minimize optimally the error (11).

References

[1] J. Bernsen. Dynamic thresholding of grey-level images. *Proceedings of the eigth International Conference on Pattern Recognition (ICPR)*, pages 1251–1255, 1986.

[2] E. Kavallieratou. A binarization algorithm specialized on document images and photos. *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR05)*, pages 463–467, 2004.

[3] Y. Li, C. Suen, and M. Cheriet. A Threshlod Selection Method Based on Multiscale and Graylevel Co-occurrence Matrix Analysis. *Proceedings of the Eighth International Conference on Document Analysis and Recognition (ICDAR05)*, pages 575–579, 2005.

[4] S. Needleman and C. Wunsch. A general method applicable to search for similarities in the amino acid sequence of two proteins. *JMolBiol.*, pages 48(3):443–453, 1970.

[5] W. Niblack. *An Introduction to Digital Image Processing*. Strandberg Publishing Company Birkeroed, Denmark, Denmark, 1985.

[6] O. D. Trier and A. K. Jain. Goal-directed evaluation of binarization methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, 17(12):1191–1201, 1995.