Binarization of Document Images Using Hadamard Multiresolution Analysis

Fu Chang, Kung-Hao Liang, Tzu-Ming Tan, and Wen-Liang Hwang
Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.

Abstract

In this article, we propose a new method that combines the use of a global threshold and a window-based scheme for computing local thresholds. The latter scheme compares the contrast of gray values within a neighborhood whose size varies with the scale of the objects being examined. To compute the scale quantity, a new wavelet model entitled Hadamard multiresolution analysis is also proposed. When the window-based scheme is applied to the areas where global threshold is likely to fail, we obtain uniformly better binary results than using a global threshold only. Significant improvements to OCR performance can also be achieved by our binary results.

Keywords: *Binarization, document image, global threshold, Hadamard multiresolution analysis, local threshold, scale, wavelet transform, window-based dynamic binarization*

1. Introduction

In the literature, binarization methods are divided into those that compute static (global) thresholds and those dynamic (local) thresholds. For most of the former methods [9-13], they look at histograms of gray values and determine a global cutting point (the static threshold) to optimize certain statistical measures, such as between-class variance, entropy, etc. Those methods do not go back to their binary outcomes to seek clues for further improvements. In a recent work of Liu and Srihari [14], however, some features reflecting the texture property of binarized objects (such as stroke widths, degree of character brokeness, etc.) are examined and are also used as criteria for selecting candidates for thresholds.

The dynamic thresholding methods often rely upon some features extracted from images for determining the values of local thresholds. Many such methods employed tools that detect edges [1-4]. The well-known operator $\nabla^2 G$ (Marr [4]) and their variants are such examples. Documents, however, contain objects that always have widths. Edge detection helps to locate the fringe of those objects. Efforts are still needed to fill in the areas surrounded by the edges thus detected. Moreover, since the detection methods may not be able to find all the edge points, the filling operations can make mistakes sometimes.

In a paper of White and Rohrer [1], one method proposed by them starts with Sobel operators to locate character boundaries. It then proceeds to traverse through all horizontal lines to fill in the pixels between each pair of edge points encountered. In a recent paper of ours [3], which is also an edge-based approach, we extend the search for edge pairs through eight half-lines radiated from each given pixel.

Edge detection methods have to also solve the problem of scales. Back in the 70s, the problem was already manifested by Rosenfeld and Thurston [5]. Essentially, it has to do with how large a neighborhood one needs to choose for computing the contrast. The appropriate size of neighborhood relates to that of objects. Thus, a good solution for edge detection requires more than comparing the contrast.

There are also methods for computing dynamic thresholds based on other types of measures, such as contrast measure [15], or weighted running average [1]. But once again, we argue that scale is an essential issue that needs to be considered. Basically, the issue is concerning how large a variation within how wide a range should be taken as a clue for the binary-value assignment.

From our experience, binarization of document images can suffer from two sources of difficulties. First, even though clear peaks exist in the histogram of gray values, a sheer cutting point between them does not sufficiently discriminate all the blacks from the whites. Wrong decisions (judged by human eyes) on significant number of pixels can associate with any level serving as the cutting point. Secondly, a more serious problem lies in the fact that documents contain characters of various sizes. It is possible that the difference in gray scale between the interior and exterior of a small-sized character may fall below that within a large-sized character. Such low degree of difference can be found between some small white regions and their surrounding strokes that form a very busy small-sized character

In this paper, we present a new approach to document image binarization. This method resorts to the use of scale information. The latter tells us how large a window we should choose in order to compute the binary value for a given pixel. The key idea is the following. If a pixel P lies in the vicinity of a character, it suffices to compare the gray value of P with those in a neighborhood that is large enough to overlap with both the interior and exterior of the character.

Thus to solve the binarization problem, we actually tackle two sub-problems. One is to group and classify pixels that express the same scale propensity. The other is to compute the binary value for each pixel using scale as the determinant for the neighborhood size. The side benefit of this approach is that we do away with edge detection as an intermediate step.

For the first sub-problem, we are able to work systematically under a framework, called Hadamard multiresolution analysis (HADMA). This framework is adopted from a theory of signal decomposition, or better known as theory of wavelet transform, developed by S. G. Mallat [6,7]. A novel feature of HADMA is that the band-filters used in this particular framework come from Hadamard basis. HADMA is found to be useful for analyzing the structure of document images [8]. Various coarse objects (text lines, text blocks, half-tone pictures) on the images can be easily detected in certain channels of the HADMA representation. The scale information can also be derived from such a representation.

This paper will proceed in the following way. Section 2 presents a brief description of the theory of wavelets, and how it leads to the framework of HADMA. In Section 3, HADMA is used for segmenting coarse objects from document images. The scales associated with them can also be calculated. Section 4 is devoted to the proposed binarization method of this paper. In Section 5, we report the results of our experimental findings. Section 6 contains a brief conclusion.

2. Hadamard Multiresolution Analysis

The important groundwork laid down by S. G. Mallat is a wavelet transform model that represents signals in terms of a family of transformed quantities [6-7]. In this model, the *2-D dyadic* wavelet represenation of a two-dimensional signal (or image, as we prefer to call in what follows) is associated with a low-pass filter h(.) and a high-pass filter g(.). Since we are dealing a two-dimensional image, the filters can be applied in two directions. There are four possibilities: we can first apply h(.) in the horizontal direction and then g(.) in the vertical direction, or h(.) in the vertical direction and then g(.) in the horizontal direction, etc. The four combinations of the two 1-D filters then form four different channels, and they are called low-passed (L), horizontal (H), vertical (V) and diagonal (D), respectively. The transformed values in these four channels can be obtained interatively as follows.

$$L_{j+1}(u,v)=\sum_x\sum_y L_j(x,y)h(x-2u)h(y-2v),$$

$$H_{j+1}(u,v)=\sum_x\sum_y L_j(x,y)h(x-2u)g(y-2v),$$

$$V_{j+1}(u,v)=\sum_x\sum_y L_j(x,y)g(x-2u)h(y-2v),$$

$$D_{j+1}(u,v)=\sum_x\sum_y L_j(x,y)g(x-2u)g(y-2v),$$

where x and u are horizontal coordinates, y and v are vertical coordinates, and j is the scale parameter. Note that L_0 denotes the orginal image.

Here, the scope of $(j+1)^{th}$-level quantities is the quarter size of that of j^{th}-level quantities (the down-sampling rate is 2 per dimension). Moreover, the quantity L_{j+1} is obtained by convolving h(.) with L_j in both horizontal and vertical directions. The other three quantities are obtained by convolving h(x)g(y), g(x)h(y), and g(x)g(y) with L_j, respectively. The three 2-D functions h(x)g(y), g(x)h(y), and g(x)g(y) are called *wavelets*.

The transformed values in each of the four channels occupy a quarter size of the original image. Placed together, they look like what is depicted in Figure 1a. The L channel is a smoothed and down-sampled version of the image. On top of it, one can further construct a new set of four channels. Keeping on constructing new channels at a higher level on top of the L channel of a given level, we get a pyramidal representation, similar to the Lapacian pyramid [7], as depicted in Figure 1b.

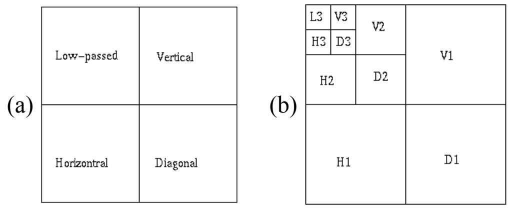

Figure 1. (a) The four channels of a one-level wavelet transform. (b) Three-level wavelet transform.

The HADMA representation is a 2-D dyadic wavelet representation which employs two Hadamard coefficients [1,1,1,1] and [1,-1,-1,1]. These coefficients are further normalized with respect to l^1 norm and are used as h(.) and g(.) filters, as described in the above.

L_{j+1} channel is obtained by the covolution of L_j with the 2-D filter h(x)h(y) and has the effect of obtaining a smoothed and down-sampled version of L_j. The H_{j+1} (V_{j+1}) channel is produced by convolving L_j with the 2-D filter h(x)g(y) (h(y)g(x)). Since h(x)g(y) (h(y)g(x)) looks like a horizontal (vertical) bar, this filter serves to detect objects on L_j that have similar shapes. As a consequence, vertical strokes would react strongly (namely, their transformed values are very positive) in some of the V channels, and horizontal strokes in some of the H channels. Diagonal strokes, on the other hand, would react in both H and V channels but their reactions are less vigorous than horizontal or vertical strokes in their corresponding channels.

3. Computation of Scales Using HADMA

First, we produce the pyramidal representation of a given document, as shown in Figure 1b.

One of our goals in this section is to find those regions such as text blocks or pictures from a given document image. To do so, we would like to work on a down-sampled version of the image. Two benefits can be derived from a down-sampled representation. (1) The scope becomes smaller. (2) Many contiguous regions become physically connected. However, the resolution level j we choose to work on has to meet the following constraint: any two heterogeneous regions in the original image should not be merged in the j^{th} resolution level. The value j is an empirical value and has to be determined by experiments. We found that j = 2 is the most appropriate level for documents produced at 300dpi.

To segment those regions from L_2 image, we do the following. First, the gray-scale histogram of the down-sampled image is used for binarization. Since at this stage the binarized outcome is not used as the final output, a static binarization scheme would meet the purpose. More words about it will be given in the next section. The black pixels in the binarized outcome are then grouped together to form connected regions.

Having decomposed the document into connected regions, we assume that each region comprises objects of the same type. Let R be such a region. Since it is a region in L_2, it will also be referred to as R_2. Now we define the corresponding regions of R in other resolution levels. The definition obeys the following principle. If R_j is a region in L_j, its corresponding region in L_{j-1} is
R_{j-1} = {q: q is in a 2×2 box in L_{j-1} that corresponds to a pixel in R_j}. On the other hand, the corresponding regions of R_j in L_{j+1} is R_{j+1} = {q: q ∈ L_{j+1} and q corresponds to a 2×2 square that contains a pixel of R_j}. Now, at each level j, let

$$H_j(R)=\frac{\sum\{H_j(u,v):(u,v)\in R_{j-1}\}}{\# R_{j-1}},$$

where the normalizing factor is the total number of pixels in R_{j-1}. Similarly, let

$$V_j(R)=\frac{\sum\{V_j(u,v):(u,v)\in R_{j-1}\}}{\# R_{j-1}}.$$

$H_j(R)$ and $V_j(R)$ are used to measure the response strength of R in the H_j and V_j channel, respectively.

Now, since the shape of a bar detector is similar to the gray-level profile of a character stroke, a horizontal (vertical) stroke of width W has strong response in some H (V) channels. Moreover, it has the strongest response in the particular H_j (V_j)

such that 2^j is closest to W. We justify the magnitude 2^j as follows. Suppose a stroke has width 2^j, its representation in L_{j-1} is 2-pixel wide. Such a width fits just well with the valley of the high-pass Hadamard filter. It therefore has very strong response in H_j or V_j, since the latter is obtained by convolving high-pass Hadamard filter with the (j-1)th-level image (the low-pass Hadamard filter serves to smooth out the responses along the orthogonal direction).

A text block contains characters of the same size and therefore of approximately the same stroke widths. Presumably, the pixels composing the horizontal strokes would pick up their strengths in some H channels and those composing the vertical strokes in some V channels. However, we do not know in advance whether the character strokes are horizontal or vertical. In fact, we do not even know exactly where they are located. So we simply sum up the response strengths of a given region R in both the H_j and V_j channels. We then combine the strengths $H_j(R)$ and $V_j(R)$ for each j, and also find the value of j at which the combined value is maximized. This value is taken as the representative scale for R. Thus, we stipulate the following definitions.

$$Strength_j(R) = H_j(R) + V_j(R),$$

$$Scale(R) = \underset{j}{aug\max}(Strength_j(R)).$$

Let us use an example to illustrate the above definitions. We pick up samples from three text blocks from a newspaper article. Each set of samples, considered as a region, contains five characters. The *Strength$_j$* values for each set are shown in Table 2. The maximum value of *Strength$_j$* for each set of samples is shown as boldface. It is seen that large-scaled samples have their maximal value at j = 3, medium-scaled samples at j = 2, and small-scaled samples at j = 1.

	Level 1	Level 2	Level 3
Large characters	2.75	3.55	**5.62**
Medium characters	3.40	**4.80**	4.37
Small characters	**5.25**	4.93	3.28

Table 2. The transformed values at different levels for characters of different sizes.

4. Window-based dynamic binarization

Scale information, among other things, is very useful for window-based dynamic binarization. The idea is the following. For each pixel p located within or near a character, we set a window centered at p, whose size is determined by the scale of the given character (or more precisely, the scale of the text block to which the character belongs). The important observation being that, within such a window, there is a fairly good mixture of background and foreground pixels. By taking advantage of this fact, we can make accurate determination of the dark/bright tone of p. Details are described as follows.

First, we resort to the static threshold T that was obtained at the step for segmenting coarse objects from the document images (cf. Section 3). The method we used is Otsu's statistic method that seeks the cutting value to maximize the variance between two resulted clusters [9]. This value divides the pixels of the image into two parts. Let M_1 denote the statistical average of the lower part (those whose gray values are less than T). We now consider the set S of all pixels p such that p lies within a 5×5 neighborhood of q with $M_1 \leqq g(q) \leqq T$, where g(q) denotes the gray value of q. For those pixels in S, we shall determine their binary values later. For the rest of the pixels, we consider them as either solid black or solid white and assign their binary values as follows.

For every pixel $p \notin S$, b(p) = 1, if $g(p)<M_1$; or b(p) = 0, if g(p)>T, where b(p) denotes the binary value of p, 1 stands for black and 0 for white.

Now, let us focus on the pixels in S. They are considered as pixels for which there is high probability for a static-threshold method to misjudge their binary values. Thus, we will further examine those pixels with a window-based binarization scheme. The window centered at each of such pixels will be chosen to be sufficiently large to cover representatives from both foreground and background. Thus, a reasonable size of the window has to be related to the scale of the objects. Note that there are some of the pixels in S whose scales have not been defined. They are those pixels whose gray values exceed T but they fall within a 5×5 neighborhood of some pixel whose scales are well-defined. If p is such a pixel and q is the center of the 5×5 neighborhood, then we simply define the scale of P as that of q.

Now, if p falls within a region whose scale has been determined as j, we will choose the window W(p) to be a neighborhood centered at p and of size 2^j+c, c being an odd number.

We determine the binary value of p by the following way. We use a static-threshold method to find the cutting point for all the gray values found within W(p). We then assign the binary value of p as black or white according to whether its gray value falls below or above this point.

5. Experimental Studies

To evaluate the binarization method, we asked human testers to scrutinize all the binary results of our method as applied to thousands of Chinese newspaper articles and business cards. Out of all the outcomes thus examined, the proposed binarization method was found to yield satisfactory quality for those documents with uniform background. Moreover, we also compare of our results on those documents with Otsu's method. We saw uniformly better performance of our method over the latter. For those documents containing small colored patches, our method can still yield acceptable results, although some drawbacks may be seen in those areas whose gray values are deviated from the average background level.

Among those testing samples, some were collected from newspapers stored in libraries for more than ten years. There are 23 of them in total. Applied to these documents, our method was seen to achieve significantly better quality than all other means. For a further comparison, we fed all the binary outcomes to a software package that has been rated as the best Chinese character recognizer. We saw uniformly higher recognition rates achieved by our results. For the same documents, the range of performance discrepancy between the two sets of binary images (ours vs. others) is more than **10%**. For an extreme case, we witnessed a **24%** discrepancy.

Let us further elaborate on one such example. A portion of the source article is shown in Figure 2a. This article came from a newspaper that dated 14 years ago. Out of this source image, we first produced a series of binary images using each possible value, ranging from 0 to 255, as global threshold. For low threshold values, the resulting images were seen to contain broken characters. As the threshold is elevated, the image suddenly jumps to one that contains smeared characters. This finding proves that no global threshold, irrespective what value being assumed, is able to yield a satisfactory binary result.

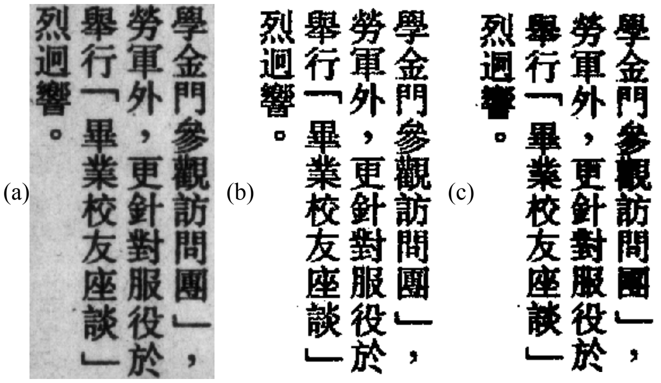

Figure 2. (a) Gray-scale image produced at 300dpi. (b) Binary image obtained by the binarization method proposed here. (c) Binary image obtained by Otsu's method.

The reason for the sudden jump is the following. The gray values of many tiny holes (small white regions) fall below the level of some stroke pixels. Thus, when the threshold is chosen below the level of tiny holes, both the holes and weak stroke pixels are classified as white, resulting in broken characters. If the threshold is raised above the level of some tiny holes, these holes as well as their surrounding strokes are classified as black, resulting in smeared characters.

In Figure 2b and 2c, we display the binarized result obtained by our method and by Otsu's method, respectively. In Figure 3, we further show the results of our binarization scheme with various window sizes. It is seen that two kinds of undesirable outcomes can result from improper choice of window sizes: either white spots emerge in large characters, or tiny holes get filled in small characters.

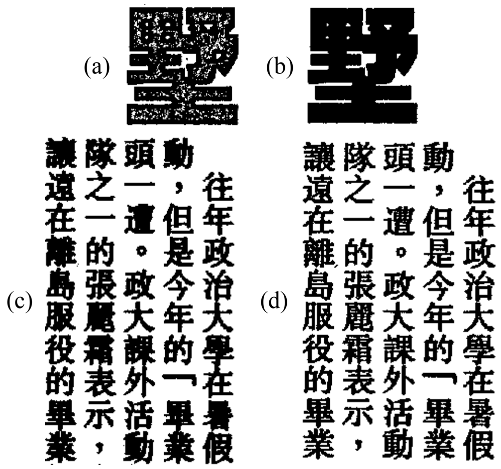

Figure 3. (a) The window sizes are set for small-scaled character, causing emergence of white spots. (b) The window sizes are set appropriately. (c) The window sizes are set for large-scaled character and characters become smeared. (d) The window sizes are set appropriately.

6. Conclusions

Two novel aspects of this paper are the following. (1) A window-based scheme is used as supplement to a global-threshold scheme in the binarization process. (2) The computation of scales is under the framework of HADMA, which is a wavelet transform model with two band-filters adopted from Hadamard basis.

The window-based scheme is found to work particularly well in the regions where objects are of small or medium scales. An adaptive use of this scheme is therefore useful for correcting the ill performance of a single global threshold.

Incidentally, the binarization work reported in this paper illustrates the value of HADMA framework for document image analysis. Basically, the multiresolution nature of such a framework is expected to be useful for documents that contain objects of various scales. The adoption of Hadamard filters add powers to this framework, since the wavelets formed by them look exactly like a stroke or a dot cluster and therefore well represent those types of objects.

References

J. M. White and G. D. Rohrer, "Image thresholding for optical character recognition and other applications requiring character image extraction," IBM J. Res. Develop. Vol. 27, No. 4, pp. 400-411, 1983.

[1] T. Pavilidis and G. Wolberg, "An algorithm for the segmentation of bilevel images," Proc. IEEE Comput. Vision Patt. Recogn. Conf., Miami Beach, pp. 570-575, 1986.

[2] W. L. Hwang and F. Chang, Character Extraction from Documents Using Wavelet Maxima, Image and Vision Computing, Vol. 16, pp. 307-315, 1998.

[3] D. Marr, Vision, Freeman, San Francisco, 1982.

[4] A. Rosenfeld and M. Thurston, "Edge and curve detection for visual scene analysis," IEEE Trans. Comput., Vol. C-20, 1971.

[5] S. G. Mallat, "A theory of multiresolution signal decomposition: the wavelet representation," IEEE Trans. Pattern Anal. Machine Intell., Vol. PAMI-11, No. 7, pp. 674-693, 1989.

[6] S. G. Mallat, "Multifrequency channel decompositions of images and wavelet models," IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-37, No. 12, pp. 2091-2110, 1989.

[7] K. H. Liang, F. Chang, T. M. Tan, and W. L. Hwang, "Multiresolution Hadamard representation and its applications to document image analysis," Proc. Intern. Conf. Multimodal Interface, pp. V1-V6, Hong Kong, 1999.

[8] N. Otsu, “A threshold selection method from gray-scaled histogram,” IEEE Trans. Systems, Man, and Cybernetics, Vol. 8, pp. 62-66, 1978.

[9] W. Tsai, "Moment-preserving thresholding: a new approach," Computer Vision, Graphics, and Image Processing, Vol. 29, pp. 377-393, 1985.

[10] G. Johannsen and J. Bille, "A threshold selection method using information measures," Proc. Sixth Intern. Conf. Pattern Recognition, pp. 140-143, Munich, Germany, 1982.

[11] J. N. Kapur, P. K. Saboo, and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," Computer Vision, Graphics, and Image Processing, Vol. 29, pp. 273-285, 1985.

[12] J. Kittler and J. Illingworth, "On threshold selection using clustering criteria," IEEE Trans. Systems, Man, and Cybernetics, Vol. 15, pp. 652-655, 1985.

[13] Y. Liu and S. N. Srihari, "Document image binarization based on texture features," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, pp. 540-544, 1997.

[14] E. Giuliano, O. Paitra, and L. Stringa, "Electronic character reading system," U.S. Patent 4,047,15, 1977.