Decompose-Threshold Approach to Handwriting Extraction in Degraded Historical Document Images

Chen Yan & Graham Leedham
School of Computer Engineering, Nanyang Technological University
N4-#2A-28 Nanyang Avenue, Singapore 639798
Email: asgleedham@ntu.edu.sg

Abstract

Historical documents contain important and interesting information. A number of techniques have previously been proposed for thresholding document images.

In this paper a new thresholding structure called the ***decompose-threshold approach*** *is proposed and compared against some existing global and local algorithms. The proposed approach is a local adaptive analysis method, which uses local feature vectors to find the best approach for thresholding a local area. Appropriate algorithm(s) are selected or combined automatically for specific types of document image under investigation. The original image is recursively broken down into sub-regions using quad-trees until an appropriate thresholding method can be applied to each of the sub-region.*

The algorithm has been evaluated by testing on 10 historical images obtained from the Library of Congress. Evaluation of the performance using 'recall' value demonstrates that the approach outperforms any existing single methods.

1. Introduction

Thresholding historical handwritten document image converts the grayscale image to binary format by separating the useful font and information from the background. Thresholding is an important step for further image analysis, such as handwriting segmentation and recognition. Historical document images are normally handwritten by ink pen tens and even hundreds of years ago. After years of storage, the documents are frequently degraded because of the different storage skills, environments, and the different quality of paper and ink.

A number of techniques have previously been proposed for thresholding document images. All previously reported thresholding methods have been demonstrated to be effective for certain classes of document images. None, however, has proven effective for all examples of 'difficult' document images. Images of historical documents offer a particularly challenging problem for the thresholding or information separation problem. Given the current state-of-the-art in computer recognition and processing of script, most historical documents are impossible to recognize automatically.

Otsu's method [1] is an early, but still popular histogram-based global threshold algorithm. It proposed a criterion for maximizing the variance of the between-class of pixel intensity to perform thresholding.

Solihin & Leedham [2] elaborated two global techniques: Native Integral Ratio (NIR) and Quadratic Integral Ratio (QIR). These two histogram shape based methods were developed from a new class: Integral Ratio, which is a new class of global thresholding technique.

Niblack [3] presented his method based on the calculation of the local mean and of local standard deviation. The threshold is decided by the formula:

$$T(x, y) = m(x, y) + K * s(x, y) \quad (1)$$

where $m(x, y)$ and $s(x, y)$ are the average of a local area and standard deviation values, respectively.

Zhang & Tan [4] proposed another improved version of Niblack's method. In comparison with the Niblack algorithm, it is good at shadow boundary detection. The threshold is determined by the formula:

$$T(x, y) = m(x, y) * \left[1 + K * \left(1 - \frac{s(x, y)}{R}\right)\right] \quad (2)$$

where K and R are empirical constants.

Bernsen [5] proposed a local thresholding technique based on neighbors.

Eikvil et al. [6] proposed an adaptive thresholding technique. Accoding to the technique, the pixels inside a small window S are thresholded on the basis of clustering of the pixels inside a large window L. The principle of the technique is that a large window L with a small window S in the center is moved across the image in a zig-zag fashion, in steps equal to the size of S. S is labeled as print or background on the basis of the clustering of the pixels inside L.

A local gradient-based thresholding algorithm was proposed by Yanowitz & Bruckstein [7]. The gradient map of the image was used to point at well-defined portions of

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

object boundaries. In Trier & Taxt (1995) [8], Yanowitz & Bruckstein's algorithm was found to be the best among eleven thresholding techniques they illustrated.

These global and local adaptive techniques have met with varying degrees of success. Whilst each has been shown effective on a particular type or class of document, none is able to produce consistently good results on the wide range of documents images that exist in the world, or on the wide range of image qualities encountered in images of historical documents.

Most existing techniques apply the same process over the whole scanned image. The result is that it is effective in some parts of the image but is less effective in others. In this paper, a new method, which selects and applies the most effective threshold method to different local regions, is proposed. The outline of the new method is:

1. Decompose image
2. Extract features from each local region
3. Local region classification
4. Repeat 1, 2 and 3 until all regions are classified
5. Threshold methods are applied to each region

The new proposed idea analyzes the feature information of the local regions with different sizes, and so accordingly applies different effective threshold methods to obtain the best result.

2. Proposed Decompose-Threshold Approach

2.1 Image Decomposition

The input image is first decomposed into four equal size sub-images if the contrast of the decomposed input image C >= Threshold, T (T is currently empirically set at a greyscale value of 180), and then the sub-image is further decomposed. It can be done in 5 steps:

1. Set the input image as sub-image.

2. Calculate the mean value of the 15-neighbours of the pixels that lie at (2*M, 2*N) coordinate of sub-image.

Where $\begin{cases} M = 1,2,...,\frac{1}{2}(row\,number\,of\ sub-image) \\ N = 1,2,...,\frac{1}{2}(column\,number\,of\ sub-image) \end{cases}$

3. Find the maximum mean value ***Maximean*** and minimum mean value ***Minimean*** in point 1.

4. C = ***Maximean – Minimean.***

If C >= T, the sub-image is decomposed into four smaller equal-sized sub-images. Practically, T =180 was the best choice for historical document images.

Repeat steps 2 to 4 until C < T, or when the sub-image reaches 64 * 64. A sub-image, which is smaller than 64 * 64, has insufficient information for further feature extraction.

2.2. Feature Extraction

The next step is to use feature vectors to extract useful information from the decomposed sub-images.

Many feature vectors have previously been used in document image binarization. Most of them were applied to printed and clear (white) background document image, which are not suitable for handling handwritten and messy background document images. New proposed Word Direction GLCM (WD-GLCM) based feature vectors are favorable for detecting information from historical document images.

A Gray Level Co-occurrence Matrix (GLCM) [9] contains information about the positions of pixels having similar gray level values. A co-occurrence matrix is a two-dimensional array, **G**, in which both the rows and the columns represent a set of possible image values.

A GLCM **Gd**[i,j] is defined by first specifying a displacement vector **d**=(dx,dy) and counting all pairs of pixels separated by **d** having gray levels i and j.

The GLCM is defined by

$$\mathbf{G_d[i,j] = n_{ij}} \qquad (3)$$

Where $\mathbf{n_{ij}}$ is the number of occurrences of the pixel values (i,j) lying at distance **d** in the image.

The co-occurrence matrix **Pd** has n*n dimensions, where n is the number of gray levels in the image.

Lam [9] described the gray-level gradient co-occurrence matrix (GLGCM). In GLGCM, $\mathbf{n_{ij}}$ is the number of occurrences of the pixel values (i,j) lying at distance **d** in the four directions: 0^o, 45^o, 90^o and 135^o.

For the handwritten document image shown in Figure 1, there are three main slant directions in handwritten words: left top to right bottom, top to bottom, and right top to left bottom. In a counter-clockwise direction, they are at 45 (or 225) degrees, 0 (or 180) degrees, and 135 (or 315) degrees.

Figure 2 shows the 8 directions of word slant in a counter-clockwise. The corresponding matrices of word directions slant can be defined as in Figure 4.

The mean direction of the input handwritten document image is determined by G0~ G7 in Figure 4.

The proposed WD-GLCM can be determined after the stroke direction is calculated as follows:

$$WDGLCM = \text{occurrences of pixel}(i,j)\ \text{lying at}:$$

$$\left[(i,j-d)\ \ ...\ \ (i,j)\ \ ...\ \ (i,j+d\right], \text{when G0};$$

$$\begin{bmatrix} ... & ... & ... & ... & (i-d,j+d) \\ ... & ... & ... & ... & ... \\ ... & ... & (i,j) & ... & ... \\ ... & ... & ... & ... & ... \\ (i+d,j-d) & ... & ... & ... & ... \end{bmatrix}, when\ G1;$$

$$\begin{bmatrix} (i-d, j) \\ \dots \\ (i, j) \\ \dots \\ (i+d, j) \end{bmatrix}, \text{when G2;}$$

$$\begin{bmatrix} (i-d, j-d) & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & (i, j) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & (i+d, j+d) \end{bmatrix}, \text{when G3.}$$

where **d** is half of the stroke width of the word in the input document image.

The co-occurrence matrix ***WD-GLCM*** has n*n dimensions, where n is the number of gray levels in the image.

- **Word Direction Based Edge Strength (DWES)**

Edge Strength can be defined based on the DW-GLCM vector as

$$WDES = \sqrt{\frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}(i-j)^2 \times WDGLCM(i, j)} = \sqrt{mean\left[(i-j)^2 \times WDGLCM(i, j)\right]} \quad (4)$$

where *i* and *j* are the coordinates of *WDGLCM*, K is the gray level of the input image.

Direction Based Edge Strength measures the gray level gradient differences in certain direction determined by the conditions of the input image. It can provide more useful information for further analysis and works more effectively than simple ES based on GLCM.

- **Word Direction Based Variance**

Word direction based variance (WDVAR) measures variability of gray value differences and hence coarseness of texture. Large value of Variance indicates large local variation. Word Direction Based Variance is defined by WDGLCM as below:

$$WDVAR^2 = \frac{1}{K-1}\sum_{i=1}^{K}\sum_{j=1}^{K}\left[WDGLCM(i, j) - \mu\right]^2 \quad (5)$$

where $\mu = \frac{1}{K^2} \bullet \sum_{i=1}^{K}\sum_{j=1}^{K} WDGLCM(i, j)$;

- **Mean-Gradient Value**

Gradient is the change of image texture along some direction in the image. From the definition, the gradient of the intensity image *I(x,y)* at location *(x,y)* defined as

$$G(x, y) = \sum_{x=0}^{i-1}\sum_{y=0}^{j-1}\frac{\left[\nabla I(x, y)/\nabla x, \nabla I(x, y)/\nabla y\right]}{x * y} \quad (6)$$

G_N is the mean-gradient value [10] of the sub-image at direction *N* given by

$$G_N(x, y) = \sum_{x=0}^{i-1}\sum_{y=0}^{j-1}\frac{\left[\nabla I(x, y)/\nabla x_N, \nabla I(x, y)/\nabla y_N\right]}{x * y} \quad (7)$$

Mean-gradient is sensitive to small variance between strokes; it can be used to detect the faint strokes between heavy strokes.

2.3 Sub-Region Classification

The above three feature vectors are used to examine the local regions and classify them into three types: *background*, *heavy strokes* or *faint strokes*. Typical examples of these three types of region are shown in Figure 5.

Background does not contain any useful information. This kind of area typically has lower values of edge strength and variance. A noise-free background also has a small mean-gradient value.

Heavy stroke areas have strong edge strength, variance and mean-gradient value. This class can be sub-divided into two sub-classes: 1) Heavy strokes only; 2) Heavy strokes with faint strokes.

Faint stroke areas contain faint strokes, which are very difficult to detect from the background. This kind of area typically has a medium value of edge strength and mean-gradient but less variance.

2.4 Appling The Thresholding Method

Different threshold methods are applied for the above three classes of sub-images.

The six methods described above [1] [2] [4] ~ [7] cannot provide ideal results for degraded historical handwritten images, especially for regions in the faint strokes class. Bernsen's method, the improved Niblack's algorithm and Yanowitz & Bruckstein's method is the most outstanding of the other exist famous thresholding method described [8]. These three thresholding methods are chosen for the heavy and faint stroke classes.

2.4.1 Appling The Thresholding Method for Heavy Strokes Class.

There are two sub-classes in *heavy stroke class*. One contains heavy strokes only; the other contains some faint connection strokes alongside heavy strokes.

Bernsen's method is applied on heavy strokes region. It works well because those regions have high contrast. The contrast threshold value for the experiment of Bernsen's method is set to 180. Practically, the range of contrast value for heavy strokes regions is 190 ~ 220 (for a 256-graysacle image).

The improved Niblack method is applied on heavy strokes region with some faint connection strokes. It is sensitive to edge information, and is able to clearly maintain the faint strokes, which are connected to heavy strokes.

2.4.2 Appling Thresholding Method for Faint Strokes Class

The sub-image in the faint strokes class contains lower edge strength because the strokes are skating over the region. The higher mean-gradient value will be detected for regions if there is more noise, thus noise removal is required. The variance will be low because of the variation of the region is low.

Noise removal and enhancement for the *faint stroke class* are needed before threshold method is applied.

To avoid the enhancement of noise, a Wiener filter was first applied. The enhancement can be divided into two steps.

1). Use a 3*3 window to enhance the image by finding the maximum and minimum gray value in the window using (8) and (9):

$Mini$ = min (elements in the window) (8)

$Maxi$ = max (elements in the window) (9)

2). Compare the value of $Pixel - Mini$ and $Maxi - Pixel$, where *Pixel* stands for pixel-value. If the former is larger, the *Pixel* is closer to the highest gray value than the lowest value in this window; hence the value of *Pixel* is set to the highest gray value ($Pixel = Maxi$). If the former is smaller, then the value of *Pixel* is set to the lowest gray value ($Pixel = Mini$).

Yanowitz & Bruckstein's method is applied to the faint stroke class; it works well on retaining detail information of handwriting.

3. Experimental Results

The proposed method has been evaluated using 10 historical images obtained from the Library of Congress. All the 10 images were available as 256-grayscale Tiff images. The images were chosen to have varying resolutions, sizes, and contrast to ensure correct comparison of performance between the algorithms. The images were characterized by high resolution of the scanned images with varying contrast of the handwriting.

3.1 Stroke Direction of Image

Stroke Direction of an input image was measured using the Direction Matrix shown in Figure 4. (There are 8 directions as shown in Figure 2) The stroke directions of the example in Figure 1 are direction 1, direction 4 and direction 3 respectively. 'Direction = 1' means the direction of word stroke is G1, which is equal to 45 degree; 'Direction = 4' means the direction of word stroke is G4, which is equal to 180 degree; 'Direction = 3' means the direction of word stroke is G3, which is equal to 135 degree.

It was observed in the experiments that 64*64 pixels is the best image size to measure the stroke direction. The result is not accurate if the size of local area is too small, and there will be too many calculations if the local size is too big.

3.2 Local Area Classification

Background is an area, which contain only noise with no information content. Background areas contain low edge strength and low mean-gradient value, but may have high variance value, which is produced by noise or useless information.

Faint stroke area include noise and faint handwritten strokes, but heavy strokes. Practically, this area contains stronger edge strength, variance and mean-gradient value than background.

An area that contains heavy strokes has the strongest edge strength, variance and mean-gradient value.

Some experimental results are shown in Table 1 and Figure 3.

3.3 Text/Background Separation Result of the Proposed Method

The proposed method was evaluated using these six thresholding algorithms on 10 historical images selected from the Library of Congress. Some experimental results of the seven methods are shown in Figure 6.

The standard measures, *recall* [11], were used to evaluate the performance of the proposed methods. Recall is defined as:

$$\text{Recall} = \frac{\text{Correctly Detected Words}}{\text{Total Words}}, \quad (10)$$

Figure 7 shows the recall value of the seven threshold methods, and the average *Recall* values are presented in percentage of the content of each image. From the table, it is apparent that the proposed *Decompose Thresholding method* produced significantly better recall results than the other individual six methods.

Of these six methods, Bernsen's method works well on clear background and high contrast historical images, but it is a contrast-based method so sensitive to the noise in the images. ETM's method uses a manual value to determine

the difference between two windows. It works well for both faint and heavy handwriting, but failed when there was a noisy background. Yanowitz's method can retain very detailed strokes but still retains useless noise points. The improved Niblack technique can retain detailed stroke but is sensitive to noise. QIR and Otsu's technique only works well on bimodal histogram images.

From an aesthetic and subjective point of view, the *Decompose-Threshold Approach* is better than other local threshold methods. It detects feature vectors of different sub-areas and then applies appropriate methods to avoid losing important useful information.

The decompose-threshold structure is highly effective for the images, which contain different conditions at different locations.

4. Conclusion

Historical images often exhibit degraded characteristic qualities after years of storage. Satisfactory thresholding results can rarely be obtained if same process is applied to the entire image.

The decompose threshold approach is effective at resolving this problem. It uses local feature vectors for analysis and hence to find the best approach for thresholding local area. Appropriate algorithm(s) is selected or combined for specific types of document image under investigation automatically. The original image is recursively broken down into sub-regions using quad-trees until an appropriate thresholding method can be applied to each of the sub-region. The approach outperforms existing single methods by measurement of 'recall' value. The future application of this technique can contribute to other difficult document images, such as cheques and newspaper images.

5. References

[1] N. Otsu, "A threshold selection method from grey level histogram", *IEEE Trans. Syst. Man Cybern.*, Vol. 9, No. 1, 1979, pp. 62-66.

[2] Y.Solihin, C.G.Leedham, "Integral ratio: a new class of global thresholding techniques for handwriting images", *IEEE Trans. PAMI*, Vol. 21, No. 8, 1999, pp. 761-768.

[3] W. Niblack, *An introduction to digital image processing*, Prentice Hall, 1986.

[4] Z. Zhang and C. L. Tan, "Restoration of images scanned from thick bound documents", *Proc. Int. conf. Image Processing*, Vol. 1, 2001, pp.1074-1077.

[5] J. Bernsen, "Dynamic thresholding of grey level images, *Proc. Int. conf. Patt. Recognition*, 1986, pp. 1251-1255.

[6] L. Eikvil, T. Taxt & K. Moen, "An adaptive method for binarization of grey level images", *NOBIM National Conference on Image Processing and Pattern Recognition*, June 1991, pp 123-131,.

[7] S.D. Yanowitz and A.M. Bruckstein, "A new method for image segmentation", *Computer Vision, Graphics and Image Processing*, 1989, Vol. 46, No. 1, pp. 82-95.

[8] O.D. Trier and T. Taxt, "Evaluation of binarization methods for document images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume: 17 Issue: 3, 1995, pp. 312 – 315.

[9] S.W. Lam, "Texture feature extraction using gray level gradient based co-occurrence matrices", *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 1, No. 14-17, Oct. 1996, pp. 267 –271.

[10] C.G. Leedham, Y. Chen, K. Takru, J. Tan and M. Li, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images", *Proc. 7th Int. Conf. on Document Analysis and Recognition*, Scotland, Vol. 2, 2003, pp 859 -865.

[11] M. Junker, R. Hoch, "On the evaluation of document analysis components by recall, precision, and accuracy", *Proc. 5th Int. Conf. on Document Analysis and Recognition*, India, 1999, pp. 713-716.

Table 1. Experiment Result of Area Classification

Class Name / Feature Name	Background	Faint Strokes	Heavy Strokes	
			With some faint strokes	Only heavy strokes
Edge Strength	1=<ES<=13	14=<ES<=40	41=<ES	
Variance	1=<V<=30	10=<V<=45	45<=V	
Mean-Gradient	1=<G<=2	3=<G<=10	3=<G<=10	10<=G

Note: ES –> Edge Strength, V –> Variance; G –> Mean-Gradient.

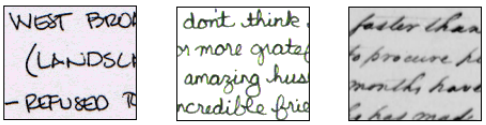


Figure 1. Three main word Directions

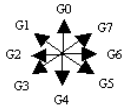


Figure 2. Eight Directions of Word

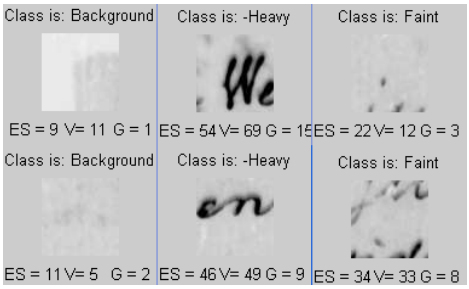


Figure 3. Local Region Classification & Feature Extraction

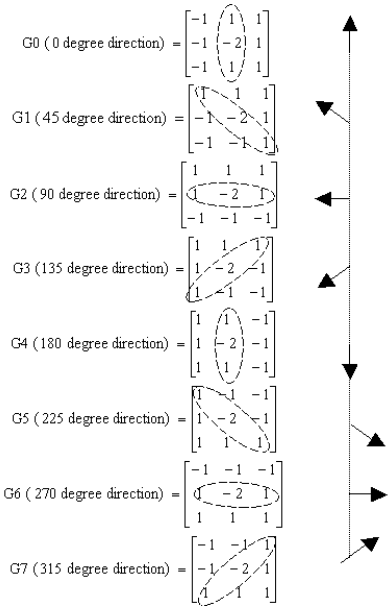


Figure 4. Direction matrices

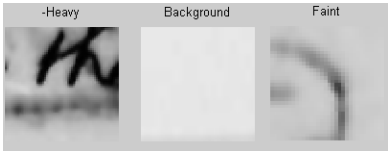
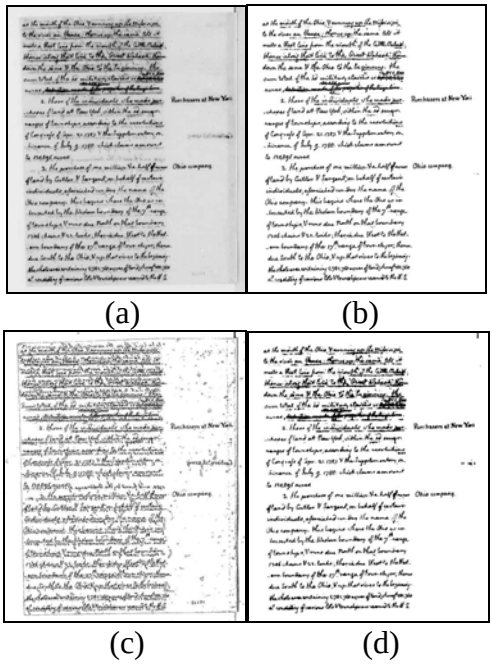


Figure 5. Examples of sub-regions containing heavy strokes, background and faint strokes.



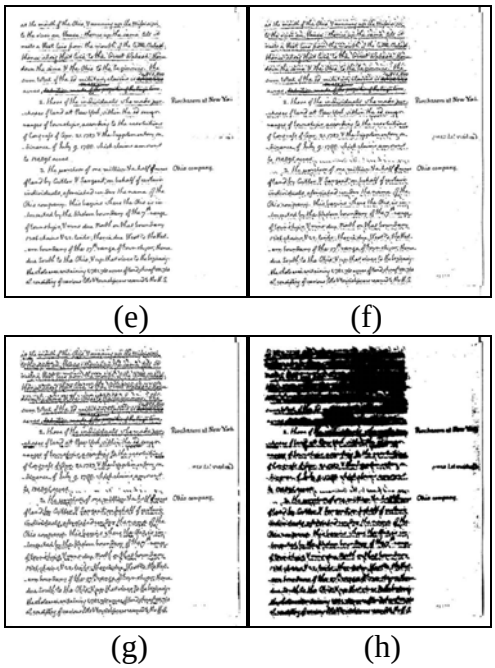


Figure 6. Experiment Results of 7 Algorithms

(a): Original Historical Image (b): Proposed Decompose Threhsolding Method (c): Bernsen's Method (d): Otsu's Algorithm (e): Improved Niblack's Method (f): Eikvil/Taxt/Moen's Method (g): Yanowitz/Bruckstein's Algorithm (h): QIR Algorithm

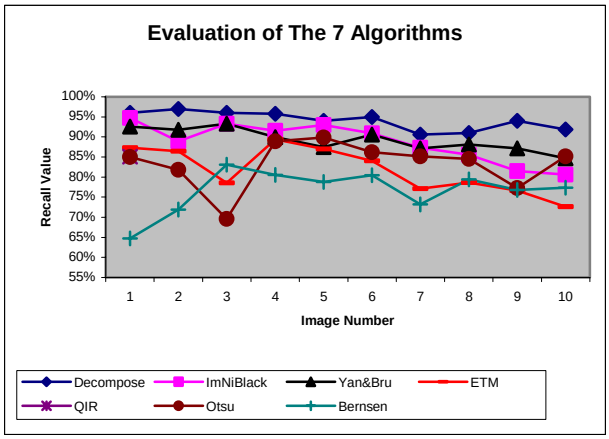


Figure 7. Evaluation by Recall Value

Decompose: Proposed Decompose Threhsolding Method;

imNiblack: Improved Niblack's Method;

Yan&Bru: Yanowitz/Bruckstein's Algorithm;

ETM: Eikvil/Taxt/Moen's Method;

QIR: QIR Algorithm;

Otsu: Otsu's Algorithm ;

Bernsen: Bernsen's Method