Evaluation of Binarization Methods for Document Images

Øivind Due Trier and Torfinn Taxt

Abstract— **This paper presents an evaluation of eleven locally adaptive binarization methods for gray scale images with low contrast, variable background intensity and noise. Niblack's method with the addition of the postprocessing step of Yanowitz and Bruckstein's method added performed the best, and was also one of the fastest binarization methods.**

Keywords— **Locally adaptive binarization, Thresholding, Evaluation, Utility maps, Document images.**

I. Introduction

THERE is currently a large and growing interest in the field of document image analysis. Numerous research groups are trying to design systems for extracting relevant information from such diverse documents as engineering drawings, maps, magazines and newspapers, forms, and mail envelopes [1–3]. In most of these systems, binarization of the scanned gray level image (labeling each pixel print or background) is done prior to further processing. This paper will focus on binarization of map images.

Maps are central to a large number of public and private agencies. An example is utility maps, which are film copies of ordinary maps with specialized information added in the form of text, symbols, lines, and networks. The availability of high quality digital geographical information systems has created the possibility for a more flexible use and maintenance of the utility maps. However, these systems demand that the maps are available in digital format, preferably in a vector representation, with the associated text information in symbolic form.

To digitize a map, it is first scanned at a high resolution, giving a gray scale image which is then binarized. Subsequently, symbol recognition and line vectorization programs are used to label symbols, text, lines and solid areas in the binary raster image. However, these algorithms require a high quality binary raster image to be able to correctly identify all the information present. Professionally drawn map originals for use in offset printing usually fulfill this requirement when global binarization methods are applied. In contrast, the print quality of utility maps is poor due to extensive manual handling and updating over time, giving areas with variable background intensity, low contrast and stochastic noise. Therefore, it is essential to find binarization methods which will correctly label all the information present, even in low contrast areas of the map, while being insensitive to variable background intensity and stochastic noise. These requirements exclude global binarization methods, and lead us to evaluate the best adaptive binarization methods available in the literature.

In our evaluation, binarization methods requiring manual inspection of subimages and fine-tuning of parameters were excluded. This is because the binarization algorithms are intended to be part of an automatic digitizing system for maps, being able to process a large number of maps per time unit. Each gray scale map image may consist of a billion or more pixels due to its large size and small symbols. A typical utility map is $1m^2$ large and is scanned at $50\mu m$ pixel size ($\approx$ 500 dpi) with 256 gray levels.

A total of seven subimages taken from five gray scale map images were used in this evaluation study. Each subimage was carefully selected to contain particularly challenging areas with low contrast, variable background intensity or stochastic noise. The eleven most promising locally adaptive binarization methods we could identify in the literature were applied to the test images. The resulting binary images were compared and evaluated using five visual criteria. We found that the postprocessing step (PS) of the Yanowitz and Bruckstein method [4] improved all the other binarization methods as well. Niblack's method [5] with PS performed the best. Preliminary results of this work have been presented earlier [6].

II. Binarization Methods

Locally adaptive binarization methods (local methods) compute a threshold for each pixel in the image on the basis of information contained in a neighborhood of the pixel. Some methods calculate a threshold surface. If a pixel (x, y) in the input image has a higher gray level than the threshold surface evaluated at (x, y), then the pixel (x, y) in the output image is labeled background, otherwise it is labeled as print. Other methods do not use explicit thresholds, but search for print pixels in a transformed image.

We tested the local methods of Bernsen [7], Chow and Kaneko [8][9], Eikvil et al. [10], Mardia and Hainsworth [11], Niblack [5], Taxt et al. [12], Yanowitz and Bruckstein [4], Parker [13], White and Rohrer's Dynamic Thresholding Algorithm [14], White and Rohrer's Integrated Function Algorithm [14], and our Improved White and Rohrer's method [15]. These methods were selected because they are frequently cited, or appeared to be promising. The first seven methods and White and Rohrer's Dynamic Thresh-

Manuscript received March 1, 1994; Revised Oct. 17, 1994.

Ø. D. Trier and T. Taxt are with the University of Oslo, Department of Informatics, P.O. Box 1080 Blindern, N-0316 Oslo, Norway. Email: trier@ifi.uio.no, torfinn@tor.pki.uib.no .

IEEECS Log number P95033.

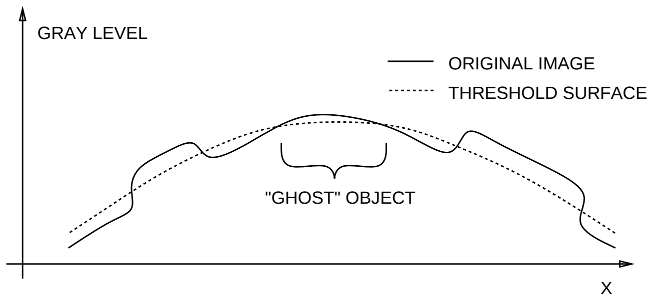

Fig. 1. Cross section through an image, illustrating how "ghost" objects might occur. Based on a figure by Yanowitz and Bruckstein [4].

olding Algorithm use explicit thresholds, or threshold surfaces, while the other three methods search for print pixels after having located the edges. All these methods are briefly described in [16].

To make sure that the global binarization was insufficient to threshold the map images properly, we also tested the four global binarization methods of Abutaleb [17], Kapur et al. [18], Kittler and Illingworth [19] and Otsu [20]. The latter two are known to be optimal under two different normality assumptions [21]. Abutaleb's method seemed to be the most promising global method, since it uses some local information to generate an additional feature for each pixel, using a two dimensional threshold to classify the pixels.

A. Postprocessing step

A postprocessing step is used in Yanowitz and Bruckstein's method and Improved White and Rohrer's method to remove "ghost" objects (Fig. 1), and can be incorporated into other methods as well. The average gradient value at the *edge* of each printed object is calculated. Objects having an average gradient below a threshold T_P are labeled as misclassified, and are removed. The main steps of the algorithm are given below:

1. Smooth the original image by a (3×3) mean filter to remove noise.
2. Calculate the gradient magnitude image G of the smoothed image, using, e.g., Sobel's edge operator [22].
3. Select a value for T_P. There is no automatic method to specify T_P for a given image, so it is specified by trial and error (see below).
4. For all 4-connected print components, calculate the average gradient of the edge pixels. Edge pixels are print pixels that are 4-connected to the background. Remove print components having an average edge gradient below the threshold T_P.

III. Experimental Methodology and Results

We present the evaluation results of the published methods first. Then, we add the postprocessing step and describe the improved results. Two of the test images are shown along with three of the 19 binarization results for each image, to give the reader a fair impression of the best methods. More results are described in [16].

A. Test Images

The maps were scanned with a SysScan Flatbed scanner, with pixel size $50\mu m$ and 256 gray levels. At this resolution, even the finest details in the maps are preserved in the gray-scale image. Of the two images shown, CABLE IMAGE (512×512 pixels, Fig. 2a) is from a scanned paper copy of a map showing a dense network of electricity cables, courtesy of Oslo Energi AS, Norway. The HYDRO IMAGE (256×256 pixels, Fig. 2d) is from a scanned hydrographic original, courtesy of the Norwegian Mapping Authorities. It shows depth numbers and contours from a part of the Norwegian coast.

B. Evaluation Procedure

As no ground truth was available, which is usually the case with real data, we had to use visual criteria when evaluating the binarization results. The result of applying a binarization method to a test image was evaluated using the following five criteria:

1. **Broken line structures.** Gaps of all sizes in lines were roughly counted. Large gaps were considered worse than small.
2. **Broken symbols, text, etc.** Symbols and text characters with gaps were roughly counted, and the degree of fragmentation was also assessed.
3. **Blurring of lines, symbols and text.** Both the number of blurred print objects, and the degree of blurring were assessed.
4. **Loss of complete objects.** The number of print objects which were completely lost was roughly counted.
5. **Noise in homogeneous areas.** Both the number and the size of noisy spots and false objects in both background and print were estimated.

A scale of 1–5 was used for each criterion. Thus, for each method, the maximum score on a single test image was 25, and the maximum total score for one method on 7 test images was 175.

The performance evaluation was a blind test. For each test image, the order of the 19 binarized images was selected at random, and independently for the seven different test images. An expert in the field of image binarization evaluated the binary images. The 1–5 scale was used as a relative scale to discriminate as much as possible on each criterion. Then, the scores on each criterion were adjusted and weighted against each other by comparing each binary image with the other binary images which originated from the same test image and scored about equally well.

Test image 7 is exceptional, since it is not a representative of map images to be digitized. Therefore, the ranking of the methods is based on the total score for test images 1–6. Still, test image 7 gives us useful insight into how the different methods behave in a very difficult situation.

An alternative evaluation method would be to apply an automatic symbol recognition and line vectorization algorithms to the binarized images to quantify the binarization

TABLE I

SCORES FOR THE BINARIZATION METHODS. THE BEST SCORES ARE SHOWN IN **boldface**. EXECUTION TIMES ARE GIVEN FOR CABLE IMAGE (512×512 PIXELS) ON A UNIX WORKSTATION (SILICON GRAPHICS INDY).

method		image number[a] 1	2	3	4	5	6	7	sum 1–7	sum 1–6	rank	execution time
locally adaptive methods												
Yanowitz/Bruckstein	YBM	17	**24**	**21**	**21**	**24**	21	18	146	128	4	98 s
Improved White/Rohrer	IWRM	**21**	20	21	18	21	**23**	19	143	124	5	6 s
Niblack	NM	20	19	19	18	20	19	19	134	115	6	1 s
Eikvil/Taxt/Moen	ETMM	18	22	19	18	20	18	13	128	115	7	15 s
Bernsen	BM	16	18	18	17	20	19	**21**	129	108	8	1 s
White/Rohrer DTA	WRM1	20	21	15	15	17	19	12	119	107	9	1.6s
Parker	PM	17	15	14	14	19	18	15	112	97	11	30 s
Chow/Kaneko[b]	CKM	14	20	15	13	16	19	7	104	97	12	66 s
Mardia/Hainsworth	MHM	9	18	13	15	15	19	5	94	89	14	14 s
Taxt/Flynn/Jain	TFJM	9	16	12	12	15	21	7	92	85	16	83 s
White/Rohrer IFA	WRM2	16	13	11	9	10	14	12	85	73	18	4 s
global methods												
Otsu	OM	9	17	15	16	16	22	7	102	95	13	0.3s
Kapur/Sahoo/Wong	KSWM	12	14	16	14	11	19	5	91	86	15	0.3s
Abutaleb	AM	8	16	13	15	15	18	5	90	85	17	1 s
Kittler/Illingworth	KIM	9	9	15	10	10	9	7	69	62	19	0.3s
modified locally adaptive methods												
Niblack	NM/PS	**23**	21	**24**	22	23	**25**	22	**160**	**138**	1	3 s
Eikvil/Taxt/Moen	ETMM/PS	21	23	22	23	21	23	14	147	133	2	18 s
Bernsen	BM/PS	19	20	23	**24**	22	24	**24**	156	132	3	3 s
Parker	PM/PS	17	18	14	17	19	19	16	120	104	10	32 s

[a] 3=CABLE IMAGE, 5=HYDRO IMAGE

[b] Nakagawa and Rosenfeld's implementation

quality. An early attempt is reported in [23].

C. Results

C.1 Published methods

Of the existing methods, Yanowitz and Bruckstein's method was the best overall, and was somewhat better than Improved White and Rohrer's method (Table I). All the local methods gave rather poor binarization results on at least one test image. None of the global methods gave reasonable binarization results on any of the test images. When applied to CABLE IMAGE, Yanowitz and Bruckstein's method resulted in a minor gap in some numerals, and a few gaps in the lines (Fig. 2e). Bernsen's method kept too much of the background noise (Fig. 2f). Applied to HYDRO IMAGE, Yanowitz and Bruckstein's method performed the best overall. Still, many numbers were attached to lines or other numerals (Fig. 2h). Of the global methods, Otsu's method performed the best, but most of the numerals were smeared out, and many more numerals were attached to other print objects than by Yanowitz and Bruckstein's method (Fig. 2i).

C.2 Performance with postprocessing step

The postprocessing step used in Yanowitz and Bruckstein's method and Improved White and Rohrer's method was added to Eikvil et al.'s, Niblack's, Bernsen's, and Parker's methods. For all the four methods, most mislabeled objects were removed by the postprocessing step. Except for Parker's method, the modified methods scored higher than their original counterparts on all the seven test images. The postprocessing step was not added to the other methods, since White and Rohrer's Dynamic Thresholding Algorithm gave no noise in the background, and the other methods had serious defects that the postprocessing step could not repair.

Niblack's method with PS was the best method overall, slightly better than Improved White and Rohrer's method (Table I).

On CABLE IMAGE, Bernsen's method with PS was the best (Fig. 2g), and the postprocessing step removed almost all the noisy spots. On HYDRO IMAGE, Niblack's method with PS (Fig. 2j) performed almost as well as Yanowitz and Bruckstein's method (Fig. 2h), but had the same shortcomings.

We selected the parameter T_P manually by trial and error. The best identified value for T_P was close to the mean value $\hat{\mu}_G$ of the gradient image G for six of the seven test images. Thus, $T_P = \hat{\mu}_G$ can not always be used.

Niblack's and Bernsen's methods were the fastest of the local methods according to execution times for CABLE IMAGE (Table I). Even with the addition of the postprocessing step, these two methods were faster than the other local methods.

IV. Discussion

We have evaluated the most promising published locally adaptive binarization methods. Our experimental results on a range of gray scale images of maps with low print quality indicates that no single method performs better than the other methods on all the images. In our evaluation scheme, Yanowitz and Bruckstein's method [4] had the best overall performance of the published methods, slightly better than our improved White and Rohrer's method [15]. Furthermore, inclusion of the postprocessing step used in the Yanowitz and Bruckstein method lead to improvement of the other binarization methods. By adding the postprocessing step, Niblack's method [5] was the best overall, followed by Eikvil et al.'s method [10] and Bernsen's method [7]. These five methods were significantly better than the other methods.

We believe that for images like HYDRO IMAGE, better binarization results will be achieved if the original map is scanned at even a higher resolution, for example at $25\mu m$ pixel size instead of $50\mu m$. Then, we will expect fewer numerals to be attached to other print objects after binarization. However, some numerals will still suffer from this problem, as they actually are (unintentionally) written attached to other objects in the original map.

More generally, even the best local method does not necessarily give sufficiently high quality binarization results which could be used for further automatic processing. The map may simply be of such a low printing quality that no available automatic binarization method will work properly on it. In such cases, direct recognition of the objects in the gray scale image may be a better alternative.

This evaluation study includes only seven test images. Still, we believe these images contain numerous challenges to establish a tentative ranking of the locally adaptive binarization methods that is applicable to a wide range of images. For document image applications, it is probably sufficient to implement a few of the best methods. If the postprocessing step is to be included, a method to automatically determine the threshold T_P must be found.

Acknowledgments

We thank Anil K. Jain, Fritz Albregtsen, Knut Liestøl and Xinli Wang for useful discussions. This work was supported by The Research Council of Norway.

References

[1] *Proceedings of the First International Conference on Document Analysis and Recognition*, Saint-Malo, France, Sept.–Oct. 1991. IEEE Computer Society Press.

[2] *Proceedings of the Second International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, Oct. 1993. IEEE Computer Society Press.

[3] Hideyoshi Tominaga (Guest ed.), "Special issue on postal processing and character recognition", *Pattern Recognition Letters*, vol. 14, no. 4, pp. 257–354, Apr. 1993.

[4] S. D. Yanowitz and A. M. Bruckstein, "A new method for image segmentation", *Computer Vision, Graphics and Image Processing*, vol. 46, no. 1, pp. 82–95, Apr. 1989.

[5] Wayne Niblack, *An Introduction to Digital Image Processing*, pp. 115–116, Prentice Hall, 1986.

[6] Øivind Due Trier and Torfinn Taxt, "Evaluation of binarization methods for utility map images", in *Proceedings of the first IEEE International Conference on Image Processing*, Austin, Texas, Nov. 1994, vol. 2, pp. 1046–1050.

[7] John Bernsen, "Dynamic thresholding of grey-level images", in *Proceedings of the 8th International Conference on Pattern Recognition*, Paris, France, Oct. 1986, pp. 1251–1255.

[8] C. K. Chow and T. Kaneko, "Automatic detection of the left ventricle from cineangiograms", *Computers and Biomedical Research*, vol. 5, pp. 388–410, 1972.

[9] Yasuo Nakagawa and Azriel Rosenfeld, "Some experiments on variable thresholding", *Pattern Recognition*, vol. 11, no. 3, pp. 191–204, 1979.

[10] Line Eikvil, Torfinn Taxt, and Knut Moen, "A fast adaptive method for binarization of document images", in *Proceedings of the First International Conference on Document Analysis and Recognition*, Saint-Malo, France, Sept.–Oct. 1991, pp. 435–443.

[11] K. V. Mardia and T. J. Hainsworth, "A spatial thresholding method for image segmentation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 919–927, 1988.

[12] Torfinn Taxt, Patric J. Flynn, and Anil K. Jain, "Segmentation of document images", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 12, pp. 1322–1329, 1989.

[13] J. R. Parker, "Gray level thresholding in badly illuminated images", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 813–819, 1991.

[14] J.M. White and G.D. Rohrer, "Image thresholding for optical character recognition and other applications requiring character image extraction", *IBM Journal of Research and Development*, vol. 27, no. 4, pp. 400–411, July 1983.

[15] Øivind Due Trier and Torfinn Taxt, "Improvement of 'integrated function algorithm' for binarization of document images", *Pattern Recognition Letters*, to appear.

[16] Øivind Due Trier and Torfinn Taxt, "Evaluation of binarization methods for document images", Tech. Rep., Department of Informatics, University of Oslo, Norway, July 1994.

[17] Ahmed S. Abutaleb, "Automatic thresholding of gray-level pictures using two-dimensional entropy", *Computer Vision, Graphics and Image Processing*, vol. 47, pp. 22–32, 1989.

[18] J. N. Kapur, P. K. Sahoo, 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.

[19] J. Kittler and J. Illingworth, "Minimum error thresholding", *Pattern Recognition*, vol. 19, no. 1, pp. 41–47, 1986.

[20] Nobuyuki Otsu, "A threshold selection method from gray-level histograms", *IEEE Trans. Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[21] T. Kurita, N. Otsu, and N. Abdelmalek, "Maximum likelihood thresholding based on population mixture models", *Pattern Recognition*, vol. 25, no. 10, pp. 1231–1240, 1992.

[22] William K. Pratt, *Digital Image Processing*, pp. 501–504, John Wiley & Sons, second edition, 1991.

[23] Øivind Due Trier and Anil K. Jain, "Goal-directed evaluation of binarization methods", in *Proceedings of NSF/ARPA Workshop on Performance Versus Methodology in Computer Vision*, Seattle, Washington, June 1994, pp. 206–217.

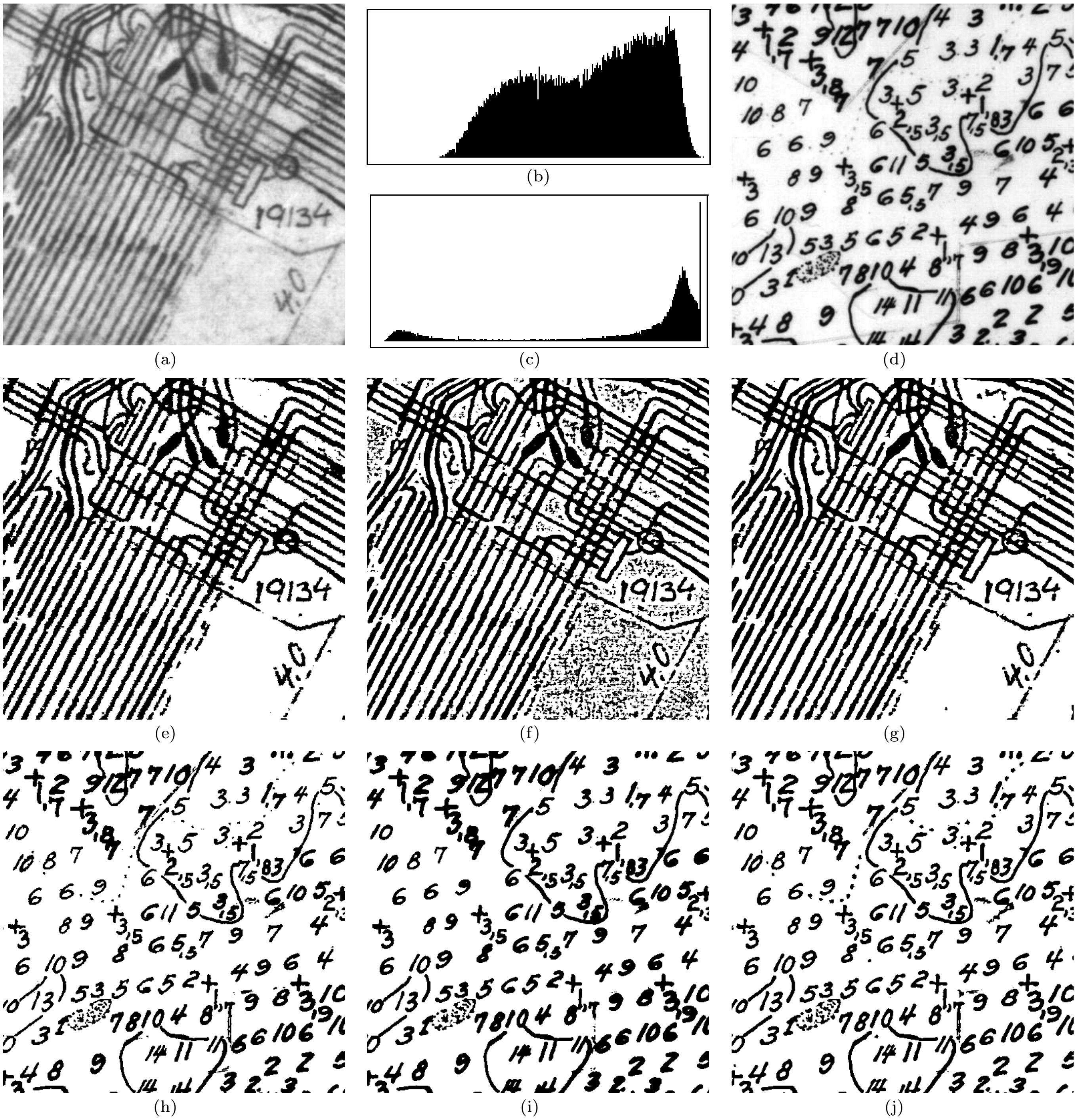

Fig. 2. CABLE IMAGE and HYDRO IMAGE. **(a)** Original CABLE IMAGE. **(b)** Histogram of CABLE IMAGE. Y-axis: [0 . . . 2410]. **(c)** Histogram of HYDRO IMAGE. Y-axis: [0 . . . 13473]. **(d)** Original HYDRO IMAGE. **(e–g)** CABLE IMAGE, binarized: **(e)** Yanowitz and Bruckstein's method with postprocessing step, $T_V = 60$. **(f)** Bernsen's method. **(g)** Bernsen's method with postprocessing step, $T_V = 70$. **(h–j)** HYDRO IMAGE, binarized: **(h)** Yanowitz and Bruckstein's method, $T_V = 100$. **(i)** Otsu's method. **(j)** Niblack's method with postprocessing step, $T_V = 100$.