Prediction of the Optical Character Recognition Accuracy based on the Combined Assessment of Image Binarization Results

Piotr Lech[1], Krzysztof Okarma[1]
[1]*Department of Signal Processing and Multimedia Engineering, Faculty of Electrical Engineering, West Pomeranian University of Technology, Sikorskiego 37, 70-313 Szczecin, Poland*
piotr.lech@zut.edu.pl

***Abstract*—In the paper the problem of reliable evaluation of the effects of image binarization is discussed in view of image recognition accuracy. Considering the Optical Character Recognition methods, typically used for document images obtained by cameras or scanners, their accuracy is strongly dependent on the results of image binarization. Unfortunately, metrics typically used for the evaluation of binarization results, such as Peak Signal to Noise Ratio, Distance Reciprocal Distortion or Misclassification Penalty Metric, are not always well correlated with the recognition accuracy of individual characters. Therefore, a novel approach related to the use of combined metric for the assessment of binarization results is proposed and verified for the binary images obtained using some popular histogram-based methods from the original images with degraded quality. For the experimental prediction of the character recognition accuracy, the popular open source engine supported by Google, known as Tesseract, has been used.**

***Index Terms*—Image analysis, image recognition, image quality, machine vision.**

I. INTRODUCTION

The role of the Optical Character Recognition (OCR) algorithms in view of image analysis and recognition methods is still important regardless of the fact that more and more sophisticated methods based on the analysis of words and phrases are included in the modern OCR systems. Since the applications area of such image based text recognition methods is still growing, mainly due to the development of mobile devices equipped with relatively cheap cameras, there is still a need to develop some fast and reliable algorithms which should be able to recognize the individual characters properly in the presence of various distortions or different lighting conditions. A proper recognition of text from the document image captured by the smartphone's camera is not always an easy task assuming unknown lighting conditions and limited available resources.

Nevertheless, one of the most relevant elements of the text recognition workflows is still the image binarization step. Since several more or less complicated algorithms can be applied for this purpose, such as the popular methods proposed e.g. by Otsu [1], Sauvola [2], Niblack [3], Rosin [4] and Kapur [5], their performance differs significantly for lower quality input images, noticeably influencing the text recognition accuracy.

Manuscript received April 5, 2015; accepted September 29, 2015.

II. EVALUATION OF BINARIZATION RESULTS

One of the unsolved issues related to this field is still the evaluation of the binary images in view of its usefulness for further text recognition. Since the process of image binarization is ambiguous and various algorithms lead to different results, there is a need of reliable comparison of the outputs of the binarization algorithms. Unfortunately, there are no "blind" methods, which are even more popular in image quality assessment area and do not require the knowledge of the original image. For the evaluation of binarization results the original "ground-truth" image has to be provided since all the metrics are calculated on the base of relatively simple comparison of binary values for corresponding pixels. Therefore, both compared images must be geometrically matched to each other in order to obtain proper results.

Considering the problem of prediction of the OCR accuracy, an ideal solution would be the use of the "blind" (no-reference) metric but the development of such one would be possible only using a large image dataset with results of the character recognition. Moreover, such a metric would be probably suitable only for a limited number of image distortions, specific binarization methods and recognition algorithms.

The first step towards such solution should be the development of a full-reference image binarization evaluation metric, similarly as for general image quality assessment purposes, which would be well correlated with character recognition accuracy using different algorithms in the presence of various distortions. Nevertheless, the verification of such metric also requires the dataset of images containing various distortions, subjected to binarization using different methods, together with the numerical results of the OCR accuracy for each obtained binary image.

For the verification of the idea proposed in the paper,

a relatively small such dataset has been prepared, consisting of three pristine images subjected to five types of distortions typical for machine printed documents. In order to simulate them the original images have been printed on the letterhead paper, on the other side of the previously printed paper, on the older paper sheet or the colour paper as well as subjected to wrinkling. The scanned images of such documents have been geometrically matched with "ground-truth" images and subjected to binarization using three popular histogram-based methods, namely Otsu, Kapur and Rosin algorithms.

Illustration of the "ground truth" images used in the experiments is shown in Fig. 1, some results of distortions are presented in Fig. 2, whereas the exemplary obtained binarization results are illustrated in Fig. 3.

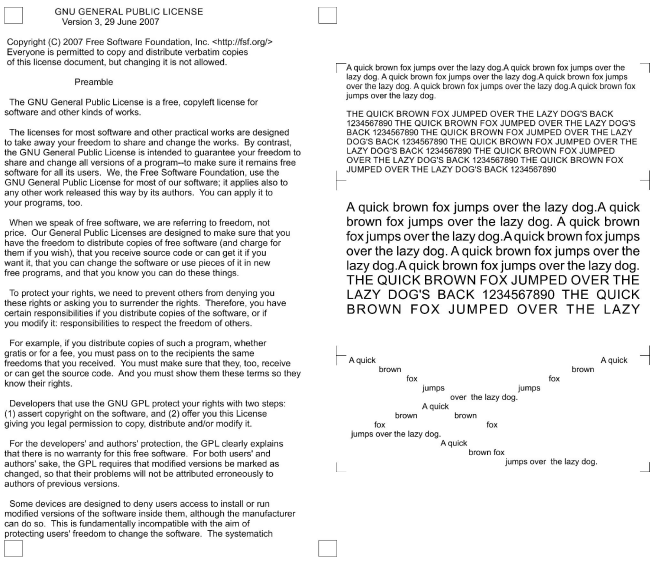

Fig. 1. Original "ground-truth" images used during the experiments (cropped).

Since there are some metrics which are often used for the evaluation of the binarization results by the comparison with the "ground-truth" binary image [6]–[8], a natural solution seems to be their application also for this purpose. Unfortunately, typically used well-known metrics which are fast to compute, such as Peak Signal to Noise Ratio (PSNR), Distance Reciprocal Distortion (DRD) [9] or Misclassification Penalty Metric (MPM) [10] turn out to be rather poorly correlated with OCR accuracy.

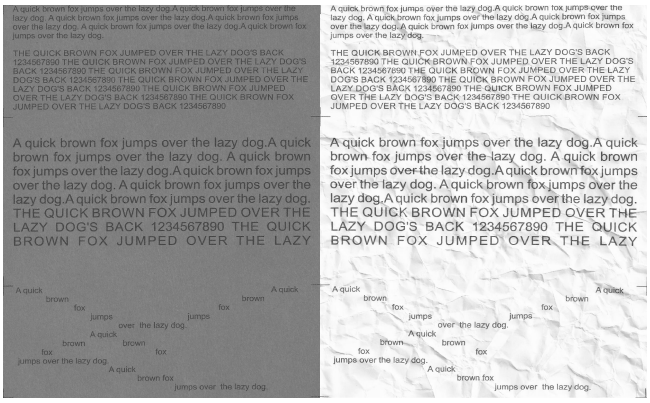

Fig. 2. Original images used during the experiments subjected to exemplary distortions.

For those reasons, we have focused on the development of the combined metric which should be better correlated with the results of the recognition of individual characters. In order to verify the validity and usefulness of the proposed approach some experiments have been conducted with the use of Tesseract [11] which is probably the most accurate open-source OCR software, developed previously in HP Labs and now supported by Google.

III. IDEA OF COMBINED METRIC FOR EVALUATION OF BINARIZATION OUTPUT

Recognition of individual characters on the binary image is strongly dependent on their shapes which may be influenced especially on the edges due to the improper choice of the threshold value. A general rule seems to be relatively simple – the higher number of pixels differing between the result of binarization and the "ground-truth" image causes more errors during the character recognition.

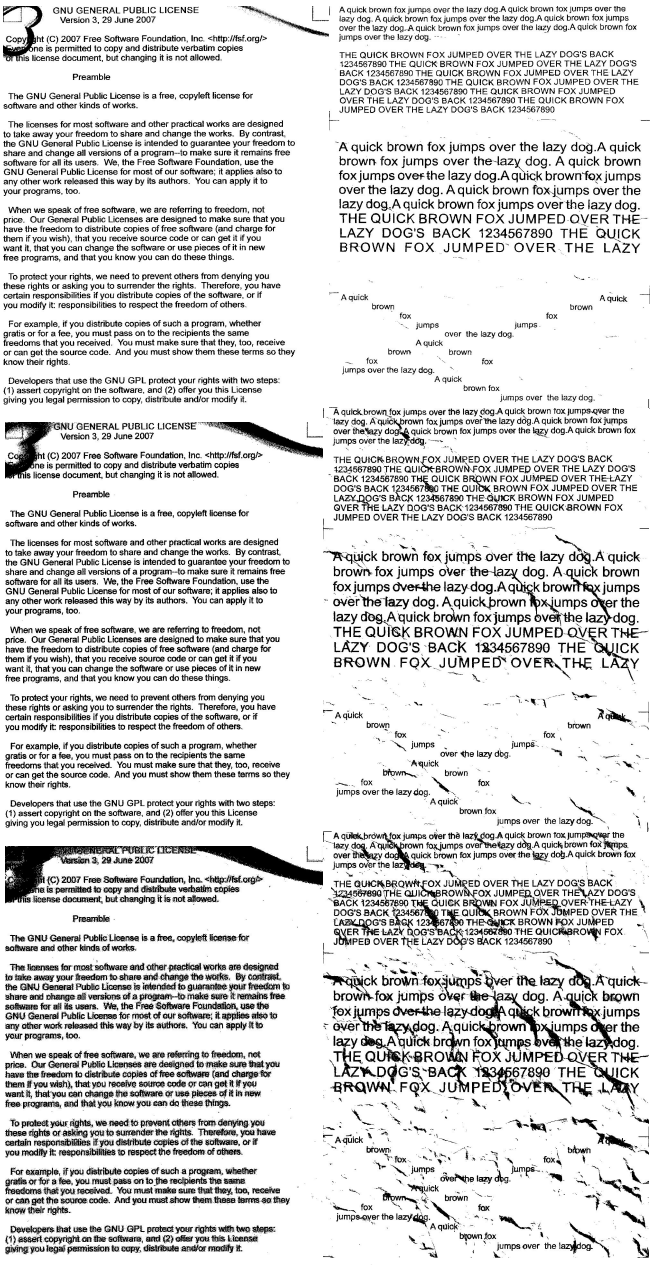

Fig. 3. Exemplary binarization results – from top to bottom: using Otsu, Kapur and Rosin algorithms.

Since most of the metrics typically used for the evaluation of the binarization algorithms are based on the similar

assumptions, three of them have been chosen for the initial verification – previously mentioned PSNR, DRD and MPM.

The results of the OCR using Tesseract engine for each test image have been compared with the proper results obtained from the original document file and the number of errors has been calculated for each of them as well as the recognition accuracy defined as

$$rec\_acc = 1 - \frac{N_{\text{err}}}{N_{\text{total}}}, \quad (1)$$

where N_{err} denotes the number of errors and N_{total} stands for the total number of characters in the text. It is worth to mention that the recognition accuracy achieved using Tesseract for all "ground-truth" images has been equal to 1.

The results of the obtained recognition accuracy as well as the values of three image binarization metrics (PSNR, DRD and MPM) have been stored in vectors consisting of 14 elements each (one image has been removed from the experiments due to improper binarization result in order to prevent its impact on the obtained results). Next, the Pearson's linear correlation coefficients (PCC) with the recognition accuracy have been calculated for each metric.

Individual metrics are defined as

$$PSNR = 10\log\left(\frac{M \times N}{\sum_{m=1}^{M}\sum_{n=1}^{N}\left(GT(m,n) - BW(m,n)\right)^2}\right), \quad (2)$$

where GT is the "ground-truth" image and BW denotes the result of binarization

$$DRD = \frac{1}{NU} \times \sum_{k=1}^{K} DRD_k, \quad (4)$$

where NU is the number of non-uniform (fully black or fully white) 8 × 8 blocks in the "ground-truth" image and K is the number of flipped pixels and for k-th flipped pixel

$$DRD_k = \sum_{i=-2}^{2}\sum_{j=-2}^{2}\left|GT_k(i,j) - BW_k(x,y)\right| \times W(i,j), \quad (5)$$

where W is 5 × 5 normalized weight matrix [9], whereas

$$MPM = \frac{1}{2 \times D} \times \left(\sum_{i=1}^{FN} d_{FN}^{i} + \sum_{j=1}^{FP} d_{FP}^{j}\right), \quad (6)$$

where D is the sum of all the pixel-to-contour distances of the ground truth object, FN and FP are the numbers of false positives and false negatives for which the distances d can be calculated respectively.

In order to increase the correlation of metrics with the OCR accuracy, the Combined Binarization Evaluation Metric (CBEM) has been proposed in the following form

$$CBEM = DRD^{a} \times MPM^{b} \times PSNR^{c}, \quad (7)$$

where a, b and c are the values of the parameters obtained by optimization. Such an idea comes from the general image quality assessment where a similar approach has been successfully applied [12]–[14] leading to significant increase of the correlation of metrics with subjective quality evaluations which are available in several dedicated databases.

IV. EXPERIMENTAL RESULTS

The maximum value of the Pearson's correlation coefficient between the Tesseract OCR accuracy and the proposed CBEM results has been obtained using the MATLAB's functions *fminsearch* and *fminunc*. The obtained values of the parameters of the combined metrics are equal to: a = -1.39, b = -0.83 and c = -4.44 leading to a significant increase of the PCC value from 0.6158 for the best single metric to 0.7145 for the CBEM. The detailed results are presented in Table I together with the result obtained for the unweighted combined metric (without optimization of its parameters).

TABLE I. PEARSON'S CORRELATION COEFFICIENTS BETWEEN THE OCR ACCURACY AND THE BINARIZATION EVALUATION METRICS.

Metric	PCC value
DRD	0.6158
MPM	0.3777
PSNR	0.5743
Unweighted CBEM	0.4221
Proposed	**0.7145**

The additional illustration of the advantages of the proposed approach is provided in Fig. 4–Fig. 7 where the scatter plots illustrating the values of the metrics and the obtained recognition accuracy for each image are presented.

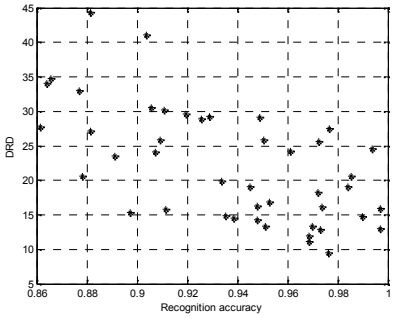

Fig. 4. Scatter plot illustrating the correlation between the DRD metric and the OCR accuracy.

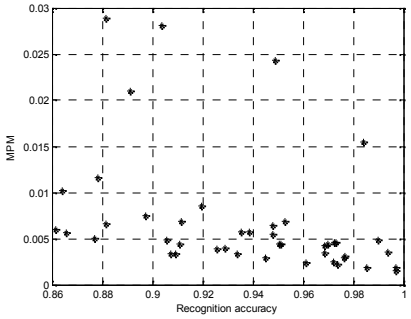

Fig. 5. Scatter plot illustrating the correlation between the MPM and the OCR accuracy.

The improved correlation of the proposed CBEM metric

with the recognition accuracy can be obtained only due to the optimization of weighting coefficients as the PCC value achieved for the unweighted version of the CBEM is lower even than for the use of single DRD or PSNR metric.

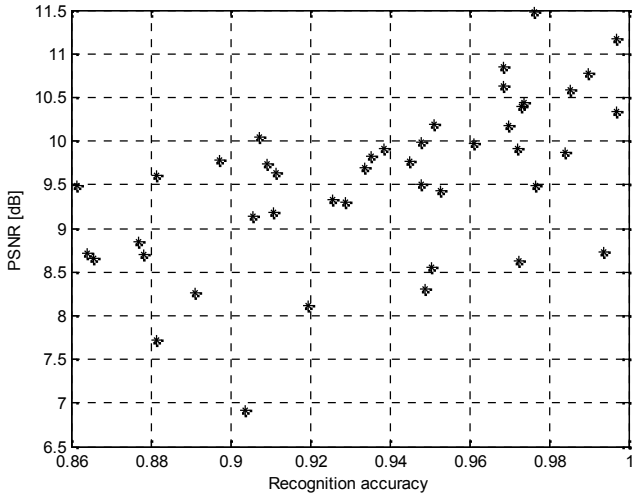

Fig. 6. Scatter plot illustrating the correlation between the PSNR metric and the OCR accuracy.

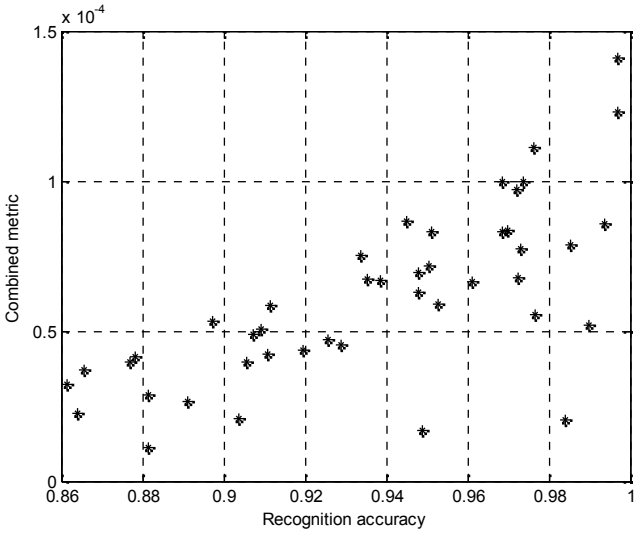

Fig. 7. Scatter plot illustrating the correlation between the proposed metric and the OCR accuracy.

Analysing the scatter plots presented in Fig. 4–Fig. 7 much more linear relationship between the proposed metric and the OCR accuracy in comparison to three single metrics can be easily determined.

V. CONCLUSIONS

The novel approach based on the application and optimization of the combined metric proposed in the paper has led to great results. Since such methods have never been applied in the OCR applications and binary image analysis, it may be an interesting stimulation for the development of new OCR algorithms for degraded quality document images.

The additional verification conducted for more demanding images available in the DIBCO'2011 dataset, widely used by the community for the verification of document image binarization algorithms, has led to similar conclusions. However, due to some troubles caused by the presence of serious degradations as well as some historical gothic font shapes, the OCR accuracy values are not representative and therefore they have not been presented in the paper.

Nevertheless, the obtained results are encouraging for further research which should concentrate on the development of a larger database which should be annotated with the results of character recognition as well as the development of some metrics (preferably no-reference) even better correlated with the OCR accuracy at least for some typical font shapes and typical distortions.

REFERENCES

[1] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, 1979. [Online]. Available: http://dx.doi.org/10.1109/TSMC.1979.4310076

[2] J. Sauvola, M. Pietikainen, "Adaptive document image binarization", *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000. [Online]. Available: http://dx.doi.org/10.1016/S0031-3203(99)00055-2

[3] W. Niblack, *An Introduction to Digital Image Processing*. Prentice-Hall, Englewood Cliffs, 1986, pp. 115–116.

[4] P. Rosin, "Unimoal thresholding", *Pattern Recognition*, vol. 34, no. 11, pp. 2083–2096, 2001. [Online]. Available: http://dx.doi.org/10.1016/S0031-3203(00)00136-9

[5] J. Kapur, P. Sahoo, A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram", *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273–285, 1985. [Online]. Available: http://dx.doi.org/10.1016/0734-189X(85)90125-2

[6] E. H. B. Smith, A. Chang, "Effect of 'ground truth' on image binarization", in *Proc. 10th IAPR Int. Workshop Document Anal. Systems*, Gold Coast, Queensland, Australia, 2012, pp. 250–254. [Online]. Available: http://dx.doi.org/10.1109/DAS.2012.32

[7] H. Z. Nafchi, S. M. Ayatollahi, R. F. Moghaddam, M. Cheriet, "An efficient ground truthing tool for binarization of historical manuscripts", in *Proc. Int. Conf. Document Anal. Recognit.*, Washington, DC, 2013, pp. 807–811. [Online]. Available: http://dx.doi.org/10.1109/ ICDAR.2013.165

[8] K. Ntirogiannis, B. Gatos, I. Pratikakis, "Performance evaluation methodology for historical document image binarization", *IEEE Trans. on Image Processing*, vol. 22, no. 2, pp. 595–609, 2013. [Online]. Available: http://dx.doi.org/10.1109/TIP.2012.2219550

[9] H. Lu, A. C. Kot, Y. Q. Shi, "Distance-Reciprocal Distortion measure for binary document images", *IEEE Signal Proc. Letters*, vol. 11, no. 2, pp. 228–231, 2004. [Online]. Available: http://dx.doi.org/10.1109/LSP.2003.821748

[10] D. P. Young, J. M. Ferryman, "PETS Metrics: On-Line Performance Evaluation Service", in *Proc. 2nd Joint IEEE Int. Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, China, 2005, pp. 317–324. [Online]. Available: http://dx.doi.org/10.1109/VSPETS.2005.1570931

[11] Tesseract-OCR engine. [Online] Available: https://code.google.com/p/tesseract-ocr/

[12] K. Okarma, "Combined image similarity index", *Opt. Rev.*, vol. 19, no. 5, pp. 349–354, 2012. [Online]. Available: http://dx.doi.org/10.1007/s10043-012-0055-1

[13] K. Okarma, "Extended Hybrid Image Similarity – combined full-reference image quality metric linearly correlated with subjective scores", *Elektronika ir Elektrotechnika*, vol. 19, no. 10, pp. 129–132, 2013. [Online]. Available: http://dx.doi.org/10.5755/j01.eee.19.10.5908

[14] T.-J. Liu, W. Lin, C.-C.J. Kuo, "Image quality assessment using multi-method fusion", *IEEE Trans. Image Processing*, vol. 22, no. 5, pp. 1793–1807, 2013. [Online]. Available: http://dx.doi.org/10.1109/TIP.2012.2236343