A Tool for Tuning Binarization Techniques

Vavilis Sokratis
Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
sokratisvav@gmail.com

Ergina Kavallieratou
Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
kavallieratou@aegean.gr

Abstract— **In this paper a user friendly tool appropriate to get user feedback for the application of binarization algorithms is presented. The human feedback is very useful in order to apply next the algorithm to similar images. The tool supports Image Selection and Display, Selection of Binarization Algorithm and Parameter Configuration, Feedback gathering and Creation of log file for further processing.**

Keywords- document image processing; binarization algorithms;user-feedback

I. INTRODUCTION

In the last years, since many systems have reached the best possible performance that a computer system can have the interaction with the user intend to improve them further. In the past, the goal was the system automation. Today, the scientific community has realized that the combination of human and computer can produce systems that combine advantages of both: the human accuracy and the computational speed. Thus user's feedback or/and other intelligent techniques are used in order to improve the results of a system. This procedure has been used in the fields of OCR, Document Processing, Information Extraction and Retrieval, Computer Vision, Natural Language processing, etc.

In the field of image processing, the CAVIAR model has been presented [1]. The distinctive aspect of the CAVIAR technology is a visible, parameterized geometrical model that serves as the human-computer communication channel. Automated algorithms segment each unknown picture, construct a visible model, and extract from the picture of the unknown object a set of features. The candidates are then automatically ranked according to the similarity of their features to those of the unknown picture. These models are constructed automatically, corrected interactively only when necessary. If one of the displayed candidates matches the unknown picture, the user simply clicks on it, thereby classifying the unknown. If not, the user can adjust the visible model. The visible model guides the system in feature extraction. Therefore whenever the visible model is adjusted, new features are extracted, and all the candidates are automatically reordered. CAVIAR flower and face recognition systems show that their accuracy is much higher than that of the machine alone, while their recognition time is much lower than that of the human alone.

As far as it concerns the user's feedback in the fields of document and OCR, a description of most of the existing systems can be found in [2].

Moreover, in [3] a method for accessing the content of Greek historical documents printed during the 17th and 18th centuries by searching words directly in digitized documents based on word spotting, without the use of an optical character recognition engine. User feedback is used in order to refine the search procedure. A word retrieval phase aims to rank the segmented words according to their similarity to the query word. From the initial ranking the words of the document that are similar to the synthetic keyword are obtained. The user selects as input query one or more correct results from the list produced after the initial word matching process. Then, a new matching process is initiated. The segmented words are ranked according to their similarity to the selected word(s) which, in this case, are not synthetic but real words of the document's corpus. The critical impact of the user feedback in the word spotting process lies upon this transition from synthetic to real data.

In [4], a framework to design an adaptive OCR system is proposed. The adaptability lies in the automatic training sample extraction and clustering with limited user interaction. The training part consists of three steps: (i) Template initialization; (ii) Iterative template refinement; and (iii) Template combination and labeling. The post-processed templates are saved into images; the user is required to prune the templates by browsing these images and then assigning a Unicode value to each template. This approach does not require the support of the ground truth text and a corpus, which is useful for the processing of noisy document images where most of the characters touch each other.

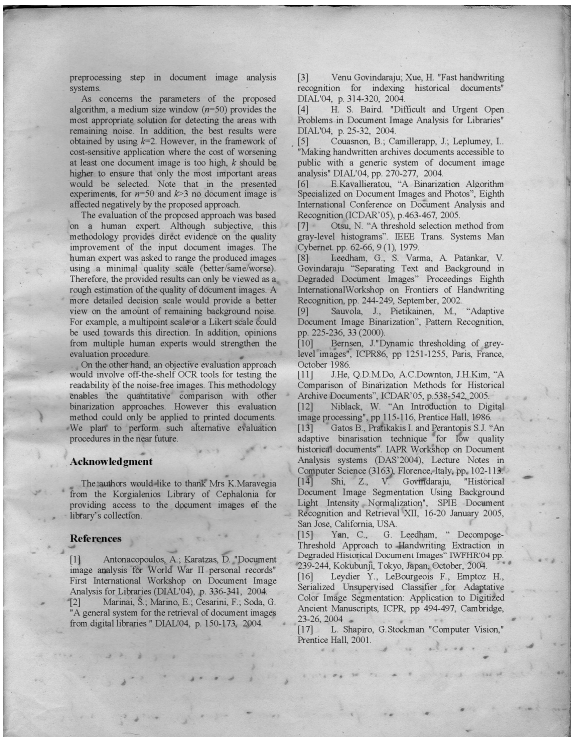
In [5], a software framework, BinarizationShop, combines a series of binarization approaches that have been tailored to exploit user assistance.

In this paper, a tool appropriate to train binarization techniques by gathering user feedback is presented. The system is appropriate to check in detail the performance of a binarization technique on a specific image. The tool tracks the user's reaction and saves the feedback in each repetition of binarization procedure. The user can afterwards study the algorithm dependence on the parameters and adapt the most appropriate for the document image.

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.10

In Section II, the motivation for the tool is explained, while the tool is fully presented in Section III. In Section IV, our conclusions are given.

II. MOTIVATION FOR THE TOOL



preprocessing step in document image analysis systems.

As concerns the parameters of the proposed algorithm, a medium size window (n=50) provides the most appropriate solution for detecting the areas with remaining noise. In addition, the best results were obtained by using k=2. However, in the framework of cost-sensitive application where the cost of worsening at least one document image is too high, k should be higher to ensure that only the most important areas would be selected. Note that in the presented experiments, for n=50 and k=3 no document image is affected negatively by the proposed approach.

The evaluation of the proposed approach was based on a human expert. Although subjective, this methodology provides direct evidence on the quality improvement of the input document images. The human expert was asked to range the produced images using a minimal quality scale (better/same/worse). Therefore, the provided results can only be viewed as a rough estimation of the quality of document images. A more detailed decision scale would provide a better view on the amount of remaining background noise. For example, a multipoint scale or a Likert scale could be used towards this direction. In addition, opinions from multiple human experts would strengthen the evaluation procedure.

On the other hand, an objective evaluation approach would involve off-the-shelf OCR tools for testing the readability of the noise-free images. This methodology enables the quantitative comparison with other binarization approaches. However this evaluation method could only be applied to printed documents. We plan to perform such alternative evaluation procedures in the near future.

Acknowledgment

The authors would like to thank Mrs K.Maravegia from the Korgialenios Library of Cephalonia for providing access to the document images of the library's collection.

References

[1] Antonacopoulos, A.; Karatzas, D. "Document image analysis for World War II personal records" First International Workshop on Document Image Analysis for Libraries (DIAL'04), p. 336-341, 2004.

[2] Marinai, S.; Marino, E.; Cesarini, F.; Soda, G. "A general system for the retrieval of document images from digital libraries " DIAL'04, p. 150-173, 2004.

[3] Venu Govindaraju; Xue, H. "Fast handwriting recognition for indexing historical documents" DIAL'04, p. 314-320, 2004.

[4] H. S. Baird. "Difficult and Urgent Open Problems in Document Image Analysis for Libraries" DIAL'04, p. 25-32, 2004.

[5] Couasnon, B.; Camillerapp, J.; Leplumey, I.. "Making handwritten archives documents accessible to public with a generic system of document image analysis" DIAL'04, pp. 270-277, 2004.

[6] E.Kavallieratou, "A Binarization Algorithm Specialized on Document Images and Photos", Eighth International Conference on Document Analysis and Recognition (ICDAR'05), p.463-467, 2005.

[7] Otsu, N. "A threshold selection method from gray-level histograms". IEEE Trans. Systems Man Cybernet. pp. 62-66, 9 (1), 1979.

[8] Leedham, G., S. Varma, A. Patankar, V. Govindaraju. "Separating Text and Background in Degraded Document Images" Proceedings Eighth InternationalWorkshop on Frontiers of Handwriting Recognition, pp. 244-249, September, 2002.

[9] Sauvola, J., Pietikainen, M., "Adaptive Document Image Binarization", Pattern Recognition, pp. 225-236, 33 (2000).

[10] Bernsen, J."Dynamic thresholding of grey-level images", ICPR86, pp 1251-1255, Paris, France, October 1986.

[11] J.He, Q.D.M.Do, A.C.Downton, J.H.Kim, "A Comparison of Binarization Methods for Historical Archive Documents", ICDAR'05, p.538-542, 2005.

[12] Niblack, W. "An Introduction to Digital image processing", pp.115-116, Prentice Hall, 1986.

[13] Gatos B., Pratikakis I. and Perantonis S.J. "An adaptive binarisation technique for low quality historical documents" IAPR Workshop on Document Analysis systems (DAS'2004), Lecture Notes in Computer Science (3163), Florence, Italy, pp. 102-113.

[14] Shi, Z., V. Govindaraju, "Historical Document Image Segmentation Using Background Light Intensity Normalization", SPIE Document Recognition and Retrieval XII, 16-20 January 2005, San Jose, California, USA.

[15] Yan, C., G. Leedham, " Decompose-Threshold Approach to Handwriting Extraction in Degraded Historical Document Images" IWFHR'04 pp. 239-244, Kokubunji, Tokyo, Japan, October, 2004.

[16] Leydier Y., LeBourgeois F., Emptoz H., Serialized Unsupervised Classifier for Adaptative Color Image Segmentation: Application to Digitized Ancient Manuscripts, ICPR, pp 494-497, Cambridge, 23-26, 2004.

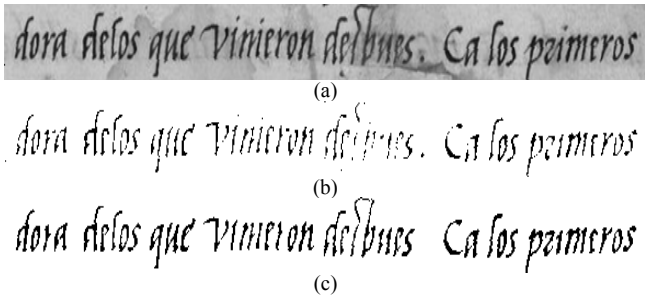
[17] L. Shapiro, G.Stockman "Computer Vision," Prentice Hall, 2001.

Figure 1. The application of the algorithm [7] to an image of the database used in [8].

The recent digitization of large collections of documents has created a necessity for massive binarization of these documents. Although many algorithms have been proposed for the document binarization task, the selection of the most appropriate one or the tuning of the parameters of these algorithms is not a simple procedure [6].

In fig.1 the application of the algorithm [7], appropriate for historical documents, is shown applied to an image of the database used in the contest [8]. Three parameters of the algorithm can be tuned in order to adapt the algorithm performance to each document case. By using the database of the contest [8], those parameters were tuned in order to give the best result for the specific database. This database includes entire documents combined with noise of historical documents.

On the other hand, applying the algorithm with the same parameters on another database with line images from Spanish historical documents, the results are not satisfactory (fig.2a,b). In fig.2c the best result after the parameter tuning is shown.



(a)

(b)

(c)

Figure 2. (a) historical documtn image, (b) binarized image by the parameters of fig.1, (c) binarized image by retuning theparameters.

On the synthetic database mentioned before, the parameters can be tuned automatically and objectively by comparing with the ground truth results, i.e. the original documents: Two sets of images have been combined by using image mosaicing techniques. The first set consists of ten document images in pdf format, including tables, graphics, columns, and many of the typical elements that can be found in a document. The second set consists of fifteen old blank noisy images, taken from a digitized document archive of the 18th century. These documents include most kinds of problems that can be met in old documents: presence of stains, background of big variations and uneven illumination, etc.

Thus, due to the philosophy of the database, the correct final color (black or white) of each pixel is known in advance (pdf files), the performance can be easily evaluated and the parameters can be tuned.

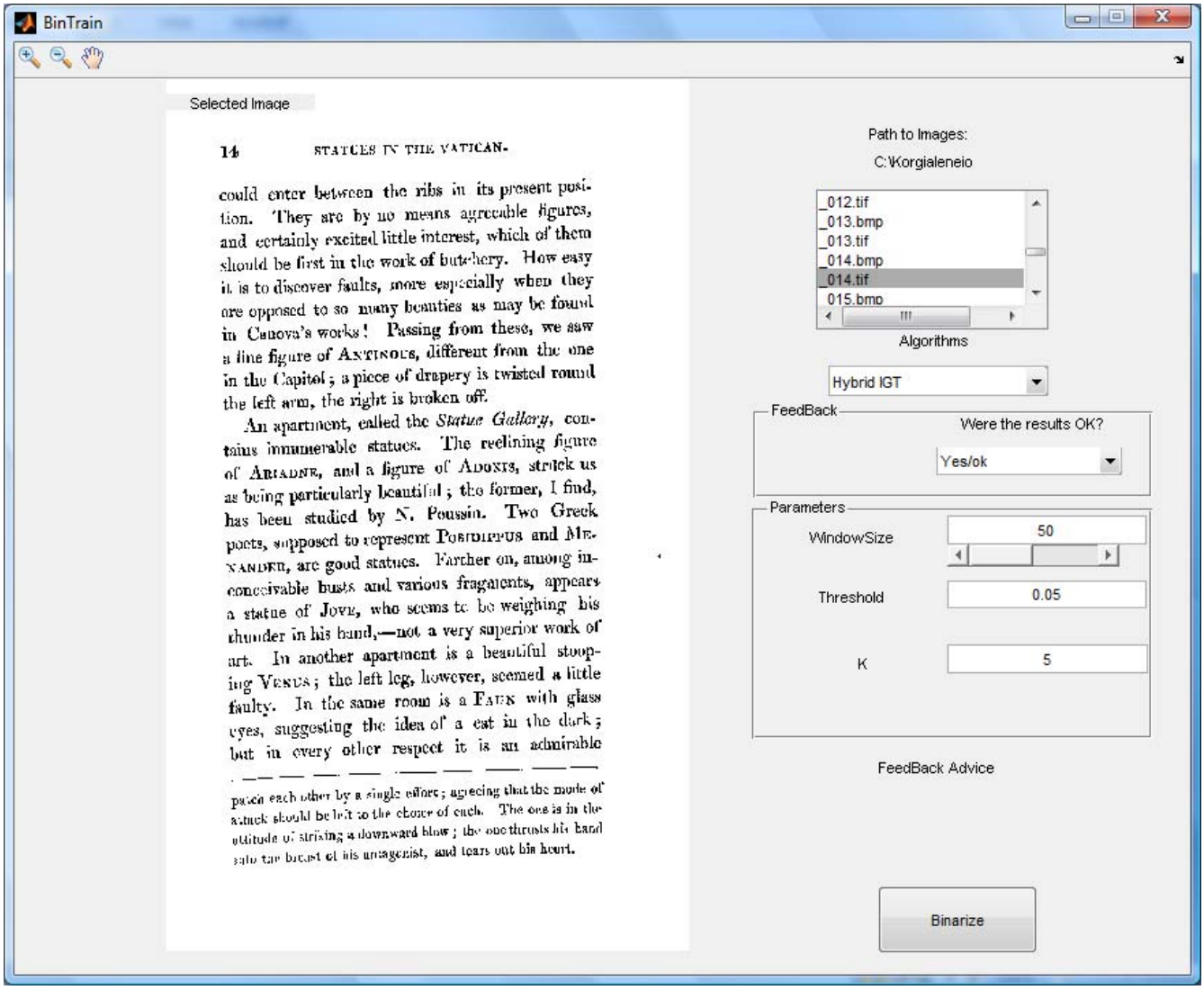


Figure 3. A screenshot of the proposed tool.

On the other hand, this is not possible in the case of real historical collections where the correct color is not known. In these cases only the human user can give an evaluation of the result by choosing the best looking result. Although such feedback is possibly more subjective, it can be very useful e.g. in the case of massive binarization of large document collections.

Such collections very often present a homogeneity, which allows the successful application of any algorithm to the entire collection, after the tuning of its parameters on just few documents. This could be easily done using the proposed tool to gather human feedback.

III. THE PROPOSED TOOL

The proposed tool (fig.3) has been implemented in Matlab because of the ease of implementation and inclusion of very complicated, mathematically, algorithms. The tool supports the functions:

1. Image Selection and Display, including zoom in, zoom out and panning,
2. Selection of binarization algorithm and parameter configuration, with dynamic adjustment of each parameter according to the feedback,
3. Feedback gathering,
4. Creation of a log file for further processing.

These functions are described below in more detail.

A. Image Selection and Dispay

The user can select the image that he wishes to process. The selected image is displayed all the times on the left side and the user can zoom in, zoom out and pan in order to examine the details and provide exact feedback. This can be very useful when the user wants to examine a binarization algorithm, because he is not obligated to open and examine

the binarized image with another tool, leading to a waste of time and effort.

B. *Selection of binarization algorithm and parameter configuration*

The user can select the binarization algorithm he wishes to use by a pop-up menu (see detail in fig.4). The menu includes all the algorithms that have been included in the tool. Moreover the user can change the parameters. This can be done, either right from the start, if he does not agree with the default values that appear in the parameter section or at any time during the successive procedure.

The parameter determination can be done either by moving the slider, when it is available depending on the algorithm, or by typing the value. When feedback is provided by the user, the tool changes the parameters depending on the feedback, according to the instructions determined in the tool. However, the user can still change them as described above.

It is of great importance, the fact that the proposed tool, offers a dynamic and automatic parameter readjustment, which can help even an amateur user. Also, the user is able to notice the changes made, in order to follow the procedure and change the values accordingly. Moreover, it would be interesting the integration of a dynamic procedure in order to readjust the parameter values, providing more intelligent and dynamically adjustable algorithms for the binarization task.

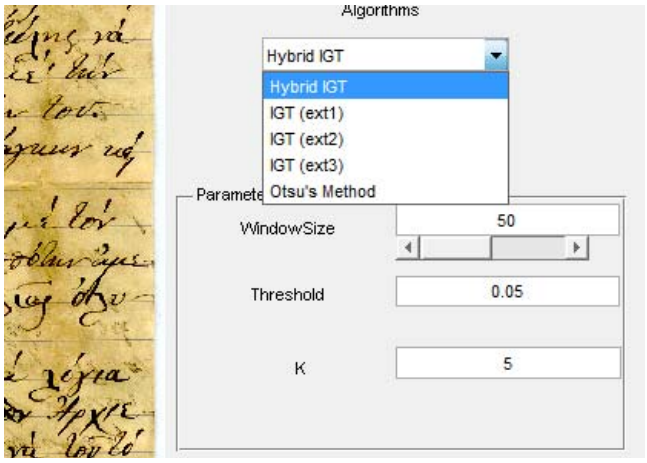


Figure 4. Selection of binarization algorithm and parameter configuration.

C. *Feedback gathering*

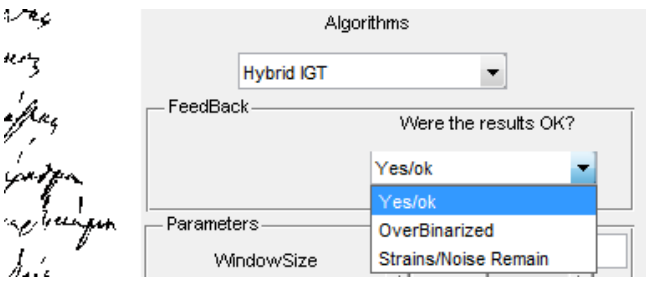


Figure 5. Feedback gathering.

After the first application of the selected algorithm to the image, the user can give his feedback as far as it concerns the results. The user is able to examine the binarized image, using the features provided by the tool and decide about the quality of the binarization according to his needs. The feedback is provided in a user friendly way by a pop-up menu (fig.5). The user can choose between Yes/Ok, OverBinarized or Stains/Noise Remain.

OverBinarized selection shows that the selected binarization algorithm, with the specific parameters, leads to an image where useful information (i.e. text, words) was lost. Stains/Noise Remain choice means that, stains or noise remain in the binarized image and the binarization procedure should me tougher.

In the last two cases the tool will suggest new values for the parameters according to the instructions that have been given during the inclusion of the algorithm in the tool. The new values proposed from the tool according to the feedback, are calculated in relation to the previous ones rather than by reducing or increasing them by a constant value. The user can either accept them or determine other parameter values. In any case the re-application of the algorithm will be done on the original image.

It should be noted that there is no limit in the number of times that this procedure can be repeated.

D. *Creation of log file*

Every time that the user pushes the *Binarize* button (fig.3), the selected binarization algorithm is applied to the original image with the last determined parameter values and a new entry is added in the log file (fig.6). It includes the name of the image, the selected algorithm, the rank of the trial, the values of the current parameters and the feedback that was provided by the user for the specific parameters.

The log file is formatted as Comma Separated Value (CSV) in order to make easier further processing of the data. Assuming that the user can use a CSV parser, or a similar tool, he can make conclusions about the appropriate values for the parameters of the algorithm and the changes needed to be done.

IV. CONCLUSION

In this paper a novel tool appropriate for getting user feedback for the application of binarization algorithms is presented. The binarization procedure, is necessary in many tasks of image processing e.g. medical images, document image processing, OCR, etc.

Many binarization algorithms of general or specific purpose have been proposed. However, the selection of the appropriate algorithm has been proved a difficult task. Moreover, many of the existing algorithms need their parameters to be determined in order to succeed the best result on a specific image or set of images. As in many cases the best result is not predetermined, the human feedback is very useful in order to apply next the algorithm to similar images.

Image: image51a.tif , Algo: Hybrid IGT , Try: 0 ,Window Size: 50 ,Threshold: 0.05 , Times: 5, Feedback: Over Binarized
Image: image51a.tif , Algo: Hybrid IGT , Try: 1 ,Window Size: 60 ,Threshold: 0.05 , Times: 5, Feedback: Over Binarized
Image: image51a.tif , Algo: Hybrid IGT , Try: 2 ,Window Size: 66 ,Threshold: 0.06 , Times: 5, Feedback: Noise Remain
Image: image51a.tif , Algo: Hybrid IGT , Try: 3 ,Window Size: 70 ,Threshold: 0.06 , Times: 5, Feedback: Noise Remain
Image: image51a.tif , Algo: Hybrid IGT , Try: 4 ,Window Size: 80 ,Threshold: 0.06 , Times: 5, Feedback: Noise Remain
Image: image51a.tif , Algo: Hybrid IGT , Try: 5 ,Window Size: 80 ,Threshold: 0.06 , Times: 5, Feedback: Noise Remain
Image: image51a.tif , Algo: Hybrid IGT , Try: 6 ,Window Size: 100 ,Threshold: 0.06 , Times: 5, Feedback: Noise Remain
Image: image51a.tif , Algo: Hybrid IGT , Try: 7 ,Window Size: 100 ,Threshold: 0.06 , Times: 2, Feedback: Noise Remain

Figure 6. The image of fig.1 cleaned by the original technique (left) and the three proposed ones.

The proposed tool provides a user friendly interface for applying a selected algorithm, and setting its parameters. Furthermore, the user can employ the features provided by the tool, in order to closely examine the resulting image. The tool also gets the user feedback and makes readjustments to the parameters accordingly, while keeping a log file for later examination and statistical study.

As future work, an easier procedure for the inclusion of new algorithms is planned, as well as the introduction of more tools that will secure more exact feedback and the automatic binarization of a document image set after the parameter tuning of a specific algorithm. As soon it is ready, the tool with example images, will be available online to the research community.

REFERENCES

[1] Jie Zou and George Nagy, Visible models for interactive pattern recognition, Pattern Recognition Letters 28 (2007) 2335–2342

[2] George Nagy and Sriharsha Veeramachaneni, Adaptive and Interactive Approaches to Document Analysis, Springer, Machine Learning in Document Analysis and Recognition, Volume 90/2008

[3] A Kesidis, E Galiotou, B Gatos, A Lampropoulos, Ioannis Pratikakis, Ioanna Manolessou, Angela Ralli, Accessing the content of Greek historical documents - Proceedings of The Third Workshop on AND '09, July23-24, 2009, Barcelona, Spain.

[4] H. Ma and D. Doermann, Adaptive OCR with Limited User Feedback, 8th Int'l Conf. Document Analysis and Recognition (ICDAR), 2005, pp 814-818 .Marte A. Ramirez-Ortegon, Raul Rojas, "Unsupervised Evaluation Methods Based on Local Gray-Intensity Variances for Binarization of Historical Documents," Pattern Recognition, International Conference on, pp. 2029-2032, 2010 20th International Conference on Pattern Recognition, 2010.

[5] Fanbo Deng, Zheng Wu, Zheng Lu, and Michael S. Brown, "BinarizationShop: a user-assisted software suite for converting old documents to black-and-white", In *Proceedings of the 10th annual joint conference on Digital libraries* (JCDL '10), pp. 255-258, 2010.

[6] Pavlos Stathis, Ergina Kavallieratou and Nikos Papamarkos, "An Evaluation Survey of Binarization Algorithms on Historical Documents", IEEE proceedings of 19th International Conference on Pattern Recognition (ICPR'08), vol. III, pp. 742-745, 2008.

[7] E. Kavallieratou, E. Stamatatos, "Improving the quality of degraded document images", IEEE proceedings of *DIAL*, pp. 340-349, Second International Conference on Document Image Analysis for Libraries (DIAL'06), 2006.

[8] Roberto Paredes, Ergina Kavallieratou, Rafael Dueire Lins, "ICFHR 2010 Contest: Quantitative Evaluation of Binarization Algorithms," 12th International Conference on Frontiers in Handwriting Recognition, pp. 733-736, 2010.