Title of the contest

Quantitative evaluation of binarization algorithms of images of historical documents with bleeding noise.

Names and contact information of organizers

Ergina Kavallieratou
Dept. of Information and Communication Systems Engineering
University of the Aegean, 83200 Karlovassi, Samos, Greece
kavallieratou@aegean.gr

Rafael Dueire Lins
Departamento de Eletrônica e Sistemas
Universidade Federal de Pernambuco
rdl@ufpe.br

Roberto Paredes Palacios
Pattern Recognition and Human Technology Group
Instituto Tecnológico de Informática
Universidad Politécnica de Valencia, 46022, Valencia, Spain
rparedes@dsic.upv.es

Short description of the contest explaining its necessity, potential, relevance

The evaluation and comparison of binarization algorithms proved to be difficult task since there is no objective way to compare the results. Several review papers have tried to compare binarization algorithms by using the precision and recall analysis of the resultant words in the foreground or by evaluating their effect on end-to-end character or word recognition performance in a complete archive document recognition system utilizing OCR.

Every work, that performed comparison, presented some very interesting conclusions. However, the problem is that in all cases, they try to use results from ensuing tasks in document processing hierarchy, in order to survey the algorithm performance. Although in many case this is the objective goal, it is not always possible. Such is the case of historical documents, where their quality, in many cases obstructs the recognition, and sometimes even the word segmentation, this way of evaluation can be proved problematic.

On the other hand, we need different evaluation technique, since the processing of historical documents is one of the hardest cases and binarization can be required for removing the noise and facilitate their appropriate presentation.

The ideal way of evaluation should be able to decide, for each pixel, if it finally has succeeded the right color (black or white) after the binarization. This is an easy task for a human observer but very difficult to do it automatically for all the pixels of several images.

The proposed method includes the experimentation on document archives made by constructing noisy images, using techniques of image mosaicing, and combining old blank historical document pages with noise-free pdf documents. After the application of the binarization algorithms to the synthetic images, it is easy to evaluate the results by comparing the resulted image with the original document as ground truth image. This way, we are able to count the exact amount of the remaining wrong pixels either on background or on foreground.

This technique has been presented and used by the organizers in SAC'07 (Lins), ICPR 2008 (Kavallieratou, Papamarkos), SAC'08 (Kavallieratou).

Short description of the data and the performance evaluation methodology that will be used in the contest

The evaluation of the binarization methods will be made on synthetic images. That is, starting from a clean document image (*doc*), which is considered as the ground truth image, noise of different types is added (*noisy* images). This way, during the evaluation, it is able to decide, objectively, for every single pixel if its value is correct comparing it with the corresponding pixel in the original image.

image	Description
doc_1	only text, variation in columns, variation in type and size of fonts
doc_2	only text, two columns, variation in type and size of fonts
doc_3	two columns, table
doc_4	two columns
doc_5	single column, figure
doc_6	single column, figure, formula
doc_7	printed and handwrittten text
doc_8	single column, figure
doc_9	single column, formulas
doc_10	single column, figure and graphics

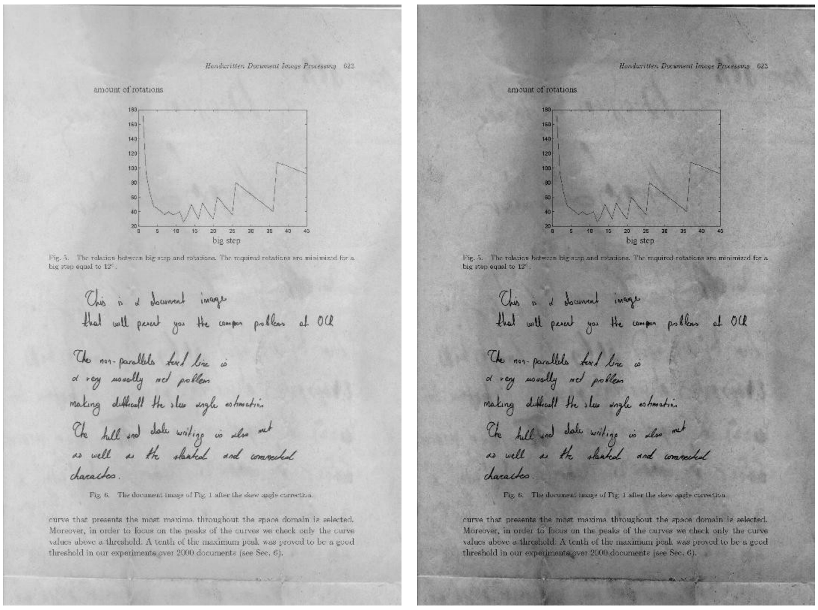
Table 1: Description of doc images.

image	description	Size
noise_1	uneven illumination, ink seepage, stains	1912x2281
noise_2	uneven illumination, ink seepage	1912x2218
noise_3	uneven illumination, ink seepage	1912x2219
noise_4	ink seepage, stains, strains	1188x889
noise_5	stains, strains, stripes	1218x1405
noise_6	uneven illumination, ink seepage, stains	1661x2335
noise_7	uneven illumination, stains	1701x2340
noise_8	uneven illumination, stains, ink seepage	2453x3502
noise_9	uneven illumination, stains	2552x3509
noise_10	background variation, stains	2552x3510
noise_11	background variation, stains	2507x3510

noise_12	uneven illumination, strains	2317x3419
noise_13	uneven illumination, strains, ink seepage	2552x3510
noise_14	uneven illumination, strains	2544x3510
noise_15	background variation, stains	949x595

Table 2: Description and size of noisy images.

Two different sets of 150 document images each of images combined by using image mosaicing techniques are used. The *doc* set consists of ten document images in pdf format, including tables, graphics, columns and many of the elements that can be found in a document. A short description of each document is given in Table 1. The *noisy* set consists of fifteen old blank images, taken from a digitized document digitized archive of the 18th century. These include most kinds of problems that can be met in old documents: presence of stains and strains, background of big variations and uneven illumination, ink seepage etc. Their description as well as their size is shown in Table 2. Samples of both sets are shown in Figure 1. The *docs* are used as target images and all the *noisy* images are resized to A4 size. Then, two different techniques for the blending are used: the maximum intensity and the image averaging approaches.



(a)

(b)

Figure 1: (a) averaging image and (b) maximum intensity image.

As we already mentioned, our intention is to be able to check for every pixel if it is right or wrong. Thus, *pixel error* will be used, that is the total amount of pixels of the image that in the output image have wrong color. A black pixel misclassified as white pixel is denoted by “b” and a white pixel misclassified as black is denoted as “w”. The total number of black and white pixels is B and W respectively. The pixel error is:

$$\frac{b+w}{B+W}$$

On the other hand, the document binarization can be considered an unbalanced problem where the number of black pixels is too much low than the number of white pixels. In this case a better measure could be the geometric-mean. The geometric-mean pixel error is:

$$\sqrt{\frac{b}{B}\cdot\frac{w}{W}}$$

The evaluation of both measures will be done by software that will be available to the participants as well as the data sets.

Short biographies of the organizers depicting their expertise in the field of the contest

Ergina Kavallieratou was born in Kefalonia, Greece, in 1973. She received her Diploma in Electrical and Computer Engineering in 1996 from the Polytechnic School of the University of Patras and her PhD in Handwritten Optical Character Recognition and Document Image processing from the same department in 2000. During the academic year 1997-1998 she was a member of the Signals, Systems and Radiocomunications Laboratory of the Dept. of Telecommunications Engineering of the Polytechnic School of Madrid, while in the years 2000 and 2001, she joined as guest researcher the Institute of Communication Acoustics of Ruhr-Universitaet Bochum, Germany. Since 2001, she teaches Informatics in Greek Open University. During the years 2002-2004, she had a position as Assistant Professor of Audio Processing in Dept. of Audio and Musical Instruments Technology in Technological Educational Institute of Ionian Islands, Greece. Since September 2004, she is a permanent Lecturer in the University of the Aegean. Her research interests include Optical Character Recognition, Document Image Analysis, Computer Vision and Pattern Recognition.

Rafael Dueire Lins was born in Recife (Brazil) in 1960. He graduated in Electrical Engineering at Federal University of Pernambuco at Recife (Brazil) in 1982. He holds a Ph.D. degree in Computing from The University of Kent, England, 1986. His research interests spans from image processing and compression to parallelism and compiler construction techniques. In 1990, Lins started the Nabuco Project for digitalizing, filtering, indexing, transcribing, compressing, storing and retrieving the images of historical documents belonging to the bequest of Joaquim Nabuco, a Brazilian writer and diplomat, one of the leaders of the freedom of black slaves in

Brazil. Lins was possibly the first researcher to address the problem of *back-to-front noise* (later named as *bleeding* or *show-through*) in the technical literature. Professor Lins is currently at the Electronics Department of the Federal University of Pernambuco (Brazil). He wrote over one hundred journal and conference papers and co-authored the best-seller Garbage Collection: algorithms for dynamic memory management published by John Wiley and Sons (1996) and translated into Chinese in 2002.

Roberto Paredes Palacios received the PhD degree in Computer Science in 2003 from the Universidad Politécnica de Valencia, Spain. From 1998 to 2000, he was with the Instituto Tecnológico de Informática working on computer vision and pattern recognition projects. In 2000, he joined the Departamento de Sistemas Informáticos y Computación of the Universidad Politécnica de Valencia, where he is an Associate Professor on the Facultad de Informática. His current fields of interest include statistical pattern recognition, multimedia retrieval, computer vision and biometric identification. In these fields, he has published more than 40 papers in conferences proceedings and journals. Dr. Paredes is the Secretary of the Spanish Society for Pattern Recognition (AERFAI) and Chair of the Technical Committee 5 (Benchmarking and Software) of the International Association for Pattern Recognition (IAPR).

Expected number of participation in the proposed contest

Since the binarization problem is a very active field the last few years, I would expect the participation of around 30 researchers.