A Hybrid Binarization Technique for Document Images

Vavilis Sokratis, Ergina Kavallieratou, Roberto Paredes,
and Kostas Sotiropoulos

Abstract. In this chapter, a binarization technique specifically designed for historical document images is presented. Existing binarization techniques focus either on finding an appropriate global threshold or adapting a local threshold for each area in order to remove smear, strains, uneven illumination etc. Here, a hybrid approach is presented that first applies a global thresholding technique and, then, identifies the image areas that are more likely to still contain noise. Each of these areas is re-processed separately to achieve better quality of binarization. Evaluation results are presented that compare our technique with existing ones and indicate that the proposed approach is effective, combining the advantages of global and local thresholding. Finally, future directions of our research are mentioned.

Keywords: Document Image Processing, Historical Document Images, Binarization Algorithm, Hybrid Algorithm.

1 Introduction

Documents can be a valuable source of information but often they suffer degradation problems, especially in the case of historical documents, such as strains, background of big variations and uneven illumination, ink seepage, etc. Binarization techniques

Vavilis Sokratis · Ergina Kavallieratou
Dept. of Information and Communication Systems Engineering
University of the Aegean
Greece
e-mail: sokratisvav@gmail.com, kavallieratou@aegean.gr

Roberto Paredes
PRHLT - Universidad Politecnica de Valencia
Spain
e-mail: rparedes@dsic.upv.es

Kostas Sotiropoulos
University of Patras
Greece
e-mail: kosotiro@upatras.gr

M. Biba and F. Xhafa (Eds.): Learning Structure and Schemas from Documents, SCI 375, pp. 165–179.
springerlink.com

should be applied to remove the noise and improve the quality of the documents. A sample document image is shown in figure 1.

Document binarization is a useful and basic step of the image analysis systems. When applied, it converts gray-scale document images to black and white (binary) ones. In the transformed binary image the background is represented by white pixels and the text by black ones. By using binarization, the problems, mentioned before, are treated in order to provide a document form more suitable for further processing.

No matter how simple and straightforward, this procedure seems, it has been proved to be a complex task. The binary document image is essential to have good quality in order to proceed to the further stages of document analysis independent whether we are interested in performing OCR, or document segmentation, or just presentation of the document after some restoration stages [1]. Knowledge can be extracted from the documents and such systems are used in many applications, from electronic libraries or museums to search engines and other intelligent systems [2, 3]. Any remaining noise, due to bad binarization, could reduce the performance of the forthcoming processing stages and in many cases could even cause their failure.

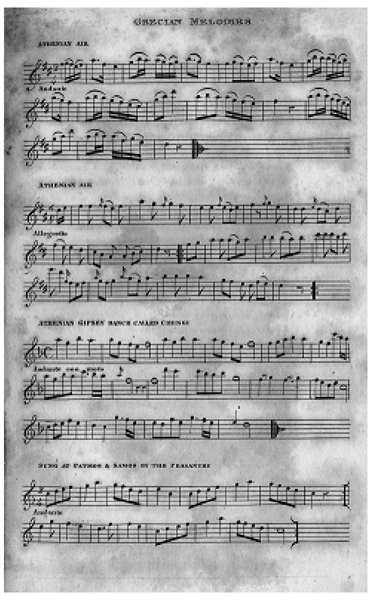


Fig. 1 Sample Document Image

In this chapter, a binarization technique that can be applied to both regular and historical document images is presented. It is a hybrid binarization approach that attempts to combine the advantages of both global and local thresholding. It is important to mention the basic information about the framework of this hybrid technique. First, a global thresholding technique is applied to the entire image.

Then, the image areas that still contain background noise are detected and the same technique is re-applied to each noisy area separately. The proposed hybrid framework is summarized as:

- Application of Global Thresholding Algorithm,
- Detection of "Noisy" Areas, and
- Application of Global Thresholding Algorithm to the detected areas.

By selecting only specific areas of the image for further thresholding, the cost of applying local thresholding is reduced. Moreover, the integral images [15] are used in order to further reduce the computational cost. Hence, by the hybrid approach, better adaptability of the algorithm is achieved in cases where various kinds of noise coexist on the same image, as in local binarization techniques, while keeping low the computational and time cost, as in global ones, since only a limited number of areas (instead of the entire image) need to be processed separately.

The rest of this chapter is organized as follows: Section 2 includes related work while the section 3 describes our approach in more detail. Section 4 presents comparative experimental results for the evaluation of the technique. Section 5 covers a discussion about the algorithm and the future work directions.

2 Related Work

Many different approaches and algorithms have been proposed for the document binarization task. Most of them can be categorized, according to the way they treat the problem, in two categories:

1. Global thresholding techniques: The pixels of the image are classified into text or background according to a global threshold. Usually, such techniques are fast. On the other hand, they are not effective in case the background noise is unevenly distributed in the entire image (e.g. presence of smear or strains) [4, 11].
2. Local thresholding techniques: The pixels of the image are classified into text or background according to a local threshold determined by their neighborhood area. Such techniques are more robust to the presence of different kinds of noise in the image. On the contrary, they are significantly more complex and time-consuming [5, 6].

Global thresholding binarization algorithms employ methods based on classification procedures, histogram, clustering, entropy and Gaussian distribution [7]. One of the oldest techniques is Otsu's [4] which calculates a global threshold based on a foreground and a background class. The threshold that minimizes the interclass variance of the thresholded black and white pixels, is selected by the algorithm. Clustering techniques are also used, e.g. methods based on K-means algorithm [8] to cluster the gray-level pixels in two clusters: background and foreground.

Such a global algorithm is the Iterative Global Thresholding (IGT) [9], an approach specifically designed for document images. This technique has the additional advantage of providing the option to maintain the image in grey-scale after the removal of background noise, a friendlier form for human readers.

Local thresholding binarization algorithms employ methods based on clustering procedures, local variation, entropy, neighborhood information, and Otsu's method [7]. Some techniques are Niblack's [6] that uses local mean and standard deviation, Bernsen's [5] which calculates local thresholds using neighbors and Sauvola's [10] which applies two different algorithms to determine a different threshold for each image and finally binarize the document image.

From a different point of view, binarization approaches can be divided as follows:

- General-purpose methods: Methods that may be applied to any image without taking into account specific characteristics of document images.
- Document image-specific methods: They attempt to take advantage of document image characteristics (e.g., background pixels is the majority, foreground pixels are in similar grey-scale tones, etc). In many cases, such methods are variations of general-purpose approaches [10].

Finally, another binarization category is the hybrid approach. Either being general or document specific binarization algorithm, these techniques combine element from both global and local thesholding techniques. They are based on the intuition that a hybrid algorithm can be fast like global thresholding ones while providing high quality binarization results like the local thresholding algorithms. Our proposal, as mentioned before, is based on this concept, and uses a global thresholding algorithm (IGT) and some noise area detection techniques.

3 Algorithm Description

In this section, the hybrid binarization approach that deals with document images is presented. As input, a grayscale document image is considered where document specific characteristics such as text or graphics are placed on the foreground and outrange over the background. The input images are described by the equation:

$$I(x, y) = r, r \in [0,1] \tag{1}$$

where x and y are the horizontal and vertical coordinates of the image, and r can take any value between 0 and 1 with 0 value representing a black pixel and 1 value a white pixel. The aim is the transformation of all the intermediate valued pixels to finally lie, on the background, pixel value 1 or, on the foreground, pixel value 0. The algorithm is based on the fact that a document image consists of few pixels of useful information placed on the foreground, such as characters, compared to the total size of the image [9]. Taking advantage of this fact, we assume that the average value of the pixel values of a document image is determined mainly by the background even if the document is quite clear of noise, which is quite helpful for a global threshold determination. Thus, a global technique may be used. In addition to this, a method for fixing specific document problems must be proposed which leads to the hybrid approach.

The algorithm first applies a global thresholding technique (the IGT), to the document image. Then, the areas that still contain noise are detected and re-processed separately. In more detail, the proposed algorithm consists of the following steps:

- Application of IGT to the document image.
- Noisy area detection (areas with remaining noise).
- Application of IGT to each detected area separately.

Next, the analysis of the above steps follows.

3.1 Application of Iterative Global Thresholding

Iterative Global Thresholding (IGT) method is both simple and effective. It selects a global threshold for a document image based on an iterative procedure. In each iteration *i*, the following steps are performed:

Average pixel value calculation (Threshold T_i).

1. Subtraction of T_i from each pixel.
2. The grayscale histogram is stretched so that the remaining pixels to be distributed in all the grey scale tones.
3. Repetition of steps 1-3 till the termination condition is fulfilled.
4. Binarization of the final image.

The calculation of the T_i, threshold used in i-th repetition, for an MxN document image, is given by the formula:

$$T_i = \frac{\sum_x \sum_y I_i(x,y)}{MxN} \tag{2}$$

where $I_i(x,y)$ is the image after the (i-1)-th repetition. Keeping in mind that 1s stand for background and 0s for foreground the formula used for the subtraction that provides the after-subtraction and before-equalization image I_s is:

$$I_s(x,y) = I_i(x,y) - T_i + 1 \tag{3}$$

In each repetition, after the subtraction, a lot of pixels are moved to the side of the background and the rest of the pixels are fading. After the subtraction step, the intensity of the image is adjusted by using the histogram and extending the values to all the grey-scale range from 0 to 1. The background pixels retain their values while the rest of the pixel values should extend from 0 to 1. The relation used for the histogram stretching is:

$$I_{i+1}(x,y) = 1 - \frac{1 - I_s(x,y)}{1 - E_i} \quad (4)$$

where I_s is given by the equation (3) and E_i is the minimum pixel value in the image I_s during the i-th repetition, just before the histogram stretching. The whole procedure is repeated the necessary times till the document image is satisfactorily cleaned. Each repetition removes more stains from the image. The number of iterations depends on the image and the intensity of any existent stains, crumples and lighting effects on the image.

The necessary amount of repetitions depends very much on the document image, as well as on the required result, thus the termination condition of the algorithm or the specification of repetition boundary is the next step. In our implementation a maximum iteration upper bound of 20 rounds is posed, because in our experiments the algorithm never exceeded it. However, the process after the first repetitions is very slow. Finally, after many experiments, it was concluded that the amount of transformed pixels in each repetition compared to the previous one is an objective measure. Thus, the iterations stop based on the following criterion:

$$|T_i - T_{i-1}| < 0.05 \quad (5)$$

Having already concluded to the appropriate final stage, the image is binarized by turning all the pixels that are not white (value 1) to black (value 0).

3.2 Noisy Area Detection

The detection of areas that need further processing is performed by using a simple method. The key idea is based on the fact that the areas that still contain background noise will include more black pixels on average in comparison with other areas, or the whole image. This is reasonable especially for document images that only include textual information.

The image is divided into segments, denoted by (S), of fixed size nxn. In each segment, the frequency of black pixels is calculated. The segments that satisfy the following criterion are, then, selected as:

$$f(S) > m + ks \quad (6)$$

where f(S) is the frequency of the black pixels in the segment S while m and s are the mean and the standard deviation of the black pixel frequency of the entire page, respectively. The selected segments form areas by connecting neighboring segments in respect to their original position in the image. The row-by-row labeling algorithm [12] is used for scanning the document by the nxn window.

The parameter k in the formula determines the sensitivity of the detection method. The higher the k, the less segments will be detected. This could mean that some of the areas that may still need further improvement will not be selected. On the other hand, a low k guarantees that all the areas that still need improvement will be selected however together with other areas in which the noise has already been removed. Moreover, the computational and time cost of the global thresholding step will increase. Therefore, an appropriate value of k should be selected to deal with this trade-off. The window segment size (nxn) is also an important factor both for the noisy areas selection and the successful re-application of IGT (as presented in the next subsection). For the area selection procedure a small window could select areas that not need further processing. If the window is really small, document information e.g. bold characters, may be marked as noisy areas because the amount of black pixel will be higher than the average of the image and this may lead to wrong clustering of more pixels to the background. In the proposed version of the hybrid algorithm k was assigned a value of 2 and the segment window size was set to 30x30 pixels.

3.3 Re-application of IGT (Local Thresholding)

The areas detected by the previously described procedure are separately reprocessed based on local thresholding. It is motivated by the belief that, reapplication of the IGT method to a smaller and "noisy" segment will further clean the specified area. The algorithm may adjust better to the characteristics of the area and lead to a higher quality result. For a given image segment, the IGT global thresholding method is applied to the corresponding area of the original image. The process stops when either the termination condition (5) is satisfied or the number of iterations exceeds the corresponding number of iterations previously required for the global thresholding on the entire image.

Taking into account that the selected regions of the image have relatively high average density of black pixels, IGT removes a lot of black pixels during the first iterations. In comparison to the application of IGT to the entire page, the background noise in the selected areas is more likely to be removed since the area is likely to be more homogeneous than the entire image. In general, this procedure tends to move more pixels of the selected areas to the background in comparison with the previous application of IGT to the entire image.

As mentioned before the window size used for noisy area selection is a critical factor for the success of the iterative thresholding binarization of the selected areas. It is obvious that a small window size forms many but rather small segments. This strategy has the advantage of processing the areas in more detail and adapting parts that contain noise. On the other hand, the resulting areas are too small to provide the necessary information for successful application of the IGT. In case of large window size, fewer but bigger areas are detected. This provides enough information to the IGT algorithm in order to effectively remove background noise, but the areas cannot be easily adapted to a specific part of the image that still contains noise. As a consequence, the final image may contain neighboring areas that have dissimilar amount of background noise.

4 Experimental Results

In order to compare the proposed technique with other ones, the comparative results presented in detail in [7] are next shown.

The evaluation of the binarization methods was 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 can be objectively decided for every single pixel if its final value is correct comparing it with the corresponding pixel in the original image. Two sets of images were combined by using image mosaicing techniques [13]. In the first case, the maximum intensity technique (max_int), the new image was constructed by picking up for each pixel in the new image, the darkest corresponding pixel of the two images. This means that in case of foreground, the doc would have a lead over the noisy, but in the background we would have the one from the noisy image since it is almost always darker than the document background that is absolutely white. This technique has a good optical result but it is not very natural as the foreground would be always the darkest, since it is not affected at all from the noise. This set permits us to check how much of the background can be detracted. However, in order to have a more natural result, we also used the image averaging technique (ave-int), where each pixel in the new image is the average of the two corresponding ones in the original images. In this case, the result presents a lighter background than that of the maximum intensity technique but the foreground is also affected by the level of noise in the image.

The intention is to check if every pixel was binarized in the right way. For the evaluation the following metrics were used: Pixel error, that is the total amount of pixels of the image that in the output image have wrong color: black if white in original document or white if black originally. Thus, the pixel error rate (PERR):

$$PERR = \frac{pixelerror}{MxN} \tag{7}$$

Also traditional measures of image quality description were used such as the square error (MSE), the signal to noise ratio (SNR) and the peak signal to noise ratio (PSNR) [14]. In tables 1 and 2 the evaluation result of several binarization algorithms using the metrics mentioned above are shown. For more details on the mentioned techniques please check [7].

Although there is a slightly better performance of the local binarization techniques vs. the global ones, the global ones based on histograms or classification techniques present almost as good results as the local ones. There is no obvious dependence of the algorithm performance on how recent the algorithm is.

During the evaluation procedure, Hybrid IGT exceeds the average performance. Although being very simple and fast, the algorithm in many cases performed better than very complex ones.

Table 1 The evaluation metrics for ave-int technique

	MSE	SNR	PSNR	PERR	PERR variat.
Johansen	1030.09	18.29947	18.49771	1.584145	0.39702
Li	1064.482	18.11257	18.31684	1.637035	0.39598
Reddi	1067.702	18.1055	18.31057	1.641987	0.407469
ALLT	1080.179	18.00511	18.20386	1.661175	0.374789
Gatos	1082.475	18.13241	18.39107	1.664706	0.950808
Vonikakis	1116.98	17.86367	18.08074	1.717771	0.416447
Otsu	1136.256	17.82513	18.03767	1.747414	0.653088
Fuz.C-means	1143.395	17.85714	18.04058	1.758393	0.521405
Bernsen	1148.187	17.75559	17.96667	1.765763	0.443693
Ramesh	1317.732	17.48979	17.65106	2.026501	1.453468
Palumbo	1388.759	16.88965	17.10351	2.135731	0.461838
Koh. SOM	1479.509	17.33808	17.57842	2.275293	11.58626
Sauvola	1493.785	16.5649	16.80586	2.297247	0.77009
Hybrid IGT	1592.008	16.22354	16.31476	2.448303	0.435930
Black Perc.	1626.66	15.93992	16.15941	2.501591	0.354353
Brink	1956.728	16.05018	16.15053	3.009194	3.234369
Kapur	1958.988	15.41409	15.69104	3.012669	1.64126
IIFA	2043.185	15.25285	15.58478	3.142154	1.827724
Yen	2080.253	15.2717	15.55781	3.199158	4.127165
Hist. peaks	2184.715	15.7486	15.91618	3.359808	15.59393
Abutaleb	4079.849	11.6206	12.08744	6.274278	1.246441
Parker	8455.937	8.187518	8.912703	13.00413	4.279259
K-means	9069.963	14.99826	15.19859	13.94842	992.2235
Kittler	14453.05	8.047554	9.957478	22.22692	639.8624
Niblack	15780.57	4.806632	6.192451	24.26846	12.07129
Riddler	15970.74	6.585206	8.091815	24.56092	213.2842
Rosenf.Kak	18277.45	3.958347	5.613259	28.10834	36.88286
Lloyd	19626.18	3.494714	5.314108	30.18251	46.62763
Mardia	19973.03	3.291748	5.205405	30.71592	34.67375
Pun	27847.65	0.950004	3.697339	42.82607	12.31683

Table 2 The evaluation metrics for max-int technique

	MSE	SNR	PSNR	PERR	PERR variat.
Johansen	1105.647	17.9324	18.1326	1.700341	0.4167
Li	1176.348	17.66227	17.86199	1.80907	0.4898
Reddi	1712.938	16.28393	16.54482	2.634276	3.4973
ALLT	1772.267	15.51629	15.73644	2.725517	0.2751
Gatos	1843.791	15.66916	15.92546	2.83551	1.3631
Vonikakis	1875.928	15.88174	16.01977	2.884933	5.1598
Otsu	2350.078	14.64512	14.99464	3.614115	2.7403
Fuz.C-means	2587.894	16.40223	16.72078	3.979844	52.695
Bernsen	2595.835	14.18978	14.54335	3.992057	3.5038
Ramesh	2795.906	14.99625	15.36254	4.299741	15.996
Palumbo	2922.703	14.58387	14.9179	4.494738	14.289
Koh. SOM	4388.948	14.85988	15.36612	6.749631	161.52
Sauvola	4404.042	11.2722	11.73721	6.772845	0.9628
Hybrid IGT	5842.581	13.26888	13.87394	8.98513	150.38
Black Perc.	6242.384	12.80606	13.44569	9.599975	157.17
Brink	6356.625	12.45118	13.08459	9.775664	138.09
Kapur	8952.282	7.901008	8.661	13.76745	4.3483
IIFA	9014.171	3.062373	4.19233	13.86262	0.1431
Yen	9285.395	11.90665	12.82858	14.27973	314.35
Hist. peaks	11824.21	11.68115	12.30377	18.1841	683.19
Abutaleb	13901.2	9.544554	11.21195	21.37825	721.94
Parker	15288.62	5.023332	6.333367	23.51191	12.132
K-means	16567.28	5.726616	7.271567	25.47832	165.41
Kittler	18423.7	1.403792	8.922133	28.33326	1399.4
Niblack	18582.36	3.881317	5.582062	28.57725	49.992
Riddler	22771.88	2.429526	4.65036	35.0202	49.996
Rosenf.Kak	23270.71	2.789177	11.63249	35.78733	2026.4
Lloyd	23486.78	-1.63902	7.439523	36.11962	1494.8
Mardia	27002.61	2.016883	5.234863	41.5265	549.26
Pun	29081.33	0.61298	3.506799	44.72331	11.337

5 Discussion – Future work

A binarization technique has been presented that combines the advantages of global and local binarization. As it has been shown the method consists of several modules, such as IGT (global thresholding), noisy area detection and others, that can be extended or changed. As future work, in order to make the algorithm more effective in terms of qualitative binarization, the following points will be further studied:

- Parameters tuning
- Changes in the global thresholding procedure
- Changes in the noisy area detection
- Post processing steps

It is expected that the proposed algorithm is able to provide better results if it is properly tuned. For example the algorithm parameters, such as k or the window size nxn, can be properly set according to the document image, in order to increase the binarization quality. This tuning can be the result of a series of experiments or a user feedback system.

Of particular interest is the dynamic determination of the parameters values by the algorithm on its own. The algorithm can collect information about the document image, like the image dimensions or the variance in the mean pixel value, and decide the proper parameter values. An interesting case could be that the dynamic selection of the threshold formula and the calculation of the scanning window used for the "noisy" area detection to be determined according to the image size, e.g. 10% of the image size on each dimension, and the average pixel value. This process can be also assisted by human feedback, in various ways, helping the algorithm by correcting inappropriate decisions.

The global thresholding procedure is one of the fundamental components of the hybrid algorithm. The overall result depends heavily on its effectiveness. Aspects that could be extended or changed concern the determination method of the threshold used for the "pixel" shifting to the background, after each iteration. An initial proposal could be the extension of the threshold by adding a variance factor this would result to this threshold formula:

$$T_i = m+ks \tag{8}$$

where T_i is the threshold, s is the standard derivation and k is a tuning parameter [6]. Another alternative could be the relation of the threshold to the noise of the image e.g. the Signal to Noise Ratio (SNR) could be used as a threshold.

Additionally, it has to be pointed out that the termination criterion of the iteration process is a critical factor of the proposed method. Currently, the criterion applied is the ratio of the pixels shifted between two successive iterations. A global boundary condition could be used instead, e.g. the amount of black pixels not to exceed the 10% of the total pixel count or by the use of SNR. In the former case, it is based on a ground truth hypothesis, that the document has a homogenized structure and the amount of black pixel ranges in a specific ratio. In the later case, the total remaining noise should be below a limit, or the reduction ratio of noise between two iterations should be satisfactory.

So far we mentioned that the detection of a problematic area in the image is based on the frequency of the black pixels inside the segment. Again the detection can be based on a ground truth belief of a global boundary e.g. the amount of black pixel doesn't exceed the 10% of the image. By setting a global boundary, the algorithm is forced to correct all the regions that break the condition. Here it is also possible to apply a detection condition based on the image noise, similarly to the termination criterion proposed before.

Although a post-processing step in the algorithm may reduce the performance by raising the total computation cost, it can improve the overall quality of the document. After the application of the hybrid binarization methodology, several filters could be applied to the image fixing specific problems. An illustrative filter could clean any black spots (salt-and-pepper noise) from the final image. This way, small black regions will be removed without destroying any letters or figures of the binarized image.

Summarizing, some alternatives that focus on key points of the method, are planned to be further studied:

Proposal 1: The threshold calculation and termination condition of the global thresholding iteration is changed. Initially, the threshold is calculated by the formula [6]:

$$T_i = m+ks \tag{9}$$

The threshold calculation includes not only the average pixel value m of the image but also the standard derivation s multiplied by the sensitivity parameter k. This way, the threshold represents a more representing barrier between background and foreground elements. In addition the termination condition is configured accordingly, to fit the new threshold, as mentioned before. All the other functionality of the algorithm remains the same.

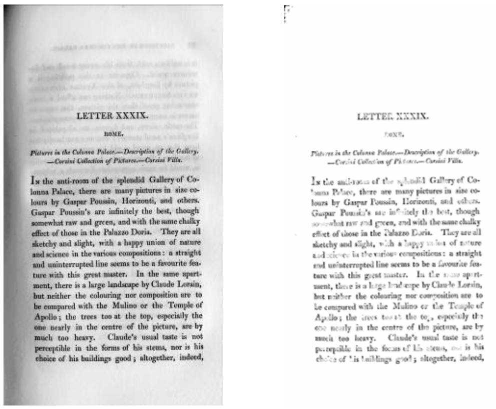
Proposal 2: The noisy area detection criterion is changed. The detection of the problematic area is based on noise detection. As a criterion the Signal to Noise Ratio (SNR) is used. For each window, the SNR of the region is calculated and then compared to the SNR of the whole image. The condition is represented by the formula:

$$\text{SNR(Region)} <= \text{SNR(Image)} \tag{10}$$

All the other functionality of the algorithm remains the same.

Proposal 3: It is a combination of the proposals 1 and 2 described before. The threshold calculation is based on the formula (9) and the segment detection on the formula (10). All the remaining functionality remains the same. By combining the two different extensions, it is expected that the method will benefit from both, providing better binarization results.

In figures 2-4 the basic hybrid algorithm and the results of the three proposals are presented for the same document image.



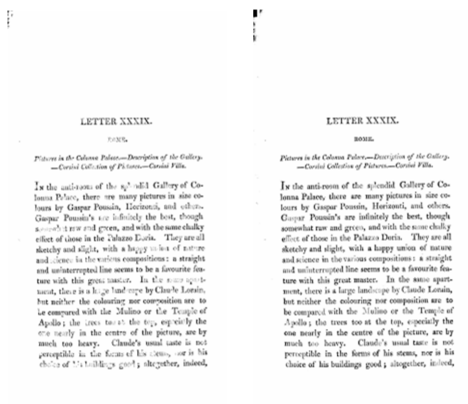
LETTER XXXIX.

ROME.

Pictures in the Colonna Palace.—Description of the Gallery.—Corsini Collection of Pictures.—Corsini Villa.

In the anti-room of the splendid Gallery of Colonna Palace, there are many pictures in size colours by Gaspar Poussin, Horizonti, and others. Gaspar Poussin's are infinitely the best, though somewhat raw and green, and with the same chalky effect of those in the Palazzo Doria. They are all sketchy and slight, with a happy union of nature and science in the various compositions: a straight and uninterrupted line seems to be a favourite feature with this great master. In the same apartment, there is a large landscape by Claude Lorain, but neither the colouring nor composition are to be compared with the Mulino or the Temple of Apollo; the trees too at the top, especially the one nearly in the centre of the picture, are by much too heavy. Claude's usual taste is not perceptible in the forms of his stems, nor is his choice of his buildings good; altogether, indeed,

Fig. 2 Document image (left) after applying the Hybrid Method (right)



LETTER XXXIX.

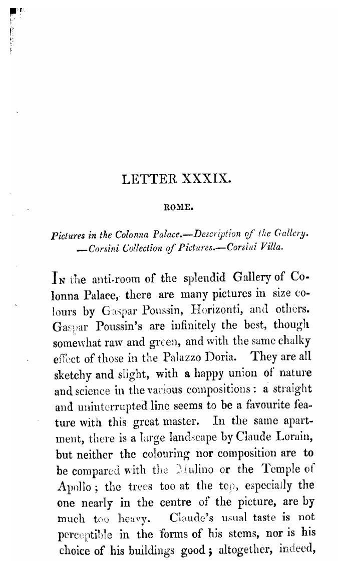
ROME.

Pictures in the Colonna Palace.—Description of the Gallery.—Corsini Collection of Pictures.—Corsini Villa.

In the anti-room of the splendid Gallery of Colonna Palace, there are many pictures in size colours by Gaspar Poussin, Horizonti, and others. Gaspar Poussin's are infinitely the best, though somewhat raw and green, and with the same chalky effect of those in the Palazzo Doria. They are all sketchy and slight, with a happy union of nature and science in the various compositions: a straight and uninterrupted line seems to be a favourite feature with this great master. In the same apartment, there is a large landscape by Claude Lorain, but neither the colouring nor composition are to be compared with the Mulino or the Temple of Apollo; the trees too at the top, especially the one nearly in the centre of the picture, are by much too heavy. Claude's usual taste is not perceptible in the forms of his stems, nor is his choice of his buildings good; altogether, indeed,

Fig. 3 Results of proposal 1 (left) and 2 (right)

The original document image presents problems such as background of uneven illumination and seepage of ink. The hybrid algorithm fixes many of the problems and makes a good quality binarized images. By a closer look, we can see that a small region in the upper left corner remains, and that some of the characters in the sides of the image are a bit faded out. Although not a big problem for a human, this fading problem may be critical to an OCR tool. Proposal 1 solves this problem by providing more clear characters while keeping the background noise in the same level with the original hybrid approach. In the case of Proposal 2, it should be mentioned that even more solid characters are provided after the binarization, but the level of background noise is higher. Finally, the combination of both extensions leads to the best results. Background noise is kept low, while the quality of the characters is better than the other experiments.



LETTER XXXIX.

ROME.

Pictures in the Colonna Palace.—Description of the Gallery.—Corsini Collection of Pictures.—Corsini Villa.

In the anti-room of the splendid Gallery of Colonna Palace, there are many pictures in size colours by Gaspar Poussin, Horizonti, and others. Gaspar Poussin's are infinitely the best, though somewhat raw and green, and with the same chalky effect of those in the Palazzo Doria. They are all sketchy and slight, with a happy union of nature and science in the various compositions: a straight and uninterrupted line seems to be a favourite feature with this great master. In the same apartment, there is a large landscape by Claude Lorain, but neither the colouring nor composition are to be compared with the Mulino or the Temple of Apollo; the trees too at the top, especially the one nearly in the centre of the picture, are by much too heavy. Claude's usual taste is not perceptible in the forms of his stems, nor is his choice of his buildings good; altogether, indeed,

Fig. 4 Result of Proposal 3

It is obvious from the results, that the change in the calculation of the threshold makes the algorithm more objective. The distinction barrier between background and foreground elements is clearer, so the characters are not fading. Also the change in the selection of the noisy areas resulted in even more clear text on the binarized images but with the drawback of background noise. This happens because after the global thresholding step, fewer areas are selected by the algorithm (only those with relatively higher noise ratio). The combined Proposal 3 pointed out the best result among the other extensions. On the one hand, the change in the threshold calculation made the global thresholding procedure and every local iteration more efficient in the removal of background noise. On the other hand, the SNR noisy area detection technique chooses only areas with high noise ratio.

More experiments on the extension of the algorithm should be performed in order to expose the strength of the proposed hybrid binarization framework. Also the algorithm is planned to be extended to a more dynamic and human assisted procedure, in order to adapt to each image characteristics and provide better binarization.

6 Conclusion

In this Chapter we presented a hybrid binarization approach aiming at the removal of background noise from document images. This way, we attempt to combine the advantages of global and local thresholding, that is, better adaptability of various kinds of noise at different areas of the same image in low computational and time cost.

The evaluation results indicate that the proposed approach provides quality results and is comparable to other global and local thresholding techniques. Initial proposals for the extension of the hybrid approach were presented, in order to increase the efficiency and the quality of the binarized images. The next step is to study more extensions and combine them with dynamic and human assisted systems. Finally, we aim to provide the evaluation of such extensions, and a comparison with other binarization methods.

References

1. Couasnon, B., Camillerapp, J., Leplumey, I.: Making handwritten archives documents accessible to public with a generic system of document image analysis. In: DIAL 2004, pp. 270–277 (2004)
2. Baird, H.S.: Difficult and Urgent Open Problems in Document Image Analysis for Libraries. In: DIAL 2004, pp. 25–32 (2004)
3. Marinai, S., Marino, E., Cesarini, F., Soda, G.: A general system for the retrieval of document images from digital libraries. In: DIAL 2004, pp. 150–173 (2004)
4. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Systems Man Cybernet. 9(1), 62–66 (1979)
5. Bernsen, J.: Dynamic thresholding of grey-level images. In: 8th Int. Conf. on Pattern Recognition, pp. 1251–1255 (1986)
6. Niblack, W.: An Introduction to Digital image processing, pp. 115–116. Prentice-Hall, Englewood Cliffs (1986)
7. Stathis, P., Kavallieratou, E., Papamarkos, N.: An evaluation technique for binarization algorithms. Journal of Universal Computer Science 14(18), 3011–3030 (2008)
8. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs (1988)
9. Kavallieratou, E.: A Binarization Algorithm Specialized on Document Images and Photos. In: 8th Int. Conf. on Document Analysis and Recognition, pp. 463–467 (2005)
10. Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern Recognition 33, 225–236 (2000)
11. Leedham, G., Varma, S., Patankar, A., Govindaraju, V.: Separating Text and Background in Degraded Document Images. In: Proceedings Eighth InternationalWorkshop on Frontiers of Handwriting Recognition, pp. 244–249 (September 2002)
12. Shapiro, L., Stockman, G.: Computer Vision. Prentice-Hall, Englewood Cliffs (2001)
13. Gottesfeld Brown, L.: A survey of image registration techniques. ACM Computing Surveys 24(4), 325–396 (1992)
14. Kitte, T.D., Evans, B.L., Daamera-Venkata, N., Bovil, A.C.: Image Quality Assessment Based on Degradation Model. IEEE Trans. Image Processing 9, 909–922 (2000)
15. Veksler, O.: Fast Variable Window for Stereo Correspondence using Integral Images. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2003, vol. 1, pp. I-556 – I-561 (2003)