Methods of Bitonal Image Conversion for Modern and Classic Documents

Costin - Anton Boiangiu, Andrei - Iulian Dvornic
Computer Science Department
"Politehnica" University of Bucharest
Splaiul Independentei 313, Bucharest
ROMANIA
Costin@cs.pub.ro, Andrei.Dvornic@yahoo.co.uk

Abstract: Bitonal conversion is a basic preprocessing step in Automatic Content Analysis, a very active research area in the past years. The information retrieval process is performed usually on black and white documents in order to increase efficiency and use simplified investigation techniques. This paper presents a number of new modern conversion algorithms which are aimed at becoming an alternative to current approaches used in the industry. The proposed methods are suitable for both scanned images and documents in electronic format. Firstly, an algorithm consisting of a contrast enhancement step, followed by a conversion based on adaptive levelling of the document is presented. Then a new multi-threshold technique is suggested as a solution for noise interferences, a common feature of scanned books and newspapers. Finally, three more approaches adapted to the particular properties of electronic documents are introduced. Experimental results are given in order to verify the effectiveness of the proposed algorithms.

Key-Words: automatic content analysis, electronic documents, bitonal conversion, information retrieval, noise, scanned images, contrast enhancement

1 Introduction

The attention paid to automatic content analysis has been increasing lately due to the need of both converting old documents to digital media and acquiring useful and on-time information in electronic and searchable format.

The goal is to develop faster and better methods for information retrieval and classification. Extraction of graphical and textual data has proven to be a difficult task because of variations in layout, orientation, size, input form (binary, grayscale and color) and quality of scanned documents. In order to cope with all this problems, most of the current techniques require image binarization before the actual features extraction. This methodology reduces the computational load and enables the use of simplified methods of analysis. The general context of a document, as well as the local one, contains relevant information that can lead to correct classification of pixels as belonging to the foreground or to the background. In this view, we propose an algorithm which solves the global issues of a document by contrast preprocessing before the actual binarization. This method allows the algorithm to focus on the local details in the later stages. Gaussian blur is used to detect the foreground pixels by comparing the blurred intensity of a pixel with its original value.

The scanning process, as well as the degradation in time of old papers, generates a large quantity of noise elements which hinders most of data analysis. Hence, avoiding noise as much as possible is one of the most important tasks during black and white conversion. The goal is to manage to differentiate between relevant and irrelevant data as safely as possible, enhancing analysis and avoiding loss of any information in the mean time. The second algorithm proposed in this paper is a conversion method which is aimed at reducing noise. The results from several masks, obtained by using *safe thresholds*, are combined in order to detect noise elements and remove them from the bitonal output. We call a *safe threshold* any color thresholding conversion method which outputs the lowest noise ratio possible, even if some document information is lost in the process.

The most popular types of binarization methods currently is use are: error dispersion in 1D/2D domain (Floyd Steinberg), threshold [5] and cell-dithering based (Order/Half Tone) [6] conversions. The main drawback of using these approaches is the fact that they are aimed at general-use applications and cannot be mapped to the specific characteristics of documents used for Automatic Content Analysis.

For example, threshold-based conversions usually fail to reach accurate result because of the fact that a single global threshold for an entire document cannot be computed. The variable illumination as well as the depreciation in time of some parts of an image demand for an adaptive approach (which in the following proposed algorithms is solved using either Gaussian blur effect or resampling).

Cell-dithering-based conversions tent to generate irregular contours. This kind of undesired quality attributes cause errors during the measurements of geometric characteristics of text entities. Apart from that, OCR algorithms usually fail dramatically in these cases, as conversion results do not follow exactly letters' contours.

Taking into consideration the error-dispersion algorithms, the main problem that baffles document analysis is the presence of dithering fields. Entire areas inside the image are filled with singular pixels of small groups of interconnected pixels, which are constantly distributed in space in order to mimic shades of grey. This fact, although very interesting in the case of artistic pictures for example, is undesired for automatic analysis because the goal is to detect a *correct* number of entities, each with a firm contour.

Apart from having their individual limitations, all this techniques are not addressing critical issues specific to electronic documents (e.g. detecting text inside images or uneven background, converting areas where both the text and background is multi-coloured, etc.). Hence, these approaches are not a viable solution for such conversions. In this view, we propose three new methods which are aimed at solving some of the problems areas on which the actual methods fail.

All of the proposed algorithms work for both *grayscale to black and white* and *color to black and white* conversions and are meant to be an alternative to the methods currently used in the field of content analysis and document digitization. In the following, all the color intensities for each color channel are considered normalized in the range [0, 1].

2 Document Binarization Using Contrast Preprocessing

Because it means reshaping the tonal foundation of an image, black and white conversion is considered a radical transformation. A variety of solutions have been developed and experts in the domain still debate which one is the best. The bottom idea is that every conversion method works reasonably well on some set of input documents, but finding the method that generates the best results depends on the desired output and the set of images to be converted.

Having all of the above in mind, the first proposed conversion algorithm takes into consideration the particular characteristics of old scanned documents and tries to yield the best possible output. The main problems encountered during the conversion of such documents are: brightness variations (caused by document degradation or poor scanner quality) and low contrast ratios. Most of the scanned images are obtained using automatic scanning devices and because of poor calibration there might be large fluctuations of light between parts of a document. As a result, finding conversion threshold levels is a very difficult task. An adaptive local approach is needed in such cases, as any global algorithms will most probable fail for some areas of the input picture.

As its name states, document binarization using contrast preprocessing is a two-step process which tries to solve the problem of conversion by first emphasizing the discrimination between foreground and background and then applying a local conversion algorithm.

2.1 Auto-stretch Contrast Preprocessing

The contrast is the measure of the difference in brightness between light and dark areas of a document. In the field of content analysis, the main issues that must be taken into consideration regarding this property of a document are: spatial variation (as a result of both subject matter and lighting) and the histogram shape. The goal is to have broad-shaped histograms which reflect a large gap between the background texture of the document and the actual content.

The auto-stretch contrast method that we propose is an image processing technique used for preparation of input documents before further black and white conversion. The algorithm works both for grayscale (8 bits per pixel) and color (24 bits per pixel) documents and the output is a similar image with the original, but with another contrast ratio. This initial stage of the conversion increases the contrast between the text and the background of a picture, ensuring a more relevant input for the actual conversion step.

The proposed algorithm is performed using two successive iterations through the input document: the first one for computing the stretch bounds and the second one for performing the actual contrast stretch.

For computing the stretch bounds, the histogram(s) of the input image are produced. In the case of color images, three individual histograms (one for each color channel) are considered whereas for gray scale images only one. A horizontal threshold level is applied to the previously computed histograms in order to separate the scarce intensity tones to the common ones. The longest horizontal segment cutting the histogram at this threshold level is appreciated and its bounds are used as references for the actual contrast stretch.

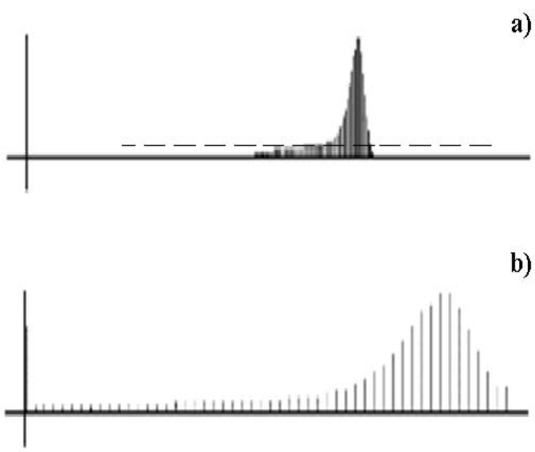


Fig. 1 Automatic contrast stretch histogram (a) input histogram (b) output histogram

If any color index is missing from the histogram, a triangular filter is applied repeatedly (before the longest segment estimation) until all the color values have a corresponding representation in the histogram.

$$I'(x,y) = \frac{I(x,y) - lEnd}{rEnd - lEnd} \quad (1)$$

where: *I'(x,y)* is the new intensity level for the individual channel; *I'(x,y)* is the original intensity level; *lEnd* and *rEnd* are the histogram-based contrast stretch bounds.

For grayscale images the values of the stretch bounds are the endpoints of the longest segment at the threshold level in the histogram. For color documents:

$$rEnd = \min(rREnd, rGEnd, rBEnd) \quad (2)$$

$$lEnd = \max(lREnd, lGEnd, lBEnd) \quad (3)$$

where *rXEnd*, *lXEnd* are the corresponding histogram-based endpoints for the *X* color channel (*X – R(ed)/G(reen)/B(lue)*).

2.1.1 Experimental Results

For testing purposes, we have used over 1000 scanned documents of old books and newspapers from the "British Library" and a histogram threshold level of 5%. Based on these experiments, a set of advantages and disadvantages of this contrast processing method have been noticed.

CHAPTER IV.

(*Continued.*)

BOTH the sisters now turned to the Captain in alarm, and the elder observed—

"You had better take the advice of Harvey; rest assured, brother, his opinion in such matters ought not to be disregarded."

"Yes," added the younger, "if, as I suspect, Mr. Birch assisted you to come here, your safety, our happiness, dear Henry, requires you to listen to him now."

"I brought myself out, and can take myself in," said the Captain positively; "our bargain went no further than to procure my disguise, and let me know when the coast was clear; and in the latter particular you were mistaken, Mr. Birch."

"I was," said the pedlar, with some interest, "and the greater is the reason why you should get back to-night—the pass I gave you will serve but once."

"Cannot you forge another?"

The pale cheek of the trader showed an unusual colour, but he continued silent, with his eyes fixed to the ground, until the young man added, with great positiveness—

Fig. 2 Effect of contrast preprocessing applied on a scanned document

Firstly, the main advantage of this algorithm is that is can be applied with success on any document, regardless of the color depth. Secondly, the method does not generate a contrast equalization of the input document, but just increases the contrast ratio between the background and the foreground. This means that if an effective black and white conversion is desired, this stage must be followed by a local conversion approach and not a global one. Finally, taking into consideration the effectiveness of the algorithm compared to the cost of it in time, we can conclude that for pictures that have acceptable contrast ratio this step might not be needed, as the increase in precision of the conversion might not compensate for the time consumed by the preprocessing stage.

The size of the documents is also an important issue to be considered when using this conversion technique, as the algorithm's time performance is directly proportional with the dimensions of the input.

2.2 Black and White Conversion Using Gaussian Blur Effect

The second step of the proposed document binarization process consists of a black and white conversion based on the Gaussian blur effect [8]. A threshold will be applied to the difference between the image obtained after the auto-stretch contrast preprocessing and a Gaussian blur version of this intermediary document:

$$B(x,y) = ((I(x,y) - I_{GB}(x,y))/2 + 0.5) > th \quad (4)$$

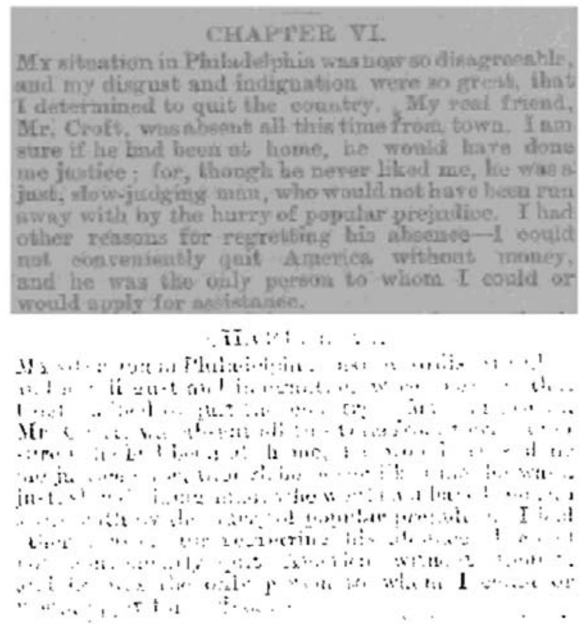
where *B(x, y)* is the binary output value for the pixel having the coordinates (x, y), *I'(x, y)* is the pixel value obtained during preprocessing (1), I_{GB} is the Gaussian blur pixel value and *th* is the decision threshold used to ensure a "*safe-distance*" from the Gaussian value to switch "safely" between black and white colors.

The visual effect of the blurring technique is a smoothness of the document, resembling that of viewing the image through a translucent screen. Hence, the basic idea behind using this approach is that areas in the image which belong to the background will have intensity values below the blur value, whereas the pixels belonging to the foreground will most probably have intensity values above it. This is the effect of the convolution between the original image and the Gaussian kernel, which yields a new image where the intensity in each pixel depends on the intensity of all pixels in a neighborhood of his.

The Gaussian Blur's linearly separable property is used in order to divide the process into two passes. In the first pass, a one-dimensional kernel is used to blur the image in only the horizontal or vertical direction. In the second pass, another one-dimensional kernel is used to blur in the remaining direction. The resulting effect is the same as convolving with a two-dimensional kernel in a single pass, but requires fewer calculations. When converting the Gaussian's continuous values into the discrete values needed for a kernel, the sum of the values will be less than 1. This will cause a darkening of the image. To remedy this, each discrete value is divided by the sum of all values.

2.2.1 Experimental Results

For the evaluation part of this algorithm, a set of 600 different scanned documents have been used. The results of the algorithm have been tested for precision and level of noise against standard methods used for decreasing color depth. Each document was rated 1-10 for the noise level and 1-10 for the precision of the conversion. The two scores were added to obtain the final score for each test. If one of the two marks were less than 3 the test was marked as unacceptable and obtained a 0 points final score.



CHAPTER VI.

My situation in Philadelphia was now so disagreeable, and my disgust and indignation were so great, that I determined to quit the country. My real friend, Mr. Croft, was absent all this time from town. I am sure if he had been at home, he would have done me justice; for, though he never liked me, he was a just, slow-judging man, who would not have been run away with by the hurry of popular prejudice. I had other reasons for regretting his absence—I could not conveniently quit America without money, and he was the only person to whom I could or would apply for assistance.

Fig. 3 Low-contrast document binarization without contrast preprocessing

The results of the previously introduced conversion method have been comparable with standard algorithms in 38% of the cases. For 15% of the input documents, at least one standard method obtained a better score than the proposed technique, while for the rest of 47% document binarization using contrast processing was the most precise conversion method.

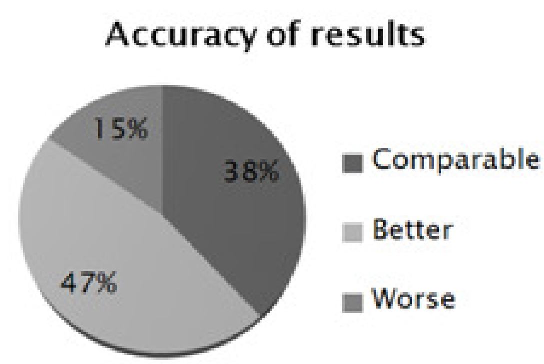


Fig. 4 Comparison vs. standard algorithm

During the experiments it was noticed that the proposed algorithm obtained constant high marks for all kind of different documents, managing to solve both noise, low-contrast and backside image showing problems. A constant threshold level of 0.43 was used during the experiments and a blur ratio of 1.5% from the sum between the height and the width of the document.

Improvements could be made in the future by finding a method of computing the radius size and threshold level based on document properties.

2.2.2 Resampling Alternative

In order to increase the speed of the algorithm, a test aimed at replacing the Gaussian blur with a resampling [9] transformation has been conducted. Comparable results have been obtained by using a *downsampling followed by upsampling* technique.

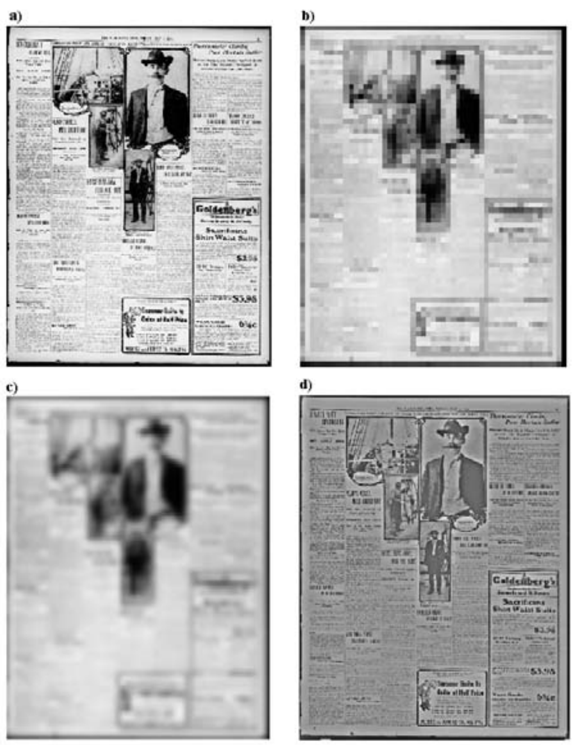
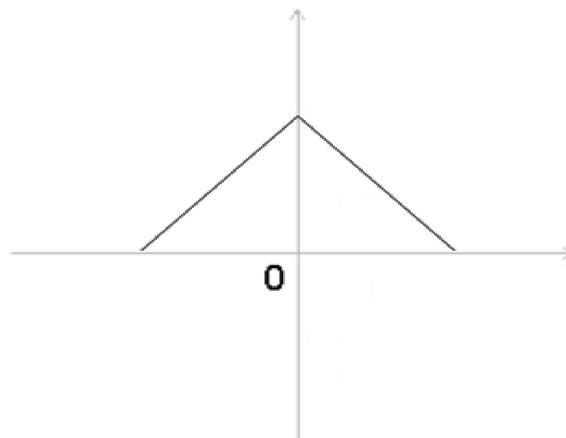


Fig. 6 Resampling technique (a) initial picture (b) Downsampled image (c) Upsampled image (d) Final result

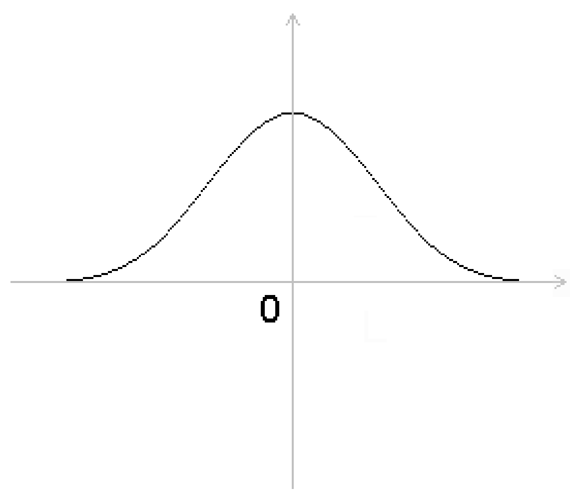
This method is based on image interpolation in two directions, trying each time to achieve a best approximation of a pixel's color and intensity based on the values at surrounding pixels. The image loses some quality both during downsampling and upsampling stages and hence the final result will be a corrupted document which resembles a blur effect.

The size of the input document was decreased until 4K primary color zones were obtained. It was noticed that an equalization effect was generated after the upsampling step, which can be of use both as a preprocessing before binarization or a post-processing aimed at image export. The following interpolation modes have been used: Lanczos, Hermite, Triangle, Mitchell, Bell and B-Spline. All of these algorithms are non-adaptive, which means they treat all pixels as equally, no matter on what the interpolation is conducted (sharp edges, smooth texture etc.). As a result, aliasing (jagged diagonal edges) cannot be avoided; this is the main drawback in using the resampling technique. A few anti-aliasing approaches have been considered as well, but the computational load was affecting the efficiency of the algorithm way too much.



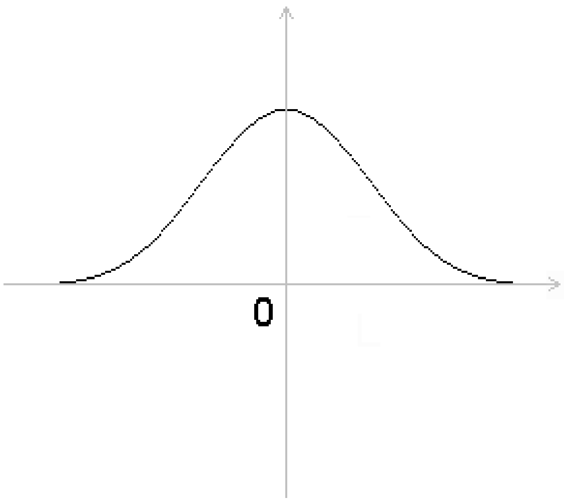
$$T(t) = \begin{cases} 1-|t|, & |t|<1 \\ 0, & otherwise \end{cases}$$

Triangle Interpolation



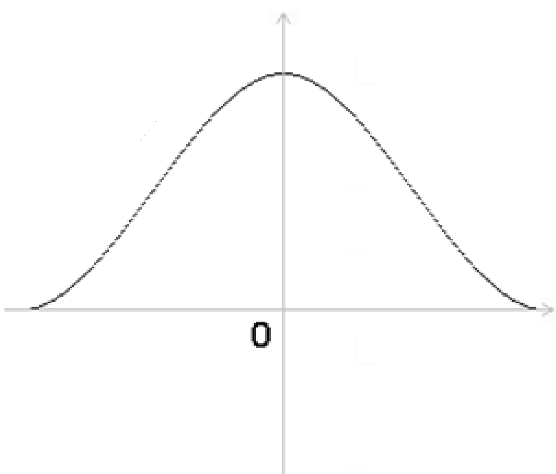
$$B(t) = \begin{cases} 0.75-t^2, & |t|<0.5 \\ 0.5*(t-1.5)^2, & 0.5 \le |t|<1.5 \\ 0, & otherwise \end{cases}$$

Bell Interpolation



$$Bs(t)=\begin{cases} \frac{1}{2}*t^3-t^2+\frac{2}{3}, & |t|<1 \\ \frac{1}{6}*(2-t)^3, & 1\le|t|<2 \\ 0, & otherwise \end{cases}$$

B-Spline Interpolation



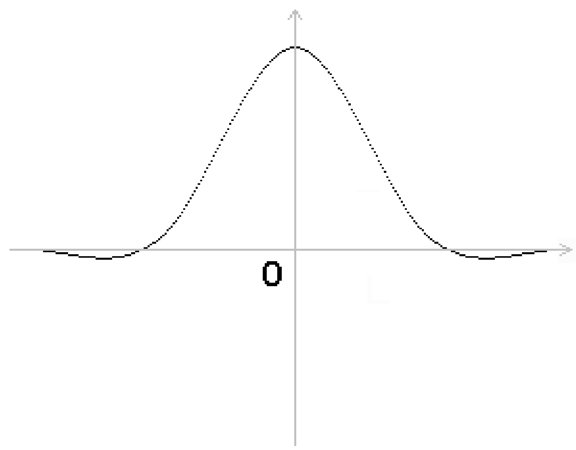
$$H(t)=\begin{cases} (2*t-3)*t^2+1, & |t|<1 \\ 0, & otherwise \end{cases}$$

Hermite Interpolation

Fig. 5 Graphical representation of some interpolation filters

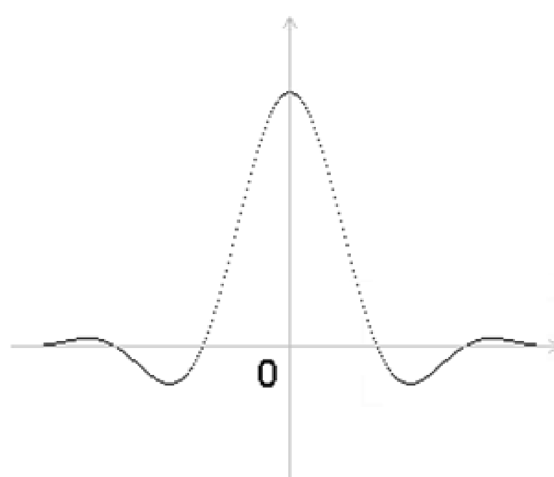
Obviously, the more you know about the surrounding pixels, the better the interpolation becomes. Even though, it was noticed that results quickly deteriorate the more the image was stretched and hence caution must be taken when trying to find a compromise between the final smoothness ratio and execution time.

Considering all the tested algorithms, B-Spline has proved to generate the best results (closest to the blur effect) for single-mode interpolation - both the upsampling and downsampling is performed using the same filter.



$$M(t)=\begin{cases} \frac{7}{6}*t^3-2*t^2+\frac{9}{8}, & |t|<1 \\ -\frac{7}{18}*t^3+\frac{1}{2}*t^2-\frac{8}{3}t+\frac{13}{9}, & 1\le|t|<2 \\ 0, & otherwise \end{cases}$$

Mitchell Interpolation



$$L(t)=\begin{cases} \sin c(t)*\sin c(\frac{t}{a}), & |t|<a \\ 0, & otherwise \end{cases}$$

Lanczos Interpolation

Fig. 7 Recommended upsampling interpolation filters

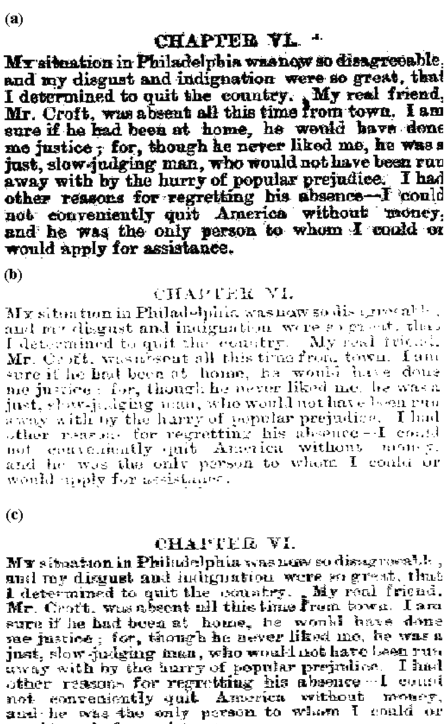
If the two stages of the resampling process are split we can consider a combination of two different filters in order to increase the accuracy of the results. In this case Lanczos and especially Mitchell (see Fig. 7) are recommended for the downsampling process. These two filters tend to emphasize the sharpness of the image due to their shape (similar to an edge-detection filter). As a result, on large areas an increase in the precision of the details-positioning is noticed at the passing between black and white pixels.

For the upsampling case, the two previously described filters can also be a viable solution. Even though, the goal is to obtain more of a "cloud effect" for the blurred neighborhoods and hence the sharp edges generated by both Mitchell and Lanczos can be harmful to the final visual perception.

Being bell-shaped filters B-Spline and Bell (see Fig. 5) are the recommended solutions in the case of downsampling, as the smoothness effect that they generate resembles a blur transformation quite accurate.

2.3 Partial Conclusions

Document digitization using contrast preprocessing tries to be a viable solution for black and white conversion in the field of document content analysis.


(a)

CHAPTER VI.

My situation in Philadelphia was now so disagreeable, and my disgust and indignation were so great, that I determined to quit the country. My real friend, Mr. Croft, was absent all this time from town. I am sure if he had been at home, he would have done me justice; for, though he never liked me, he was a just, slow-judging man, who would not have been run away with by the hurry of popular prejudice. I had other reasons for regretting his absence—I could not conveniently quit America without money, and he was the only person to whom I could or would apply for assistance.

(b)

CHAPTER VI.

My situation in Philadelphia was now so disagreeable, and my disgust and indignation were so great, that I determined to quit the country. My real friend, Mr. Croft, was absent all this time from town. I am sure if he had been at home, he would have done me justice; for, though he never liked me, he was a just, slow-judging man, who would not have been run away with by the hurry of popular prejudice. I had other reasons for regretting his absence—I could not conveniently quit America without money, and he was the only person to whom I could or would apply for assistance.

(c)

CHAPTER VI.

My situation in Philadelphia was now so disagreeable, and my disgust and indignation were so great, that I determined to quit the country. My real friend, Mr. Croft, was absent all this time from town. I am sure if he had been at home, he would have done me justice; for, though he never liked me, he was a just, slow-judging man, who would not have been run away with by the hurry of popular prejudice. I had other reasons for regretting his absence—I could not conveniently quit America without money, and he was the only person to whom I could or would apply for assistance.

Fig. 8 Comparison between binarization methods (see Fig. 3 for input document)
(a) Proposed technique
(b) Standard threshold
(c) Halftone-Cell Dithering

The proposed technique focuses on problems that are most common in scanned images representing old books and newspapers, but can be successfully applied on other documents as well.

Future improvements can be made by adding other preprocessing steps or replacing the contrast stretch with a more efficient method. In addition to this, increasing the speed of the algorithm can be taken into consideration as the performance of this procedure is not very high, especially because of the time-consuming Gaussian blur stage. The smaller the Gaussian function radius we choose, the faster the algorithm will be, but a value too small considered for this parameter will lead to a high-noise sensitivity of the conversion and increasingly poor results.

3 Noise-free Binarization Using "safe masks"

It is well known that noise is one of the main causes of failure in OCR algorithms [7]. Since most of the state-of-the-art page segmentation algorithms report textual noise regions as text-zones, the OCR accuracy decreases in the presence of textual noise (OCR system usually outputs several extra characters in these regions). Removing such entities can thus help increasing the OCR accuracy.

The following proposed algorithm is a multi-stage binarization technique which is proper for conversions where the noise aspect is very sensitive. This includes both documents which have large background variations and regular images on which text extraction must be very precise.

There are four different masks which are used simultaneously in order to obtain the final binary output. We call a mask a simple threshold-based conversion method used in this case as an intermediary black and white transformation. The parameters of all four masks are considered as "safe", which means that their output contains the least amount of noise possible, no matter the loss of relevant information from the original document. The basic idea behind this technique is that each foreground pixel will be detected eventually by at least one of the masks. Hence, the final bitonal image will contain all foreground areas, even though each individual mask misclassifies a series of points.

The considered masks are the following: a threshold applied to the grayscale transformation of the image; a threshold applied to the K component of the CMYK color model; a threshold applied to each component of the RGB color model using the minimum average intensity of the three color RGB channels and a variable threshold applied computed based on the hue of the document.

3.1 Algorithm outline

The algorithm starts by first computing each individual mask. Then the *least accurate* of these intermediary conversions is determined, along with the superposed conversion containing the sum of all other masks.

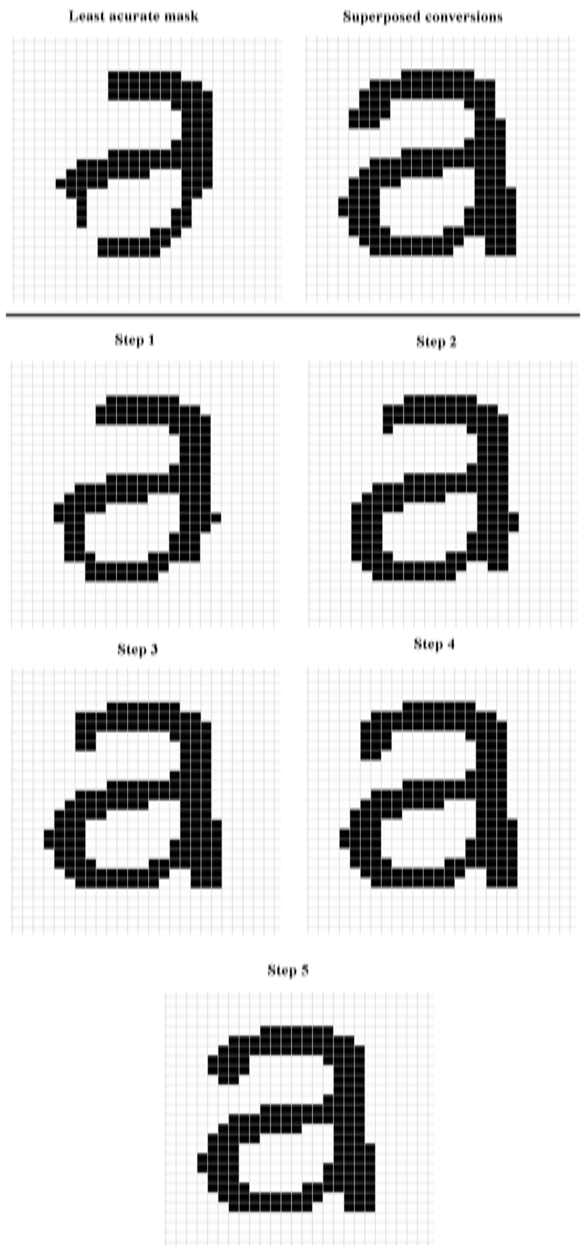


Fig. 9 Object reconstruction starting from *least accurate* mask

The *least accurate* conversion is considered the one with the least black vs. white pixel ratio. During this o stage, an additional threshold of 50% from the most effective mask (the one with the greatest ratio) is considered in order to rule out erroneous intermediary conversions. By this we mean conversions that might have detected only a very small number of black pixels, as a result of particular features of the input document. Due to the large number of masks used, this method is robust and can be applied successfully on a large variety of documents without readjusting its internal parameters.

Finally, starting from the *least accurate mask,* undetected foreground pixels are added recursively to it, provided that they are detected in one of the other masks (are present in the superposed conversion) and are neighboring a foreground pixel in the current intermediary image (see Fig. 9). The final conversion is obtained when no more pixels can be added to the intermediary image.

The basic idea behind this technique is foreground reconstruction. Starting from an initial conversion which is not very accurate, the algorithm is able to regenerate all entities. The fact that the starting point is one of very low accuracy is an advantage due to the fact that a large amount noise is undetected and as a result is not taken into consideration further.

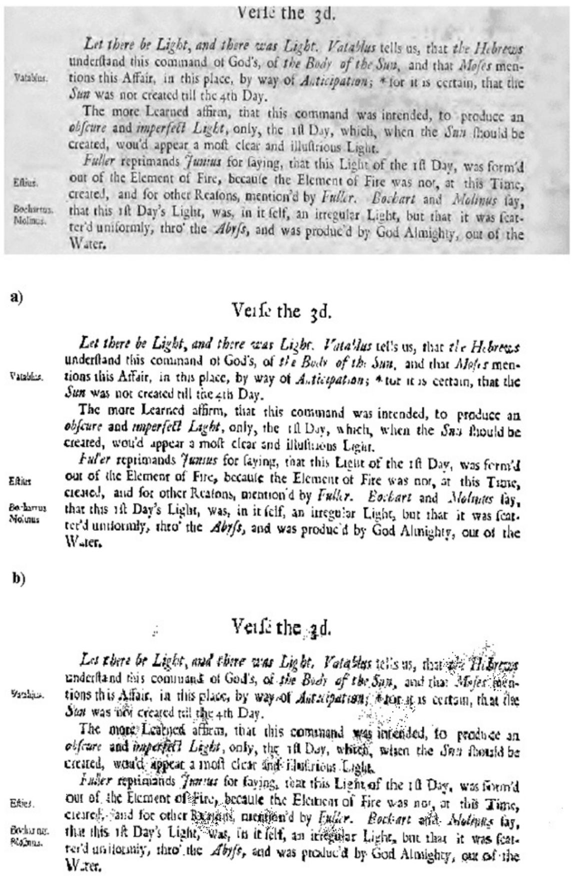
3.2 Experimental Results

Several experimental tests have been conducted using the proposed noise-free conversion method. A set of 500 scanned old documents were evaluated and some safe threshold levels have been determined.

Comparisons were made with both global and adaptive-local conversion techniques. It was noticed that the new method reduces considerably the amount of noise in the conversion because of two reasons. Firstly, some level of noise was eliminated from all masks due to the “safe” characteristic of them. Apart from that, other noise entities were excluded from the output because no pixel from such an entity was detected by the least accurate mask and hence the entity could not be restored.

A number of causes that lead to the failure of the proposed binarization algorithm have been detected as well and are subject of future improvements. This includes missing characters or parts of characters (because detection failed in all of the considered masks) and failure to eliminate large noise elements (at least one pixel from such entities was signaled in the least accurate mask and lead to their entire reconstruction).

The execution time is an important feature that must be taken into consideration when choosing this conversion method; despite the fact that this approach looks simple, a large number of operations are necessary before the final result is reached.



Verſe the 3d.

Vatablus. *Let there be Light, and there was Light. Vatablus* tells us, that *the Hebrews* underſtand this command of God's, of *the Body of the Sun*, and that *Moſes* mentions this Affair, in this place, by way of *Anticipation*; * for it is certain, that the *Sun* was not created till the 4th Day.

The more Learned affirm, that this command was intended, to produce an *obſcure* and *imperfect Light*, only, the 1ſt Day, which, when the *Sun* ſhould be created, wou'd appear a moſt clear and illuſtrious Light.

Eſtius. *Fuller* reprimands *Junius* for ſaying, that this Light of the 1ſt Day, was form'd out of the Element of Fire, becauſe the Element of Fire was not, at this Time, created, and for other Reaſons, mention'd by *Fuller*. Bochartus. Molinus. *Bochart* and *Molinus* ſay, that this 1ſt Day's Light, was, in it ſelf, an irregular Light, but that it was ſcatter'd uniformly, thro' the *Abyſs*, and was produc'd by God Almighty, out of the Water.

a)

b)

Fig. 10 Conversion results for (a) noise-free technique (b) standard thresholding technique

4 Binarization of Electronic Documents

In the case of electronic documents (like PDFs and online newspapers) the focus of attention during the binarization process is shifted from aspects like contrast or noise to translating the colour combinations of texts, images and backgrounds into a meaningful black and white version. To this goal three different approaches for black and white transformation of electronic documents have been tried.

The first method is the "outside-in" local technique. This algorithm starts from the outer white background and decides to convert everything which neighbours it black; then everything that neighbours the previously detected area is set to white and so on until the whole document is processed. This approach works better than the threshold-based global ones, avoiding the disappearance of letters printed with colours that are scarce in the picture. Apart from that, using this technique cancels the risk of errors generated by the same colour playing both the role of background and foreground in different parts of the document.

Fig. 11 Conversion Results a) Input b) "Alternate-colour" method c) "Outside-in" method d) "Edge-detection" method

The "alternate colours" approach is also a local conversion method which tries to solve the binarization decision using horizontal scan lines. This iterative process takes each pixel at a time (on each individual horizontal line in the input document) and decides its final value based on the previous made decisions. Conversion to white or black is alternated each time a colour shift occurs. Whenever a white or black pixel is encountered in the original document, the output colour is automatically set to that value, no matter the alternation rule. This algorithm manages to solve some cases in which the "outside-in" method would fail, like bi-coloured areas in which the background and foreground are combinations of two distinct colours. A threshold is used in order decide when a colour shift has occurred and output colour must be changed.

The third and final proposed method is an edge-detection based technique. This algorithm uses a threshold in order to decide if a pixel is part of an edge between foreground and background or not. In this way, the problems resulting from background variations can be reduced by adjusting the threshold. Research is still carried out in order to find a way to decide which of the detected contours are belonging to foreground and how they could be filled up with black pixels in order to generate a more relevant conversion.

4 Final Conclusions

Due to the increasing interest in document content conversion, there is the need of new image binarization methods that can cope with problem areas from a large variety of documents. The new approaches that have been proposed try to solve problems like: low contrast ratio, noise, backside image showing and electronic document conversion.

Document binarization using contrast preprocessing focuses the attention on performing a preparation of the image before the actual conversion. By stretching the contrast of the input document, a larger gap between background and foreground data is created and, as a result, the probability of success of a local threshold-based conversion is increased. The method is robust to lighting variations and backside image showing since the Gaussian blur effect removes fine image details and noise.

This paper also introduces a new method for noise-free conversions. The algorithm tries to basically reconstruct the conversion from a series of puzzle pieces (the so-called "masks" used in the algorithm).

Finally, three new approaches aimed at conversion of modern digital documents have been presented. Test cases have shown that all methods manage to reach their goals to some extent, are easy to implement and some of them might also be used in domains which are outside their designated area of interest. Further research must be conducted in order to improve current methods, as well as develop new techniques that can cope with all the challenges of document analysis (especially in the case of electronic documents).

References:

[1] L. M. Sheikh, I. Hassan, N. Z. Sheikh, R. A. Bashir, S. A. Khan, and S. S. Khan, "An Adaptive Multi-Thresholding Technique for Binarization of Color Images", *WSEAS Transactions on Information Science and Applications*, Issue 8, Vol. 2, 2005.

[2] C. A. Boiangiu, and A. I. Dvornic, "Bitonal Image Creation for Automatic Content Conversion", *Proceedings of the 9th WSEAS International Conference on Automation and Information (ICAI'08),* 2008, pp. 454-460.

[3] F. Wenzel, and R.R. Grigat, "A Framework for Developing Image Processing Algorithms with Minimal Overhead", *Proceedings of the 5th WSEAS International Conference on Signal, Speech and Image Processing,* 2005, pp. 185-190.

[4] R. Dobrescu, M. Dobrescu, S. Mocanu, and Se.Taralunga, "Development platform for parallel image processing", *Proceedings of the 6th WSEAS International Conference on Signal, Speech and Image Processing,* 2006, pp. 31-36.

[5] J. Sauvola, and M. Pietikainen, "Adaptive document image binarization", *The Journal of the Pattern Recognition Society*, Elsevier Science Ltd., 2000, pp. 225-236.

[6] F. Chang, "Retrieving information from document images: problems and solutions", *International Journal on Document Analysis and Recognition*, 2001, pp. 46-55.

[7] S.W. Lee, D.J. Lee, and H.S. Park, "A New Methodology for Gray-Scale Character Segmentation and Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 12, 1996, pp. 1045-1050.

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

[9] S.Fischer, *"Digital Image Processing: Skewing and Thresholding"*, Master of Science thesis, University of New South Wales, Sydney, Australia, 2000.

[10] M. Kamel, and A. Zhao, "Extraction of BinaryCharacter / Graphics Images from Grayscale Document Images", *CVGIP*, Vol.55, No.3, 1993, pp.203-217.

[11] R. P. Loce, and E. R. Dougherty. *Enhancement and Restoration of Digital Documents – Statistical Design of Nonlinear Algorithms*, SPIE Optical Engineering Press, 1997.

[12] B. Chen, and L. He, "Fuzzy template matching for printing character inspection", *WSEAS Transactions on Circuits and Systems,* Issue 3, Vol. 3, 2004.

[13] M.I. Rajab, "Feature Extraction of Epiluminescence Microscopic Images by Iterative Segmentation Algorithm", *WSEAS Transactions on Information Science and Applications,* Issue 8, Vol. 2, 2005.