International Journal of Computer Applications (0975 – 8887)
Volume 44– No.22, April 2012

Foreground Estimation in a Degraded Text document

I Bhuvana Chandra
Department of Computer Science Engineering, Amrita School of Engineering Coimbatore.
A-310, AgasthyaBhavanam,
Amrita School of Engineering,
Ettimadai,Coimabtore-641105

K NagarjunaVarma
Department of Computer Science Engineering, Amrita School of Engineering Coimbatore.
A-311, AgasthyaBhavanam,
Amrita School of Engineering,
Ettimadai,Coimabtore-641105

Gireesh Kumar T
Department of Cyber Security, Amrita School of Engineering Coimbatore.

Department of Cyber Security,
Amrita School of Engineering,
Ettimadai,Coimabtore-641105

ABSTRACT
In this paper an attempt is made to retrieve the text region alone from a degraded text document. For doing that, four different filters are used for noise removal in the text document. Later document binarization is done using thresholding. Three different thresholding techniques are implemented for foreground-background separation.Then candidate region is selected and features are extracted. The features are then fed to an SVM to classify text and non-text regions. The proposed approach is implemented and tested on various hand written and machine printed degraded text documents.

Keywords
Image filtering, Wavelet Decomposition, Feature Extraction, Document Binarization.

1. INTRODUCTION
Text region recognition from a document with a lot of background noise, dark spots and non-uniformly illuminated text documents is a challenging task. This is because setting a proper threshold for an image is a tedious job and there are different thresholding techniques. Local or global adaptive thresholding methods can be applied for binarization. In this paper four filtering techniques and three thresholding techniques are applied on degraded document and a comparative study is done. Mean, Median, Gaussian and LOG filters are used for background noise removal is done and their corresponding effects on the image are shown. Thresholding of the image is done using Savoula formula, Simple Image Statistics and luminance value is done and the results are recorded. After the foreground pixels of the degraded document are separated from its background pixels. The text region is detected using an SVM classifier. The final text document with only text regions is obtained..

2. METHODOLOGY
The proposed methodology for foreground estimation in a degraded text document is represented in the flow chart in figure 1.

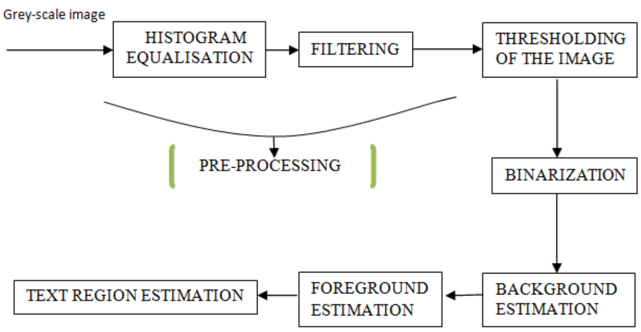


Fig. 1. Flow chart for foreground estimation of a degraded document

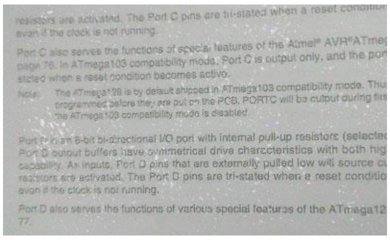
3. TYPESET TEXT
3.1 Histogram Equalization (Pre-Processing)
Consider a grey scale input image I. The preprocessing is done so that the image histogram equalization increases the detail in the image by enhancing its global contrast. Histogram equalization of given image I is obtained by the following procedure[8]:

- Reading the pixel values of I(x) into {x} in ascending order without any repetitions.
- Recording the number of occurrences of each {x}.
- Obtaining the CDF (Cumulative Distribution Function) of pixels.

The CDF must be normalized to [0,255] (for an 8-bit gray scale image). The general histogram equalization formula is given by:

$$h(x) = \left(\frac{cdf(x) - cdf_{min}}{no.\ of\ pixels - cdf_{min}} \times (L-1) \right)$$

Here L is 256.


resistors are activated. The Port C pins are tri-stated when a reset condition
even if the clock is not running.

Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the port
stated when a reset condition becomes active.

Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thus
programmed before they are put on the PCB, PORTC will be output during first
the ATmega103 compatibility mode is disabled.

Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selected
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.

Port D also serves the functions of various special features of the ATmega12
77.

Fig 2(a)

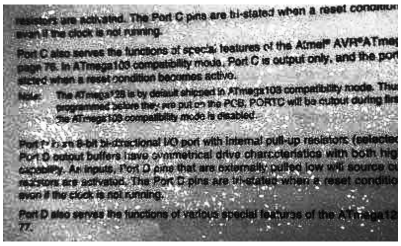


Fig 2(b)

Fig. 2. (a) Original image (b) After Histogram equalization

3.2 Filtering (Pre-Processing)

Filtering removes the basic noises such as salt, pepper and gaussian noises. The accuracy of the filter may change on the basis of the type of filter we decide to choose [6].

3.2.1 Mean Filter

Mean filter is usually used for softening of an image i.e. it is used to reduce the contrast difference between the neighbouring pixels. In this method, a window of relatively small size (3x3 in this case) is taken and its mean value is calculated. The value at the centre of the windowis replaced with the calculated mean value. This is done through the application of a simple convolution operation.

The Mean filter may not be effective in the case of small images i.e. images where the letter size is less than 10x10 pixel.

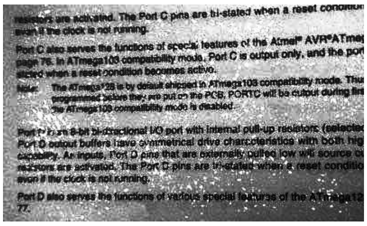


Fig 3(a)

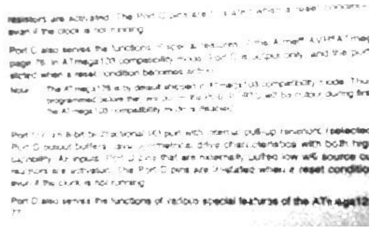


Fig 3(b)

Fig. 3.(a) Original image (b) After applying mean filter for small images

If we consider larger images i.e. letter size greater than 15x15 pixels as shown,

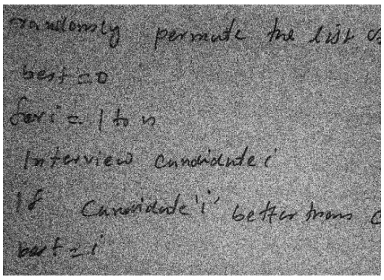


Fig.4(a)

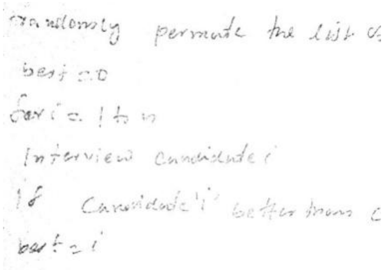


Fig.4(b)

Fig. 4(a) Original image (b)After applying mean filter for large images

3.2.2 Median filter

Median filter is very effective in reducing salt and pepper noise. It also has an edge preserving nature in addition to this noise reduction. In this method, a window of relatively small size (3x3 in this case) with odd number of pixels is taken and the median value of the window is calculated. The mean value which was calculated earlier is replaced with the median value at the center of the window.

Example:For small images the input and output images are as shown below:

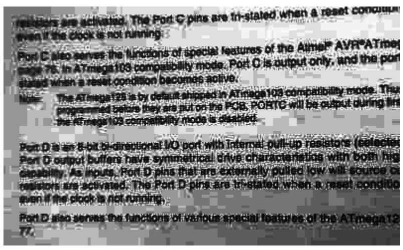


Fig. 5(a) Original image

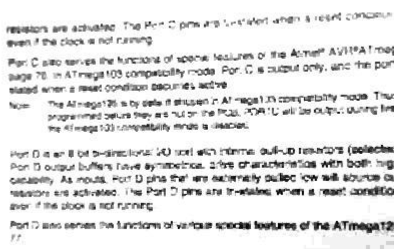


Fig 5. (b) After applying median filter for small images

From the given images it can be inferred that even the median filter does not give good results for small images. For large images the input and output images are as shown below:

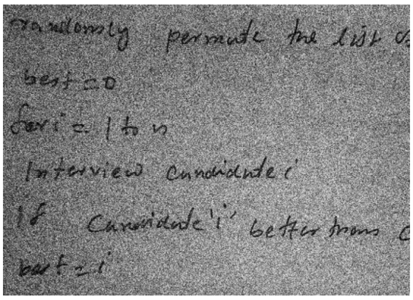


Fig.6(a) Original image

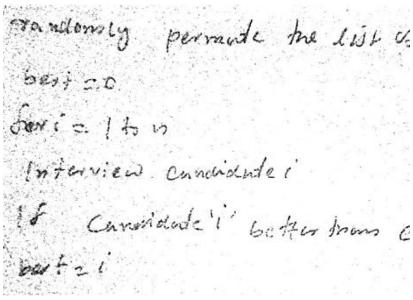


Fig.6(b) After applying median filter for large images

3.2.3 Gaussian filter

Gaussian filter uses the principle of linear convolution where the convolution matrix is obtained from the Gaussian function defined as:

$$g(x,y) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-(x^2+y^2)/2\sigma^2}$$

Here σ represents the standard deviation. The quality of the image can be changed by altering the σ value as required. Choosing the σ value is the tedious job for the user. Here, the Gaussian filter is applied for different σ values.

When σ=1,

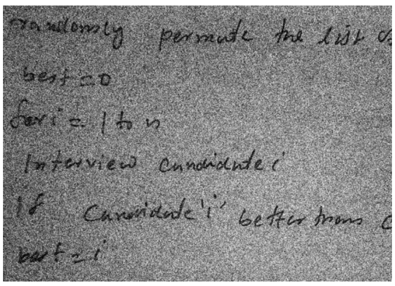


Fig. 7(a) Original image

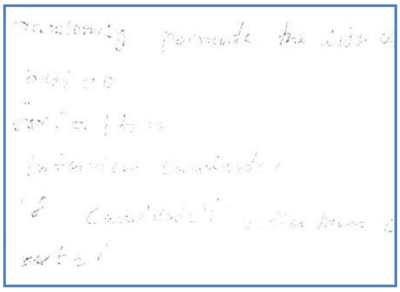


Fig.7(b) After applying gaussian filter at σ=1

When σ=3,

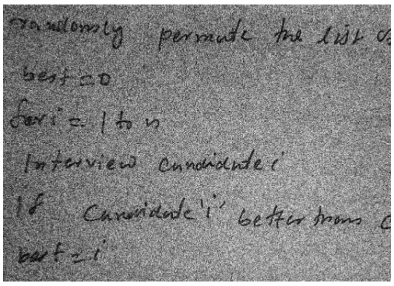


Fig. 8(a) Original image

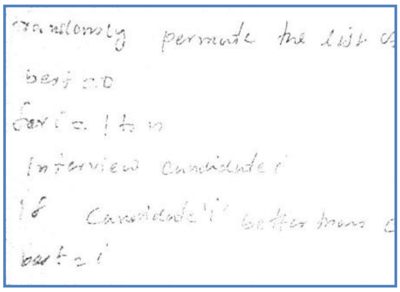


Fig.8(b) After applying gaussian filter at σ=3

When σ=5,

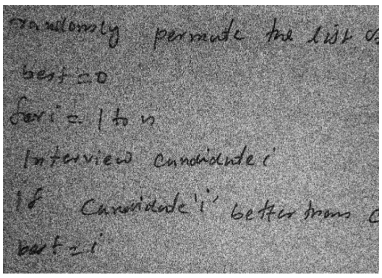


Fig. 9(a) Original image

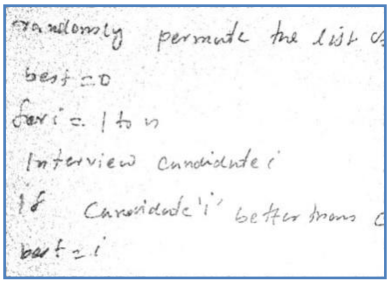

randomly permute the list o
best = 0
for i = 1 to n
Interview candidate i
If candidate 'i' better than c
best = i

Fig. 9(b) After applying gaussian filter at σ=5

When σ=7,

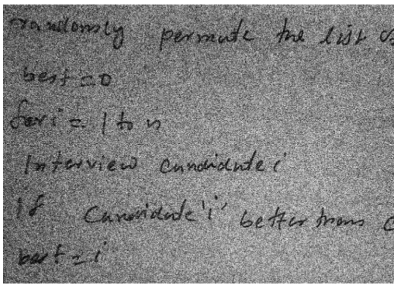

randomly permute the list o
best = 0
for i = 1 to n
Interview candidate i
If candidate 'i' better than c
best = i

Fig 10(a) Original image

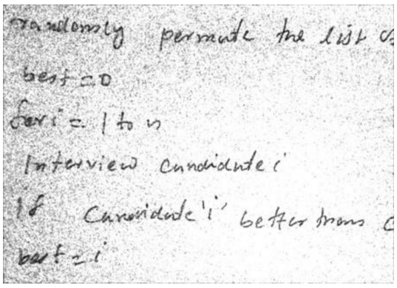

randomly permute the list o
best = 0
for i = 1 to n
Interview candidate i
If candidate 'i' better than c
best = i

Fig 10(b) After applying gaussian filter at σ=7

Hence, from the above sequence of images for different σ it can be inferred that as the σ increases, the clarity increases but at the same time the noise also increases. So, a particular σ must be found for each image in case of the Gaussian filter. The Gaussian filter thussmoothens the image.

3.2.4 Laplacian Of Gaussian (LOG)

The Laplacian of an image highlights the regions of rapid intensity change. The Laplacian is generally used for edge detection. Any smoothening filter like the Gaussian filter is applied before applying the Laplacian. The log filter has been applied by using a simple convolution mask. It is defined by the following function.

$$\mathrm{LoG}(x,y) = -\frac{1}{\pi\sigma^4} \times \left(1 - \frac{x^2+y^2}{2\sigma^2}\right) \times e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Here σ is taken to be 1.4,

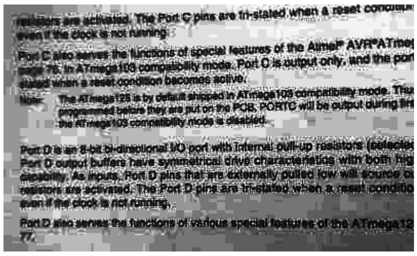

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selecte
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig. 11(a) Original image

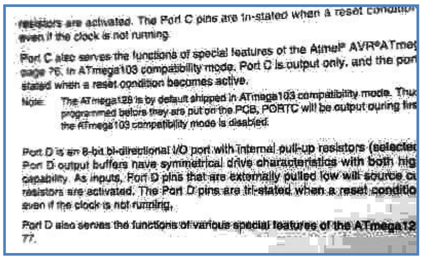

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selecte
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig. 11(b) After applying laplacian

After testing with different images, it is found that the Log filter can also be applied to images of small sizes and is better than other filters in terms of noise removal in most cases.

3.3 Thresholding

Thresholding of an image is necessary for binarization of an image and also for the background and foreground separation which is to be done [4]. Here three different thresholding techniques [3] are applied on a gray scale document, and the results are recorded. The elements of the matrix S(x, y) are assigned the values as follows:1 if the corresponding pixel is a foreground pixel and 0 if it is a background one.

$$S(x,y) = \begin{cases} 0 \text{ if } h(x,y) > T(x,y) \\ 1 \text{ if } h(x,y) \le T(x,y) \end{cases}$$

3.3.1 Local adaptive thresholding using Savoula formula

The local threshold T(x, y) of a pixel (x, y) is calculated using local mean μ(x, y) and standard deviation σ(x, y) for a window of user defined size n x n centered at (x, y). The threshold surface T(x, y) is calculated by the formula,

$$T(x,y) = \mu(x,y)\left[1 + k\left(\frac{\sigma(x,y)}{R} - 1\right)\right]$$

Where R is the maximum value of standard deviation (which is 128 for an 8-bit gray scale document) and k can take any positive value in the range of [0.2, 0.5]. Fig 3.3.1 shows an example of a binarized image using Savoula formula.

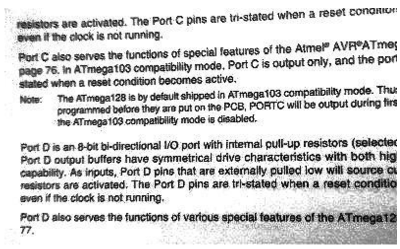
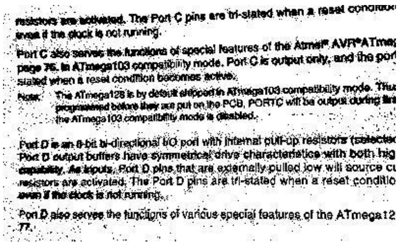

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selecte
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig.12 (a)Input grayscale image



resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the port
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selectec
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig.12(b) Binarized image applying Savoula method

3.3.2 Global Adaptive Thresholding Using Simple Image Statics

Simple Image Statistics (SIS) is a global thresholding technique which is implemented by the algorithm shown in fig. 13.

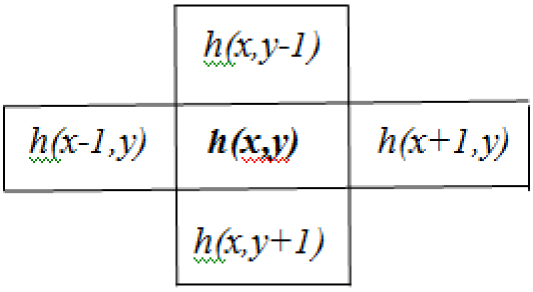


Fig.13 Global Adaptive Thresholding, using Simple Image Statics

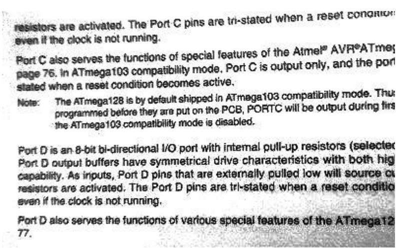
For each h(x, y), the algorithm [3] performs the operations shown below:

- Compute ex=Abs[h(x+1 ,y) – h(x-1, y)]
- Compute ey=Abs[h(x, y+1) – h(x, y-1)]
- Compute weight=max(ex,ey)
- Get the Global threshold (T) as :

$$\text{Total Weight} = \left(\sum_{y=0}^{\dim_y}\sum_{x=0}^{\dim_x} \text{weight}(h[x,y])\right) \Big/ \left(\dim_x \times \dim_y\right)$$

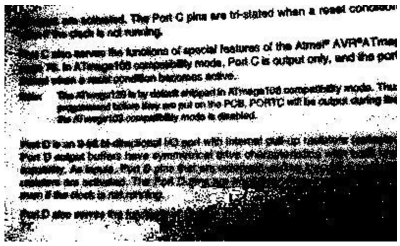
$$\text{Total} = \left(\sum_{y=0}^{\dim_y}\sum_{x=0}^{\dim_x} \text{weight}(h[x,y]) \times h(x,y)\right) \Big/ \left(\dim_x \times \dim_y\right)$$

$$\text{Global threshold (T)} = \text{Total}/\text{Total Weight}$$



resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the port
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selectec
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig.14(a)Input grayscale image



are activated. The Port C pins are tri-stated when a reset conditio
the clock is not running.
C also serves the functions of special features of the Atmel® AVR®ATmeg
76. In ATmega103 compatibility mode, Port C is output only, and the port
when a reset condition becomes active.
The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.

Fig.14 (b) Binarized image applying SIS

3.3.3 Local Adaptive Thresholding using the Luminance value

This is a global single-stage thresholding technique [3]. It uses the luminance (the brightest point) and the mean intensity of an image to calculate its threshold. For an 8-bit gray scale image the luminance can be taken as 255 (we have already done this as specified previously). But generally, for manytext documents, the luminance value will be less than 255, but it is better to calculate the luminance value of the image. The luminance value (G) can be calculated as,

$$G = F_{max}(g)$$

Where, F_{max} is a function to find out the maximum gray point in the image g. The mean intensity of the image g is calculated as

$$M = \left(\sum_{y=0}^{dim_y}\sum_{x=0}^{dim_x} (h[x,y])\right) \Big/ \left(dim_x \times dim_y\right)$$

Where M is the mean, dim_x anddim_y are the dimensions of the image and h(x,y) is the original grayscale image. Local deviation (i.e. difference between the luminance value and mean value) is calculated as

$$D = G - M$$

The difference between the mean value (i.e. initial separation point) and the local deviation gives the exact separation point. Thus,

$$\text{Threshold} = \text{Abs}(M - D)$$

4. BACKGROUND ESTIMATION

In this stage, the approximate background of the image (i.e. the approximate pixel values of the pixels found behind the foreground) is calculated. The approximate background region B(x, y) is calculated using the formula [5]

$$B(x,y) = \begin{cases} \dfrac{\left(\sum_{r=x-dx}^{x+dx}\sum_{c=y-dy}^{y+dy} (h(r,c)(1-S(r,c))\right)}{\left(\sum_{r=x-dx}^{x+dx}\sum_{c=y-dy}^{y+dy} (1-S(r,c)\right)}, & if\ s(x,y) = 1 \\ h(x,y) & , if\ s(x,y) = 0 \end{cases}$$

Where s(x,y) is the binarized image after thresholding and dx×dy is selected in such a way that it covers atleast two image characters. Here, s(x,y)=0 represents background region and s(x,y)=1 represents foreground pixels.

The above formula is suggested by [1]. This formula does not give a good result when the denominator is zero, i.e. the whole window is in the dark region. The example for the rough estimation of the background using the above formula is shown in figure 16.

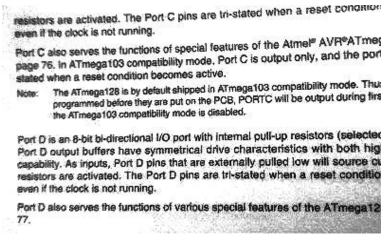

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selectec
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig 16(a) The image obtained after filtering

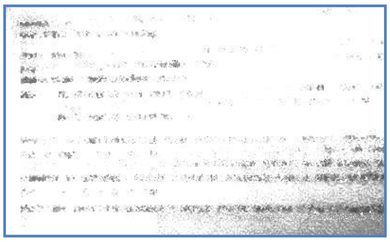


Fig 16 (b) Rough estimation of the background of image

5. FOREROUND ESTIMATION

In this step, final thresholding is done by taking into consideration, the estimated background region and the processed image h(x, y).

$$T(x,y) = \begin{cases} 1 & \text{if } h(x,y) - B(x,y) > d(B(x,y)) \\ 0 & \text{otherwise} \end{cases}$$

The text areas are located if the distance of the processed image h(x,y) from the calculated background B(x, y) is greater than a threshold d(B(x, y)). The threshold is calculated for each and every pixel so as to preserve the information even in the dark regions. The threshold is given as

$$d\big(B(x,y)\big) = q\left(\frac{1-p_2}{1+\exp\left(\left(\frac{-4B(x,y)}{b(1-p_1)}\right)+\frac{2(1+p_1)}{(1-p_1)}\right)}+p_2\right)$$

Where, $$b = \frac{\sum_x \sum_y (B(x,y))(1-S(x,y))}{\sum_x \sum_y (1-S(x,y))}$$

The above sigmoid function is recommended by [2]. For a degraded document, q, p1 and p2 are assumed to be 0.65, 0.55 and 0.85 respectively [3]. After the T(x, y) is calculated, the pixels in the processed image h(x, y) which have thecorresponding pixel values of T(x, y) are considered as the foreground pixels.

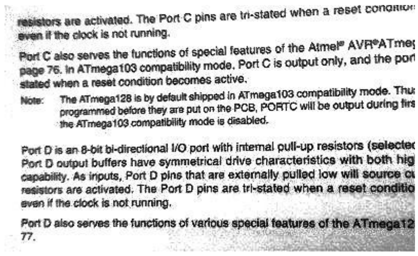

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selectec
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig.17(a)The image obtained after filtering

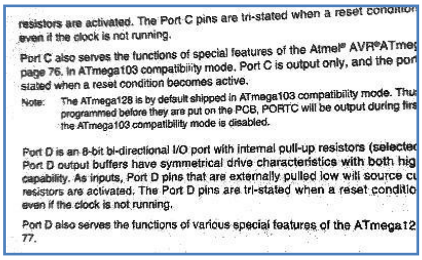

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selectec
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig.17(b) Rough estimation of the foreground of image

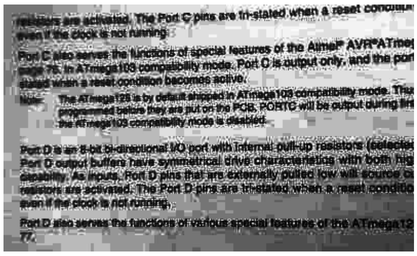

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selectec
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig.18(a)Input image

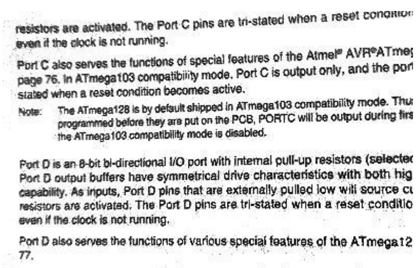

resistors are activated. The Port C pins are tri-stated when a reset conditio
even if the clock is not running.
Port C also serves the functions of special features of the Atmel® AVR®ATmeg
page 76. In ATmega103 compatibility mode, Port C is output only, and the por
stated when a reset condition becomes active.
Note: The ATmega128 is by default shipped in ATmega103 compatibility mode. Thu
programmed before they are put on the PCB, PORTC will be output during firs
the ATmega103 compatibility mode is disabled.
Port D is an 8-bit bi-directional I/O port with internal pull-up resistors (selectec
Port D output buffers have symmetrical drive characteristics with both hig
capability. As inputs, Port D pins that are externally pulled low will source cu
resistors are activated. The Port D pins are tri-stated when a reset conditio
even if the clock is not running.
Port D also serves the functions of various special features of the ATmega12
77.

Fig.18(b)Output image

6. CONCLUSION

After applying different filtering and thresholding techniques to the different set of images (too a degraded text document) with different letter sizes, the following results are observed:

- Applying mean and median filters to images with small letter size gives an irrelevant result (with disfigured text).

- Applying a gaussian alone results in smoothing of the edgesif σ is chosen properly (if σ is not chosen properly it makes binarization difficult).
- Applying its laplacian and thresholding using Savoula formula is more efficient than the other methods employed(in fig.18).

In this paper the different filtering and thresholding techniques are applied and tested using SNR ratio.

7. REFERENCES

[1] Gatos B., I. Pratikakis, S. J. Perantonis, "Adaptive Degraded Image Binarization", Journal of Pattern Recocnition, 39 (2006).

[2] Sauvola J., M. Pietikainen, "Adaptive Document Image Binarization", Pattern Recognition 33 (2000).

[3] Yahia S. Halabi, Zaid SA"SA, FarisHamdan, Khaled Haj Yousef, Modeling Adaptive Degraded Document Image Binarization andOptical Character System, European Journal of Scientific Research, ISSN 1450-216X Vol.28 No.1 (2009).

[4] WafaBoussellaa, AbderrazakZahour, Adel Alimi, A Methodology for the Separation of Foreground/Background in Arabic Historical Manuscripts using Hybrid Methods, Journal of Universal Computer Science, vol. 14, no. 2 (2008)

[5] B. Gatos, I. Pratikakis, S.J. Perantonis, Adaptive degraded document image binarization, The journal of the pattern recognition society, pg.4,5

[6] AbhijitMitra, Restoration of Noisy Document Images with an Efficient Bi-Level Adaptive Thresholding, International Journal of Computational Intelligence Volume 2 Number 2

[7] T Kasar, J Kumar and A G Ramakrishnan, Font andBackground Color Independent Text Binarization

[8] K.S. Sim *, C.P. Tso, Y.Y. Tan, Recursive sub-image histogram equalization applied to gray scale images