www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issues 7 July 2016, Page No. 17328-17332

Binarization And Post Processing of Binarized Document Images for Background Text Removal

Aparna Patil[1], Prof. Deepak Gupta[2]

[1]PG Scholar, Computer Department,
Siddhant College of Engineering,
Savitribai Phule Pune University,
Pune, Maharashtra, India
aparnapatil16@gmail.com

[2]Assistant Professor, Computer Department,
Siddhant College of Engineering,
Savitribai Phule Pune University,
Pune, Maharashtra, India
deepak _ gpt@yahoo.com

Abstract: ***There exists an assortment of proposal and books which are been composed long years back. So some of them which are essential for us we should protect them in terms of their degradation. In several instance, these reports are being debased because of some common causes such as discriminated color illusion or ink sipping from background to foreground etc. In order to isolate the content from those corrupted images, such document images need to be processed under efficient binarization methods. In this paper we will build up the framework that can fit for isolating the content from the debased image and few calculations are done such as gray scaling and local thresholding. Image contrast reversal, Edge estimation, Image bimodal binarization and Post Processing binarized images are incorporated into proposed system. Subsequent to utilizing these all techniques proposed system becomes ready to partition out the frontal area accumulation from back ground debasements.***

Keywords: Image adaptive contrast, document images, document image processing, pixel classification, degraded document binarization.

1. Introduction

For researchers all across the terrene, the Image processing domain is reputable and a popular zone of interest. The transaction of customizing or fabricating a neoteric image or an old image is known as image processing by making use of PC calculations on computerized images. The content seems to be facile to pierce and straightforward because of imaging innovation. With the image and its preparing, the preponderance of our general activities is associated. Verifiable reports are being safeguard because of putting away them into an image group. So that our cutting edge can easily be ready to see those old archives. Because of the high background and foreground variety, the partition of content from inadequately debased report images is a troublesome assignment between the archive background and also the closer view content of different document images. Because of non-uniformly debased old image has turned into a muddled archive.

In few cases the images get corrupted because of some regular issues. Facsimile can be corrupted physically to decrease the nature of image. To recoup these report images there onus to be an effective system with the goal that it can be changed over into the discernible arrangement. To meliorate things and stringent resurrection of such report; we have proposed the modish image binarization method. In the four phase of report investigation; the Binarization of image is performed and to partition the closer view content from the archive background is its primary capacity. For recuperating document image with the assistance of preparing assignments, for example, Contrast Enhancement a right archive image binarization strategy is essential. To discover the real content strokes of the image this system utilizes the gray scale technique.

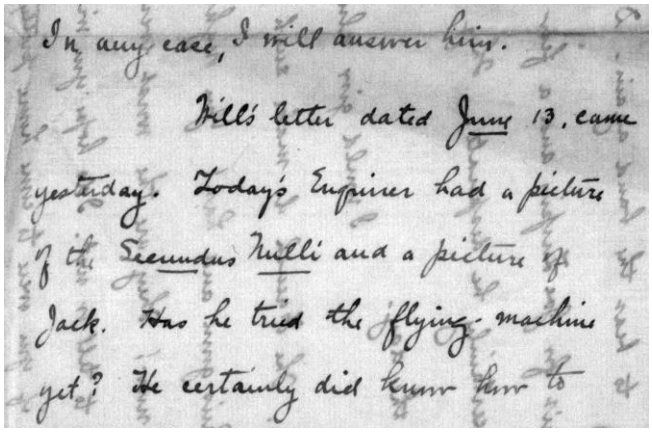

Figure 1: Example of degraded image

2. Literature Survey

Numerous procedures have been composed for testimony image binarization. Intricacy of the current strategy is more costly. For huge images the subsequent binarization maneuver is moderate. It doesn't left the background profundity and low complexity without evident loss of helpful data caused by non-uniform brightening, shadow, spread or smear. The regnant skeleton is not ready to create precise and clear output. Some background debasements might contains in this output.

Table 1: Comparison of Various Methods

Methods	PSNR	NRM	MPM
OTSU	8.98	20.41	172.68
SAUV	12.62	21.56	27.97
NIBL	9.59	9.52	105.17
BERN	5.71	18.86	183.35
GATO	10.40	21.89	36.57
LMM	10.76	17.50	72.08
BE	3.54	40.78	370.15
PROPOSED METHOD	14.50	13.82	14.81

Preceding any treatment of the printed substance from the enactment of the photo the chronicle image can be performed the substance must be segregated. A couple thresholding estimations have as of now been proposed furthermore, are for the uttermost splinter used as a piece of document taking care of. At thresholding troublesome documents none have been shown intense where the establishment and frontal zone are non-uniform. The usage of three overall thresholding counts (Otsu's, Kapur's entropy and Solihin's quadratic fundamental extent (QIR)) as the main stage in a multi-stage thresholding computation for use in debased file images we investigate in this paper.

For troublesome files as they impel to over-edge the photo, in this manner losing a critical constituent of the significant information, it is construed that Otsu's what more is; Kapur's computations don't capacity commendably. In disengaging the draining strand and establishment in these photos, leaving an extent of undecided, soft, pixels for later planning in a subsequent stage. The QIR computation is more correct [1].

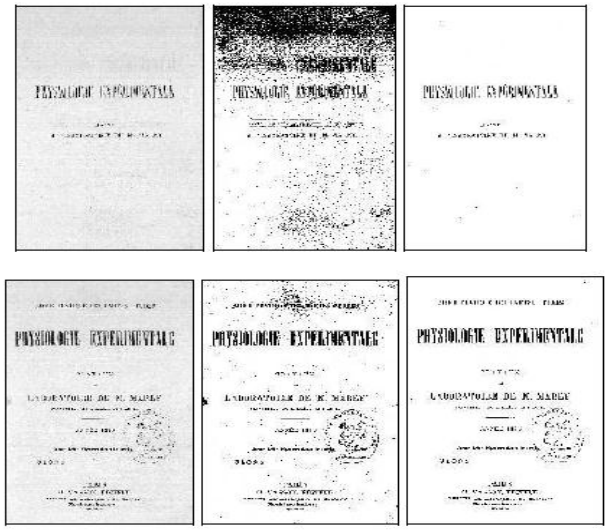

Figure 2: Flow of Entropy system

To pick edges at the bottoms of valleys on the photo's histogram is a comprehended heuristic for dividing a photo into gray level subpopulations. Exactly when the subpopulations spread, valleys may not subsist, but instead it is frequently still possible to portray awesome edges at the `shoulders' of histogram tops. To concavities on the histogram both valleys and shoulders relate, and to find awesome confident edges this prescribes it should be possible by separating the histogram's concavity structure. Histogram concavity examination as an approach as far as possible determination is investigated and its execution on a furtherance of turmoil of histograms of infrared images of tanks is appeared [2].

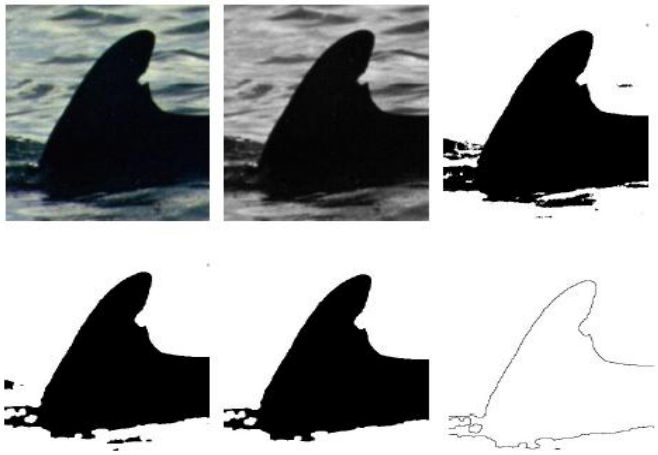

Figure 3: Histogram analysis

In perspective of a water stream demonstrate this paper kibitz an area adaptable thresholding technique, in which a photo surface is premeditated as a three-gray tensional (3-D) domain. We pour water onto the scene surface to focus characters from establishments. Water spurt sagging to the lower districts of the domain and fills valleys. By then, to the measure of filled water for character extraction the thresholding methodology is associated, in which the proposed thresholding procedure is associated with dull level document images including characters and establishments. The dominion of locally adaptable thresholding shows by the proposed system in light of a water stream model. The proposed procedure outputs capable adaptable thresholding results for binarization of document images shows by PC multiplication with built and real file image [3].

The differentiation estimation of the content background and the content are ascertained by it. There are two unique ways to deal with discover the edge which are delicate choice strategy (SDM) and content binarization technique (TBM). The abilities of SDM has clamor sifting and following of sign, To independent content parts from background of the image the TBM is utilized, because of uneven light or commotion which is in terrible conditions group. Finally, the output of these two calculations consolidated together. Future exploration ought to take legitimate approaches to benchmark uses the outcomes against ground and truth measures are vital for the calculation determination procedure and headings. A very much characterized execution assessment demonstrates which capacities of the calculation still need refinement and for a given circumstance which abilities are adequate [4].

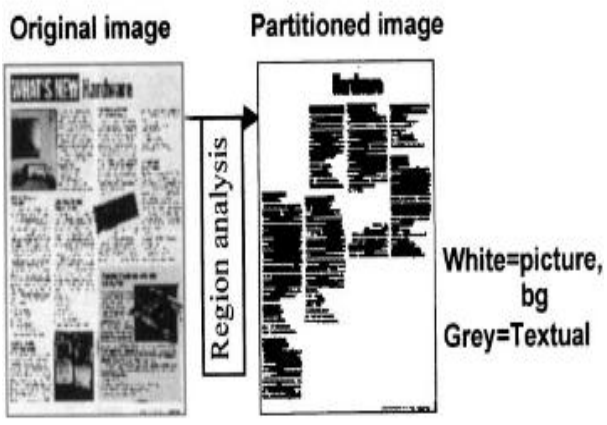

Figure 4: Example of region partitioning for algorithm

(SDM/TBM) selection.

This approach manifests another flexible strategy and change of corrupted documents. by the customer the implemented system does not require any parameter tuning and as a result of shadows, non-uniform light, low difference, immense sign ward hullabaloo, spread and strain can deal with the defilements which happen. We make after a couple of specific strides: a pre-taking care of system using a low-pass Wiener channel, an unforgiving estimation of frontal region areas, an establishment surface figuring by presenting neighbouring establishment intensities, a consolidating in order to thresholding the figured establishment surface with the principal image while combining image up-testing ultimately a post-get ready endeavour with a particular finished objective to emend the way of substance regions and ensure stroke system. After wide examinations, on different undermined document images our framework indicated unrivalled execution against four without a doubt comprehended techniques [5]

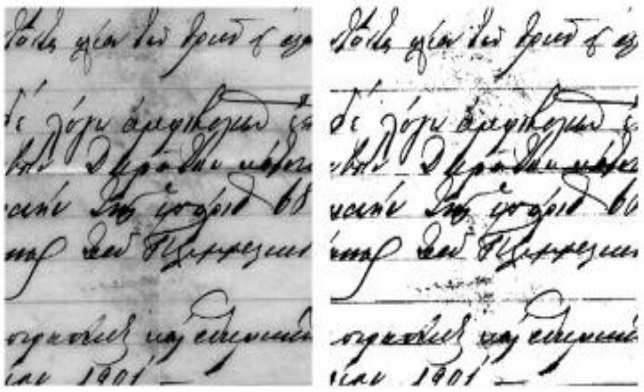

Figure 5: Adaptive Thresholding

For the development of searchable propelled music libraries the Optical music affirmation (OMR) structures are promising instruments in light of covered Markov models Utilizing a flexible OMR system for in front of timetable music prints, to upgrade affirmation precision we impact a modify partition assessment metric. With new named get ready Standard results are figured and test sets drawn from an alternate social affair of prints. In perspective of this appraisal methodology we present two trials. That first happened in a tremendous change to the component extraction limit for these photos. The second is a target composed examination of a couple of renowned adaptable binarization estimations, which are often evaluated just subjectively. For a couple pages Precision in wrinkled by as much as 55%, and for further research the tests prescribe a couple of turnpikes [6].

3. Proposed System

3.1 Modules of the System

3.1.1 Module of Contrast Image

Contradiction is the dissimilitude in fluorescence and/or shielding that makes a thing clear. In visual impression of this present reality, Contrast is the refinement in the color and intensities of the article and distinctive things within the same field of perspective.

Here we are going to use adaptable multifaceted nature which is responsibility of the two systems. Starting one is the neighborhood image contrast; it is just the inversion of the genuine image contrast. It simply has a converse effect image. Second one is neighborhood image inclination. In that we are modifying gradient level of background pixels. Incline of image is an assortment in the agreement level.

3.1.2 Module to find the edges

For revelation of the edges of each pixel we are using gray, the differentiated image is further match with dark scale edge recognition chart. This will convey the selvage of the pixel around the forefront content. Pixel having two sections, related pixels and non-related pixels. A related pixel is the zone around substance stroke. Likewise, a non-related pixel is the corrupted pixel. We get the stroke edge pixels of the archive message authentically from multifaceted nature image advancement. The fabricated differentiation image contain a sensible bi-particular sample.

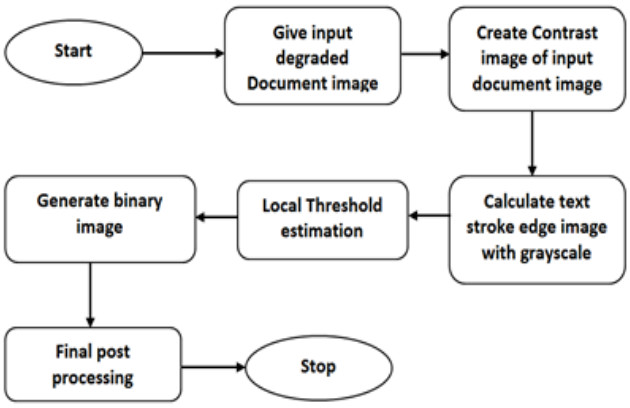

Figure 6: System Architecture

3.1.3 Local threshold Estimation

The recognized text stroke from edge content identification framework is surveyed in this system. Here we are making segment of pixels into two sorts. We are tapping one limit esteem. Contingent on that utmost pixels are named frontal region pixels and background pixels.

3.1.4. Module to convert into binary

The edge assessed image is then changed over into paired configuration i.e. 1 and 0. The images pixels are non-related pixels are exhibited by '0' and image pixels are related pixels are appeared by '1'. As the 0's are a piece of background so they are expelled from image. By then we get only the substance strokes. The created contrast image clears a sensible bi- modular illustration.

3.1.5 Post Processing Module

Binarization makes segment in image. The segment exhibits some background pixels. So we use post preparing to keep up a key separation from those corruptions. Besides, gives back an unmistakable image which involves genuine substance. We can without quite a snip of a stretch watch the modification in Output image and data image. Output image contain spotless and proficient substance.

4. Algorithms

4.1 Luminance Gray Scale Algorithm

1. Capture the red, green, and blue values of a pixel.
2. Use numerical formula to convert those numbers into a single gray value.
3 Redeem the original red, green, and blue values with the new gray value.
4 Get the Grayscale image by using following formula.

Gray = (Red * 0.2126 + Green * 0.7152 + Blue * 0.0722)

4.2 Post Processing Algorithm

1. Find out all the interface segments of the stroke edge pixels in Edg.
2. Remove those pixels that don't interface with different pixels.
3. for Each remaining edge pixels (i, j): do
4. Get its neighborhood sets: (i − 1, j) and (i + 1, j); (i, j −1) and (i, j +1).
5. if The pixels in the same sets have a place with the same class (both content or background) then
6. Assign the pixel with lower force to closer view class (content), and the other to background class.
7. end if.
8. end for.
9. Remove single-pixel ancient rarities along the content stroke limits after the report thresholding.
10. Store the new paired result.

5. Mathematical Model

The mathematical formulation of the proposed system can be done on the basis of Deterministic Finite Automata (DFA) and Context Free Grammar (CFG) to represent the entities and the various transitions taking place between the entities. The DFA for the proposed system comprised of 5 major components as mentioned below:

$S = \{\sum, \lambda, \delta, \psi, F\}$

Where,

$\sum$ = Represents the entities which can be image.

λ = Represents the sequence of operations.

δ = Represents the operation or transitions being Performed for bringing the project from one state to another.

Ψ = Collection of all states of proposed system.

F = Final state of the proposed system which depicts the output.

The proposed system states are as mentioned below:

q0- Input accepting state.

q1- Contrast Image Construction.

q2- Text Stroke Edge Detection.

q3- Local Threshold Estimation.

F- Post processing.

The proposed system goes from one state to another State as follows:

q0→q1 where → Contrast Image Construction

$$\delta_{q0 \to q1} \text{(Contrast Image Construction)}$$

q1→q2 where → Text Stroke Edge Detection

$$\delta_{q1 \to q2} \text{(Text Stroke Edge Detection)}$$

q2→q3 where → Local Threshold Estimation

$$\delta_{q2 \to q3} \text{(Local Threshold Estimation)}$$

q3→F where → Post processing

$$\delta_{q3 \to F} \text{(Post processing)}$$

5.1 Performance Metrics and Measures

PSNR = $^{10}\log_{10}(I_{(i,j)} - J_{(i,j)})^2/MSE$

Where I = Image(1), J = Image(2)

i = width, j= height.

MSE = $[(I_{(i,j)} - J_{(i,j)})/\text{total no.of pixels}]^2$

Where j= height, i = width.

6. Experimental Results

The proposed system presented in the paper works in a modular approach thereby making the system work in a sequential manner with output of first module to be considered as input to the second module. The outputs of the implemented modules of the proposed system are as follows:

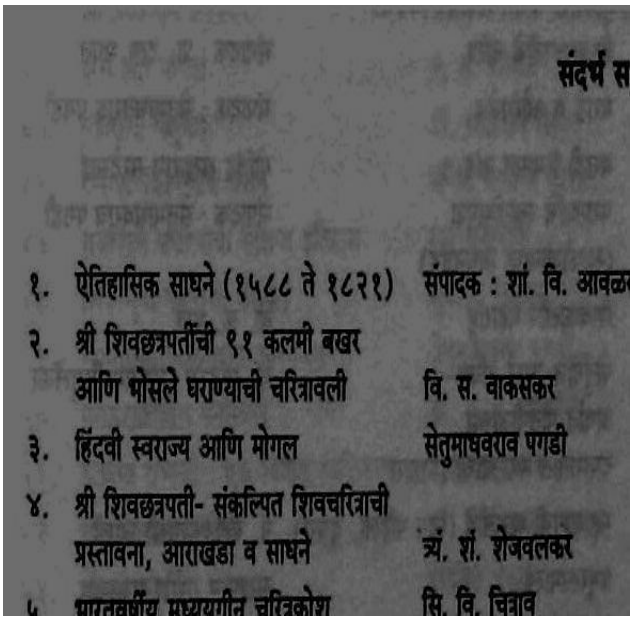

Figure 7: Input image

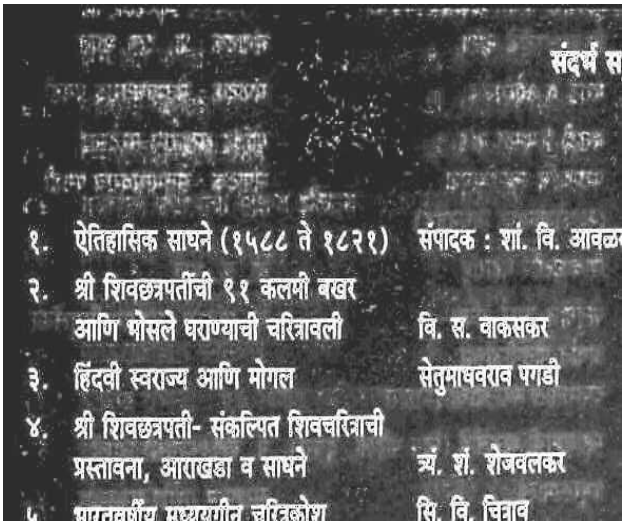

Figure 8: Contrast image

Figure 9: Text Stroke Edge detected image

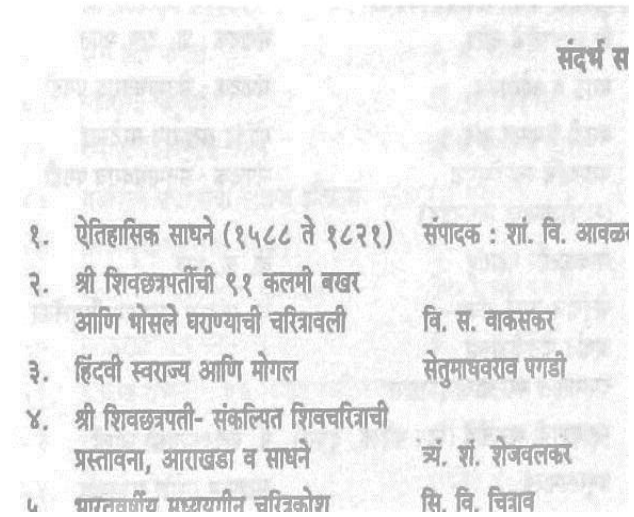

Figure 10: Binarized image

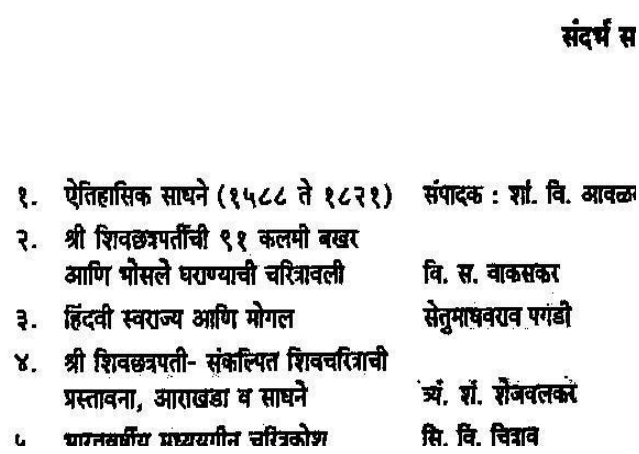

Figure 11: Post Processed Image

7. Conclusion

As per experimental results obtained can conclude that this strategy can make more proficient output than other existing procedures. This can pirouette out to be exceptionally helpful to recover unique information from debased documents. This paper utilizes gray scale edge detection strategy to make edge guide or outskirt around the content. At long last framework produces image containing just forefront content. Toward the end we are going to assess the effectiveness parameter of our framework. . In our framework we are uprooting the Canny's edge detection calculation. So that the effectiveness of framework increases by reducing the entanglement of working on same image for more than once.

Acknowledgment

I would like to thank the Principal for giving him the opportunity to work on this project and guide Prof. Deepak Gupta for his government and knowledge without which this paper would not be possible. He provided me with valuable advice which helped me to accomplish writing this paper.

References

[1] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images", in Proc. Int. Conf. Document Anal.Recognit., vol. 13. 2003, pp. 859–864.

[2] A. Rosenfeld and P. De la Torre, ''Histogram concavity analysis as an aid in threshold selection'', IEEE Trans. Syst. Man Cybern. SMC-13,231–235 1983!.

[3] I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarization based on topographic analysis using a water flow model", Pattern Recognit., vol. 35, no. 1, pp. 265–277, 2002.

[4] J. Sauvola and M. Pietikainen, "Adaptive document image binarization", Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.

[5] L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images", in Proc. Int. Conf. Document Anal. Recognit., Sep. 1991, pp. 435–443.

[6] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods", IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 12,pp. 1191–1201, Dec. 1995.

[7] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "ICDAR 2011 document image binarization contest (DIBCO 2011)", in Proc. Int. Conf. Document Anal. Recognit., Sep. 2011, pp. 1506–1510.

Author Profile

Aparna Patil received the B.E. and Pursuing M.E. degrees in Computer Engineering from Siddhant College of Engineering, Pune. In academic year 2015-16.

Prof. Deepak Gupta received the B.E and M.Tech degrees in Information Technology from S.A.T.I., Vidisha, Madhya Pradesh, in 2002 and 2007, respectively. Now he is working as assistant Professor in Siddhant College of Engineering, Pune.