IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

An Efficient Recovery of Degraded Document Images using Binarization Technique

Umesh B. Sangule[1], Dr. A. B. Pawar [2]

PG Student, Computer Engineering, SRES's College of Engineering, Kopargaon, India [1]

Associate Professor, Computer Engineering, SRES's College of Engineering, Kopargaon, India [2]

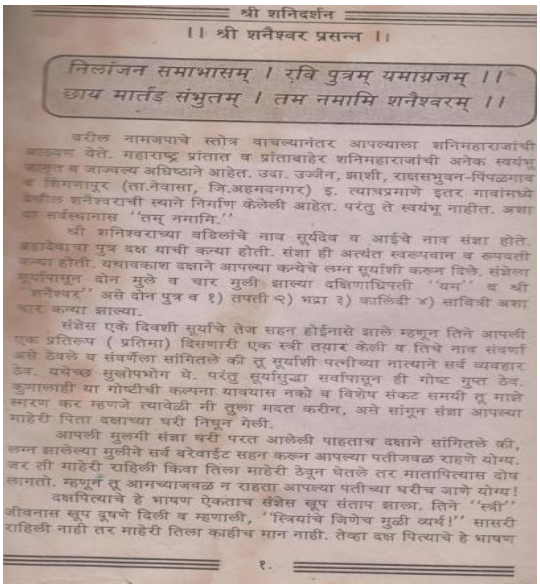
Abstract: There exists a variety of proposal and books which are been composed long years back. So some of them which are essential for us we should protect them in terms of their degradation. In any case, these reports are being debased because of some common causes such as discriminated color illusion, or ink sipping from background to foreground etc. Because of such debasement a hefty portion of the report images are not in lucid arrangement. In order to isolate the content from those corrupted images, such document images need to be processed under efficient binarization methods. For this reason we will build up the framework that can fit for isolating the content from the debased image. Proposed system eradicates the Cannys' edge recognition calculation in its framework. In this paper, for isolating the frontal area and the background of the image, use of 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 applying these all techniques proposed system becomes ready to partition out the frontal area content from back ground debasements.

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

I. INTRODUCTION

For researchers all over the world, the Image processing domain is well known and a popular zone of interest. The process of modifying or creating a new image or an old image is known as image processing by making use of PC calculations on computerized images. The content seems to be easy to see and straightforward because of imaging innovation. With the image and its preparing, the majority of our general activities are 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 some cases the images get corrupted because of some regular issues. Images can be corrupted physically to decrease the nature of image. To recoup these report images there ought to be an effective system with the goal that it can be changed over into the discernable arrangement. To improve things and precise recuperation of such report we have proposed the new 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.


श्री शनिदर्शन

।। श्री शनैश्वर प्रसन्न ।।

नीलांजन समाभासम् । रवि पुत्रम् यमाग्रजम् ।।
छाय मार्तंड संभुतम् । तम नमामि शनैश्वरम् ।।

वरील नामजपाचे स्तोत्र वाचल्यानंतर आपल्याला शनिमहाराजांची आठवण येते. महाराष्ट्र प्रांतात व प्रांताबाहेर शनिमहाराजांची अनेक स्वयंभू जागृत व जाज्वल्य अधिष्ठाने आहेत. उदा. उज्जैन, झाशी, राक्षसभुवन-पिंपळगाव व शिंगणापूर (ता.नेवासा, जि.अहमदनगर) इ. त्याचप्रमाणे इतर गावांमध्ये देखील शनैश्वराची स्थाने निर्माण केलेली आहेत. परंतु ते स्वयंभू नाहीत. अशा सर्वस्थानास ''तम् नमामि.''

श्री शनिश्वराच्या वडिलांचे नाव सूर्यदेव व आईचे नाव संज्ञा होते. ब्रह्मदेवाचा पुत्र दक्ष याची कन्या होती. संज्ञा ही अत्यंत स्वरूपवान व रूपवती कन्या होती. यथावकाश दक्षाने आपल्या कन्येचे लग्न सूर्याशी करून दिले. संज्ञेला सूर्यापासून दोन मुले व चार मुली झाल्या दक्षिणाधिपती ''यम'' व श्री ''शनैश्वर'' असे दोन पुत्र व १) तपती २) भद्रा ३) कालिंदी ४) सावित्री अशा चार कन्या झाल्या.

संज्ञेस एके दिवशी सूर्याचे तेज सहन होईनासे झाले म्हणून तिने आपली एक प्रतिरूप (प्रतिमा) बिसणारी एक स्त्री तयार केली व तिचे नाव सवर्णा असे ठेवले व संवर्णेला सांगितले की तू सूर्याशी पत्नीच्या नात्याने सर्व व्यवहार ठेव. यथेच्छ सुखोपभोग ये. परंतु सूर्यापासून ही गोष्ट गुप्त ठेव. कुणालाही या गोष्टीची कल्पना यावयास नको व विशेष संकट समयी तू माझे स्मरण कर म्हणजे त्यावेळी मी तुला मदत करीन, असे सांगून संज्ञा आपल्या माहेरी पिता दक्षाच्या घरी निघून गेली.

आपली मुलगी संज्ञा घरी परत आलेली पाहताच दक्षाने सांगितले की, लग्न झालेल्या मुलीने सर्व बरेवाईट सहन करून आपल्या पतीजवळ राहणे योग्य. जर ती माहेरी राहिली किंवा तिला माहेरी ठेवून घेतले तर मातापित्यास दोष लागतो. म्हणून तू आमच्याजवळ न राहता आपल्या पतीच्या घरीच जाणे योग्य! दक्षपित्याचे हे भाषण ऐकताच संज्ञेस खूप संताप झाला. तिने ''स्त्री'' जीवनास खूप दूषणे दिली व म्हणाली, ''स्त्रियांचे जिणेच मुळी व्यर्थ!'' सासरी राहिली नाही तर माहेरी तिला काहीच मान नाही. तेव्हा दक्ष पित्याचे हे भाषण

१.

Fig 1: Shows example of degraded image

II. LITERATURE SURVEY

Numerous procedures have been produced for document image binarization. Intricacy of the current strategy is more costly. For huge images the subsequent binarization procedure 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 current framework is not ready to create precise and clear output. Some background debasements might contains in this output.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Table 1 Comparison of various methods

Methods	PSNR	NRM	MPM
OTSU	17.51	9.77	1.35
SAUV	15.96	16.31	1.96
NIBL	15.73	19.06	1.06
BERN	8.57	21.18	115.98
GATO	15.12	21.89	0.41
LMM	17.83	11.46	0.37
BE	18.14	9.06	1.11
PROPOSED METHOD	20.12	6.14	0.25

A. G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, "Thresholding algorithms for text/background segmentation in difficult document images"[1]

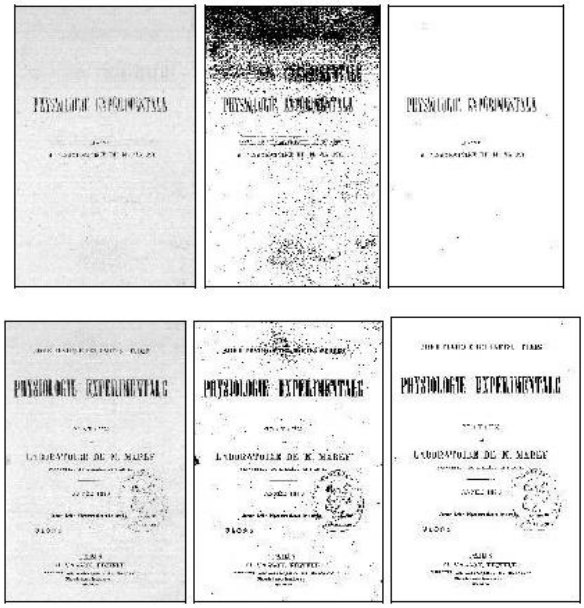


Fig 2: Flow chart of Entropy system

Preceding any treatment of the printed substance from the establishment 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 most part used as a piece of document taking care of. At thresholding troublesome document s 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 tend to over-edge the photo, in this manner losing a critical part of the significant information, it is construed that Otsu's what more is, Kapur's computations don't capacity commendably. In disengaging the bleeding edge 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.

B. Rosenfeld and P. De la Torre, ''Histogram concavity analysis as an aid in threshold selection" [2]
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 exist, 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 course of action of histograms of infrared images of tanks is appeared.

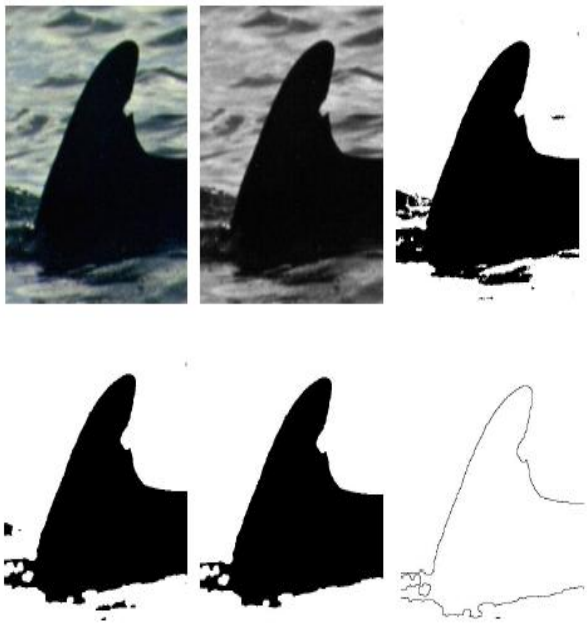


Fig 3 Histogram analysis

C. I.-K. Kim, D.-W. Jung, and R.-H. Park, "Document image binarization based on topographic analysis using a water flow model" [3]
In perspective of a water stream demonstrate this paper proposes an area adaptable thresholding technique, in which a photo surface is considered as a three-gray ensional (3-D) domain. We pour water onto the scene surface to focus characters from establishments. Water spills down 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 property 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 images.

DOI 10.17148/IJARCCE.2016.5693

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

D. J. Sauvola and M. Pietikainen, "Adaptive document image binarization," [4]

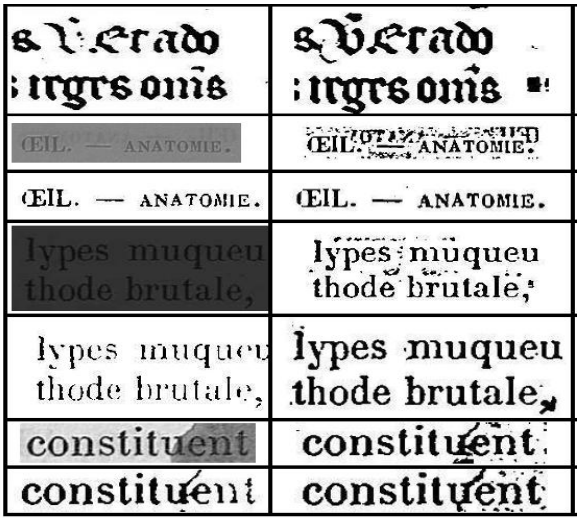


Fig. 4 Sample degraded image

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.

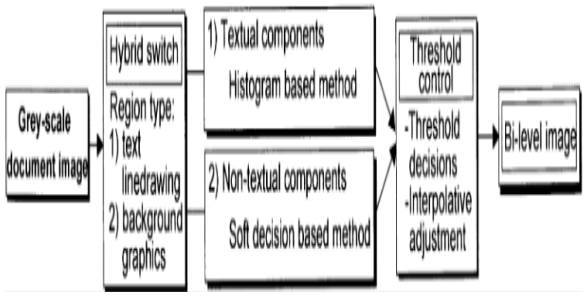


Fig 5. Overview of the binarization algorithm

E. L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images," [5]

For the binarization this paper shows another flexible strategy and change of corrupted document s. by the customer the proposed 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 neighboring 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 endeavor with a particular finished objective to improve the way of substance regions and ensure stroke system.

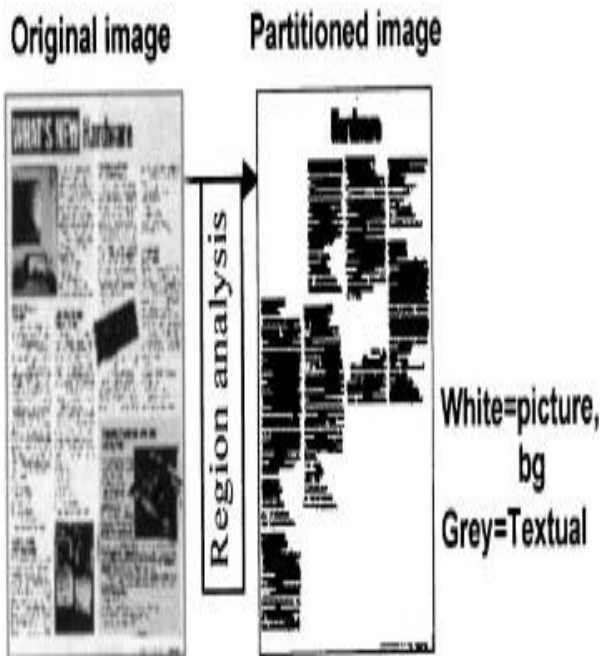


Fig 6 Example of region partitioning for Algorithm

After wide examinations, on different undermined document images our framework indicated unrivalled execution against four without a doubt comprehended techniques.

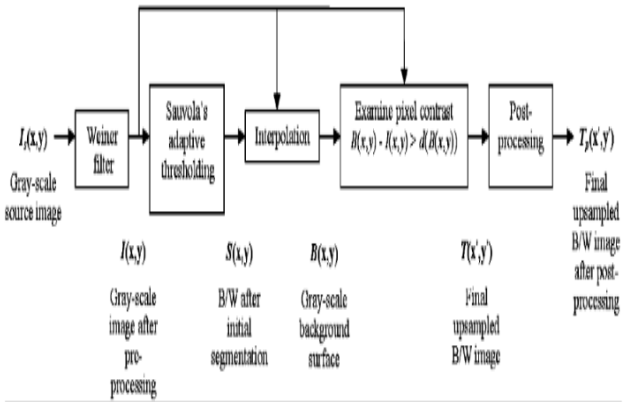


Fig 7. Block diagram

F. O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarization methods," [6]

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.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

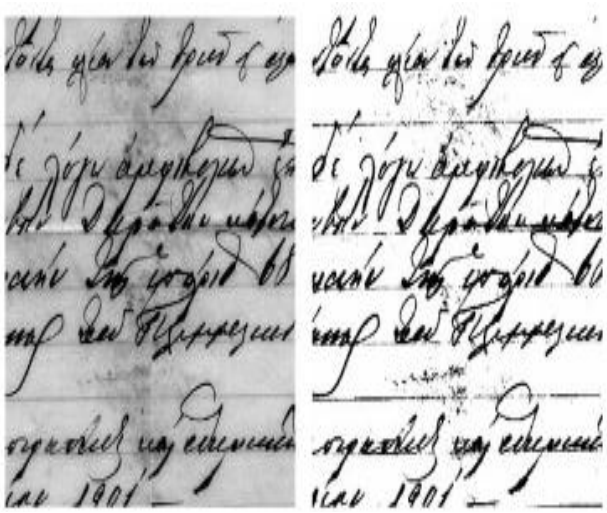


Fig 8. Adaptive Thresholding

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.

III. PROPOSED SYSTEM

A. Modules

• Module of Contrast Image:
Contrast is the distinction in luminance and/or shading 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 have 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.

- 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 border 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.

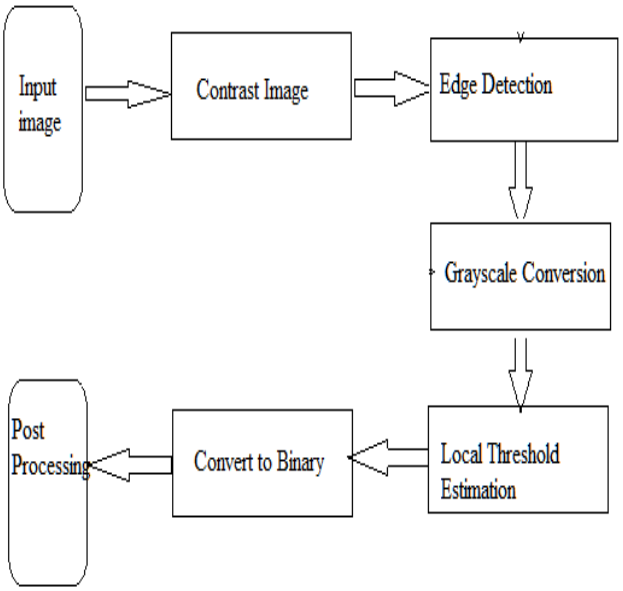


Fig. 9 System Architecture

• Local threshould 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 picking one limit esteem. Contingent on that utmost pixels are named frontal region pixels and background pixels.

• Module to convert into binary:
The edge assessed image is then changed over into paired configuration i.e. 1 and 0. The image pixels are non-related pixels are exhibited by 0 and image pixels are related pixels are appeared by 1. As the 0z are a piece of background so they are expelled from image. By then we get only the substance strokes. The created contrast image clear a sensible bi- modular illustration.

• 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 that corruptions. Besides, gives back an unmistakable image which involve genuine substance. We can without quite a bit of a stretch watch the modification in Output image and data image. Output image contain spotless and proficient substance.

B. Algorithms

• Luminance Gray scale Algorithm
In my proposed system, I have chosen to go with the original ITU-R recommendation which is the historical precedent. This formula, sometimes called Luma, looks like this:

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

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

• Text Stroke Edge Detection Algorithm

1. Get the width of input image as well as height of input image I
2. for Each Row of input image i =1 to height of input image Edg do
3. Scan the image from left to right to determine edge pixels that meet the required criteria: an) its name is 0 ((background); b) the following pixel is named as 1(edge).
4. Check the pixel values in I of those pixels chose in Step 3, and uproot those pixels that have a lower force than the accompanying pixel by it in the same line of I.
5. Match the staying nearby pixels in the same column into sets, and figure the separation between the two pixels in pair.
6. end for
7. Construct a histogram of those calculated distances.
8. Use the most much of the frequently occurring distance as the assessed stroke edge width EW

• 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.

C. Expected Outputs

The input image should be degraded. The contrast image should be generate after giving to the input image as an input to the contrast image module. The text stroke image should be generate after processing the text stroke detection module.

The binarization image should be generate after local threshold estimation module whose background should be white and foreground text should be black. The cleared image should be produce after post processing module.

IV. 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 output of the implemented modules of the proposed system are as follows:

A. Input Image:

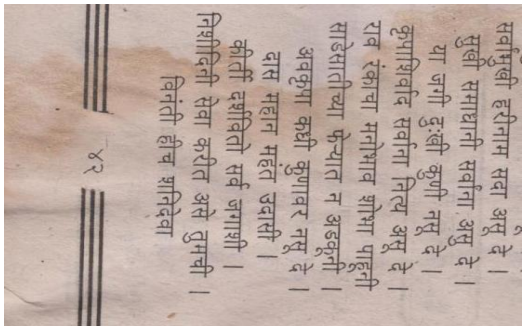


Fig. 10. Input image (a)

B. Contrast Module Output:

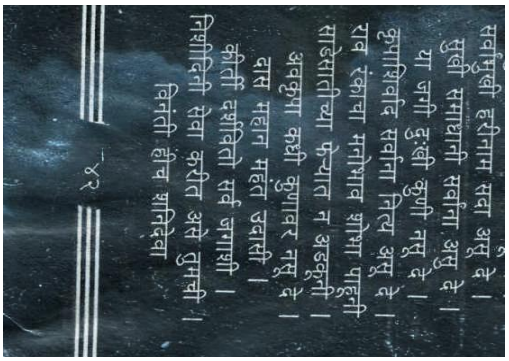


Fig. 11. Contrast image (a)

C. Text Stroke Detected Image:

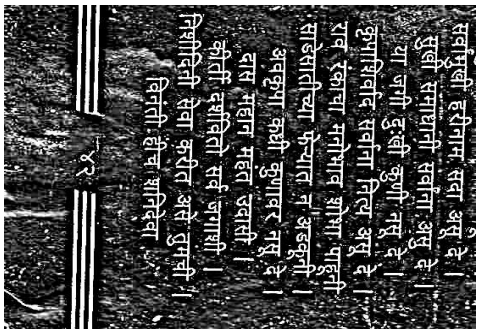


Fig. 12 Text Stroke Edge detected image (a)

D. Binary Output Image:

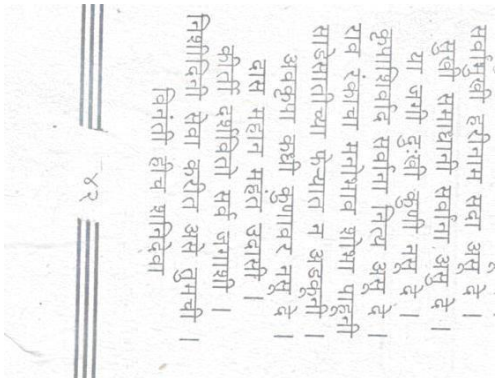


Fig. 13 Binarized image (a)

DOI 10.17148/IJARCCE.2016.5693

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

E. Post Processing Output Image:

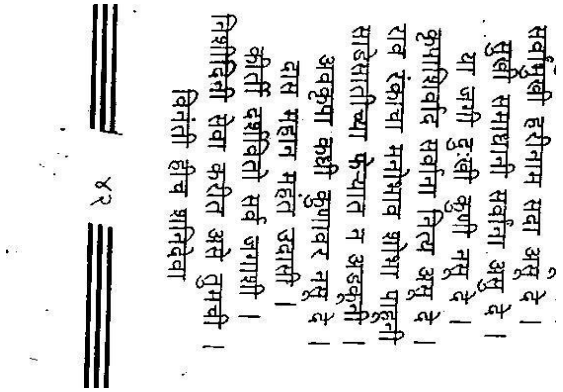


Fig. 14 Post Processed image (a)

F. PSNR Values Obtained:

Mean Square Error (MSE)	Signal to Noise Ratio (SNR)	Peak Signal to Noise Ratio (PSNR)
0.13*(10^-14)	18.57	4.31

Fig. 15 a. PSNR value image (a).

Mean Square Error (MSE)	Signal to Noise Ratio (SNR)	Peak Signal to Noise Ratio (PSNR)
0.19*(10^-14)	19.87	4.47

Fig. 15 b. PSNR value image (b).

V. CONCLUSION

As per experimental results obtained can conclude that this strategy can make more proficient output than other existing procedures. This can turn 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 complexity of working on same image for more than once.

ACKNOWLEDGMENT

I would like to take this opportunity to express my heartfelt thanks to my guide **Dr. A. B. Pawar** for his esteemed guidance and encouragement, especially through difficult times. His suggestions broaden my vision and guided me to succeed in this work. I am also very grateful for his guidance and comments while designing part of my research paper and learnt many things under his leadership.

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] Rosenfeld and P. De la Torre, ''Histogram concavity analysis as an aid in threshold selection" Int. J. Document Anal. Recognit., vol. 13, no. 4, pp. 303–314, Dec. 2010.

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

[4] J. Sauvola and M. Pietikainen, "Adaptive document image binarization," J. Electron. Imag., vol. 13, no. 1, pp. 146–165, Jan. 2004.

[5] L. Eikvil, T. Taxt, and K. Moen, "A fast adaptive method for binarization of document images," in Proc. Int. Workshop Document Anal. Syst., Jun. 2010, pp. 159–166.

[6] O. D. Trier and A. K. Jain, "Goal-directed evaluation of binarizationmethods," Int. J. Document Anal. Recognit., vol. 13, no. 4, pp. 303–314, Dec. 2010

[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.

BIOGRAPHIES

Umesh B. Sangule received the B.E. degree in Information Technology from University of Pune, Pune, Maharashtra, India in 2014. He is currently pursuing the M.E. degree in Computer Engineering in SRES Sanjivani College of Engineering, Kopargaon, Savitribai Phule Pune University, Pune, Maharashtra, India. His current research interests include Image Processing and wireless networking, sensor networks.