Recovery of badly degraded Document images using Binarization Technique

Prof. S. P. Godse, Samadhan Nimbhore, Sujit Shitole, Dinesh Katke, Pradeep Kasar

Computer Science Engineering Department, Pune University, Sinhgad Academy of Engineering, Kondhwa(bk), Dist- Pune-48 , Maharashtra, India.

Abstract- Recovering of text from badly degraded document images is a very difficult task due to the very high inter/intra-variation between the document background and the foreground text of different document images. In this paper, we propose a robust document image binarization technique that addresses these issues by using inversion gray scale image contrast. The Inversion image contrast is a done by first converting the input image to invert image and then finding the contrast of the inverted image to differentiate text and background variation caused by different types of document degradations. In the proposed technique, an adaptive contrast map is first constructed for an input degraded document image. The contrast map is then converted to grayscale image so as to clearly identify the text stroke from background and foreground pixels. The document text is further segmented by a local threshold that is estimated based on the intensities of detected text stroke edge pixels within a local window. The proposed method is simple, robust, and involves minimum parameter tuning. Several challenging bad quality document images also showthe superior performance of our proposed method, compared with other techniques.

Index Terms- Image contrast,gray scale image, document analysis, document image processing, degraded document image binarization, pixel classification.

I. INTRODUCTION

DOCUMENT Image Binarization is performed in the preprocessing stage for document analysis and it aims to segment the foreground text from the document background. A fast and accurate document image binarization technique is important for the ensuing document image processing tasks such as optical character recognition (OCR). Though document image binarization has been studied for many years, the thresholding of degraded document images is still an unsolved problem due to the high inter/intra-variation between the text stroke and the document background across different document images. As illustrated in Fig. 1, the handwritten text within the degraded documents often shows a certain amount of variation in terms of the stroke width, stroke brightness, stroke connection, and document

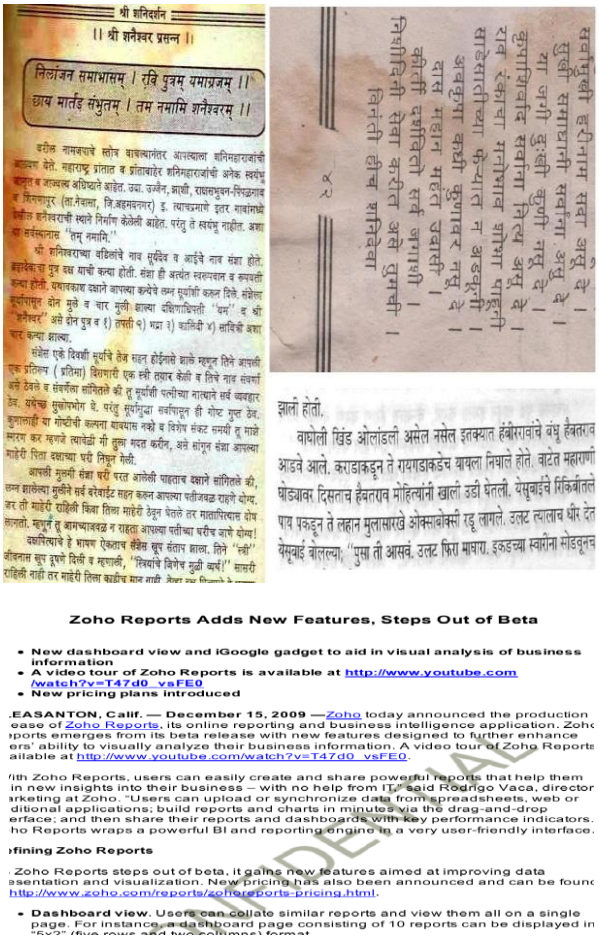

Fig. 1.*Threedegraded document images (a)–(d) Images from various degraded documents aretaken from Internet.*

Background . In addition, historical documents are often degraded by the bleedthrough as shown in Fig. 1(a) and (c) where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts as shown in Fig. 1(b). These different types of document degradations tend to induce the document thresholding error and make degraded document image binarizationa big challenge to most state-of-the-art techniques.This paper presents a document binarization technique that extends our previous local maximum-minimum method and the method used in the latest DIBCO 2011. The proposed method is simple, robust and capable of handling different types of degraded document images with minimum parameter tuning. It makes use of the inversion image contrast

that converts the input image into invert image and then converts the invert image into contrast image and therefore is tolerant to the text and background variation caused by different types of document degradations. In particular, the proposed technique addresses the over-normalization problem of the local maximum minimum algorithm. At the same time, the parameters used in the algorithm can be adaptively estimated. The rest of this paper is organized as follows. Section II first reviews the current state-of-the-art binarization techniques. Our proposed document binarization technique is described in Section III. Then experimental results are reported in Section IV to demonstrate the superior performance of our framework. Finally, conclusions are presented in Section V.

II. RELATED WORK

Many thresholding techniques have been reported for document image binarization. As many degraded documents do not have a clear bimodal pattern, global thresholding is usually not a suitable approach for the degraded document binarization. Adaptive thresholding, which estimates a local threshold for each document image pixel, is often a better approach to deal with different variations within degraded document images. For example, the early window-based adaptive thresholding techniques, estimate the local threshold by using the mean and the standard variation of image pixels within a local neighborhood window. The main drawback of these window-based thresholding techniques is that the thresholding performance depends heavily on the window size and hence the character stroke width. Other approaches have also been reported, including background subtraction, texture analysis, recursive method, decomposition method, contour completion, Markov Random Field, matched wavelet, cross section sequence graph analysis, self-learning, Laplacian energy user assistance, and combination of binarization techniques. These methods combine different types of image information and domain knowledge and are often complex. The local image contrast and the local image gradient are very useful features for segmenting the text from the document background because the document text usually has certain image contrast to the neighboring document background. They are very effective and have been used in many document image binarization techniques . In Bernsen'spaper [14], the local contrast is defined as follows:

$$C(x,y) = I\max(x,y) - I\min(x,y) \quad (1)$$

where $C(x,y)$ denotes the contrast of an image pixel (x,y),$I\max(x,y)$ and $I\min(x,y)$ denote the maximum and minimum intensities within a local neighborhood windows of (x,y),respectively. If the local contrast $C(x,y)$ is smaller than a threshold, the pixel is set as background directly. Otherwise it will be classified into text or background by comparing with the mean of $I\max(x,y)$ and $I\min(x,y)$. Bernsen's method is simple, but cannot work properly on degraded document images with a complex document background. We have earlier proposed a novel document image binarization method by using the local image contrast that is evaluated as follows :

$$C(x,y) = \frac{I\max(x,y) - I\min(x,y)}{I\max(x,yj) + I\min(x,y) + €} \quad (2)$$

where €is a positive but infinitely small number that is added in case the local maximum is equal to 0. Compared with Bernsen's contrast in Equation 1, the local image contrast in Equation 2 introduces a normalization factor (the denominator) to compensate the image variation within the document background. Take the text within shaded document areas such as that in the sample document image in Fig. 1(b) as an example. The small image contrast around the text stroke edges in Equation 1 (resulting from the shading) will be compensated by a small normalization factor (due to the dark document background) as defined in Equation 2.

III. PROPOSED METHOD

This section describes the proposed document image binarization techniques. Given a degraded document image, an inversion contrast map is first constructed and the text stroke edges are then detected through the grayscale conversion of contrast image. The text is then segmented based on the local threshold that is estimated from the detected text stroke edge pixels. Some post-processing is further applied to improve the document binarization quality.

A. Contrast Image Construction

The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many non-stroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image needs to be normalized (inverted) to compensate the image variation within the document background. In our earlier method, The local contrast evaluated by the local image maximum and minimum is used to suppress the background variation as described in Equation 2. In particular, the numerator (i.e. the difference between the local maximum and the local minimum) captures the local imagedifference that is similar to the traditional image gradient. The denominator is a normalization factor that suppresses the image variation within the document background. For image pixels within bright regions, it will produce a large normalization factor to neutralize the numerator and accordingly result in a relatively low image contrast. For the image pixels within dark regions, it will produce a small denominator andaccordingly result in a relatively high image contrast. However, the image contrast in Equation 2 has one typical limitation that it may not handle document images with the bright text properly. This is because a weak contrast will be calculated for stroke edges of the bright text where the denominator in Equation 2 will be large but the numerator will be small. To overcome this over-normalization problem, we convert the input degraded image into invert image where the image color pixels are inverted according to 256 bitmap colors. The inverted image is then converted to contrast image to get clear differentiation between background and foreground pixels.

The contrast is found as follows:

The proposed binarization technique relies more on inverted image and avoid the over normalization problem of our previous method.

Section IV. Fig. 2 shows the contrast map of the sample document images in Fig. 1 (b) and (d) that are created by using local image gradient, local image contrast and our proposed method in Equation 3, respectively. For the sample document with a complex document background in Fig. 1(b), the use of the local image contrast produces a better result as shown in Fig. 2(b) compared with the result by the local image gradient as shown in Fig. 2(a)(because the normalization factors in Equation 2 helps to suppress the noise at the upper left area of Fig. 2(a)). But for the sample document in Fig. 1(d) that has small intensityvariation within the document background but large intensityvariation within the text strokes, the use of the local image contrast removes many light text strokes improperly in the contrast map as shown in Fig. 2(b) whereas the use of local image gradient is capable of preserving those light text strokesas shown in Fig. 2(a). As a comparison, the adaptive combination of the local image contrast and the local image gradient in Equation 3can produce proper contrast maps for document images with different types of degradation as shown in Fig. 2(c). In particular, the local image contrast in Equation 3 gets a high weight for the document image in Fig. 1(a) with high intensity variation within the document background whereas the local image gradient gets a high weight for the document image inFig. 1(b).

Fig. Shows Contrast Images constructed.

B. Text Stroke Edge Pixel Detection

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. The constructed contrast image has a clear bi-modal pattern, where the inversion image contrast computed at text stroke edges is obviously larger than that computed within the document background. We therefore detect the text stroke edge pixel candidate by using Otsu's global thresholding method. For the contrast images in Fig. 2(c), Fig. 3(a) shows a binary map by Otsu's algorithm that extracts the stroke edge pixels properly. As the local image contrast and the local image gradient a reevaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels. So,Before applying the Otsu's global thresholding algorithm, we first convert the image into grayscale as the gray scale version of the contrast image has the efficient variation between background and the foreground pixels. In the final text stroke edge image map, we keep only pixels that appear within the high contrast image pixels in gray scale image. The gray scale conversion helps to extract the text stroke edge pixelsaccurately as shown in Fig.

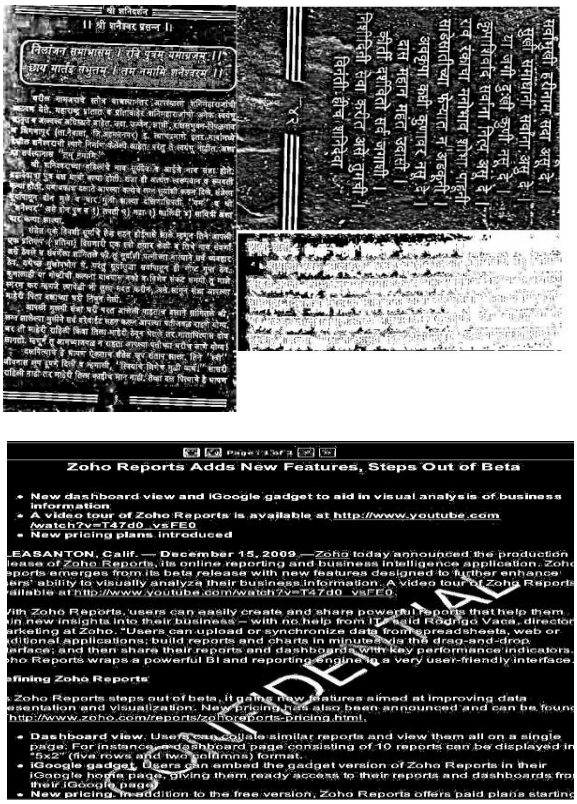

Fig. 3.Edge map for above images after grayscale of the sample document images.

C. Local Threshold Estimation

The text can then be extracted from the document background pixels once the high contrast stroke edge pixels are detected properly. Two characteristics can be observed from different kinds of document images: First, the text pixels are close to the detected text stroke edge pixels. Second, there is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels. The document

image text can thus be extracted based on the detected text stroke edge pixels as follows:

$$R(x,y) = \begin{cases} 1 & I(x,y) \le E_{\text{mean}} + \frac{E_{\text{std}}}{2} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Where *E*mean and *E*std. are the mean and standard deviation of the intensity of the detected text stroke edge pixels within a neighborhood window *W*, respectively. The neighborhood window should be at least larger than the stroke width in order to contain stroke edge pixels. So the size of the neighborhood window *W* can be set based on the stroke width of the document image under study, *EW*, which can be estimated from the detected stroke edges [shown in Fig. 3(b)]as stated in Algorithm 1.Since we do not need a precise stroke width, we justcalculate the most frequently distance between two adjacent edge pixels (which denotes two sides edge of a stroke) in horizontal direction and use it as the estimated stroke width. First the edge image is scanned horizontally row by row and the edge pixel candidates are selected as described in step 3.If the edge pixels, which are labelled 0 (background) andthe pixels next to them are labeled to 1 (edge) in the edge map (*Edg*), are correctly detected, they should have higher intensities than the following few pixels (which should be the text stroke pixels). So those improperly detected edge pixels are removed in step 4. In the remaining edge pixels in the same row, the two adjacent edge pixels are likely the two sides of a stroke, so these two adjacent edge pixels are matched to pairs and the distance between them are calculated in step 5. After that a histogram is constructed that records the frequency of the distance between two adjacent candidate pixels. The stroke edge width *EW* can then be approximately estimated by using the most frequently occurring distances of the adjacent edge
pixels as illustrated in Fig. 4.

Algorithm 1 Edge Width Estimation

Require: The Input Document Image *I* and Corresponding Binary Text Stroke Edge Image *Edg*
Ensure: The Estimated Text Stroke Edge Width *EW*
1: Get the width and height of I
2: for Each Row x = 1 to height in Edg do
3: Scan from left to right to find edge pixels that meet the following criteria:
a) its label is 0 (background);
b) the next pixel is labeled as 1(edge).
4: Examine the intensities in I of those pixels selected in Step 3, and remove those pixels that have a lower intensity than the following pixel next to it in the same row of I .
5: Match the remaining adjacent pixels in the same row into pairs, and calculate the distance between the two pixels in pair.
6: end for
7: Construct a histogram of those calculated distances.
8: Use the most frequently occurring distance as the estimatedstroke edge width EW.

D. Post-Processing

Once the initial binarization result is derived from Equation5 as described in previous subsections, the binarization result can be further improved by incorporating certain domain knowledge as described in Algorithm 2. First, the isolated foreground pixels that do not connect with other foregroundpixels are filtered out to make the edge pixel set precisely.Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text). One pixel of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operatorsas described in [4].

Algorithm 2 Post-Processing Procedure

Require: The Input Document Image I , Initial Binary Result B and Corresponding Binary Text Stroke Edge Image Edg
Ensure: The Final Binary Result B*f*
1: Find out all the connect components of the stroke edge pixels in Edg.
2: Remove those pixels that do not connect with other pixels.
3: for Each remaining edge pixels (x,y): do
4: Get its neighborhood pairs: (x − 1, y) and (x + 1, y);
(x, y − 1) and (x, y + 1)
5: if the pixels in the same pairs belong to the same class (both text or background) then
6: Assign the pixel with lower intensity to foreground
class (text), and the other to background class.
7: end if
8: end for
9: Remove single-pixel artifacts [4] along the text stroke boundaries after the document thresholding.
10: Store the new binary result to B*f* .

D. Post-Processing

Once the initial binarization result is derived from Equation 5 as described in previous subsections, the binarizationresult can be further improved by incorporating certain domain knowledge as described in Algorithm 2. First, the isolated foreground pixels that do not connect with other foreground pixels are filtered out to make the edge pixel set precisely. Second, the neighborhood pixel pair that lies on symmetric sides of a text stroke edge pixel should belong to different classes (i.e., either the document background or the foreground text). One pixel of the pixel pair is therefore labeled to the other category if both of the two pixels belong to the same class. Finally, some single-pixel artifacts along the text stroke boundaries are filtered out by using several logical operators as described in [4].

Algorithm 2 Post-Processing Procedure

Require: The Input Document Image I , Initial Binary Result B and Corresponding Binary Text Stroke Edge Image Edg
Ensure: The Final Binary Result B*f*
1: Find out all the connect components of the stroke edge pixels in Edg.
2: Remove those pixels that do not connect with other pixels.
3: for Each remaining edge pixels (x,y): do
4: Get its neighborhood pairs: (x − 1, y) and
(x + 1, y);
(x, y − 1) and (x, y + 1)
5: if The pixels in the same pairs belong to the same class (both text or background) then

6: Assign the pixel with lower intensity to foreground class (text), and the other to background class.
7: end if
8: end for
9: Remove single-pixel artifacts [4] along the text stroke boundaries after the document thresholding.
10: Store the new binary result to Bf .

IV. EXPERIMENTS AND DISCUSSION

A few experiments are designed to demonstrate the effectiveness and robustness of our proposed method. We first analyze the performance of the proposed technique on public datasets for parameter selection. Due to lack of ground truth data in some datasets, no all of the metrics are applied on every images.

A. Parameter Selection

The γ increases from 2^{-10} to 2^{10} exponentially and monotonically as shown in Fig. 5(a). In particular, α is close to 1 when γ is small and the local image contrast C dominates the adaptive image contrast Ca in Equation 3. On the other hand, Ca is mainly influenced by local image gradient when γ is large. At the same time, the variation of α for different document images increases when γ is close to 1. Under such circumstance, the power function becomes more sensitive to the global image intensity variation and appropriate weights can be assigned to images with different characteristics.

Data set improves significantly when γ increases to 1. Therefore the proposed method can assign more suitable α to different images when γ is closer to 1. Parameter γ should therefore be set around 1 when the adaptability of the proposed technique is maximized and better and more robust binarization results can be derived from different kinds of degraded document images. Fig. 6 shows the thresholding results when W varies from EW to $4EW$. Generally, a larger local window size will help to reduce the classification error that is often induced by the lack of edge pixels within the local neighbourhood window. In addition, the performance of the. Proposed method becomes stable when the local window size is larger than $2EW$ consistently on the three datasets. W can therefore be set around $2EW$ because a larger local neighborhood window will increase the computational load significantly.

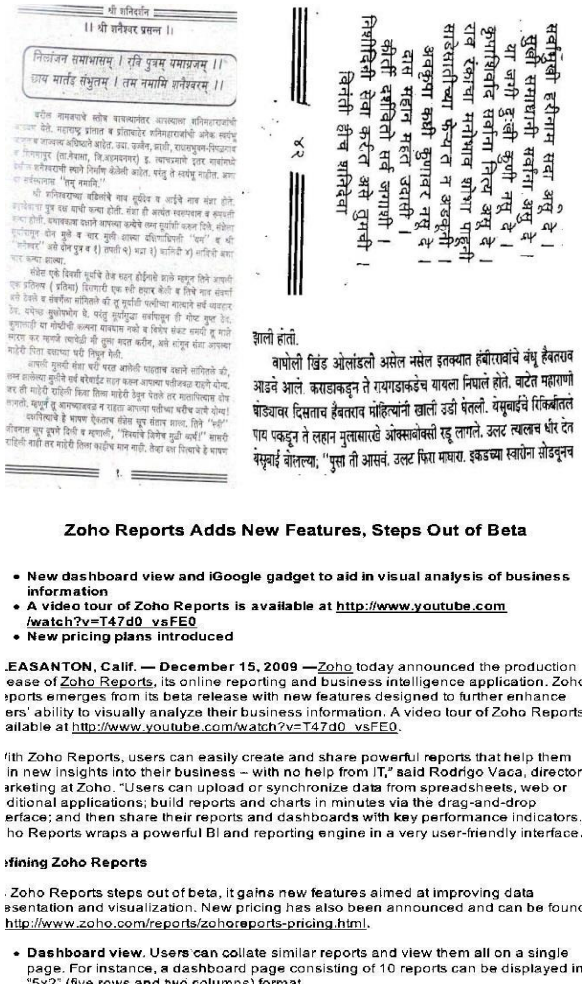

Fig .shows recovered images by proposed method.

V. CONCLUSION

This paper presents an adaptive image contrast based document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The proposed technique is simple and robust, only few parameters are involved. Moreover, it works for different kinds of degraded document images. The proposed technique makes use of the local image contrast that is evaluated based on the local maximum and minimum. The proposed method has been tested on the various datasets.

ACKNOWLEDGMENT

We would like to thank Prof. S. P. Godse for helping us in making this project.

REFERENCES

[1] O. D. Trier and T. Taxt, "Evaluation of binarization methods for document images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 17, no. 3, pp. 312–315, Mar. 1995.

[2] J. Kittler and J. Illingworth, "On threshold selection using clustering criteria," IEEE Trans. Syst., Man, Cybern., vol. 15, no. 5, pp. 652–655, Sep.–Oct. 1985.

[3] 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.

[4] Bolan Su, Shijian Lu, and Chew Lim Tan, Senior Member, IEEE," Robust Document Image Binarization Technique for Degraded Document Images", IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 22, NO. 4, APRIL 2013

[5] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "ICDAR 2009 document image binarization contest (DIBCO 2009)," in Proc. Int. Conf. Document Anal. Recognit., Jul. 2009, pp. 1375–1382

[6] 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.

[7] I. Pratikakis, B. Gatos, and K. Ntirogiannis, "H-DIBCO 2010 handwritten document image binarization competition," in Proc. Int. Conf. Frontiers Handwrit. Recognit., Nov. 2010, pp. 727–732.

[8] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," Int. J. Document Anal. Recognit., vol. 13, no. 4, pp. 303–314, Dec. 2010.

[9] B. Su, S. Lu, and C. L. Tan, "Binarization of historical handwritten document images using local maximum and minimum filter," in Proc. Int. Workshop Document Anal. Syst., Jun. 2010, pp. 159–166.

[10] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," J. Electron. Imag., vol. 13, no. 1, pp. 146–165, Jan. 2004.

AUTHORS

First Author – Prof. S. P. Godse, Computer Science Engineering Department, Pune University, Sinhgad Academy of Engineering, Kondhwa(bk), Dist- Pune-48 , Maharashtra, India., Email: sachin.gds@gmail.com

Second Author – Samadhan Nimbhore, Computer Science Engineering Department, Pune University, Sinhgad Academy of Engineering, Kondhwa(bk), Dist- Pune-48 , Maharashtra, India., Email: samnimbhoresai@gmail.com

Third Author – Sujit Shitole, Computer Science Engineering Department, Pune University, Sinhgad Academy of Engineering, Kondhwa(bk), Dist- Pune-48 , Maharashtra, India., Email: sujit2602@gmail.com

Fourth Author – Dinesh Katke, Computer Science Engineering Department, Pune University, Sinhgad Academy of Engineering, Kondhwa(bk), Dist- Pune-48 , Maharashtra, India., Email: katkedinesh44@gmail.com

Fifth Author – Pradeep Kasar, Computer Science Engineering Department, Pune University, Sinhgad Academy of Engineering, Kondhwa(bk), Dist- Pune-48 , Maharashtra, India., Email: prdpkasar@gmail.com