RESEARCH ARTICLE OPEN ACCESS

Text Extraction and Document Image Binarization Using Sobel Edge Detection

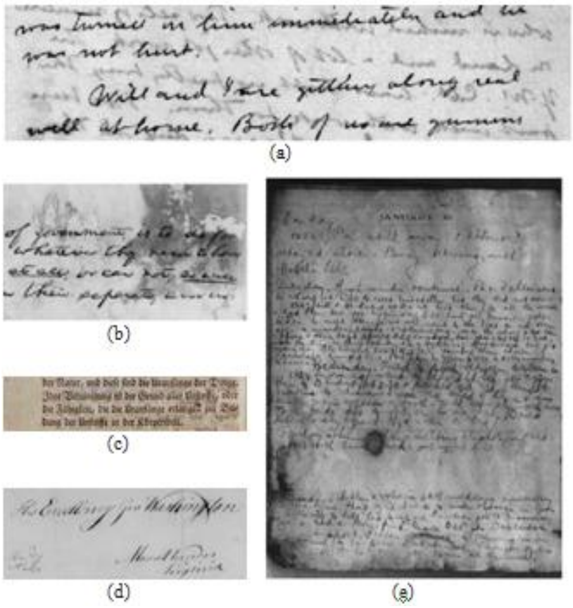
G. Santhanaprabhu[1*], M. E., B. Karthick[2], M.E., P. Srinivasan[3], M.E., R. Vignesh[4], M.E., K. Sureka[5], M.E.,
[1,5]Student/Dept of Applied Electronics
[2,3,4]Assistant Professor/ECE Jayam College of Engineering and Technology, Dharmapuri, Tamilnadu, India.

Abstract
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).So the novel document image binarization technique by using adaptive image contrast. The Adaptive Image Contrast is a combination of the local image contrast and the local image gradient. The proposed technique, an adaptive contrast map is first constructed for an input degraded document image. The contrast map is then binarized and combined with Sobel edge detector .The Sobel edge detection algorithm is the one of the most commonly used methods for edge detection. The document text is further segmented by using local threshold estimation based on the intensities and post processing.

Key Words - Adaptive image contrast, Document analysis, Document image processing, Degraded document image binarization, sobel edge detector.

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.

Fig. 1. Five degraded document image examples (a)-(d) are taken from DIBCO series datasets and (e) is taken from Bickley diary dataset.

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 background. In addition, historical documents are often degraded by the bleed-through as illustrated 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 illustrated in Fig. 1(e). These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge to most state-of-the-art techniques.

The recent Document Image Binarization Contest (DIBCO) [1], [2] held under the framework of the International Conference on Document Analysis and Recognition (ICDAR) 2009 & 2011 and the Handwritten Document Image Binarization Contest (H-DIBCO) [3] held under the framework of the International Conference on Frontiers in Handwritten Recognition show recent efforts on this issue. We participated in the DIBCO 2009 and our background estimation method [4] performs the best among entries of 43 algorithms submitted from 35 international research groups. We also participated in the H-DIBCO 2010 and our local maximum-minimum method [5] was one of the top two winners among 17 submitted algorithms. In the latest DIBCO 2011, our proposed method achieved second best results among 18 submitted algorithms.

This paper presents a document binarization technique that extends our previous local maximum-minimum method [5] 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 adaptive image contrast that combines the local image contrast and the local image gradient adaptively 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 mini-mum algorithm [5]. At the same time, the parameters used in the algorithm can be adaptively estimated.

II. PROPOSED METHOD

This section describes the proposed document image binarization techniques. Given a degraded document image, an adaptive contrast map is first constructed and the text stroke edges are then detected through the combination of the binarized adaptive contrast map and the sobel edge map. 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 gradient needs to be normalized to compensate the image variation within the document background.

$$\mathbf{C(i,j) = I_{max}(i,j) - I_{min}(i,j)} \qquad \mathbf{\ldots\ldots\ldots(1)}$$

where *C(i, j)* denotes the contrast of an image pixel *(i, j)*, Imax*(i, j)* and Imin*(i, j)* denote the maximum and minimum intensities within a local neighborhood windows of *(i, j)*, respectively. If the local contrast *C(i, j)* 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 Imax*(i, j)* and Imin*(i, j)*. 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 [5] by using the local image contrast that is evaluated as follows [41]:

$$\mathbf{C(i,j)} = \frac{\mathbf{I_{max}(i,j) - I_{min}(i,j)}}{\mathbf{I_{max}(i,j) - I_{min}(i,j) + £}} \qquad \mathbf{\ldots\ldots\ldots.(2)}$$

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. In our earlier method [5], 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 image difference 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 and accordingly 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 combine the local image contrast with the local image gradient and derive an adaptive local image contrast as follows:

$$\boldsymbol{C_a (i, j) = \alpha C (i, j) + (1 - \alpha)(I_{max}(i, j) - I_{min}(i, j)).. (3)}$$

where *C (i, j)* denotes the local contrast in Equation 2 and *($I_{max}(i, j) - I_{min}(i, j)$)* refers to the local image gradient that is normalized to [0, 1]. The local windows size is set to 3 empirically. *α* is the weight between local contrast and local gradient that is controlled based on the document image statistical information. Ideally, the image contrast will be assigned with a high weight (i.e. large *α*) when the document image has signiÞcant intensity variation. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. The proposed binarization technique relies more on image gradient and avoid the over normalization problem of our previous method [5]. So that the proposed binarization technique depends more on the local image contrast that can capture the intensity variation well and hence produce good results. Otherwise, the local image gradient will be assigned with a high weight. The proposed binarization technique relies more on image gradient and avoid the over normalization problem.

The mappings from document image intensity variation to α by a power function as follows:

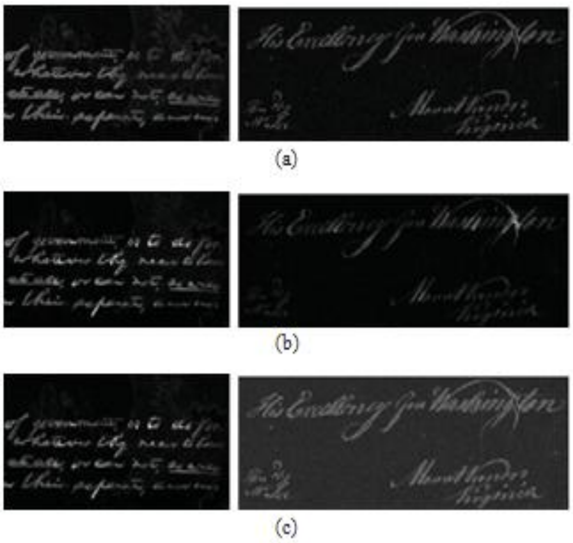
$$\boldsymbol{\alpha} = \frac{\mathbf{std}^{\gamma}}{\mathbf{128}} \qquad \textbf{...(4)}$$

Where 'Std' denotes the document image intensity standard deviation, and γ is a pre-defined parameter. The power function has a nice property in that it monotonically and smoothly increases from 0 to 1 and its shape can be easily controlled by different γ. γ can be selected from [0,∞], where the power function becomes a linear function when γ = 1.

Therefore, the local image gradient will play the major role in Equation 3 when *γ* is large and the local image contrast will play the major role when *γ* is small.

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 [43], local image contrast [5] and our proposed method in Equation 3, respectively.

For the sample document with a complex document back-ground 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 intensity variation within the document background but large intensity.



(a)

(b)

(c)

Fig. 2. Contrast Images constructed using (a) local image gradient [42], (b) local image contrast [5], and (c) our proposed method of the sample document images in Fig. 1(b) and (d), respectively.

Variation 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 strokes as shown in Fig. 2(a). As a comparison, the adaptive combination of the local image contrast and the local image gradient in Equation 3 can 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 in Fig. 1(b).

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 [5], where the adaptive 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 are evaluated 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. The binary map can be further improved through the combination with the edges by Sobel edge detector, because Sobel edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image. In addition, sobel edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as

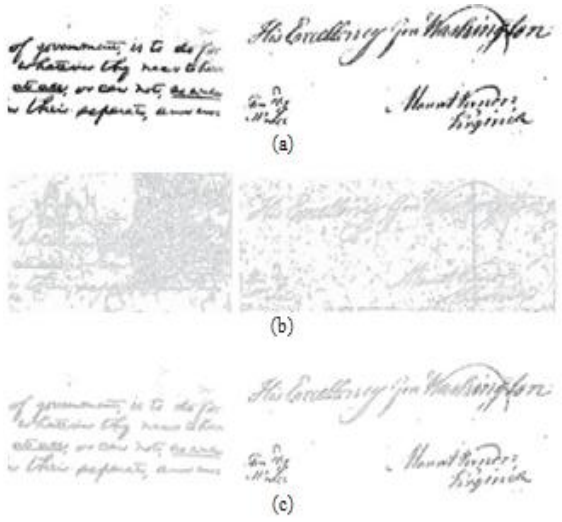


Fig. 3. (a) Binary contrast maps, (b) sobel edge maps, and their (c) combined edge maps of the sample document images in Fig. 1(b) and (d), respectively.

shading [44]. It should be noted that sobel edge detector by itself often extracts a large amount of non-stroke edges as illustrated in Fig. 3(b) without tuning the parameter manually. In the combined map, we keep only pixels that appear within both the high contrast image pixel map and sobel edge map. The combination helps to extract the text stroke edge pixels accurately as shown in Fig. 3(c).

C. Local Threshold Estimation

The text can then be extracted from the document back-ground pixels once the high contrast stroke edge pixels are detected properly. Two characteristics can be observed from different kinds of document images [5]: 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 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, *E W* , which can be estimated from the detected stroke edges [shown in Fig. 3(b)].

III. EXPERIMENTS AND DISCUSSION

A few experiments are designed to demonstrate the effec-tiveness and robustness of our proposed method. We Þrst analyze the performance of the proposed technique on pub-lic datasets for parameter selection. The proposed technique is then tested and compared with state-of-the-art methods over on three well-known competition datasets: DIBCO 2009 dataset [1], H-DIBCO 2010 dataset [3], and DIBCO 2011 dataset [2]. Finally, the proposed technique is further evaluated over a very challenging Bickley diary dataset [37].

The binarization performance are evaluated by using F-Measure, pseudo F-Measure, Peak Signal to Noise Ratio (PSNR), Negative Rate Metric (NRM), Misclassification Penalty Metric (MPM), Distance Reciprocal Distortion (DRD) and rank score that are adopted from DIBCO 2009, H-DIBCO 2010 and DIBCO 2011 [1],[3]. Due to lack of ground truth data in some datasets, no all of the metrics are applied on every images.

Table I
Evaluation Results of The Dataset of DIBCO 2009

Methods	F-Measure(%)	PSNR	NRM (X10^{-2})	MPM(X10^{-3})	Rank Score
OTSU[12]	78.72	15.34	5.77	13.3	196
SAUV[18]	85.41	16.39	6.94	3.2	177
NIBL[19]	55.82	9.89	16.4	61.5	251
BERN[14]	52.48	8.89	14.29	113.8	313
GATO[21]	85.25	16.5	10	0.7	176
LMM[5]	91.06	18.5	7	0.3	126
BE[4]	91.24	18.6	4.31	0.55	101
Proposed Method	93.5	19.85	3.7	0.4	100

Table II
Evaluation Results of The Dataset of DIBCO 2010

Methods	F-Measure(%)	PSNR	NRM (X10^{-2})	MPM(X10^{-3})	Rank Score
OTSU[12]	85.27	17.51	9.77	1.35	188
SAUV[18]	75.3	15.9	16.31	1.96	225
NIBL[19]	74.1	15.73	19.06	1.06	263
BERN[14]	41.3	8.57	21.18	115.9	244
GATO[21]	71.99	15.12	21.8	0,41	284
LMM[5]	85.49	17.83	11.46	0.37	216
BE[4]	86.41	18.14	9.06	1.11	202
Proposed Method	92.03	20.12	6.14	0.25	178

Table III
Evaluation Results of The Dataset of DIBCO 2011

Methods	F-Measure(%)	PSNR	DRD	MPM($X10^{-3}$)	Rank Score
OTSU[12]	82.22	15.77	8.72	15.64	412
SAUV[18]	82,54	15.78	8.09	9.20	403
NIBL[19]	68.52	12.76	28.31	26.3	362
BERN[14]	47.28	7.92	82.28	136.54	664
GATO[21]	82.11	16.04	5.42	7.13	353
LMM[5]	85.56	16.75	6.02	6.42	516
BE[4]	81,67	15.59	11.24	11.40	376
Proposed Method	87.8	17.56	4.84	5.1	307

is also shown in Table III. Although it does not reach the lowest ranking score, our proposed technique produces good results on all the testing images, which is reflected on the high F-measure score.

D. Discussion

As described in previous sections, the proposed method involves several parameters, most of which can be automatically estimated based on the statistics of the input document image. This makes our proposed technique more stable and easy-to-use for document images with different kinds of degradation. The superior performance of our proposed method can be explained by several factors. First, the proposed method combines the local image contrast and the local image gradient that help to suppress the background variation and avoid the over-normalization of document images with less variation. Second, the combination with edge map helps to produce a precise text stroke edge map. Third, the proposed method makes use of the text stroke edges that help to extract the foreground text from the document background accurately. But the performance on Bickley diary dataset and some images of DIBCO contests still needs to be improved, we will explore it in future.

IV. CONCLUSION

This paper presents an adaptive image contrast based docu-ment 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 pro-posed 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. Experiments show that the proposed method outperforms most reported document binarization methods in term of the F-measure, pseudo F-measure, PSNR, NRM, MPM and DRD.

REFERENCES

[1] B. Gatos, K. Ntirogiannis, and I. Pratikakis, ÒICDAR 2009 document image binarization contest (DIBCO 2009),Ó in *Proc. Int. Conf. Document Anal. Recognit.*, Jul. 2009, pp. 1375Ð1382.

[2] I. Pratikakis, B. Gatos, and K. Ntirogiannis, ÒICDAR 2011 document image binarization contest (DIBCO 2011),Ó in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 1506Ð1510.

[3] I. Pratikakis, B. Gatos, and K. Ntirogiannis, ÒH-DIBCO 2010 hand-written document image binarization competition,Ó in *Proc. Int. Conf. Frontiers Handwrit. Recognit.*, Nov. 2010, pp. 727Ð732.

[4] S. Lu, B. Su, and C. L. Tan, ÒDocument image binarization using back-ground estimation and stroke edges,Ó *Int. J. Document Anal. Recognit.*, vol. 13, no. 4, pp. 303Ð314, Dec. 2010.

[5] S. Lu, B. Su, and C. L. Tan, ÒDocument image binarization using back-ground estimation and stroke edges,Ó *Int. J. Document Anal. Recognit.*, vol. 13, no. 4, pp. 303Ð314, Dec. 2010. B. Su, S. Lu, and C. L. Tan, ÒBinarization of historical handwritten document images using local maximum and minimum Þlter,Ó in *Proc. Int. Workshop Document Anal. Syst.*, Jun. 2010, pp. 159Ð166.

[6] G. Leedham, C. Yan, K. Takru, J. Hadi, N. Tan, and L. Mian, ÒCompar-ison of some thresholding algorithms for text/background segmentation in difÞcult document images,Ó in *Proc. Int. Conf. Document Anal. Recognit.*, vol. 13. 2003, pp. 859Ð864.

[7] M. Sezgin and B. Sankur, ÒSurvey over image thresholding techniques and quantitative performance evaluation,Ó *J. Electron. Imag.*, vol. 13, no. 1, pp. 146Ð165, Jan. 2004.

[8] O. D. Trier and A. K. Jain, ÒGoal-directed evaluation of binarization methods,Ó *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 12, pp. 1191Ð1201, Dec. 1995.

[9] O. D. Trier and T. Taxt, ÒEvaluation of binarization methods for document images,Ó *IEEE Trans. Pattern Anal. Mach.*

Intell., vol. 17, no. 3, pp. 312Ð315, Mar. 1995.

[10] A. Brink, ÒThresholding of digital images using two-dimensional entropies,Ó *Pattern Recognit.*, vol. 25, no. 8, pp. 803Ð808, 1992.

[11] J. Kittler and J. Illingworth, ÒOn threshold selection using clustering criteria,Ó *IEEE Trans. Syst., Man, Cybern.*, vol. 15, no. 5.pp. 652Ð655, Sep.ÐOct. 1985.

[12] N. Otsu, ÒA threshold selection method from gray level histogram,Ó *IEEE Trans. Syst., Man, Cybern.*, vol. 19, no. 1, pp. 62Ð66, Jan. 1979.

[13] N. Papamarkos and B. Gatos, “A new approach for multithreshold selection,” *Comput. Vis. Graph. Image Process.*, vol. 56, no. 5, pp. 357–370, 1994.

[14] J. Bernsen, “Dynamic thresholding of gray-level images,” in *Proc. Int. Conf. Pattern Recognit.*, Oct. 1986, pp. 1251–1255

[15] 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.

[16] 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.

[17] J. Parker, C. Jennings, and A. Salkauskas, “Thresholding using an illumination model,” in *Proc. Int. Conf. Doc. Anal. Recognit.*, Oct. 1993, pp. 270–273.

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

[19] W. Niblack, *An Introduction to Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1986

[20] J.-D. Yang, Y.-S. Chen, and W.-H. Hsu, “Adaptive thresholding algorithm and its hardware implementation,” *Pattern Recognit. Lett.*, vol. 15, no. 2, pp. 141–150, 1994.

[21] Y. Liu and S. Srihari, “Document image binarization based on texture features,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 540–544, May 1997.

[22] M. Cheriet, J. N. Said, and C. Y. Suen, “A recursive thresholding technique for image segmentation,” in *Proc. IEEE Trans. Image Process.*, Jun. 1998, pp. 918–921.

[23] B. Gatos, I. Pratikakis, and S. Perantonis, “Adaptive degraded document image binarization,” *Pattern Recognit.*, vol. 39, no. 3, pp. 317–327, 2006.

SANTHANAPRABHU. G has obtained his B.E degree in Electronics & Communication Engineering from Annai Teresa College of Engineering, Villupuram in 2012. Currently he is doing M.E in Applied Electronics at Jayam College of Engineering and Technology, Dharmapuri. Presently he is involving in developing a Text Extraction and Document Image Binarization Using Sobel Edge Detection. He has published more than two research papers in national and international conferences. His special areas of interest are Image Processing, Digital Signal Processing, Signals and System and Digital Electronics.

KARTHICK.B has obtained his B.E degree in Jayam College of Engineering And Technology, Dharmapuri. He received his M.E degree from Jayam College of Engineering & Technology, Dharmapuri. He published more than five research papers in various National and International conferences. He worked as a Developer in the software field and then at present he is working as Assistant Professor in Department of Electronics And Communication Engineering in Jayam College of Engineering And Technology, Dharmapuri. He has participated in various National level workshops and Seminars at various colleges.

SUREKA.K has obtained her BE degree in Electronics and Instrumentation Engineering from Velammal Engineering College, Chennai in 2011. Currently she is doing her ME in Applied Electronics at Jayam College of Engineering and Technology, Dharmapuri. Presently she is involving in developing a automated method for identification and classification of retinal blood vessels to identify the diseases in retina. She has published more than two research papers in national and international conferences. Her special areas of interest are Image processing, Control system and Measurements & Instruments.

VIGNESH.R has obtained his BE degree in Electronics and Communication Engineering from Jayam College of Engineering and Technology. He received his ME in Applied Electronics from Jayam College of Engineering and Technology. He published more than four research papers in various national and international conferences/journals. At present he is working as Assistant Professor in the department of Electronics and Communication Engineering in Jayam College of Engineering and Technology, Dharmapuri. He has participated in various national level workshops and seminars at various colleges.

SRINIVASAN.P has obtained his M.E degree in K.S.R College of Technology, Thiruchengode. He received his B.E degree from Maha College of Engineering, Salem. He published more than 3 papers in various National and International Conferences. International Conferences. And he is working as Assistant Professor in Department Of Electronics And Communication Engineering at Jayam College of Engineering And Technology, Dharmapuri. He has participated in various national workshops and seminars at various colleges. and also he is an active membership of IEEE.