SCENE TEXT DETECTION IN VIDEO IMAGES BY USING MULTI-SPECTRAL FUSION BASED APPROACH

Nagasudha D[1], Madhaveelatha Y[2]
[1]Research Scholar, Dept of ECE JNTUCEH, Hyderabad, Andhra Pradesh, India.
[2]Dept of ECE MRECW Hyderabad, Andhra Pradesh, India

ABSTRACT

Scene text detection from video as well as natural scene images is challenging due to the variations in background, contrast, text type, font type, font size, and so on. Besides, arbitrary orientations of texts with multi-scripts add more complexity to the problem. The proposed approach introduces a new idea of convolving Laplacian with wavelet sub-bands at different levels in the frequency domain for enhancing low resolution text pixels. Then, the results obtained from different sub-bands (spectral) are fused for detecting candidate text pixels. We explore maxima stable extreme regions along with stroke width transform for detecting candidate text regions. Text alignment is done based on the distance between the nearest neighbour clusters of candidate text regions. In addition, the approach presents a new symmetry driven nearest neighbour for restoring full text lines.

KEYWORDS: *Text detection, character recognition, text extraction, scene image, text alignment, stroke width transform*

I. INTRODUCTION

With the advances in multimedia technology, content based indexing and text extraction in images plays a major role. It contains different contents in it such as caption, text, scene etc. Among all the contents of the image text is found to be one of the most important features to understand the image contents. Text in images can be used as indexing purpose, document processing [1,2],video content summary[3-5],video retrieval[6]and video understanding[7].The text information can be extracted in two stages: Text detection and text recognition. Text detection detects the text regions as external regions of an image and text recognition retrieves the text information from these external regions[8].Text extraction from images have many useful applications in document analysis, detection of vehicle license plate, diagrams, maps, charts etc. Key word based image searching, content based retrieval, name plates street signs text based video indexing and document retrieving etc. Images can be broadly classified into document images, caption text images and Scene text images. A document image usually contain text and few graphical components. These are acquired by scanning journal, printed document, degraded document images, handwritten document and book cover etc. Caption text is also known as overlay text which is artificially superimposed on the video/image at the time of editing and it describes the content of the image/video .Scene text appears within the scene which is then captured by the recording device i.e. text is present in the scene when the image or video is shot.

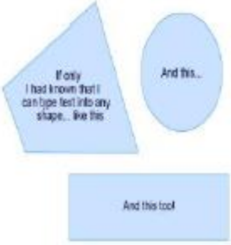

If only I had known that I can type text into any shape... like this
And this...
And this too!

Fig 1: document text image

Fig 2: colored text image

Fig 3: caption text image

Fig 4: scene text image

Retrieving the contents from images is very challenging because when extracting text with variation in fonts, size, color, alignment, orientation, illumination and back ground .Problem of text extraction is very difficult because of these deviations.

II. RELATED WORK

For text detection in images a large number of approaches have been developed, which are classified into two categories: Region base d approaches and texture based approaches.

1. Region based approaches: Region based methods use the properties of the color or gray scale in a text region or their differences with the corresponding properties of the back ground. This method uses a bottom up approach by grouping small components into larger components until all regions are identified in the image. A geometric analysis is needed to merge text components using spatial arrangement of the components so as filter out non text components and mark the boundaries of the text regions.

Shiva kumara et al[9] proposed an edge based technique for text detection in images with text present in the horizontal direction. The frame was segmented into 16 non over lapping blocks. Mean and median filter and edge analysis was used to identify the candidate text blocks. Using block growing method, the complete text block was obtained. Finally based on the vertical and horizontal bar feature , the true text regions are detected. In shiva kumara et al[10],filters and edge analysis were used for initial text detection. The straightness and cursiveness edge features were used for false positive elimination.

Text detection using a cascade Adaboost classifier with HOG and multi scale local binary pattern feature was proposed by Pan et al[10].Text localization was done using window grouping technique. Within each located text line,local binarization is done to extract candidate CCs and non text CCs are filtered using Markov Random field model and MLP in order to get the final text line.

2. Texture based approaches: Texture based approaches use distinctive properties of the text that separate them from the background. The techniques based on Gabor filters, Wavelet, FFT, spatial variance etc. can be used to detect the texture properties of a text region in an image. These approaches are used in complex background and expensive.

Connected component based approaches are fast and good for images which have high contrast texts and plain background just like methods in the document image analysis. To improve the performance of text detection mixture of texture based and region based approaches are used.

Zhaon et al. [11] used wavelet transform and sparse representation with discriminative dictionaries for text detection. Shivakumara et al [12,13]also used haar wavelet in both the works. In [14] they also used color features along with Wavelet-Laplacian method to detect text. Shivakumara et al [15] used wavelet median moment feature with k means clustering to obtain text pixels.

The wavelet transform is used in different fields such as image processing, signal processing, video compression and so on. In the frequency domain an image is decomposed into different components and computed with low pass and high pass filters and by one dimensional discrete waveform transform (DWT) an image is divided into two parts that is coarse and detailed information. The texture features are extracted from discrete wavelet transform.

III. PROPOSED APPROACH

The proposed approach consists of four steps. In the first step, we propose a idea of convolving Laplacian with wavelet sub-bands at different levels in the frequency domain to enhance text pixels through a fusion concept, which results in multi-spectral fusion. The fused images are subjected to fuzzy k-means clustering to classify the Candidate Text Pixels (CTP). Since video and natural images are complex and text has large variations in font, font size, colour, orientation, etc., conventional methods such as the non-fuzzy k-means clustering, the Max-Min clustering method and the adaptive thresholding technique do not work well. The reason is that the conventional k-means clustering produces inconsistent results because of random guess selection for clustering, the Max-Min clustering method is not accurate enough due to the lack of discriminations between text and non-text values, and the adaptive thresholding technique does not give good results because it is hard to decide threshold values dynamically for different situations.

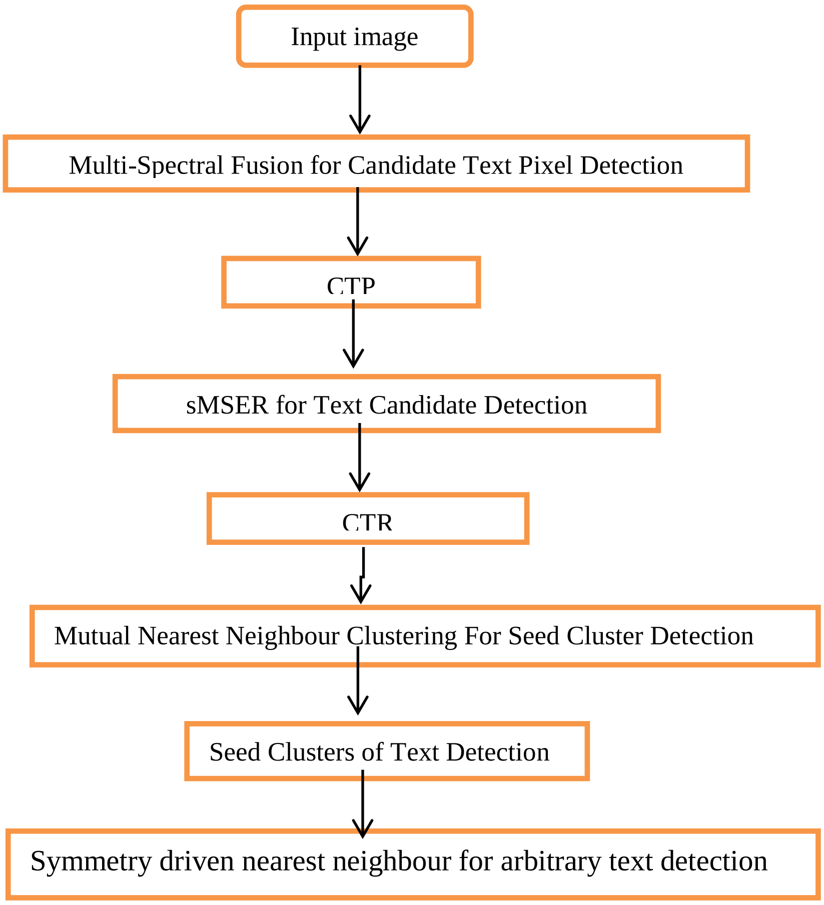

Input image
Multi-Spectral Fusion for Candidate Text Pixel Detection
CTP
sMSER for Text Candidate Detection
CTR
Mutual Nearest Neighbour Clustering For Seed Cluster Detection
Seed Clusters of Text Detection
Symmetry driven nearest neighbour for arbitrary text detection

Fig5: Flow chart of proposed approach

We prefer to use fuzzy K-means clustering, which classifies text and non-text pixels based on the probability of either text or non-text pixels with a membership function but not direct values. It is noticed that Laplacian helps in distinguishing text pixels as it gives high positive and negative values for text pixels and low values for non-text pixels. In the same way, Fourier in the frequency domain provides high coefficients for text pixels and low coefficients for non-text pixels. Therefore, we propose to combine Laplacian with wavelet sub-bands for better enhancement because wavelet sub-bands have both low and high pass filters, while Fourier behaves either as a low pass or a high pass filter to eliminate noisy pixels in text detection. To tackle the problems of multi-size and multi-contrast texts, we combine Laplacian with wavelet sub-bands at different levels through fusion.

For candidate text pixels, we explore Maximally Stable External Regions (MSER) to group the candidate text pixels into text regions. MSER has been successfully used for classifying text and non-text components in the past but in this work, we propose to modify MSER as sMSER along with Stroke Width Transform (SWT) to cluster candidate text pixels as text regions, which we call Candidate Text Regions (CTR). Later we introduce Mutual Nearest Neighbour Clustering (MNN) for the CTR image to group candidate text regions that belong to the same text line. The proposed approach compares the geometrical properties of CTR before grouping them into a single one. The output of this step is said to be seed clusters that represent a text line. Sometimes, this step may eliminate text components due to mismatching. So in the further step, we present a new symmetry driven nearest neighbour process for each seed cluster to restore missing text clusters, which results in a full text line of any orientation. The symmetry is defined as the distance between two nearest neighbouring components.

Discrete wavelet transform (DWT):

In numerical analysis and functional analysis, a discrete wavelet transform (DWT) is any wavelet transform for which the wavelets are discretely sampled. As with other wavelet transforms, a key advantage it has over Fourier transforms is temporal resolution: it captures both frequency and location information (location in time).

Types of DWT :

- Haar wavelets
- Daubechies wavelets
- The dual-tree complex wavelet transform (DℂWT)
- undecimated wavelet transform

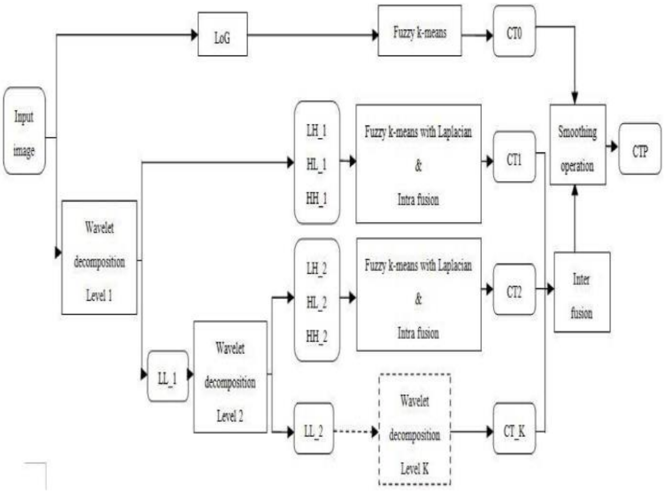

LoG
Fuzzy k-means
CT0
Input image
LH_1
HL_1
HH_1
Fuzzy k-means with Laplacian & Intra fusion
CT1
Smoothing operation
CTP
Wavelet decomposition Level 1
LH_2
HL_2
HH_2
Fuzzy k-means with Laplacian & Intra fusion
CT2
Inter fusion
LL_1
Wavelet decomposition Level 2
LL_2
Wavelet decomposition Level K
CT_K

Fig6: Flow diagram of Laplacian-Wavelet sub-bands at different levels for candidate text pixels.

Properties of DWT: The Haar DWT illustrates the desirable properties of wavelets in general. First, it can be performed in $O(n)$ operations.

- It captures not only a notion of the frequency content of the input, by examining it at different scales, but also temporal content, i.e. the times at which these frequencies occur.
- Combined, these two properties make the Fast wavelet transform (FWT) an alternative to the conventional fast Fourier transform (FFT).
- Due to the rate-change operators in the filter bank, the discrete WT is not time-invariant but actually very sensitive to the alignment of the signal in time.

Fuzzy k-means clustering:

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space and it uses cluster centres to model the data. *k*-means clustering tends to find clusters of comparable spatial extent.

Given a set of observations ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into $k\ (\leq n)$ sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K centre). In other words, its objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

sMSER: CTP of the input frame still contains some non-text candidate pixels due to background complexity. To eliminate such non-text candidate text pixels, we propose to explore MSER(Mean Square Error) along with the stroke width property because it is true that MSER is a successful approach for studying the characteristics of connected components, especially for text detection and recognition in video and natural scene images. However, the use of MSER for the whole image is not advisable because it sometimes largely misclassifies non-text components as text ones due to complex background and low resolution. This leads to poor performance. To avoid this problem, we apply MSER for only the candidate text pixels. Since CTP is a binary image and MSER requires gray information, the proposed approach extracts gray values in the input image corresponding to the candidate text pixels in the CTP image. Even then, conventional MSER still gives poor results because of background colour variations within character components. Therefore, we perform a mean filter operation over the gray values of the input image that correspond to the edges of the CTP in the CTP image. This operation smoothens edges irrespective of multiple colours in a single component. In other words, it is a simple averaging filter, which performs 3×3 mask operations for every pixel in the CTP image. The output of smoothing is considered asthe input for MSER, which we call sMSER.

IV. Result And Discussion

We evaluated this proposed approach on the images for near about 20 scene images and about 20 images for mixed document images. All the images of the natural scene and document images were captured using a digital camera under different luminosity conditions. It is evident from the results that the performance of the proposed algorithm is satisfactory on both types of images.

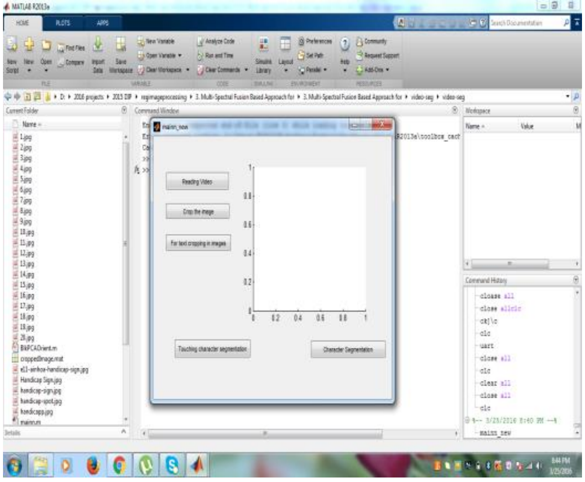


Fig 7 (a)

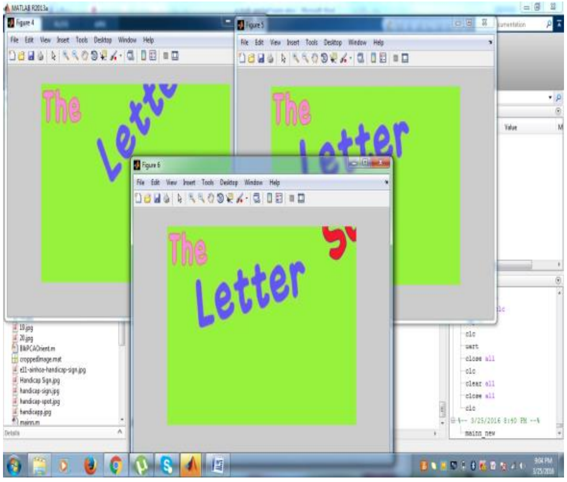


Fig 7 (b)

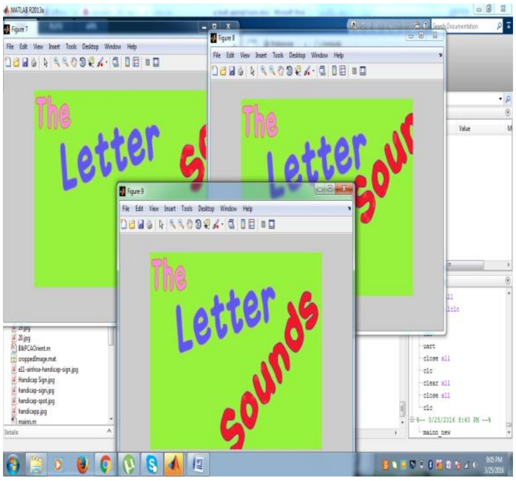


Fig 7 (c)

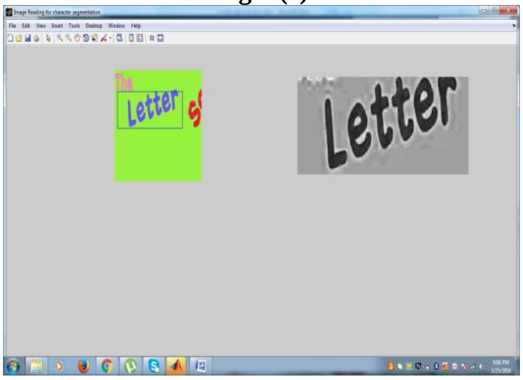
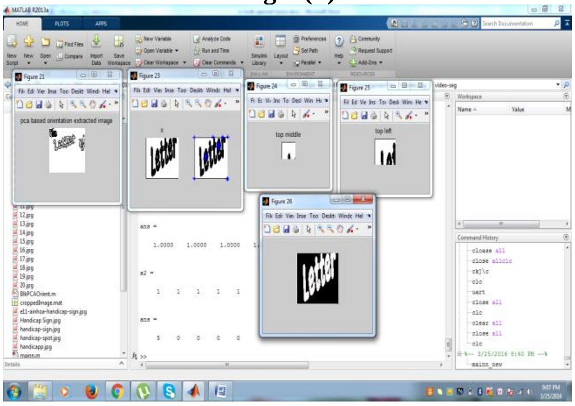


Fig 7 (d)



```
ans =

   1.0000    1.0000    1.0000

a2 =

     1     1     1     1     1

ans =

     5     0     0     0     0
```

Fig 7 (e)

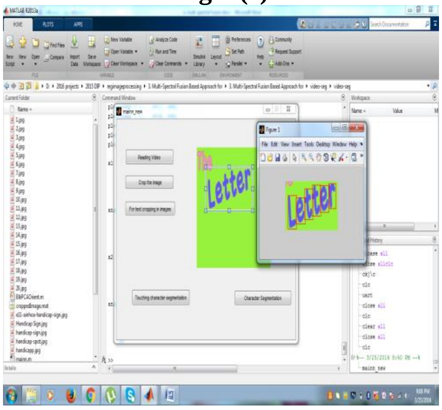


Fig 7 (f)

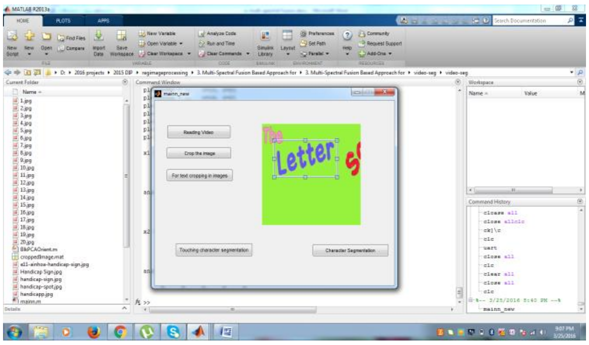


Fig 7 (g)

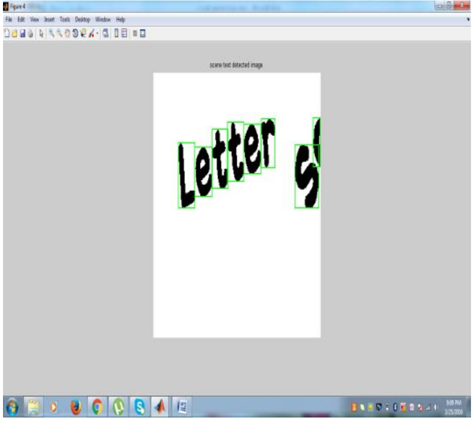


Fig 7 (h)

Fig 7 a)input frame b) Text extraction c)eMSeR natural scene data d)cropped text image e-f-g)proposed approach results h)character segmentation

$$\text{Precision} = \frac{\text{Correctly localized texts}}{\text{Correctly localized texts+False positives}} \times 100$$

$$\text{Recall} = \frac{\text{Correctly localized texts}}{\text{Correctly localized texts+False Negatives}} \times 100$$

V. CONCLUSION

Text localization in a natural image with complex background is a difficult, challenging and important problem. In this paper a novel method is proposed for text detection and localization using wavelet based features of edge images corresponding to input natural scene images. The proposed method uses Discrete wavelet Transform, Fuzzy K clustering and laplacian for segmenting and classifying the text regions. The proposed method has yielded the precision and recall rates of 78.4% and 88.24%respectively.The localized text regions may contain multiple lines of text of different scripts that need to be segmented for character extraction, which will be considered in future work.

REFERENCES

[1] jain,K., and Yu, B., “document representation and its application to page decomposition,” IEEE Trans Pattern And Machine Intell.,vol.20 pp. 294-308,Mar. 1998.

[2] Jain, K., and Yu, B., “Automatic Text Location in Images and Video Frames”,Pattern recognition,31(12) 2055-2076,1998.

[3] Jain, K., and Zhong, Y., “Page Segmentation using Texture Analysis”, Pattern Recognition , 29 (5) 743-770, 1996.

[4]Chitrakala Gopalan and D.Manjula,”Contourlet Based Approach for Text Identification and Extraction from Heterogeneous Textural images”, International Journal of Computer Science and Engineering 2(4) pp.202-211,2008

[5]Crandali, D., Antani, S., and Kasturi, R. “Robust Detection of Stylized Text Events in Digital Video”, Proceedings of International Conference on Document Analysis and Recognition, pp. 865-869,2001.

[6]Serra,J.,"Image Analysis and Mathematical Morphology" New York,Academic,1982
[7] Fa,K.C., Wang,L.S and Wang, Y.K.," Page segmentation and identification for intelligent signal Processing " Signal Process.,vol,45,pp.329-346,1995.
[8]Kim,H.K.," Efficient Automatic Text Location Method and content Based Indexing and structing of video Database", Journal of Video database", Journal of visual Communication and image Representation 7(4) 336-344,1996.
[9]p Shivakumara, W Huang, C.L Tan, "Video text detection based on filters and edge features"ICME,2009,pp 514-517
[10]Yi-Feng Pan,X Hou, and C.L Liu, "A robust system to detect and localize texts in Natural Scene images,"DAS,2008 pp35-42
[11]M.Zhao ,S.Li and J Kwok," Text detection in images using sparse representation with discriminative dictionaries", Image and Vision Computing vol28,2010,pp 1590-1599.
[12]P.shivakumara, T.Q Phan, and C.L Tan ,"A Robust Wavelet transform based technique for video text detection"ICDAR,2009,pp1285-1289.
[13]P.Shivakumara, T.Q Phan,C.L Tan "New Wavelet and color features for text detection in Video",ICPR,2010,pp 3996-3999.
[14] P Shivakumara,A Dutta,C.L Tan,U.Pal, "A new Wavelet Median –Moment based method for Multi-oriented Video Text Detection,"DAS,2010,pp 279-286.
[15]Shilpi Rani, Rakesh kumar Yadav" An efficient method for text extraction from color images" IJCA Volume 96-No,13 June 2014.

AUTHORS BIOGRAPHY

D. Naga Sudha obtained her Bachelor's degree in Electronics and Communication Engineering from Nagarjuna University and M.Tech degree in Digital systems and computer Electronics from JNTU Anathapur. Currently she is working as an Assistant Professor in Electronics and communication engineering in JNTUHCEJ.. Her research interest includes document image processing, pattern recognition and pattern classification. She is a life member of Indian Society for Technical Education. She has published more than 10 papers in International Journals, International conferences and national conferences.

Y. Madhaveelatha received BTech. (Electronics and Communication Engineering) degree and M.Tech from the JNTUH and Ph.D. degree from JNTUH in 2009. She has Total of 14 years of experience at various levels of Academic and Administrative Positions. She has published more than 50 papers in various International journals, International conferences and National conferences. He is a member of IEEE, Life member of ISTE, Fellow member of IETE and IMAP. Her current research interests are Digital Image Processing, Document Image Processing, Digital Signal Processing, Wavelet transform, Bio-medical signal processing, wireless communication networks.