A Shape Feature based Identification of a Complex Document

Shridevi Soma
Dept. of Comp. Sc. & Engg.
P.D.A College of Engineering
Gulbarga

B.V. Dhandra
Dept. of Computer Science
Gulbarga University
Gulbarga

ABSTRACT
Identification of ownership of any complex document is a challenging task in the field of Document Image Processing. There are many ways in which the constituent parts of the whole document image is used to know the ownership. These constituent parts might be seal, logo, signature, letter number, name of the organization written with different font type and size, In this paper a system is devised to identify the document using Reference Number part of the letter, which is normally found in any official letters, normally it contains the information such as name of the organization, name of the department, academic year and letter number. The segmented part of this reference number is used for defining the feature set that contains total of 17 features out of which 10 are shape features and 07 are Hu's moment invariant features. The Support Vector Machine classifier with RBF kernel is used for pattern matching. The proposed algorithm is experimented on a data set of letters from Gulbarga University Gulbarga. The experimental results have shown the average recognition accuracy of 85.42%.

General Terms
Document Image Processing, Pattern Recognition

Keywords
Connected Component Labeling, Moment Invariant, SVM.

1. INTRODUCTION
Understanding and Analysing the document image for various applications is the objective of the research carried out in the document image processing. Document Image analysis can be categorized as 'Textual Processing', where the region of interest will be the text component (either printed or written) of the document image and 'Graphics Processing' where the region of interest will be image objects like graphs, pictures, stamps, logos, etc. One of the challenging task in document Image processing is to identify the ownership of the document. This identification may be carried out by extracting the either textual image or graphical image present in a complex document image. Segmentation of the complex document and categorizing the segmented image is also one of the challenging research area. In the proposed work an official documents image of Gulbarga University, Gulbarga are considered as a Data set, and every document is identified for name of the department from that letter is received. An area based segmentation is used for selecting the region of interest that is name of the department present in the Reference number of every official letter image.

Exhaustive amount of work has been carried out on complex document image analysis and recognition. [1] Sachin Grover et. al. proposed a system based on edge features to extract text from the colour document images. A 3×3 Sobel operator was used to detect the sharp edges of the text, the results were prominent for high sensitivity and low false alarm rate, but the shortcoming with this method was when the gradient of intensities of text and image are quite similar. Yen-Lin Chen [2] developed a system that decomposes a document image into distinct object planes to separate homogeneous objects such as textual region, non-textual region and background texture. To provide effectiveness on text removal and inpainting process, the author used an adaptive inpainting neighborhood adjustment scheme to obtain textual region of the text lines to produce a clear non-text restored background image. An algorithm has been proposed by G. Rama Mohan Babu et. al. for extraction of text from the heterogeneous images, authors used morphological operations, non linear filter and thresholding for extraction of the edges, from the threshold text candidate regions were obtained and then an unique labeling is done for each text candidate. The non-text region was eliminated using variance operation. [4] Lukas Neumann and Jiri Matas developed a method for text localization and recognition in real world images, their work includes novel features such as : a departure from a strict feed-forward pipeline that is replaced by a hypothesis-verification frame-work simultaneously processing multiple text line hypothesis; use of synthetic fonts to train the algorithm eliminating the need for time-consuming acquisition and labeling of real-world training data and the use of MSERs which provides robustness to geometric and illumination conditions. Performance of the proposed method was tested on ICDAR 2003 dataset and observed optimal results. Most frequent problems of the proposed method in text localization are individual letters were not being detected as MSERs in the projections used, invalid text line formation or invalid word breaking. [7,8,12] research has been carried out on basic building blocks of document image processing system which modifies pictures to improve them, extract information and change their structure. Morphology based approach was much commonly used by many developers for the extraction of the text from the document images [13,14]. From the discussion it is clear that the recognition of document is a challenging problem for document authenticity, indexing, retrieval etc.

From the works carried out from the past it is observed that recognition of official document using spatial features like morphological, geometric and structural features suitable for official document of limited to only extraction of the text rather than recognition based on extracted text from the document image. Hence, proposed work in this paper identifies the ownership of the document based on letter number, which is a segmented part of the official document.

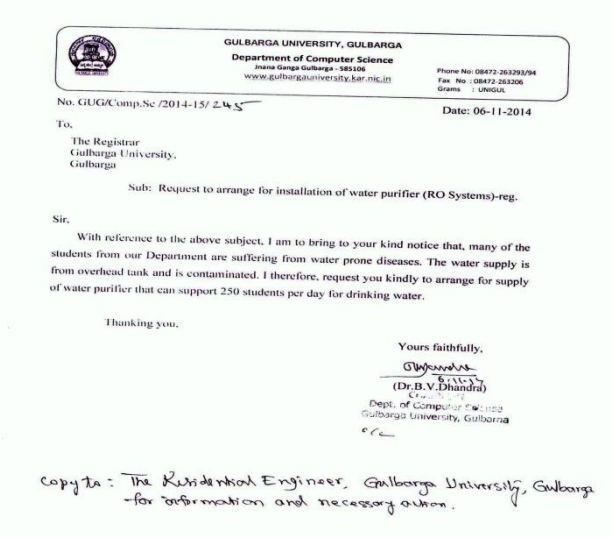
The rest of this paper is organized as follows. Section 2 deals with Data collection and Preprocessing, details of Feature extraction method and Algorithm for Document Identification. The Experimental Procedure and Discussion is

provided in Section 3 and Section 4 summarizes the paper in the form of Conclusion and Future scope.

2. DOCUMENT IDENTIFICATION

2.1 Data Collection

The proposed algorithm for identification of document image using segmented letter number is evaluated on the image dataset of Official letters of Gulbarga University, Gulbarga. The hard copy of the various printed complex documents containing Letter number are collected. These documents are then scanned by HP Scanjet G2410 with 300×300 DPI and segmented manually and stored in a 'tiff' file format. These segmented images may be colour or binary images and contain noise due to quality of scanner and printer, hence they are made noise free using morphological operations and then colour conversion from RGB to Grey Scale is carried out. The noise free grey scale images are then used for feature extraction. A Total of 13 classes of official letters from different department are considered for algorithm evaluation. From total official letter images each class 70% of the images were trained and 100% of the images were tested against trained images. Figure 1 shows the sample of official document.



GULBARGA UNIVERSITY, GULBARGA
Department of Computer Science
Jnana Ganga Gulbarga - 585106
www.gulbargauniversity.kar.nic.in
Phone No: 08472-263293/94
Fax No : 08472-263206
Grams : UNIGUL

No. GUG/Comp.Sc /2014-15/ 245 Date: 06-11-2014

To,
The Registrar
Gulbarga University,
Gulbarga

Sub: Request to arrange for installation of water purifier (RO Systems)-reg.

Sir,

With reference to the above subject, I am to bring to your kind notice that, many of the students from our Department are suffering from water prone diseases. The water supply is from overhead tank and is contaminated. I therefore, request you kindly to arrange for supply of water purifier that can support 250 students per day for drinking water.

Thanking you.

Yours faithfully,

(Dr.B.V.Dhandra)
Chairman
Dept. of Computer Science
Gulbarga University, Gulbarga

cc

Copy to : The Assistant Engineer, Gulbarga University, Gulbarga
-for information and necessary action.

Fig – 1 Sample image of Official Document

2.2 Pre-processing

The acquired image may be degraded due to the diversity in the quality of paper, ink, dust and the scanner machine used. Hence, it is essential to perform processing on the obtained image. Pre-processing is composed of sequence of steps used to generate an enhanced version of the input image. The rate at which the image is pre-processed affects the accuracy rate of classification and identification.

Different Stages

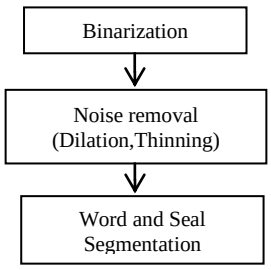


Fig 2 : Stages of Pre-processing

These are explained below

Binarization: Pre-processing always commences with this stage. The conversion of color and grayscale images to binary format is known as 'Binarization'.

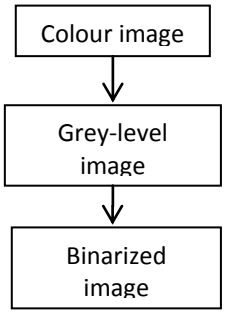


Fig 3: Process of Binarization

The color image is converted to grayscale image. The grayscale image is transformed to binary image by 'Otsu's method'. The color image cannot be directly converted to binary image. Otsu's method is based on a discriminant analysis which is unsupervised and non-parametric method. This method assumes, the image to be thresholded contains two classes of pixels foreground pixels and background pixels. Foreground pixels are represented by 1's and background pixels by 0's. The optimum threshold is calculated for separating the foreground and the background information.

Noise Removal: The noise is eliminated by morphological operations and rule-based approach. Morphological operations simplify image data and preserve the essential characteristics of the object shape. These operations are described by a structuring element. Horizontal dilation is performed. The extent of thickening is controlled by a flat line structuring element which is a set of point coordinates. Thinning is performed to fill the holes which are unfilled by dilation. It is a process of reducing the binary valued image regions to lines that approximates skeletons of the regions, so that further analysis is facilitated.

Segmentation

Segmentation of document images has to be initiated, as an early stage of document analysis and processing. Segmentation is the process of partitioning the image into homogenous regions distinct from each other. It is a critical step in image processing and computer vision systems. The document image is partitioned into several segments to locate printed text, handwritten text and seal which are the different components of the document image.

Rule based method for Word Segmentation

Rule based system should expose in a comprehensible way knowledge hidden in data, providing logical justification for drawing conclusions. Rules are used to support decision making in classification. The logical rules like if the threshold area is greater than the area of other image objects then the index of the image is 1. The threshold area is set to 81. The higher height and lower height is set to 100 and 10 respectively. The lower width is set to 10. The text, numerals and special symbols are identified and segmented within the specified range.

Connected Component Labeling

A connected component in a binary image is a set of pixels that form a connected group. For example, the binary image below has three connected components. Connected

component labeling is the process of identifying the connected components in an image and assigning each one a unique label as shown in the figure.

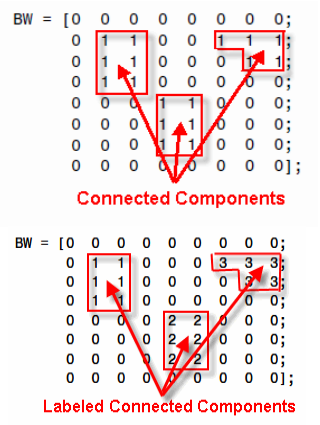


Fig 4: Connected components and Labeled connected components

The pixels labeled 0 are the background pixels and the pixels labeled 1 are the foreground pixels. In the figure the pixels labeled 1 is the first object, the pixels labelled 2 is the second object and so on. After assigning the label, a label matrix is generated. The input image is of the unsigned integer and non-sparse. The output is the binary image which is logical. All white pixels are represented by 1 and black pixels by 0. The neighborhood specifies the type of connectivity. Here the connectivity is 8. The number of connected objects, the size of the image and the number of pixels belonging to each connected component are identified. The connected neighborhood is symmetrical about its centre element. The label matrix is built to visualize the connected components. The size of the label matrix depends on the size of input image and the structure of the connected components. The first object is made up of pixels labeled 1, the second object is made up of pixels labeled 2 and this process is continued until all the objects are labeled which form the connected components. Then for each connected component a bounding box is formed.

2.3 Feature Extraction

In the proposed work total of 17 features are used to describe the image, out of which 10 are shapes features and 07 are Hu's moment Invariant features. Based on the shape of connected component of the letter number and non-letter number part of complete document image the following local features are extracted:

Shape Features:

1. **Area**: It is defined as actual number of white pixels in the region.

2. **Perimeter**: It is defined as the distance around the boundary of the region.

3. **Form factor**: The pattern of scattering white pixels in an image within the bounding box.

$$\frac{4\pi \times Area}{Perimeter^{2}} \tag{1}$$

4. **Major Axis**: It is defined as the length of the major axis of the ellipse that has the same normalized second central moments as the region.

5. **Minor Axis**: It is defined as the length of the minor axis of the ellipse that has the same normalized second central moments as the region.

6. $$Rooundness = \frac{4 \times Area}{pi \times MajorAxis^{2}} \tag{2}$$

7. **Compactness** : Compactness is an indication of solidness and convexity. It is given by ratio of the object to the area of a circle with the same perimeter. The maximum value possible is 1 which is for solid circle, image object having complicated boundaries will have lower values. This feature is calculated using Eq.3

$$Compactness = \frac{4\pi \cdot Area_{image}}{Perimeter_{image}} \tag{3}$$

8. **Density**: Density is defined as the area of white pixels within the bounding box. It is the ratio between area of white pixels within the bounding box and the area of bounding box which is given by:

$$Density = \frac{Area\ of\ white\ pixels\ within\ bounding\ box}{Area\ of\ bounding\ box} \tag{4}$$

9. **Mean of black pixel at each line(BPEL)** :

$$BPEL = \sum \frac{Number\ of\ Black\ Pixels\ of\ each\ line}{Width\ of\ Bounding\ Box} \tag{5}$$

10. **Vertical projection variance**: The vertical projection of white pixels within the bounding box is found and then the variance of only the vertical coordinates of the vertical projection profile is computed.

A feature vector is derived from the mean of above ten features.

Moment Invariant Features:

Shape of an object is the characteristic surface configuration as represented by the contour. Shape recognition is one of the modes through which human perception of the environment is executed.. Hu invariants moment are a set of nonlinear functions, which are invariant to translation, scale, and orientation and are defined on normalized geometrical central moments. Hu introduced seven moment invariants based on normalized geometrical central moments up to the third order. Since the higher order moment invariants have resulted higher sensitivity, a set of eight moment invariants limited by order less than or equal to four seems to be proper in most applications. Having normalized geometrical central moments of order four and the lesser ones, seven moment invariants $(\varphi_1 - \varphi_7)$ introduced by Hu and can be computed using equations given below.

$$\varphi_1 = \eta_{02} + \eta_{02} \tag{6}$$

$$\varphi_2 = (\eta_{02} - \eta_{02})^2 + 4\eta_{11}^2 \tag{7}$$

$$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{8}$$

$$\begin{aligned}\varphi_5 = {} & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\ & [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]\end{aligned} \tag{9}$$

$$\begin{aligned}\varphi_6 = {} & (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ & + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})\end{aligned} \tag{10}$$

$$\varphi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) [3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (11)$$

2.4 Algorithm

Algirithm-1 gives the flow of Testing, Training and Classification procedure used in the proposed work.

Algorithm1: Identification of Document Image

Input*: Colored or Binary images of Official letters of GUG*

Output*: Identification of the letter document of particular Department*

Method *:Shape and Moment Invariant Features*

Feature vector of size *: 10 Shape Features and 7 Moment Invariant Features.*

Train Phase:

Start

Step 1 : Input the preprocessed document image.

Step 2 : Perform Area based Segmentation to segment the Letter No. of the Document Image

Step 3 : Extract Shape and Moment Invariant Feature of segmented Image .

Step 4 : Store all the Features as a feature vector in train library with labels.

End

Test Phase:

Start

Step 1 : Input the preprocessed document image.

Step 2 : Perform Area based Segmentation to segment the Letter No. of the Document Image

Step 3 : Extract Shape and Moment Invariant Feature of segmented Image .

Step 4 : Store all the Features as a feature vector in test library with labels.

Step 5 : Identify the Official Document by label of the test letter No. using SVM Classifier.

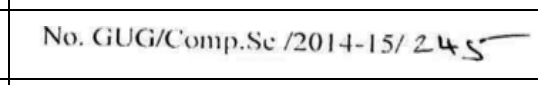
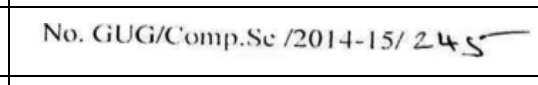
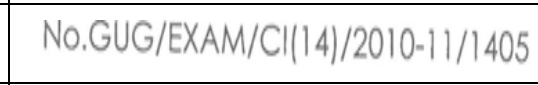
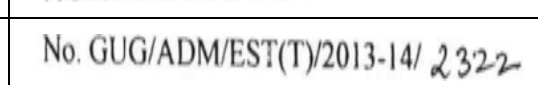
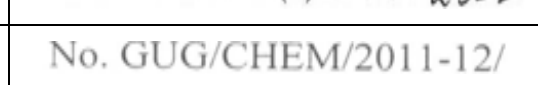
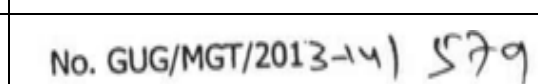
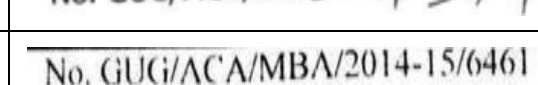
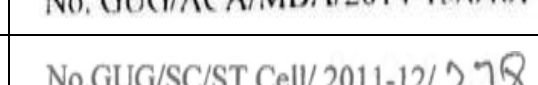
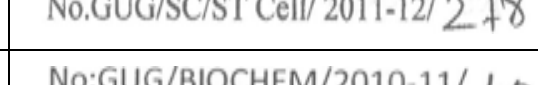
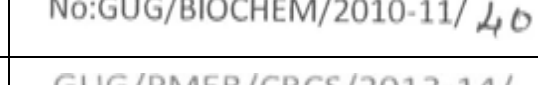
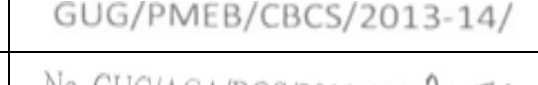
End

3. EXPERIMENTAL PROCEDURE AND DISCUSSION

A simulation model was developed in MATLAB programming language to implement the system and to analyze its simulation performance. In the quest for finding the best classification procedures. This paper analyzes a machine learning techniques using SVM classifier that resulted in the optimum identification rate.

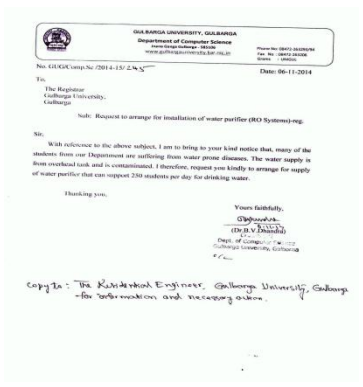
The training phase which establishes the lining up of the images is carried out as the current step. Some of the trained images of the segmented letter number images are shown in Table-1.

Table-1: Trained Images

S.No.	Trained Image
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	

The next step is the testing phase which helps in the extraction of the letter number from the document by establishing a series of phases which include Preprocessing, Colour conversion, Segmentation and finally feature matching using SVM Classifier.

Table 2 : Sample of Test Images

S.No.	Sample Test Document Images three Dept. of GUG
1	

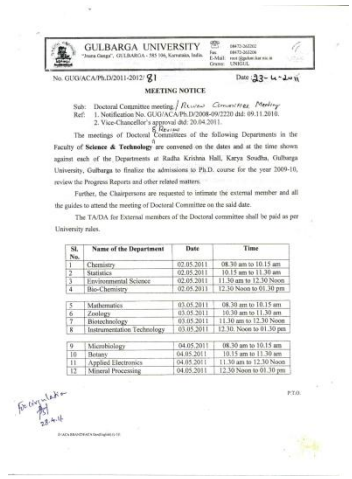
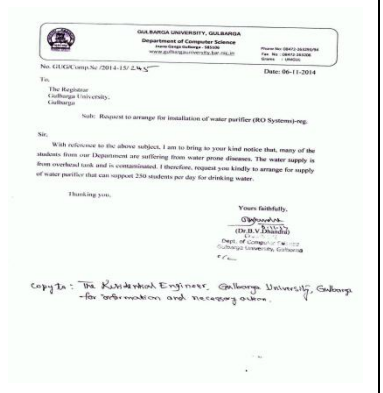
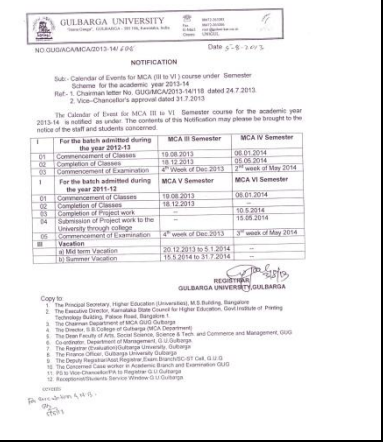
2	
3	
4	

Table-3 gives the accuracy of the algorithm for individual class of document belonging to different departments. D1 to D13 represent document of 13 different departments.

Table 3 : Identification Rate of Documents of 13 Departments

Dept. Type	% accuracy by SVM
D1	100%
D2	70%
D3	100%
D4	75%
D5	67%
D6	60%
D7	75%
D8	100%
D9	100%
D10	58%
D11	100%
D12	67%
D13	73%

4. CONCLUSION AND FUTURE WORK

The processing of complex documents is a challenging task for the purpose of identifying the ownership distinct document images of various department in an automated approach. In the proposed work, a system is developed to identify the ownership of a document using letter number present in the image of official letter. An area based segmentation method is used first to train the Letter Number part of complete official letter image and then for testing towards the query image of whole document. The combined features from the simple shape descriptors and moment invariant descriptor resulted into the optimal identification accuracy of 85.42%.

The future scope of the work is to increase the accuracy of the system by combining the image object like logo of the University/Institution and name of the Department of the document image.

5. REFERENCES

[1] Sachin Grover, Kushal Arora and Suman K Mitra, "Text Extraction from Document Image using Edge Information", IEEE India Council Conference, 2009.

[2] Yen-Lin Chen, "Automatic Text Extraction, Removal and Inpainting of Complex Document Images", International Journal of Innovative Computing, Information and Control, ISSN 1349-4198, pp 303-327, 2012.

[3] G.Rama Mohan Babu, P.Srimaiyee and A.Srikrishna, "Text Extraction From Hetrogenous Images using Mathematical Morphology", JATIT, 2010.

[4] Liukas Neumann and Jiri Matas, "A Method for text Localization and Recognition in real-world images", 10th Asian Conference on Computer Vision, Queenstown, New Zealand, 2010.

[5] B.V.Dhandra, Shridevi Soma, Rashmi T, Gururaj M, Classification of Document Image Components, International Journal of Engineering Research and Technology, Vol.2, Issue 10, October 2010, page 1429-1439.

[6] B.V Dhandra, Mallikarjun Hangarge, On Seperation of English Numerals from Multilingual Document Images, International Journal of Multimedia(JM), Vol.2, No.6

Nov. 2007, Academy Publisher, Oulu, Finland, page 26-33, ISSN: 1796-2048.

[7] Shazia Akram, Mehraj-Ud-Din Dar, Aasia Quyoum, "Docment Image Processing - A Review", International Journal of Computer Applications (0975-887), Vol. 10-No.5, Nov. 2010.

[8] Rangachar Kasturi, Lawrence O'Gorman and Venu Govindaraju, "Document image analysis:A primer", Sadhana Vol.27, Part 1, Feb 2002, pp 3-22.

[9] Robert M. Haralick, "Document Image Understanding : Geometric and Logical Layout", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 4, 1994, pp 384-390.

[10] Dengsheng Zhang, Guojun Lu, "Review of Shape representation and description techniques", Pattern Recognition Society, Elsevier Ltd., 2004, pp 1-19.

[11] Dimitri A Lisin, M. Mattar, M. Blaschko, M. Benfield, E. Learned-Miller, "Combining Local and Global Image Features for Object Class Recognition," Proceedings of IEEE Workshop on Learning in Computer Vision and Pattern Recognition (in conjunction with CVPR), San Diego, California, June, 2005.

[12] C.P. Sumathi, T. Santhanam and G.Gayathri Devi, "A Survey on various approaches of Text extraction in Images", International Journal of Computer Science and Engineering Survey(IJCSES), Vol. 3, No. 4, Aug 2012.

[13] R. Chandrasekaran, R.M. Chandrasekaran, "Morphology based Text Extraction in Images", IJCST, Vol. 2, Issue 4, 2011.

[14] Md. Shorif Uddin, Tenzila Rahman, Umme Sayma Busra and Madeena Sultana, "Automated Extraction of Text from Images using Morphology Based Approach", IJEI, Vol. 1, No. 1, Aug. 2012.

[15] Digital Image using MATLAB by Rafael C. Gonzales, Richard E. Words and Steven L Eddins, Low Price Edition, India.