Robust Text Extraction for Automated Processing of Multi-Lingual Personal Identity Documents

Pushpa B R [#1], Ashwin M [#2], Vivek K R [#3]
[1,2,3] Department of Computer Science
[1,2,3]Amrita School of Arts and Sciences,Mysuru Campus
[1,2,3]Amrita Vishwa Vidyapeetham, Amrita University,India
[1]preeths1@gmail.com
[2]ashwinmoothedath93@gmail.com
[3]vivekdaskr@gmail.com

***Abstract*—Text extraction is a technique to extract the textual portion from non-textual background like images. It plays an important role in deciphering valuable information from images. Variation in text size, font, orientation, alignment, contrast etc. makes the task of text extraction challenging. Existing text extraction methods focus on certain regions of interest and address characteristics like noise, blur, distortion and variations in fonts makes text extraction difficult. This paper proposes a technique to extract textual characters from scanned personal identity document images. Current procedures keep track of user records manually and thus give way to inefficient practices and need for abundant time and human resources. The proposed methodology digitizes personal identity documents and eliminates the need for a large portion of the manual work involved in existing data entry and verification procedures. The proposed method has been experimented extensively with large datasets of varying sizes and image qualities. The results obtained indicate high accuracy in the extraction of important textual features from the document images.**

Keywords - Text extraction, Edge detection, Canny operator, Gaussian filter, Morphological operations

I. INTRODUCTION

Text extraction is the retrieval of textual content from non-textual background in images and videos.Text extraction is a challenging area of focus in the domain of computer vision due to low image contrast of widely used acquisition scanners and complex background matter containing the combination of textual and non-textual graphics. It has a wide range of applications such as document analysis, detection of vehicle license plate, keyword based image search, identification of parts in industrial automation, content based retrieval, text based video indexing, video content analysis, page segmentation, document retrieval [1]. Steady research is being done on text extraction in images due to the rapidly growing availability of multimedia documents.

Text orientation, alignment of text, overlapping of letters makes the extraction of textual characters a challenging proposition. This proposed method seeks to address these issues and produces comparably higher accurate results than other techniques. Considering a scenario of the existing ration card system, the manual upkeep of user records makes it difficult to verify and thus time consuming to validate user credentials like name, address and other informative fields. Moreover, the user needs to renew the ration card after a certain period of time (i.e. five years). By digitizing the ration cards the manual work as well as the renewal of the ration card can be reduced. The proposed technique will automatically detect the textual characters from personal identity documents such as ration cards and allow for the efficient processing of user information.

The proposed work seeks to extract the textual characters from scanned personal identity documents. The techniques used are described in various sections and the proposed methodology achieves promising results for successful implementation of the same. The paper is divided into various sections. Section II gives a brief discussion about the various related works in this domain. Section III proposes the novel methodology and its work flow. Section IV presents the various experimental results and statistical parameters that were utilized for the evaluation of the results. The final section summarizes the content of the paper and proposes areas of future work.

II. RELATED WORK

Laurence Likforman-Sulem et.al [2] described an automatic way of extracting names from a document image. There are two approaches used, first is the image-based analysis exploits visual clues to select the regions of interest in the document and second is the textual-based analysis searches for name patterns and low-level word textual features. Both analyses are combined at the word level through a neural network fusion scheme. Daniel Loprest et.al [3] formulated an efficient technique to locate the “WWW” text in a image. The author describes a procedure based on clustering color space followed by a connected components analysis. For character

recognition technique like polynomial surface and fuzzy n-tuple classifiers are used. G. Rama Mohan Babu et.al [4] proposed a work which is insensitive to noise, skew and text orientation. The method has considered the fact that edges are reliable features of text regardless of color or intensity, layout, orientation etc. The edge detection operation is performed using the basic operators of mathematical morphology. Using edges the algorithm has tried to find out text has been labeled to identify different components of the image. Once components are identified, the variance is found for each connected component considering their gray levels. Then the text is extracted by selecting those connected components whose variance is less than a particular threshold value.

S P Chowdhury et.al [5] proposed a work to extract text from a camera grabbed image contains the huge amount of metadata about a scene.The approach proposed is the combination of the color based segmentation and the spatial distribution based patterns. Authors merged these two and segmented the color image by horizontal direction or vertical direction and matched the pattern of this segmented portion in the vertical direction or vice-versa. Claudio Antonio Peanho et.al [6] presented a work on efficient solution in dealing with heterogeneous documents containing regions with varying scanning quality, diverse layouts and multiple fonts and the reconstruction of the semantics of each field in a document. Initially, text lines and ruled lines are computed based on a threshold image; then the ruled lines are grouped into sub-forms and the text lines are converted to text using an OCR; then the obtained text is matched with a keyword dictionary consisting of different spellings for keywords, the corresponding correct spelling, incorrect keywords representing typical OCR errors and segments collected in advance from sample documents; the results are matched with each document model in a database consisting of word models and their logical relationships. Rohini K. Srihari et.al [7] described a general model for multimodal information retrieval that addresses the issues like users information need, expressing information need through composite, multimodal queries, and determining the most appropriate weighted combination of indexing techniques in order to best satisfy information need. The focus is on improving precision and recall in a MMIR system by optimally combining text and image similarity.

Partha Pratim Roy et.al [8] proposed a novel method to segment text lines and the method is based on the foreground and background information of the text components. First, individual components are detected and grouped into character clusters in a hierarchical way using size and positional information. Next, the clusters are extended in two extreme sides to determine potential candidate regions. Finally, a candidate region is used to extract individual lines. Shijian Lu et.al [9] described a scene text extraction technique that automatically detects and segments texts from scene images. Three text specific features are designed over image edges with which a set of candidate text boundaries is detected. For each detected candidate text boundary, one or more candidate characters are then extracted by using a local threshold which is estimated based on the surrounding image pixels.The real characters and words are identified by a support vector regression model that is trained using bags-of-words representation. Jui-Chen Wu et.al [11] propose a work in order to detect skewed text lines, a moment-based method is then used for estimating their orientations. According to the orientation, an x-projection technique can be applied to extract various text geometries from the text-analogue segments for text verification. However, due to noise, a text line region is often fragmented to different pieces of segments. Therefore, after the projection, a novel recovery algorithm is then proposed for recovering a complete text line from its pieces of segments. After that, a verification scheme is then proposed for verifying all extracted potential text lines according to their text geometries.

In Seop Na et.al [12] proposed a novel method for extraction of salient objects based on image clustering and saliency map from natural scene images. This method is a combination of image clustering, saliency map generation and automatic initialization. First, a graph based clustering method is applied to split the input image into regions. Second, a saliency map of the input image is generated using the contrast among split regions. From the split regions and generated saliency map, an adaptive threshold is defined, which classify the split regions into foreground and background. After that, the initial mask for object detection is determined using the classified foreground and background clusters and saliency values. A grab-cut with our initial mask is applied to extract the objects of interest. Chung-Ming Tsai et.al [13] proposed system includes color transformation, background color determination, objects extraction by top-down method, and objects classification without parameters and algorithm can be run in an embedded environment.This proposed system can process different kinds of color documents which include name cards, magazines, color receipts, and text books. Dineshkumar et.al [14] proposed an hand written character recognition framework uses forward neural network. Same characters are given as input to neural network with different set of neurons in hidden layers and the accuracy to recognize Sanskrit letters are calculated and compared. The results of the proposed work has good accuracy to recognize the letters than other handwritten character recognition systems

III. METHODOLOGY

This section discusses the proposed methodology, giving a brief outline of the processes involved and the algorithm. The first phase is to pre-process the image in order to remove noisy regions of the image and to remove the document boundaries encapsulating the textual regions of the document. After pre-processing the image, next step is to remove the photographic images from the binarized image by using edge detection

technique. The resultant image contains only textual characters present in it. Next phase is Segmentation, in this line segmentation and character segmentation techniques are applied to segment the textual characters. The characters are stored in a separate file. Fig.1 depicts the methodology of the proposed work to extract the textual characters from the scanned images of various personal identity documents.

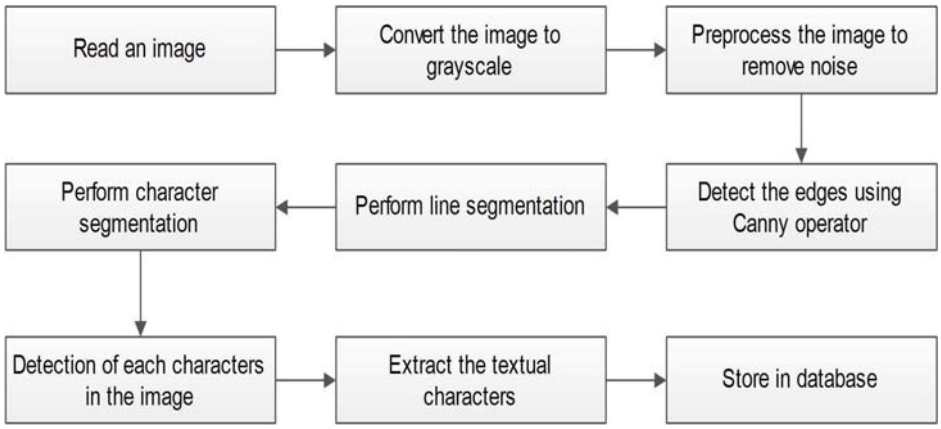

FIG 1. PROPOSED WORK FLOW FOR TEXT EXTRACTION

The general methodology of the proposed work may be sequenced as follows.

1. Read the scanned personal identity document image.
2. Convert to gray-scale and apply an erosion morphological operation on the image
3. Convert the resultant image from step 2 to binarized form
4. Enhance the image (apply median filter) for better identification and clarity of text.
5. Apply edge detection operator to detect edges.
6. Remove large dark regions from the image as per threshold definitions.
7. Remove photographic images.
8. Highlight textual characters using bounding boxes.
9. Apply line and character segmentation techniques to extract required text.

Table 1 depicts the various notations used in the proposed work to extract the textual characters from the image.

TABLE 1 NOTATIONS

Notations	**Descriptions**
I	Input image
Ig	Grayscale image
ℰ(Ig)	Eroded image
B (ℰ(Ig))	Binarized image
C(Ic)	Image with highlighted text
D(Ic)	Image with photographs removed

The detailed algorithm of the proposed work to extract relevant textual characters from the scanned image is as follows

1. Read Image I
2. Convert I to grayscale

   I → Ig

   Preprocess the image to remove noise

   % Apply morphology on grayscale image
3. Apply erosion to grayscale image

   Ig → ℰ(Ig)

   % Binarize the image

4. Binarize (erode_image, threshold) ƹ(Ig) → B (ƹ(Ig))
5. Apply median filtering on the binarized image % To remove noise
   % Detect edges in the image
   Apply Canny edge operator on image B (ƹ(Ig)) to get C(Ic)
   % Edges are detected and highlighted
6. Remove photographic images from C(Ic) Here C(Ic) has 0,1 values
   % this is done based on connected components.
   % removes all connected components (objects) that have fewer than P pixels C(Ic) →D(Ic)
7. Segmentation of each textual characters from the image
   [m n]= size(D(Ic))
   % m and n are size of the image m - rows, n - columns
8. Count Black pixels and white pixels in (D(Ic))
   Here D(i,j) has 0,1 values
9. Segment the characters from the images
   Line segmentation and character segmentation is performed
   % Perform Line segmentation
   for i=1:m
   %Check for black pixels
   If H(i) equals 1
   % store the whole line to variable
   Line(p) =i
   Print
   Line(p)
   p=p+1
   done
   % Perform Character segmentation
   % here each lines are taken ie Line(p)
   for s=1:p
   for i=:n
   %Count for white pixels column wise
   If H(i) equals 1
   Character(ch)=i
   Print
   Character(ch)
   ch=ch+1
   done
   done

IV. EXPERIMENTATION AND RESULTS

1) *Pre-Processing*: In this phase a sample ration card image is used as a personal identity document which is one out of several images in the datasets considered. Here pre-processing techniques are applied to remove noise in order to enhance image clarity. Initially the scanned image is converted to grayscale and subsequently binarized. Apply the median filter to remove noise and to enhance the intensity of the image. Fig. 2 shows the sample image of ration card

FIG.2A. SAMPLE RATION CARD IMAGE

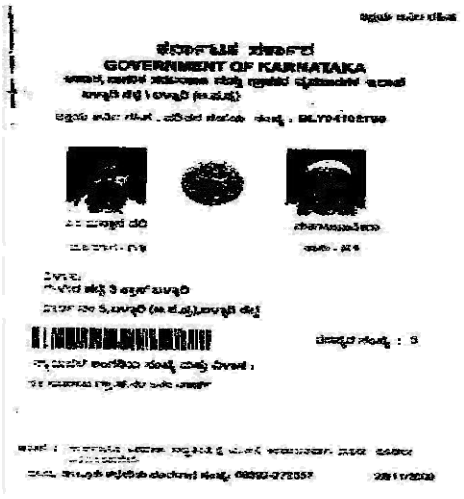

FIG.2B. BINARIZED IMAGE

a) *Converting scanned image to grayscale* : The scanned digital image would in most cases be a color image. The color images follow the RGB model - red, green and blue light are added together in various ways to produce a broad array of colors[10]. Grayscale images are compressed in nature due to the loss of color information, leading to better utilization of storage space. Grayscaling facilitates the process of erosion that is applied subsequently. The erosion morphological operation is applied to the grayscale image to increase the intensity of black pixels present in the image.

b) *Binarization* : The grayscale image is subjected to binarization using predefined thresholds. The pixels having intensity greater than the threshold value are changed to 0 (black) and the pixels with intensity value less than the threshold intensity value are changed to 1 (white). Fig.2b is the binarized image of the given ration card.

c) *Edge detection*: Edges are those features of an image that correspond to object boundaries. Edges are pixels where pixel intensities change abruptly. Edge detection is a technique to extract useful structural information from different vision objects and it reduces the amount of data to be processed. Textual data especially contain more edges than those present in non-textual areas. [11]

The general criteria for edge detection include:

1. Accurate detection of as many possible edges from the given image must be performed.
2. The edges detected by the operator should accurately localize on center of the edge.
3. The edge of an image should be marked once so that noise in the image should not create false edges.

Canny edge detection: Canny smooths the image using an advanced Gaussian filter. It is applied to remove noise. Compute gradient magnitude and direction of each pixel of the smoothed image. Based on the value of these parameters the edges are identified in the images.The first step is to apply Guassian filter and is calculated by Eq.1.

$$B_{i,j} = \frac{1}{2\pi\sigma^2}\left[exp\left[-\frac{\left(i-\left(k+1\right)\right)^2+\left(j-\left(k+1\right)\right)^2}{2\sigma^2}\right]\right], i, j = 1\ldots\left(2k+1\right) \qquad \text{(Eq.1)}$$

This operation is applied on the binarized image.

Larger the size of the threshold value, the greater the edge detection sensitivity.The noisy regions in the images are not detected as edges in the image. The next step is to find intensity gradient of the image. The gradient magnitude and direction can be calculated with a variety of different edge detection operators, and the choice of operator influences the quality of results. The edge detection operators are Roberts, Sobel and Prewitt.

The edge gradient and direction is calculated based on Eq.2

$$G = \sqrt{G_x^{\,2} + G_y^{\,2}} \qquad \text{(Eq.2)}$$

Where Gx is the horizontal direction and Gy is the vertical direction. Based on these parameters the edges are identified in the image. The below Fig.3a depicts the detection of the edges in the image.

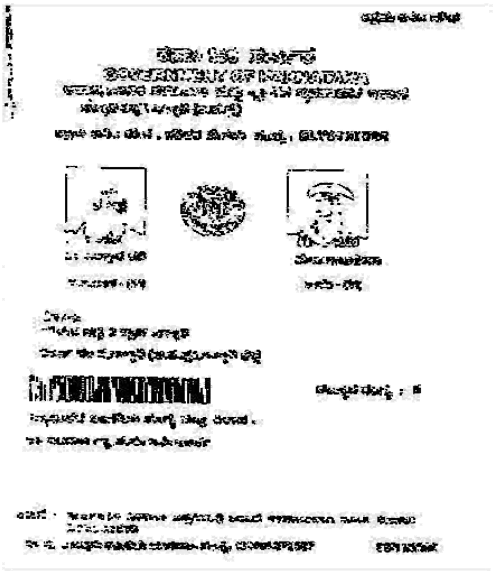

Fig 3a. Canny with HOG vertical edge detection

Fig 3b. Bounding box encapsulates image features

Fig 3b. demonstrates bounding boxes in order to highlight the edges. Histogram of Gradient is an edge based descriptor used in computer vision and image processing for the purpose of object detection. The image is divided into small regions called cells. The local appearance and shape of the object is obtained by distribution of intensity gradient and edge directions in HOG descriptors.

d) *Removal of photographic image*: The photographic images are removed by checking the count of black pixels present in the binarized image. This is done using connected components which have more number of black pixels. If the black pixel count exceeds the threshold permissible value, then the region is discarded. Thus the photographic image is also discarded. Fig.4 shows the photographic images being removed from the ration card and the textual characters in that image are highlighted.

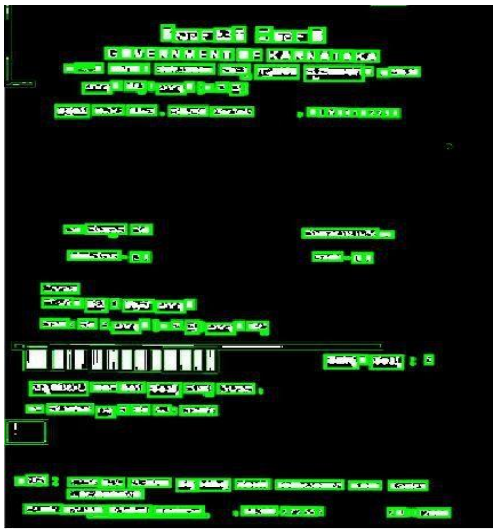

Fig. 4. Image after removal of photographic image

2) *Extraction of textual characters*: Line segmentation and Character segmentation techniques are applied in this phase to extract textual characters from the binarized image.

Segmentation: Image segmentation is the process of dividing an input image into multiple parts. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.Segmentation methods can also be applied to edges obtained from edge detectors. This is achieved by using edge detector operator for the given image. Here we used canny edge detection method to detect the edges accurately and thus make possible the extraction of text.

a) *Line segmentation:* The given binary image is traversed horizontally to analyze the pixel values. If all the pixels in the given line are white, then the line is not considered for feature extraction. If a line contains

black pixels in a contiguous fashion, then the complete line of text is extracted. This extracted text is stored in a separate file for future analysis.

b) *Character segmentation*: The segmented image is used to extract each character present in the line. Here, the image is traversed vertically and pixels are analyzed. Each line contains black and white pixels. The character segmentation extracts only black pixels that are encountered in each line for feature extraction and white pixels are ignored. The character segmented images are stored separately. Fig.5 shows text segmentation of the binarized image.

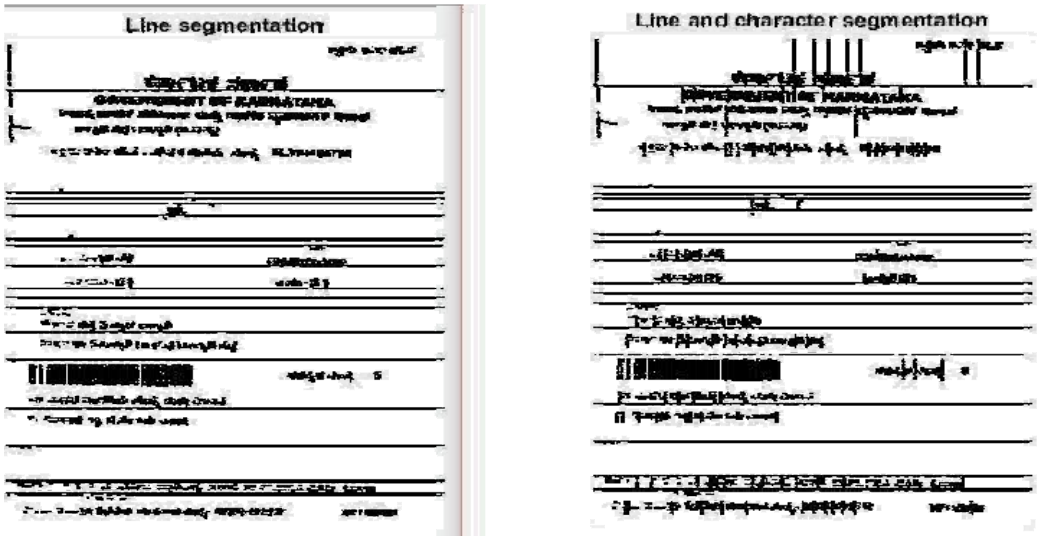

Fig.5. Resultant image of Line and Character segmentation

From each line segmented image the intensity between each of the characters is detected. If a space exists between two characters then the intensity value in between the characters will be one and thus each character is segmented. Fig 6 shows the textual characters extracted from the segmented image.

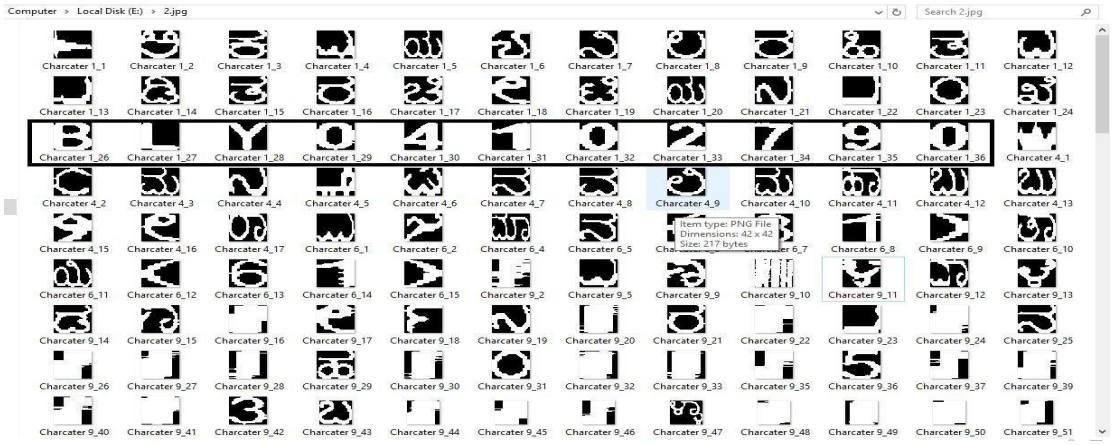

Fig. 6 Extracted textual characters.

The experimental results show that the proposed methodology extracts the textual characters from the personal identity document image. Results show that the proposed work takes less time to extract text from image. The accuracy of the proposed methodology is given by finding sensitivity and specificity. They are calculated using the following equations

$$Sensitivity = \frac{TP}{P} \qquad \text{(Eq.3)}$$

Where TP stands for True Positives and is used denote correctly segmented characters, P stands for number of positives and is used to denote number of discernible characters present in the image

$$Specificity = \frac{TN}{N} \qquad \text{(Eq.4)}$$

Where TN stands for True Negatives and is used denote non correctly segmented characters, N stands for number of negatives and is used to denote the number of non-discernible characters present in the image

$$Accuracy = Sensitivity\frac{P}{(P+N)} + Specificity\frac{N}{(P+N)} \quad \text{(Eq.5)}$$

TABLE 2 EXPERIMENTAL STATISTICS

Experimental Datasets	Sensitivity	Specificity	Accuracy
Dataset 1	0.9090	0.8888	0.9081
Dataset 2	0.9117	0.8333	0.9346
Dataset 3	0.9024	0.8334	0.8936

Table 2 depicts the experimental statistics of the proposed algorithm to extract textual characters from personal identity documents. From each dataset Sensitivity,Specificity,Accuracy are calculated. Dataset1 consists of 50 random images and an accuracy of 90.81% is achieved in extracting the textual characters. Dataset2 consists of 70 random images and an accuracy of 93.46% is obtained in extracting the textual characters. Dataset3 consists of poor scan quality image compared with other datasets. This dataset consists of 100 images and an accuracy rate of 89.63% was obtained. The overall methodology gives 91.21 % accuracy rate to extract the textual characters from the images.

V. CONCLUSION

Today there are numerous text extraction techniques available and they address varied challenges in extracting text from images. The proposed methodology extracts the textual characters from personal identity documents. The algorithm is experimented with various random sample datasets. This work reduces the manual labour involved in current systems and helps in digitizing the process of handling personal documents. The approach is efficient in extracting textual characters from any type of personal identity document image and the methodology is independent of font size and alignment of text. The work can be further extended to accommodate the extraction of handwritten text in certain personal identity documents.

REFERENCES

[1] Sumathi, C. P., Santhanam, T., & Devi, G. G. (2012). “A survey on various approaches of text extraction in images”. International Journal of Computer Science and Engineering Survey, 3(4), 27.
[2] Laurence Likforman-Sulem - Pascal Vaillant - Aliette de Bodard de la Jacopie`re, ”Automatic name extraction from degraded document images”,Volume 9(2-3), ISSN:1433-755X,19.
[3] Daniel Loprest, Jiangying Zhou, “Locating and Recognizing Text in WWW Images” ,Volume 2(2-3),ISSN: 1573-7569,17
[4] Babu, G., Srimaiyee, P., & Srikrishna, A. (2010). “Text extraction from Hetrogenous images using Mathematical Morphology”. Journal of Theoretical & Applied Information Technology,16.
[5] Chowdhury, S. P., Dhar, S., Das, A. K., Chanda, B., & McMenemy, K. (2009, July). “Robust extraction of text from camera images”. In Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on (pp. 1280-1284). IEEE.
[6] Peanho, C. A., Stagni, H., & da Silva, F. S. C. (2012). “Semantic information extraction from images of complex documents”. Applied Intelligence, 37(4), 543-557.
[7] Srihari, R. K., Zhang, Z., & Rao, A. (2000). “Intelligent indexing and semantic retrieval of multimodal documents”. Information Retrieval, 2(2-3), 245-275.
[8] Partha Pratim Roy, Umapada Pal, Josep Lladós,”Text line extraction in graphical documents using background and foreground information”,IJDAR(2012) Vol:15, ISSN:227–241 DOI 10.1007/s10032-011-0167-3,15
[9] Lu, S., Chen, T., Tian, S., Lim, J. H., & Tan, C. L. (2015). “Scene text extraction based on edges and support vector regression”. International Journal on Document Analysis and Recognition (IJDAR), 18(2), 125-135.
[10] Sumathi, C. P., Santhanam, T., & Devi, G. G. (2012). “A survey on various approaches of text extraction in images”. International Journal of Computer Science and Engineering Survey, 3(4), 27.
[11] Yang, J., Price, B., Cohen, S., Lee, H., & Yang, M. H. (2016). Object Contour Detection with a Fully Convolutional Encoder-Decoder Network.arXiv preprint arXiv:1603.04530.
[12] Na, I. S., Le, H., Kim, S. H., Lee, G. S., & Yang, H. J. (2015). “Extraction of salient objects based on image clustering and saliency”. Pattern Analysis and Applications, 18(3), 667-675.
[13] Wu, J. C., Hsieh, J. W., & Chen, Y. S. (2008). “Morphology-based text line extraction”. Machine Vision and Applications, 19(3), 195-207.
[14] Tsai, C. M., & Lee, H. J. (2011). ”Efficiently extracting and classifying objects for analyzing color documents”. Machine Vision and Applications, 22(1), 1-19.
[15] R. Dineshkumar and J. Suganthi (2015) “Sanskrit Character Recognition System using Neural Network” , Indian Journal of Science and Technology, Vol 8(1), 65–69, January 2015

AUTHOR PROFILE

Pushpa B R has completed Master's degree in Computer Applications at Vishvesvaraya Technological University, Belgaum, Karnataka and currently working as a Faculty in the Department of Computer Science at Amrita Vishwa Vidyapeetham, Mysore Campus. Her area of interests is Cryptography and Network Security.

Ashwin M currently pursuing Master's degree in Computer Applications at Amrita Vishwa Vidyapeetham, Amrita University, Mysuru Campus and completed Bachelor's degree in Computer Applications at Amrita Vishwa Vidyapeetham, Amrita University, Mysuru Campus. His area of interest is Computer Vision.

Vivek K R currently pursuing Master's degree in Computer Applications at Amrita Vishwa Vidyapeetham, Amrita University, Mysuru Campus and completed Bachelor's degree in Computer Applications at Amrita Vishwa Vidyapeetham, Amrita University, Mysuru Campus.His area of interest is Computer Vision.