A Unified Framework for Information Extraction from Newspaper Images

Jitesh Kumar and Sanjay Kumar Dubey

Amity School of Engineering & Technology, Amity University
Sector-125, Noida – 201301 (U.P.), India

Abstract

Nowadays Newspapers are very common source of information which is easily available to all. It consists of all sorts of news like social news, political news and lots of advertisements. These advertisements/announcements are concentrated on some specific page. This paper proposes a system that can extract contact information like email address, website address and telephone number from newspaper advertisements regarding job, contract, biding and other announcements of company. Proposed system will be able to store old advertisements details for future references. It is very easy for human being to spot the words in an image but it takes lots of computation for a computer to extract and separate these words. This paper explains the necessary steps which are required to recognize optical characters like segmentation, smoothing, image processing and neural network implementation for image recognition.

Keywords: *Optical Character Recognition, Neural Network, Image Processing*

1. Introduction

Newspaper contains advertisements regarding job, contracts, biding and other private/government company announcements. Huge Number of newspaper published every day and each newspaper contains different advertisements/announcements which may be local or national. For a single person it is very difficult to keep track of news content. Here we are building a system which extracts the contact information of these announcements so that person can easily fetch the relevant website, email and contact number.

The most difficult part of this system is optical character recognition. In today's world there is lots of system available to recognize optical character but they have certain factors which affect the accuracy of system such as font type and distortion of character in image due to noise. The resolution of image is also a factor which affects the accuracy of system. Optical Character Recognition (OCR) defined as the process of translating images of handwritten, typewritten, or printed text into a format understood by machines for the purpose of editing, indexing/searching, and a reduction in storage size [1].

Optical Character Recognizer invoke many technique to break down the image into single character image such as segmentation, converting colour image into black and white image, separating line of text and finally separating each character from lines. These Character images then subjected to another intelligent system which recognizes the character it represents.

2. Background

Humans are always trying to build machine that exhibits intelligence like human reading ability. However, over the last six decades, machine reading capability has grown from dream to reality. To incorporate reading ability into machine, Optical character recognition has become most successful technique. This technique is applicable in the field of pattern recognition and artificial intelligence. Many commercial systems have been developed since 1950. After so much effort and hard work machines are still not able to complete the reading abilities.

The first character recognizer was built in the middle of the 1940s that focused on machine-printed text. Some of the recognizer dealt with handwritten text or symbols. In 1950s, commercial character recognizers were available for Latin languages. The first true OCR reading machine was installed at Reader's Digest in 1954 [2]. This machine was used to convert typewritten sales reports into punched cards for input to the computer. In 1980s, statistical and structural methods developed which gives character recognition a new direction. These methods break down the images into small parts to recognize shape of character image. After 1990, many complex methodologies were developed such as neural networks, hidden Markov models and natural language processing techniques which reduces the significant error in the OCR system.

In 2009 an article published which deals with the work done by the National Library of Australia [3]. This newspaper digitization program purpose was to identify and removing the error of the OCR system for improving the accuracy of system in large scale newspaper digitization. During this program several solutions were identified, applied and tested. Now the digitised material is available in the Australian Newspapers Digitisation Program beta service. This published article explains the methods to achieve accuracy, factor that affect OCR accuracy and it also explain the testing methods. In this article the accuracy of system is tested many times by modifying the approach. The accuracy of OCR in this context is very important because of huge amount of document.

Nowadays huge number of hardware and software based system are available which are application oriented such as identifying vehicle number from vehicle number plate, reading handwritten text on the envelope, etc.

3. STEPS INVOLVED

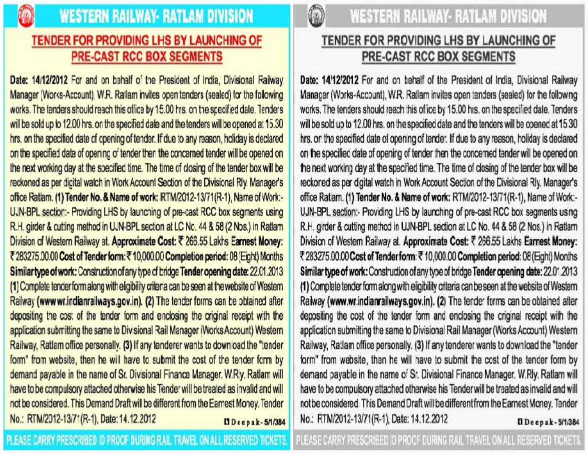
3.1 STEP 1: Segmentation

Segmentation is a process of separating different segment of digital image so that analysis of image will be easier. As we know that news articles are separated either by continuous line or relatively large blank space. Segmentation techniques are available in wide verity; some of them are Thresholding, Region Based Technique, Clustering Technique, Split and Merge Technique and Histogram-based methods. Thresholding is the simplest method in which image is classified using threshold value. This technique can be implemented using global thresholding or local thresholding. In global thresholding method only one threshold value is used for entire image where as in local thresholding method select different threshold for different region.

Region-based technique based on the colour intensity. This method divides the image into resign. Division is based on some rules like intensity of grey level image's pixel should be same for one resign [5]. K-means Clustering method split image into K group or cluster. This K cluster division can be random or based on some heuristics. Each pixel is then assigned to cluster so that the distance between pixel and centre of cluster minimizes. By averaging the pixel of cluster we re-compute centre of cluster. These two processes of assigning and re-computing centre of cluster continue till convergence is achieved. This method's accuracy depends on the number of cluster (K) used and the initial set of cluster [6]. Histogram-based method is very efficient in compare to other segmentation method. This method needed only one time scan of each pixel in image. In our example image we can see that heading is in red colour and other text is in black colour so here region-based technique will work fine. This case is not always true so we need other method to separate the heading. Here we can see that all advertisements have heading at the top and its character size is bigger than other text. So we need to identify the lines with height higher than average height of line.

3.2 STEP 2: Binary Image

Binary image contains only two colours black and white. Colour of character does not play important role in this context. Even it makes difficult to separate text in case of background colour variation. In case of newspaper articles either it contains white or some light colour background. For removing the colour from the text first we need to find the threshold value of pixel so that it can be converted into black and white image. The Method we are using here is Otsu's Threshold Selection Method [4]. The Otsu's algorithm works on two classes of pixel foreground and background. By calculating the optimum threshold separating those two classes so that their combined spread is minimal. Figure 1 shows the example image taken for calculating threshold. Example image's threshold is 147. This threshold is used to convert from grey scale image to binary image. Figure 2 shows the binary image generated by applying 147 threshold value to the image.



WESTERN RAILWAY- RATLAM DIVISION

TENDER FOR PROVIDING LHS BY LAUNCHING OF
PRE-CAST RCC BOX SEGMENTS

Date: 14/12/2012 For and on behalf of the President of India, Divisional Railway Manager (Works-Account), W.R. Ratlam invites open tenders (sealed) for the following works. The tenders should reach this office by 15.00 hrs. on the specified date. Tenders will be sold up to 12.00 hrs. on the specified date and the tenders will be opened at 15.30 hrs. on the specified date of opening of tender. If due to any reason, holiday is declared on the specified date of opening of tender then the concerned tender will be opened on the next working day at the specified time. The time of closing of the tender box will be reckoned as per digital watch in Work Account Section of the Divisional Rly. Manager's office Ratlam. **(1) Tender No. & Name of work:** RTM/2012-13/71(R-1), Name of Work:- UJN-BPL section:- Providing LHS by launching of pre-cast RCC box segments using R.H. girder & cutting method in UJN-BPL section at LC No. 44 & 58 (2 Nos.) in Ratlam Division of Western Railway at. **Approximate Cost:** ₹ 266.55 Lakhs **Earnest Money:** ₹ 283275.00.00 **Cost of Tender form:** ₹ 10,000.00 **Completion period:** 08 (Eight) Months **Similar type of work:** Construction of any type of bridge **Tender opening date:** 22.01.2013 **(1)** Complete tender form along with eligibility criteria can be seen at the website of Western Railway **(www.wr.indianrailways.gov.in)**. **(2)** The tender forms can be obtained after depositing the cost of the tender form and enclosing the original receipt with the application submitting the same to Divisional Rail Manager (Works Account) Western Railway, Ratlam office personally. **(3)** If any tenderer wants to download the "tender form" from website, then he will have to submit the cost of the tender form by demand payable in the name of Sr. Divisional Finance Manager. W.Rly. Ratlam will have to be compulsory attached otherwise his Tender will be treated as invalid and will not be considered. This Demand Draft will be different from the Earnest Money. Tender No.: RTM/2012-13/71(R-1), Date: 14.12.2012

Deepak-5/1/384

PLEASE CARRY PRESCRIBED ID PROOF DURING RAIL TRAVEL ON ALL RESERVED TICKETS.

WESTERN RAILWAY- RATLAM DIVISION

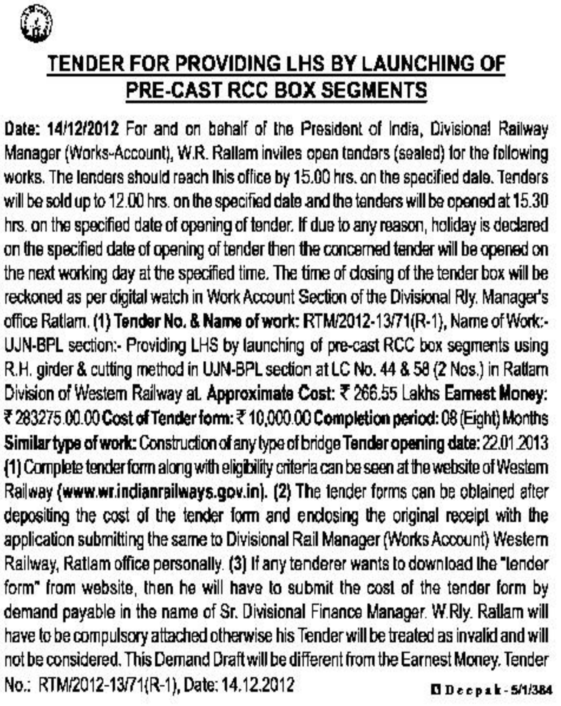
TENDER FOR PROVIDING LHS BY LAUNCHING OF
PRE-CAST RCC BOX SEGMENTS

Date: 14/12/2012 For and on behalf of the President of India, Divisional Railway Manager (Works-Account), W.R. Ratlam invites open tenders (sealed) for the following works. The tenders should reach this office by 15.00 hrs. on the specified date. Tenders will be sold up to 12.00 hrs. on the specified date and the tenders will be opened at 15.30 hrs. on the specified date of opening of tender. If due to any reason, holiday is declared on the specified date of opening of tender then the concerned tender will be opened on the next working day at the specified time. The time of closing of the tender box will be reckoned as per digital watch in Work Account Section of the Divisional Rly. Manager's office Ratlam. **(1) Tender No. & Name of work:** RTM/2012-13/71(R-1), Name of Work:- UJN-BPL section:- Providing LHS by launching of pre-cast RCC box segments using R.H. girder & cutting method in UJN-BPL section at LC No. 44 & 58 (2 Nos.) in Ratlam Division of Western Railway at. **Approximate Cost:** ₹ 266.55 Lakhs **Earnest Money:** ₹ 283275.00.00 **Cost of Tender form:** ₹ 10,000.00 **Completion period:** 08 (Eight) Months **Similar type of work:** Construction of any type of bridge **Tender opening date:** 22.01.2013 **(1)** Complete tender form along with eligibility criteria can be seen at the website of Western Railway **(www.wr.indianrailways.gov.in)**. **(2)** The tender forms can be obtained after depositing the cost of the tender form and enclosing the original receipt with the application submitting the same to Divisional Rail Manager (Works Account) Western Railway, Ratlam office personally. **(3)** If any tenderer wants to download the "tender form" from website, then he will have to submit the cost of the tender form by demand payable in the name of Sr. Divisional Finance Manager. W.Rly. Ratlam will have to be compulsory attached otherwise his Tender will be treated as invalid and will not be considered. This Demand Draft will be different from the Earnest Money. Tender No.: RTM/2012-13/71(R-1), Date: 14.12.2012

Deepak-5/1/384

PLEASE CARRY PRESCRIBED ID PROOF DURING RAIL TRAVEL ON ALL RESERVED TICKETS.

Fig. 1 Image of tender notice and grey scale image obtained from coloured image.



TENDER FOR PROVIDING LHS BY LAUNCHING OF
PRE-CAST RCC BOX SEGMENTS

Date: 14/12/2012 For and on behalf of the President of India, Divisional Railway Manager (Works-Account), W.R. Rallam inviles open tenders (sealed) for the following works. The lenders should reach lhis office by 15.00 hrs. on the specified dale. Tenders will be sold up to 12.00 hrs. on the specified date and the tenders will be opened at 15.30 hrs. on the specified date of opening of tender. If due to any reason, holiday is declared on the specified date of opening of tender then the concerned tender will be opened on the next working day at the specified time. The time of closing of the tender box will be reckoned as per digital watch in Work Account Section of the Divisional Rly. Manager's office Ratlam. **(1) Tender No. & Name of work:** RTM/2012-13/71(R-1), Name of Work:- UJN-BPL section:- Providing LHS by launching of pre-cast RCC box segments using R.H. girder & cutting method in UJN-BPL section at LC No. 44 & 58 (2 Nos.) in Ratlam Division of Western Railway at. **Approximate Cost:** ₹ 266.55 Lakhs **Earnest Money:** ₹ 283275.00.00 **Cost of Tender form:** ₹ 10,000.00 **Completion period:** 08 (Eight) Months **Similar type of work:** Construction of any type of bridge **Tender opening date:** 22.01.2013 **(1)** Complete tender form along with eligibility criteria can be seen at the website of Western Railway **(www.wr.indianrailways.gov.in)**. **(2)** The tender forms can be oblained after depositing the cost of the tender form and enclosing the original receipt with the application submitting the same to Divisional Rail Manager (Works Account) Western Railway, Ratlam office personally. **(3)** If any tenderer wants to download lhe "lender form" from website, then he will have to submit the cost of the tender form by demand payable in the name of Sr. Divisional Finance Manager. W.Rly. Rallam will have to be compulsory attached otherwise his Tender will be treated as invalid and will not be considered. This Demand Draft will be different from the Earnest Money. Tender No.: RTM/2012-13/71(R-1), Date: 14.12.2012

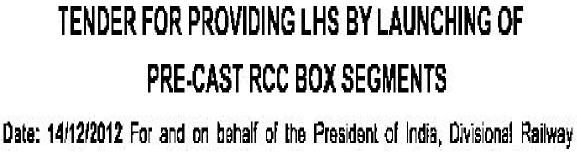
Deepak-5/1/384

Fig. 2 Binary image by applying threshold value 147.

3.3 STEP 3: Image processing

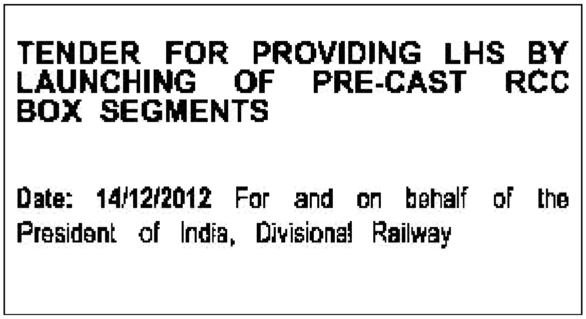
At this point we have generated binary image of segments which is ready to partition into corresponding line and then character image. Here we can use simple scanning of image horizontally looking for either blank line horizontally or lines with few pixels, if there is some noise then we need to find the connected component of that pixel if there is no components connected to it than it can be eliminated. Image of lines extracted from example Image is shown in figure 3. Now from each line we need to separate each word. This can be achieved by scanning image of line and observing the blank vertical line. As the space can be observed after character and word so we need to find the spacing length of each word in image. For this purpose we can calculate the space between all characters and by averaging it we can say that if the separation is 10 % higher than average then it is word separation. Figure 4 shows the word image separated from lines.

In case of underlined word we need to detect it and carefully measure the position of underlines we can separate character. As shown in figure 5 the character captured from line is not distinct. Some characters are combined with other so they are cropped together from the image. Now we need to analyse these image to find the connected component so that two or more character cropped together will be identified. In case of character joint together by some black colour pixel then we need smoothing process to separate them. As we can see the in “Date” word “Da” is combined because “D” and “a” have no gap of vertical pixel but they are not connected. So in this case the connected component analysis is enough to separate them. Now for “te” the separation of characters is not visible so we need to apply the smoothing process.



TENDER FOR PROVIDING LHS BY LAUNCHING OF
PRE-CAST RCC BOX SEGMENTS
Date: 14/12/2012 For and on behalf of the President of India, Divisional Railway

Fig. 3 Image of lines separated from segment.



TENDER FOR PROVIDING LHS BY LAUNCHING OF PRE-CAST RCC BOX SEGMENTS

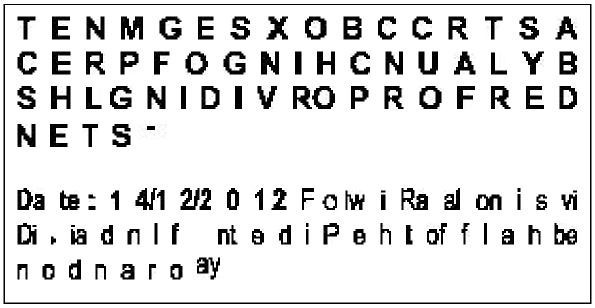
Date: 14/12/2012 For and on behalf of the President of India, Divisional Railway

Fig. 4 Character image Extracted from image of line as shown in figure 3.

3.4 STEP 4: Image Smoothing

The image captured from scanning process may have certain amount of noise. It depends on the scanner resolution. It also possible that the threshold value used for generating binary image may broke the character so that it is no longer recognizable. This may result in poor performance of the OCR system. This can be eliminated by using filters to smooth the digital image of characters. The smoothing is a process of filling and removing holes, breaks and gaps in digital image of character. Lots of efficient smoothing filters are available like low-pass filter, recursive filter, diffusion technique etc. [7]. The most common technique uses a small pixel set called window and applies rule over these pixel and then go for other set of pixel. Figure 6 shows how smoothing will work for both filling and thinning.

Here scaling is also required because input size of neural network should be constant. Each character image captured will need to scale properly so that position of each pixel will relatively same as the unscaled picture. For this purpose we will use bicubic interpolation technique [8].



T E N M G E S X O B C C R T S A
C E R P F O G N I H C N U A L Y B
S H L G N I D I V RO P R O F R E D
N E T S -

Da te : 1 4/1 2/2 0 1 2 F o lw i Ra al on i s vi
Di , ia d n l f nt e d i P e h t of f l a h be
n o d n a r o ay

Fig. 5 Image of Character produced from word.

3.5 STEP 5: Character recognition using neural network

Recently, the neural network implementation drastically changes the efficiency and accuracy of OCR system. For character recognition we are considering a multi-layer perceptron, this network contains hidden layer in between input and output layer. These layers contains neuron cell, one layer neuron interconnect with other layer neuron. A feature vector applied to the network at the input layer. Each element (Cell) of the layer computes a weighted sum of its input and transforms it into an output by a nonlinear function. During training operation each connection weight are adjusted so that desired output result obtained. Neural network approach has adaptive nature which makes it suitable for recognition.

Figure 7 shows the neurons and their connectivity with other. In case of Image recognition the number of input neurons is the number of pixel present in the input images. For example if we are using 20x20 pixel image of each character then we have to build a neural network with 400 input neurons. The number of output neurons will be the number of character image we are using for training of network. For example if we are taking 5 image of each alphabet character (Capital letter only) then total of 26x5 images is used for training of neural network. So total of 130 output neurons will be needed.

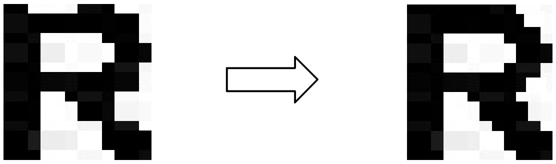


Fig. 6 Smoothing of character reduces the bumps in the image by filling and removing some pixel

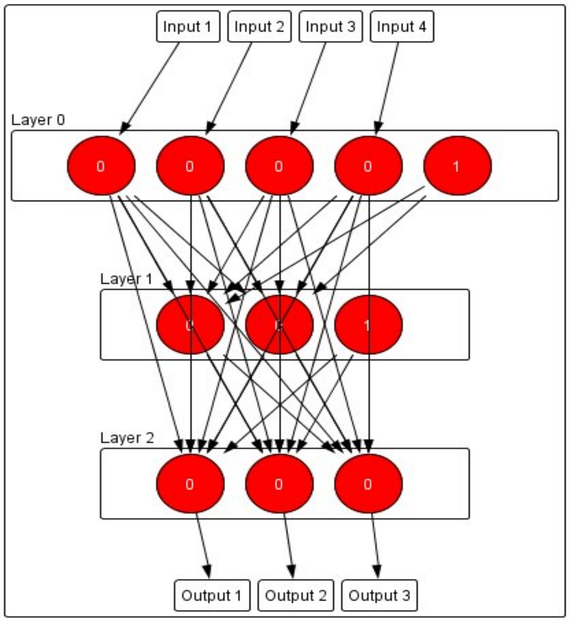


Fig. 7 A multilayer perceptron with 4 input neurons, 2 middle layer neurons and 3 output layer neurons.

3.6 STEP 6: Extracting information

After extraction of characters from image we will use GATE (General Architecture for Text Engineering) tool for annotation and tagging [9] [10]. GATE tool also provide the facility of natural language processing which will help in identifying the relevant information [11]. This text processing system separate the heading of the article, telephone number, email address and website address. All four information has some feature on the basis of which the text processing unit will distinguish it. For example if we are looking email address then it must contains '@' symbol along with that it will contain one '.' followed by some word like 'com', 'in' etc. If we are looking for website address then it will have more than two dots in between. If we are looking for contact information it must be a continuous digit of length 10. Here Gate tool is more appropriate because of huge amount of text need to analyse.

4. Overview of System

Proposed system will have mainly three parts namely Image processing unit, Character recognition unit and Text processing unit. First unit the Image processing unit will deals with all the required operation for extracting character image. It will incorporate the logic of segmentation, scaling, binary image conversion and smoothing. This unit is most important in this whole system. Any error generated by this unit will directly affect the final output of the system. The whole system accuracy will be driven by this unit.
Second unit of this system is character recognition which will use the neural network approach. Neural network accuracy for recognizing image is depends on the dataset provided for training of the neural network. We will use 10 and more images of each character and all these characters will be extracted from different newspaper itself.

Third unit is text processing unit which deals with the text extracted from the neural network. At this point we have to separate relevant text from other. For this purpose we are using GATE tool. Information Extraction (IE) functionality of this tool pulls the facts and structured information from large text collection. This unit will first try to find all the tokens from the text and then find the appropriate token. After gathering these data it will be stored for future references. The processes involve in this system is shown in the figure 8.

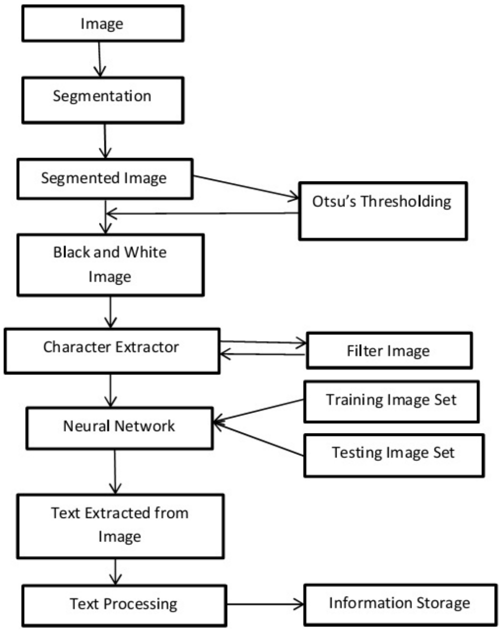


Fig. 8 Block diagram of operations execute by the system.

5. Conclusions and Future Work

In order to get effective system we need to implement efficient method for each step. Operations like Smoothing Segmentation, Image Binarization and image recognition performed on the input image can drastically affect the final output of the system so these operations needed to be good enough for maximum accuracy. Our proposed system is only intended to extract the contact information related to advertisement.

After achieving desired accuracy this system can be used for other purpose like grouping of advertisement according to place, job, biding, product etc. Using natural language processing it can be used to understand the nature of advertisements like buying, selling etc. As all the available technique of optical character recognition is still developing so in future proposed system will gain more accuracy and efficiency.

References

[1] Kurt Alfred Kluever, “Introduction to Articial Intelligence OCR using Articial Neural Networks.”, February 18, 2008

[2] Line Eikvil, “Optical Character Recognition”, December 1993, Norsk Regnesentral, P.B. 114 Blindern, N-0314 Oslo

[3] Holley, Rose (April 2009). "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". D-Lib Magazine. Retrieved 5 January 2011.

[4] Nobuyuki Otsu (1979). "A threshold selection method from grey-level Jan.histograms". IEEE Trans. Sys., Man and Cybernetics, IEEE Transactions on, 1979.

[5] Rafael C. Gonzalez, Richard E. Woods, ―Digital Image Processing□, Second Edition, Addison-Wesley, 2002.

[6] Pham, Dzung L.; Xu, Chenyang; Prince, Jerry L. (2000). "Current Methods in Medical Image Segmentation". Annual Review of Biomedical Engineering 2: 315–337

[7] Fadoua DRIRA, Franck LEBOURGEOIS, Denoising textual images using local/non-local smoothing Filters: A comparative study”, 2012 International Conference on Frontiers in Handwriting Recognition.

[8] R. Keys, “Cubic convolution interpolation for digital image Signal processing"(1981). IEEE Transactions on Signal Processing, Acoustics, Speech, and Processing 29 (6): 1153–1160

[9] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva (2013) Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics. PLoS Comput Biol 9(2): e1002854. doi:10.1371/journal.pcbi.1002854 — http://tinyurl.com/gate-life-sci/

[10] H. Cunningham, Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science. 15 April 2011.

[11] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, July 2002.

Jitesh Kumar is completed B.Tech (CSE) from Netaji Subhash Engineering College in 2010. He is pursuing M.Tech (CSE) from Amity University, Noida (UP), India. His research areas include Image Processing and Neural Network.

Sanjay Kumar Dubey is Assistant Professor and Proctor in Amity University, Noida (UP), India. His research areas include Human Computer Interaction, Object Oriented Software Engineering and Soft Computing. He has published more than 74 research papers in reputed National & International Journals. He is member of IET and ACM. He has submitted his Ph. D. (CSE) thesis in Amity University Uttar Pradesh, India.