Retrieval of Degraded Character in the Document

Aman Parkash Singh[1]
Pursuing M.Tech, Department of Electronics and Communication Engineering, ACET, Amritsar, Punjab, India

Sandeep Kaushal[2]
Department of Electronics and Communication Engineering, ACET, Amritsar, Punjab, India

***Abstract*— Optical character recognition is the mechanical or electronic translation of images of hand written, typewritten or printed text (usually capture by a scanner) into machine editable text. We have many documents written in the form of books or any historical documents. The printed document gets degrade due to time because ink used while printing get rubbed or corrupted automatically not all the characters in the image but some portion of the character get degraded. We consider those document in which one or more character are missing. For retrieval of degraded document we used Hidden Markove model(HMM) based segmentation confidence and some image processing tools. Based on the segmentation confidence we determine we determine whether they belong to the missing character or not for this the position of these character can be estimated. Various degraded documents have been used to test the proposed method.**

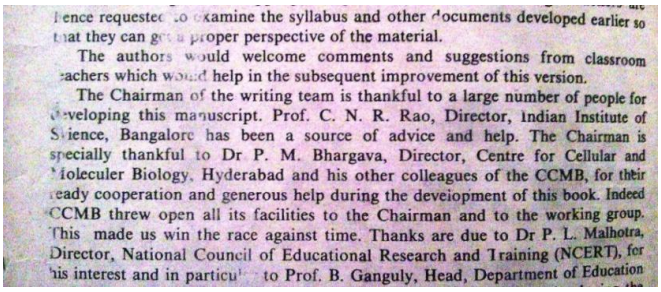
***Keywords*— Binarization, Image processing, Missing character and Segmentation.**

I. INTRODUCTION

Historical documents are important properties there are so many languages we choose English.

Some research works are don on the old corrupted documents for retrieval of the missing character in his work characters are bold and upper case and very little amount of degraded. In this paper our aim is to retrieve character that are corrupted more than 50 % or we can say that half portion of the character is missing and difficult to identify the complete character. It may be corrupted from upper side, lower side, left and right side.

We are also tried to find the characters which is completely disappear in that word. However if character is missing more than half then we use it as a clue to search in a database containing all the existing word. This method work like dictionary and the steps that we used to find word from dictionary are also used in this method


hence requested to examine the syllabus and other documents developed earlier so that they can get a proper perspective of the material.

The authors would welcome comments and suggestions from classroom teachers which would help in the subsequent improvement of this version.

The Chairman of the writing team is thankful to a large number of people for developing this manuscript. Prof. C. N. R. Rao, Director, Indian Institute of Science, Bangalore has been a source of advice and help. The Chairman is specially thankful to Dr P. M. Bhargava, Director, Centre for Cellular and Molecular Biology, Hyderabad and his other colleagues of the CCMB, for their ready cooperation and generous help during the development of this book. Indeed CCMB threw open all its facilities to the Chairman and to the working group. This made us win the race against time. Thanks are due to Dr P. L. Malhotra, Director, National Council of Educational Research and Training (NCERT), for his interest and in particular to Prof. B. Ganguly, Head, Department of Education

Figure 1.

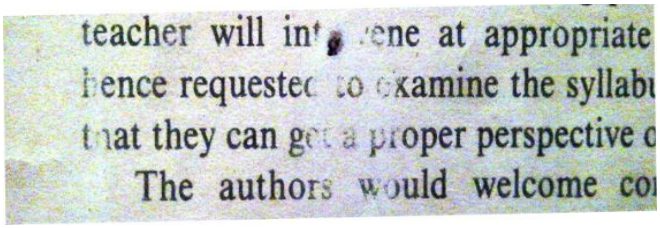

teacher will int ene at appropriate
hence requested to examine the syllabu
that they can get a proper perspective o
The authors would welcome co

Figure 2

Some scanned documents with missing characters are shown in fig.1 and fig.2 In this character are degraded and we are not able to predict the what exactly the missing character can be used in place of missing character.

I. LITERATURE REVIEW

There has been a growing interest in the development of methods for detecting, localizing and segmenting text from image. Here we present a method on detection, localization extraction and retrieval of missing character.

The first step is to locate the candidate regions that may contain a complete word in a scanned document. The second step is to segment the words from the proper regions. The steps` are similar to the steps of vehicle license plate recognition (VLPR) thus the technique used in VLPR can be referred to. For license plate localization, according to[1], the related methods can be categorized into four main categories, i.e. binary image processing [2]-[5], gray-level processing[6],[7], color processing[8]-[10], normalization[11], connected components[12], Neural network implementation[13]. However, considering the lower computational complexity and robustness, binary image processing methods ate most commonly used. These methods are bared on the fact that the change in brightness in the plate region is more remarkable and more frequent than elsewhere. In these methods, edge detection and mathematical morphology are often combined . Generally speaking, these methods may not work so well for missing character. As the discrete wavelet transform(DWT) can provide both the vertical and horizontal gradients of the image, Wag et al.[6] proposed to use the one level 5/3 DTW to locate the car plate. This method can deal with some complex environment such as low contrast images and dynamic-range problems. Al-Hmoz and Challa [7] also proposed to use various detection thresholds to detect the car plate. These detection results were then fused within Bayesian framework to give a final result. However, this method assumed that the plate region should

form a connected area under some threshold value, but in the container code images' cases, this assumption does not hold. Other local binarization methods, such as calculating threshold for each image block or even for every pixel [14], are employed to improve the segmentation performance. In classifier-based methods, the character segmentation is modeled as a global optimization problem [15]. After the bounding box of each character being obtained, the arrangement of the characters can be known in the character segmentation step for license plate. If the arrangement matches with the plate format, the segmentation results are defined as valid plate license characters.

II. THE PROPOSED APPROACH

The develop algorithm is divided into four group.

1. Image conversion
2. Locating the Region of Interest
3. Segmentation
4. Distance Measurement

1. Image Conversion

We are dealing with images which can be classified as 8-bit type, which means that 256 different colors. We know that the almost all documentation work are printed on black and white format but while scanning this image it may be in the form of RGB sot it is important to convert the image from RGB to Gray and apply a median filter to reduce the noise. After this binarization process is applied to obtain character in each candidate region. As the obtained image may be affected by noise a local binarization method is suitable The binarization resulet is shown in fig 1.

2. Locating the Region of Interest

By cropping the image we select only those words from the whole image which have some missing characters. The localization of characters are shown in figure 3.

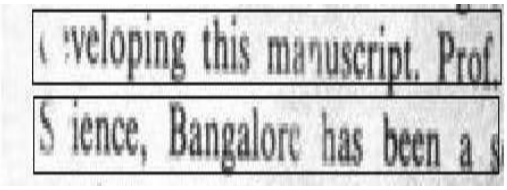


Fig. 3 Region group of characters

3 Segmentation

After the candidate region have been obtained the adjacent region are first grouped. As in fig. 3. When the objects in each region are extracted their position in the original image can also be obtained. Now we have to find the number of character in a word. This is the main challenging task because if last character of a word and the first character of the second word in the same line is missing then it is difficult to identify the number of character in a word. For this challenge we use the method that we follow to find the words form the dictionary and predict the missing characters. In fig. 4 we can see that in segmentation "requeste" is appear and next O and X are appear so in this case we are not sure about the no. of characters in a word and here dictionary helps us to predict the possibility of no of characters .

Figure 4. Segmented character

4 Distance measurement

After locating the character we measure the distance for feature comparison

a) Height
b) Width
c) Distance between character with in a word
d) Distance between words.

Based on this distance measurement we can find the distance measurement we can find the possibility of number of characters in a word and based on this we apply align mode of character

IV. HMM BASED SEGMENTATION CONFIDENCE

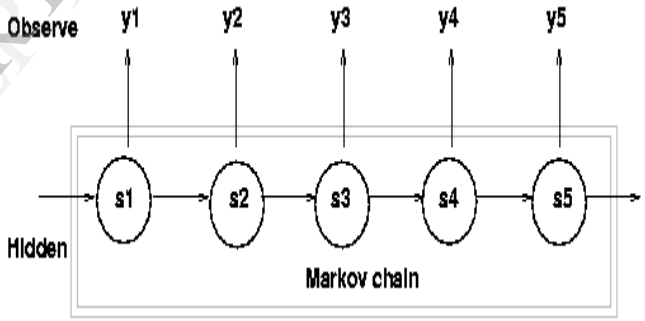


Figure no.5 shows Hidden Markov Model chain

The HMM chain for five character is shown in fig3 where $Y=(y_1,y_2,\ldots\ldots y_5)$ are observed sequence and $S=(s_1,s_2,\ldots\ldots s_5)$ are hidden states. When the no. when we get the maximum no of possibility of characters then we use $Y=(y_1,y_2\ldots\ldots\ldots y_n)$ and $S=(s_1,s_2\ldots\ldots s_{n)}$

We first consider the situation in which 18 characters are segmented out under the situation the HMM output is represented as follows

$$P(Y)=\sum_S P\left(\frac{Y}{S}\right)P(S)$$

$$S=(s_1,s_2\ldots\ldots s_n)$$

Here we use four features to describe each object spatial position i.e. object width V_w , height V_H, the horizontal center point (to measure distance between words in the same character) V_{DX1}, distance between two words V_{DX2} as shown in fig 6.

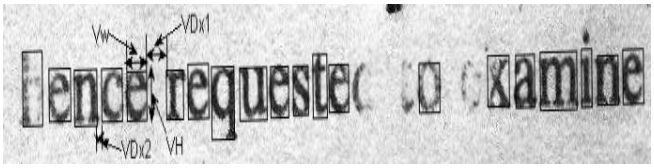


Fig. 6 Representation of four feature

V. Situation of missing character

By measuring the distance between character and using dictionary we can find the approximate number of characters in a single word using the maximum no of character of possibility we can find the missing character. Let us consider there is 23 characters in a single line and 3 characters are missing the remaining character should still be extracted and the missing character positions can be estimate. Based on the Segmentation confidence it is natural to propose that the missing character should be placed to maximize the segmentation confidence. To achieve this the spaces where the missing character can be placed should first be detected.

Given two neighboring character , how many character can be inserted into their space is calculated based on the distance between them.

i.e. $n_i = V_{Ds1}/\mu D_s - 1$

If $n_i > 0$, the space is marked as a possible insertion position with the attribute n_i which indicates how many character can be inserted into the space. At the beginning or at the end of a text line an arbitrary number of characters can be inserted.

Take a fig. 6 for eg. We have only 5 characters in first group and in second group we find the 9characters based on the distance measurement between characters and word. Out of 23 characters . Estimate where the three missing characters are of course the head and tail of the candidate line can always possibly place the three missing objects. From the distance analysis there is a space that can accommodate two objects between the 13th character "e" and 14th character "o" **.** Based on this HMM is crated

Training Phase

The purpose of the training phase is to obtain the deviation parameters. For each align mode, we select at least 20 images in which all characters appear, to train the parameter. However, to make the value of the total probability more in line with our intuitive ,in our experiment, we multiply the value of ζ_{si}^2 by 2.5. The parameters are set to to allow each component probability $\exp(-|Vsi-\mu s|^2/2\sigma^2_{si})$ to be roughly 0.6 at its mean deviation

Results

S No	No of Missing Character	No of Images	No of corrected images	Result
1	1	20	19	95%
2	2	20	18	90%
3	3	30	27	90%
4	0	20	19	95%

VI. CONCLUSION

The HMM-based segmentation confidence has been proposed to describe the probability of the segmentation results belonging to the character. The proposed method can work not only under the situation that all the no. of characters are segmented out but also under the situation that some characters are missing, which is seldom considered in other literatures. With the existing characters being recognized and each missing characters being represented by a wildcard character (such as "*"), the result can serve as a clue to search in a characters in database to obtain candidate character. Considering the computational efficiency of the algorithm and the trustiness of the estimated result, we mainly focused on tackling the situations wherein no more than three characters are missing. For those images with too many characters missing, we can also try to extract the characters from the other side of the character, such as top-view image or back-view image, which may contain more complete information. It is also one

of our future directions that provides more robust ACR result from document images of multiple-views. In this phase, the segmentation confidence can also give a clue of the characters' segmentation quality. Another thing to be mentioned is that this work can be compatible with newly emerging characters' align modes, through introducing these align modes' corresponding sets of parameters into the framework. This work can also be applied to other

applications, such as license plate recognition by changing the number of HMM states and redefining the characters' align modes.

REFERENCES

[1] C.-N. E. Anagnostopoulos, I. E. Anagnostopoulos, I. D. Psoroulas, V. Loumos, and E. Kayafas, "License plate recognition from still images and video sequences: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 9, no. 3, pp. 377–391, Sep. 2008.

[2] B. Hongliang and L. Changping, "A hybrid license plate extraction method based on edge statistics and morphology," in *Proc. ICPR*, 2004, pp. 831–834.

[3] D. Zheng, Y. Zhao, and J. Wang, "An efficient method of license plate location," *Pattern Recognit. Lett.*, vol. 26, no. 15, pp. 2431–2438, Nov. 2005.

[4] Y. R. Wang, W. H. Lin, and S.-J. Horng, *Fast License Plate LocalizationUsing Discrete Wavelet Transform*. Berlin, Germany: Springer-Verlag, 2009, ser. Lecture Notes in Computer Science, pp. 408–415.

[5] R. Al-Hmouz and S. Challa, "License plate localization based on a probabilistic model," *J. Mach. Vision. App.*, vol. 21, no. 3, pp. 319–330, Apr. 2010.

[6] C.-N. I. Anagnostopoulos, I. E. Anagnostopoulos, E. Kayafas, and V. Loumos, "A license plate recognition system for intelligent transportation system applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 3, pp. 377–392, Sep. 2006.

[7] T. D. Duan, T. L. H. Du, T. V. Phuoc, and N. V. Hoang, "Building an automatic vehicle license plate recognition system," in *Proc. Int. Conf. Comput. Sci. (RIVF)*, 2005, pp. 59–63.

[8] X. Shi, W. Zhao, and Y. Shen, *Automatic License Plate Recognition System Based on Color Image Processing*. Berlin, Germany: Springer-Verlag, 2005, ser. Lecture Notes in Computer Science, pp. 1159–1168.

[9] R. O'Malley, E. Jones, and M. Glavin, "Rear-lamp vehicle detection and tracking in low-exposure color video for night conditions," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 2, pp. 453–462, Jun. 2010.

[10] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen, "Automatic license plate recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 1, pp. 42–53, Mar. 2004.

[11] Akhito KITADAI and Masaki NAKAGAWA, "Simalarity Evaluation and Shape Extraction for character Pattern Retrieval to Support Historical Documents", *IEEE 10th International Workshop on Document Analysis Systems*, 978-0-7695-466-2/12, 2012.

[12] Mohanad Alata and Mohammad Al-Shabi, "*Text Detection and character Recognition Using Fuzzy Image Processing*", Journal of Electrical Engineering, vol.57, NO. 5, 258-267, 2006.

[13] Gaurav Kumar and Pradeep Kumar Bhatia ,"*Neural Network based Approach for Recongnition of Text Images*", International Journal of Computer Application(0975-8887), vol. 62, No. 14, Jan 2013

[14] D. Llorens, A. Marzal, V. Palazon, and J.M. Vilar, *Car License Plates Extraction and Recognition Based on Connected Components Analysis and HMM Decoding*. Berlin, Germany: Springer-Verlag, 2005, ser. Lecture
Notes in Computer Science, pp. 571–578.

[15] V. Franc and V. Hlavac, *License Plate Character Segmentation Using Hidden Markov Chains*. Berlin, Germany: Springer-Verlag, 2005, ser. Lecture Notes in Computer Science, pp. 385–392.