AN OPTICAL CHARACTER RECOGNITION SYSTEM FOR TAMIL NEWSPRINT

K.H.Aparna, Sumanth Jaganathan, P.Krishnan, V.S.Chakravarthy
Department of Electrical engineering,
IIT Madras,
Chennai-600036
schakra@ee.iitm.ernet.in

ABSTRACT

We present an early version of a complete Optical Character Recognition (OCR) system for Tamil newsprint. All the standard elements of OCR process like deskewing, preprocessing, segmentation, character recognition and reconstruction are implemented. Experience with OCR problems teaches that for most subtasks involved in OCR, there is no single technique that gives perfect results for every type of document image. We have used the ability of artificial neural networks to learn arbitrary input/output mappings from sample data for solving the key problems of segmentation and character recognition. Text segmentation of Tamil newsprint poses a new challenge owing to its italic-like font type; problems that arise in segmenting such text are discussed. The final document is reconstructed in HTML document format

1. INTRODUCTION

The OCR problem in its entirety includes preprocessing steps of skew correction, binarization and noise removal, segmentation of the image into blocks and classification of the blocks into text, tables, graphics, and line diagrams etc. and finally reconstruction of the original document.

Even a mechanical placement of document on the scanner bed normally introduces a few degrees of skew. An algorithm for skew estimation based on Hough transform is described in [1]. Other methods proposed for detecting skew include projection profile analysis [2], image gradient analysis [3], morphological transforms [4] and correlation between lines at a fixed distance [5].

Binarization is the process of converting the gray scale images to binary images by comparing each pixel value with a threshold. Ostu [6] proposed a method for threshold selection using gray scale histogram.

The next important step of document image analysis is segmenting the document page into text and non-text regions. Wang and Srihari [7] in their analysis of newspaper image documents employ the following methods. For page segmentation homogeneous rectangular blocks are first segmented out of the image using methods such as run length smearing algorithm (RLSA) and recursive X-Y cuts (RXYC) which perform well only on documents with rectangular layouts. Pavlidis and Zhou [8] described a class of techniques based on smeared run length codes that divides a page into gray and nearly white parts. Segmentation is then performed by finding connected components either by the gray elements or of the white, the latter forming white streams that partition a page into blocks of printed material. Their classification method is based on across-scan-line correlation method. Page segmentation and classification based on texture analysis using neural networks is described in [9].

The rapid spread of computer literacy and usage in the ‘90s in India had resulted in a growing interest in OCR in Indian languages. An approach to recognition of printed Tamil characters based on condensed run method and symbolic run method is discussed in [10]. A syntactic pattern analysis system with an embedded picture language has been designed for the purpose of recognition of Devanagari script in [11], where contextual constraints are used to arrive at the correct interpretation. A rule-based contextual post-processor for the recognition of Devanagari characters is developed in [12]. An overview of segmenting machine printed characters when touching and broken characters are encountered is discussed by Lu [13], but the cases of touching italic characters has not been dealt with.

All the above works on OCR deal with recognition of isolated characters. Chaudhuri and Pal in [14] and [15] describe a complete OCR system for printed Bengali documents, which uses structural feature-based tree classifier for character recognition.

The present paper deals with a complete OCR system for printed Tamil documents. The block diagram shown in Figure 1 gives the various steps involved in our approach.

The paper is organized as follows. In Section 2 preprocessing of the document image is described. Section 3 explains the segmentation of the page into blocks. Section 4 deals with Tamil character recognition and reconstruction of the document image. Finally the paper concludes with a discussion in Section 5.

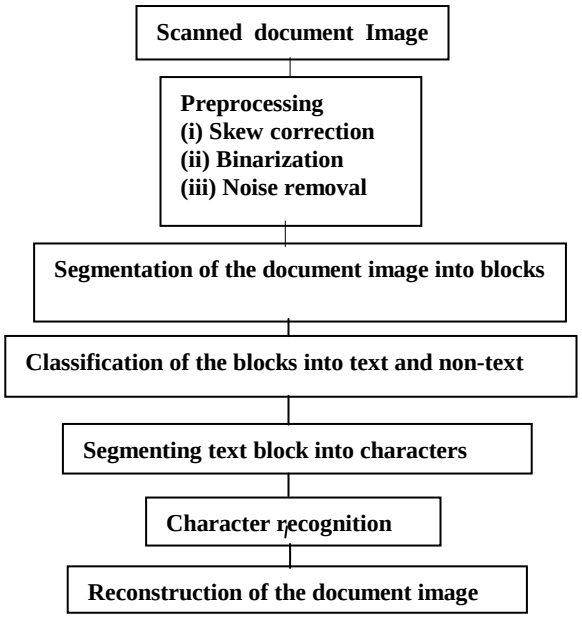

Figure 1. Steps involved in complete OCR for Tamil documents

2. PREPROCESSING

The document image obtained by scanning a hard copy magazine document as a black & white photograph at 300 dpi using a flat-bed scanner is represented as a two dimensional array. A document of size 8.27 X 11.69 inches scanned at 300 dpi would yield an image of 3324 X 2466 pixels.

Preprocessing stage consists of four steps: compression, skew correction, binarization and noise removal.

2.1. Image size reduction:

Some of the image analysis techniques of text recognition, skew detection, page segmentation and classification are applied on scaled down images. Such reduction not only increases speed of processing, but also gives more accurate results for specific tasks. For scaling down an image by half, a window of 2 X 2 pixels in the parent image is replaced by a single pixel whose value equals the median of the 2 X 2 window. The image obtained by scaling down the original document image by ¼ is referred to as *doc1by4*.

2.2 Text and non-text recognition:

Finding text regions in the document image is essential for skew estimation. For finding the text part we use the Radial Basis Function neural network (RBFNN) [16]. The network is trained to distinguish between text and non-text (non-text includes graphics, titles, line drawings). The input patterns for training the RBF neural networks are 20 Gabor filter [17] responses, with five each in horizontal, vertical and on both diagonal directions. The neural network has two outputs, one for text and the other for non-text. The network is presented with Gabor responses calculated from 40 X 40 windows of *doc1by4* images

The *doc1by4* image and the region marked as text by the neural network are shown in Figure 2.

Figure 2. Neural network output after text recognition of doc1by4

From Figure 2 it is evident that although most of the text part is recognized correctly there are a few spaces where text is recognized as non-text and vice versa. Therefore, for a perfect text, non-text block recognition we will use this output in later stages.

2.3. Skew Correction:

For skew angle detection Cumulative Scalar Products (CSP) of windows of text blocks with the Gabor filters at different orientations are calculated. Orientation with maximum CSP gives the skew angle. Alignment of the text line is used

as an important feature in estimating the skew angle. We calculate CSP for all possible 50X50 windows on the text recognized image (from *doc1by4* image) and the median of all the angles obtained gives the skew angle. The skew angle for the document in Fig. 2 (left) is found to be 0.5 degrees.

2.4. Binarization:

Binarization is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by thresholding. In the present case a global threshold of 158 is chosen.

2.5. Noise removal:

The noise introduced during scanning or due to poor quality of the page has to be cleared before further processing. For this the document is scanned for noise using a moving 5 X 5 window. If all nonzero pixels in the window are confined to the central 3 X 3 section, all those pixels are set to 0.

3. SEGMENTATION

Segmentation of a document image involves two steps: determination of page layout (segmenting the page into blocks containing a single document item) and classification of the blocks.

3.1. Page Segmentation:

When the binarized image of the document is observed we find that, if all the wide and long white spaces are removed without touching the white spaces between text lines, the page can be segmented into blocks.

Boundaries of the segmented blocks are found by "contour following" (see Figure 3). Figure 4 shows all the block coordinates that are stored. This process works well for different kinds of layouts and is fairly consistent and accurate. However, the procedure is not perfectly problem-free. For example, on occasions text and graphics are grouped together when a text part exists between a text region and non-text region.

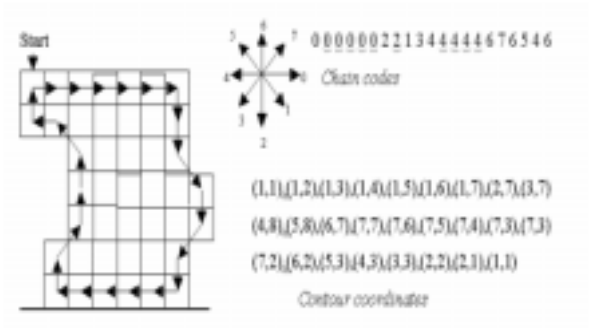

Figure 3: Contour following

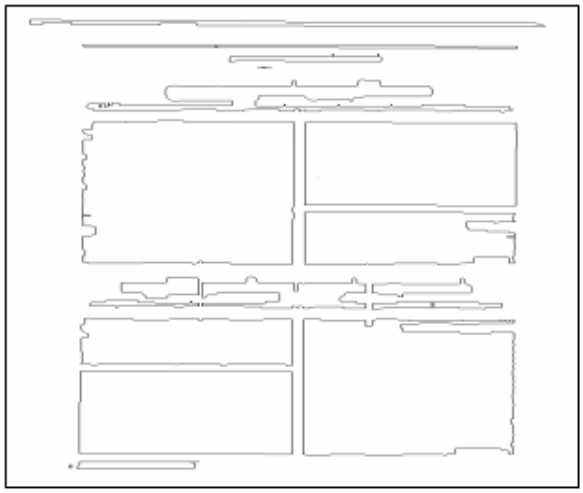

Figure 4: Segmented blocks of the image in Figure 2.

3.2 Classification of the blocks:

The blocks obtained from the segmentation stage of Section ***3.1***, and the result obtained from text recognition stage of section ***2.2*** are combined to give an accurate delineation of text borders. In the case of figure 2, all the four text blocks are extracted and passed on for character recognition; the four image blocks are stored as image files.

4. TAMIL CHARACTER RECOGNITION

The text blocks have to be initially segmented into lines, words and characters.

For the text block segmentation and character recognition the inverted binarized document (i.e. 0 for background and 1 for foreground) is being taken. For character recognition the original document (without any size reduction) is used.

4.1 Line, word and character segmentation:

For optical character recognition, the text blocks are segmented into lines, lines into words and then into individual characters.

(i) Line Segmentation:

For segmentation of text blocks into lines the horizontal projection on the y-axis is made use of. The best threshold value is chosen by trial and error.

(ii) ***Word and character segmentation***

Since the font used in Tamil newsprint is typically italic like, with the characters oriented at 79.21^0 with the horizontal, for segmenting the line into words and characters inclined projection is taken on the text line.

The segmentation is accurate if we have enough space between characters. If the characters are too close to each other or touching then segmenting becomes difficult.

For extracting characters that are too close but non-touching, connected-component extraction method is employed, in which components are segmented not by separation in one dimension but based on their connectedness.

4.2 Recognition of characters

A Radial basis function (RBF) neural network is trained for the recognition of characters. The full set of 157 characters including isolated Tamil characters, English numerals and punctuation marks are taken for training the neural network.

The characters are placed at the center of a 52 X 52 window and the input patterns to the RBF neural network are obtained from the response of 40 Gabor filters with 10 along each of four directions. The RBF neural network has 157 outputs each output corresponding to an alphabet.

The trained neural network is used for the recognition of the segmented characters.

Figure 5 shows the text part given for recognition and the output of the neural network for character recognition in HTML format

When the text block having a few touching characters (Figure 5) is sent for character recognition, 94% recognition rate is obtained. In general, for other documents the recognition rate varied from 85 to 90 percent depending on the touching characters present in the text part.

4.3 Reconstruction of the document image

Finally the recognized text blocks, represented in a suitable symbolic code, and non-text blocks, represented as image files, are put together to reconstruct the original document in HTML format (Figure 6).

Fig 5. Reconstructed text part (in HTML format)

Fig 6: Document reconstructed in HTML format

5. CONCLUSIONS AND DISCUSSIONS

We present a complete OCR system for Tamil newsprint. The system includes the full suite of processes from skew correction, binarization, segmentation, text and non-text block classification, line, word and character segmentation and character recognition to final reconstruction. Only the reconstruction step is currently done manually. Neural networks are applied to 2 subtasks: 1) text block identification, and 2) character recognition.

In the entire OCR process, the toughest challenges are faced in document segmentation and character recognition. Document segmentation is recognized as a hard problem and it may not possible to formulate a single algorithm, which works with all kinds of documents. Our approach gave reasonable

segmentation results with the class of document images chosen in the present work. The applicability of the technique for a larger class of Tamil newsprint is yet to be seen. Moreover, we have to overcome the errors in segmentation, which occur when the text and non-text are too close.

Currently characters of only a single font and font size are being recognized. To handle a larger variety of fonts we propose to train a separate neural network for each font. We assume that font type used in a given newsprint sample is known as prior knowledge. The case of touching characters presents a serious difficulty in character segmentation. This problem will be taken up as part of our future extensions of the current system.

An important lacuna in our present system is absence of a suitable document model. A document is treated as a collection of disparate items without any logical structure connecting all of them. Hence reconstruction is done by manually composing the document items. Future efforts will be focused on embedding various document components into a logical structure. This will help to automatize the reconstruction process also.

6. REFERENCES

[1] Bin Yu and A.K.Jain, "A robust and fast skew de detection algorithm for generic documents", *Pattern Recognition*, Vol. 29, No. 10,pp 1599-1629 (1996).

[2] A.Bagdanov and J.Kanai, "Projection profile based Skew estimation Algorithm for JBIG compressed images," *Intl' Conf. On Document Analysis and Recognition*, August 18-20, Ulm, Germany, pp 401-405, (1997).

[3] J.Sauvola and M.Pietikainen, "Skew angle detection Using Texture direction Analysis" *In Proc, of the 9th Scandinavian Conference on Image Analysis*, pp 1099-1106, Uppsala, Sweden, June, (1995).

[4] S.chen and R.M. Haralick, "An automatic algorithm for Text Skew estimation in document images using recursive morphological transforms", *In Proc. of the First IEEE International Conference on Image Processing*, pp 139-143, Austin, Texas, (1994).

[5] Hong Yan, "Skew correction of Document Images Using Interline Cross correlation", *CVGIP* Vol.55, No. 6, November, pp. 647-656, (1982).

[6] N.Ostu, "A threshold selection method from gray scale Histograms", *IEEE Trans on man Cybernet.*, pp 62-66, (1979).

[7] D.Wang and S.N.Srihari, "Classification of newspaper image blocks using texture analysis", *Comput. Vision Image Graphics Process*. 47, pp 327-352 (1989).

[8] T. Pavlidis and J. Zhou, " Page Segmentation and Classification", *CVGIP* Vol. 54, No. 6, pp 484-496, November (1992).

[9] A.K. Jain and Y. Zhong, "Page segmentation Using Texture analysis", *Pattern Recognition*, Vol. 29,No.5, pp. 743-770, (1996).

[10] G.Siromoney, R. Chandrasekaran and M. Chandrasekaran, "Machine recognition of printed Tamil characters", *Pattern Recognition*, vol. 10 (1978).

[11] R.M.K.Sinha and H.Mahabala, "Machine recognition of Devanagiri script", *IEEE Trans. Syst. Man Cybern*. *SMC-9*,435-441 (1979).

[12] R.M.K. Sinha, "Rule based contextual post-processing for Devanagiri text recognition", *Pattern Recognition* vol. 20, pp 475-485 (1987).

[13] Y. Lu, "Machine printed character recognition-An overview", *Pattern recognition*, Vol.28, No 1, pp 67-80, (1995).

[14] B.B. Chaudhuri and U.Pal, "A complete printed Bangla OCR system", *Pattern Recognition*, Vol. 31, No. 5, pp 531-549, (1998).

[15] B.B.Chaudhuri and U. Pal, "An OCR system to read two Indian language scripts: Bangla and Devanagari (Hindi)", *Intl' Conf. On Document Analysis and Recognition*, August 18-20, Ulm, Germany, pp1011-1015, (1997).

[16] J. E. Moody, C. J. Darken, "Fast Learning in Networks of Locally Tuned Processing Units," *Neural Computation* Vol.1, pp. 281-294, (1989).

[17] J.G.Daugman, "Uncertainty relation for resolution in space, spatial frequency, andentation optimized by two dimensional visual cortial filters", *J.Opt. Soc. Am.A/Vol. 2*,No. 7, , pp 1160-1169, July (1985).