7

A Complete OCR System Development of Tamil Magazine Documents

K.H. Aparna and V.S. Chakravarthy,

Department of Electrical engineering, IIT Madras,
e-mail: ee01m03@ee.iitm.ernet.in, schakra@ee.iitm.ernet.in

ABSTRACT

We present an early version of a complete Optical Character Recognition (OCR) system for Tamil magazine documents. 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 recognition of touching and close characters are discussed. Character recognition efficiency varied from 94 to 97% for this particular type of font. The grouping of blocks into logical units and the determination of read order within each logical unit helped us in reconstructing automatically the document image in editable format. Final reconstruction is done in HTML format.

1. INTRODUCTION

A magazine that is published weekly or fortnightly covers the top stories of that period and they form a valuable source of information. But the storing of the old magazines poses a challenge, as it becomes a difficult task to avoid degradation of the paper. An efficient way will be to store the documents in electronic form. Manual data conversion has several disadvantages of speed, cost and accuracy. Indian languages have an added disadvantage that each letter has a combination of keys. The text blocks have to be manually selected and should be given as an input, if it is a pure text OCR system. Presently, most of the currently available OCR systems for Tamil, the fourth widely used and spoken in Indian languages, are for pure text. This forms the motivation for the present work: to develop a complete OCR system for Tamil without any manual intervention.

The block diagram shown in Figure 1 gives the various steps involved in the approach of complete OCR system development.

2. PREPROCESSING

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 is discussed in Section 5. Finally the paper concludes with a discussion in Section 6.

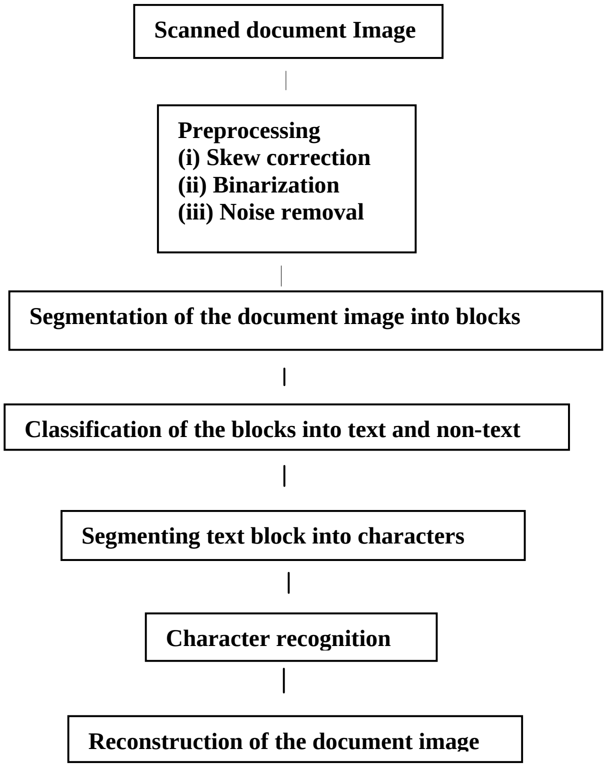

Scanned document Image
Preprocessing
(i) Skew correction
(ii) Binarization
(iii) Noise removal
Segmentation of the document image into blocks
Classification of the blocks into text and non-text
Segmenting text block into characters
Character recognition
Reconstruction of the document image

Figure 1. Steps involved in complete OCR for Tamil documents

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, the nearest neighbor interpolation method is used. 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) [1]. 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 [2] 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.

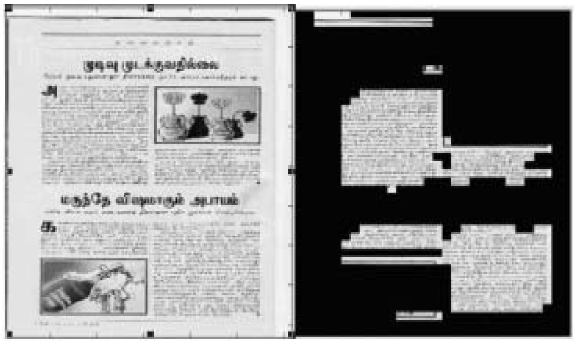


Figure 2. Neural network output after text recognition of doc1by4

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. The skew angle for the document in Figure. 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. The binary document image allows the use of fast binary arithmetic during processing, and also requires less space to store. Here for binarizing the *doc1by4* image the threshold is calculated using Ostu's method [3].

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 AND CLASSIFICATION

In the segmentation process the de-skewed, binarized and noise removed doc1by4 image is segmented into rectangular blocks of isolated regions. And in the classification process the blocks are classified as text, titles and images.

3.1 Page Segmentation

Our approach in page segmentation is a slight modified version of segmentation method of Palvidis and Zhou [4]. When the skew corrected binarized image of the compressed document is observed, we find that if all the wide and long white spaces are removed (excluding the white spaces between text lines) the page can be segmented into blocks. Figure 3 shows the images segmented into blocks. Each white region represents a block. The contour coordinates of all the blocks are stored.

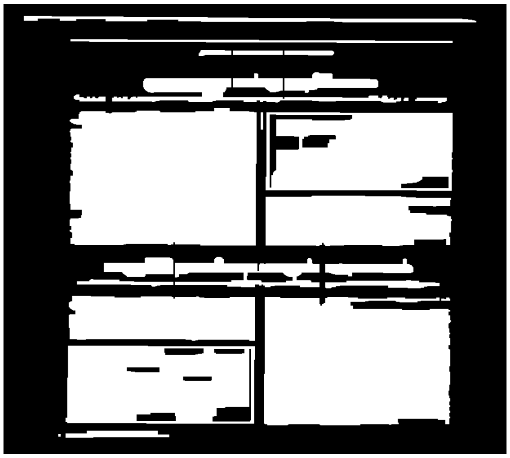


Figure 3: Document image with all the long and wide white spaces removed

3.2 Classification of the blocks

The classification of the blocks into text and non-text involves in comparing each block of the block-segmented image of Figure 3 with the corresponding region of the text-recognized image of Figure 2 The ratio of amount of text present in the region to the total area of the block classifies the block into text or non-text. The ratio ranges for classifying are given in Table 1.

TABLE 1: RATIO RANGES FOR TEXT/NON-TEXT CLASSIFICATION

Possibility	Ratio
Text	0.5 – 1
Text/non-text merge	0.3 – 0.5
Non-text	< 0.3

The second possibility in Table 1, of text/non-text merge is solved by the technique of run length smearing algorithm (RLSA), which smears off the merge part thus segmenting it into independent text and non-text. Now, the non-text is segmented into titles and images based on the feature that height of the title block is less compared to that of the height of image block. All the title and image blocks are stored as .jpg files, which are used in reconstruction.

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 docu-ment (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.

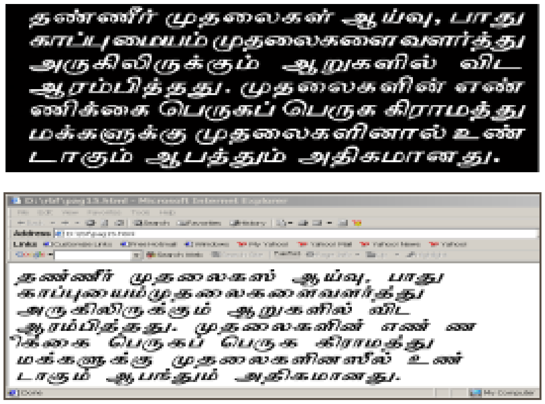
For extracting characters that are too close but non-touching, connected-component extraction method is employed, in which components are segmented not by separation in one dimension but based on their connectedness. Segmentation and recognition of touching character string involves in segmenting the string at different intervals followed by recognition.

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



தண்ணீர் முதலைகள் ஆய்வு, பாது
காப்பு மையம் முதலைகளை வளர்த்து
அருகிலிருக்கும் ஆறுகளில் விட
ஆரம்பித்தது. முதலைகளின் எண்
ணிக்கை பெருகப் பெருக கிராமத்து
மக்களுக்கு முதலைகளினால் உண்
டாகும் ஆபத்தும் அதிகமானது.

தண்ணீர் முதலைகள் ஆய்வு, பாது
காப்புமையம்முதலைகளைவளர்த்து
அருகிலிருக்கும் ஆறுகளில் விட
ஆரம்பித்தது. முதலைகளின் எண் ண
ிக்கை பெருகப் பெருக கிராமத்து
மக்களுக்கு முதலைகளினாலில் உண்
டாகும் ஆபாத்தும் அதிகமானது.

Figure 4. Reconstructed text part (in HTML format)

When the text block having a few touching characters (Figure 4) is sent for character recognition, 94% recognition rate is obtained. In general, the recognition rate varied from 90 to 97 percent.

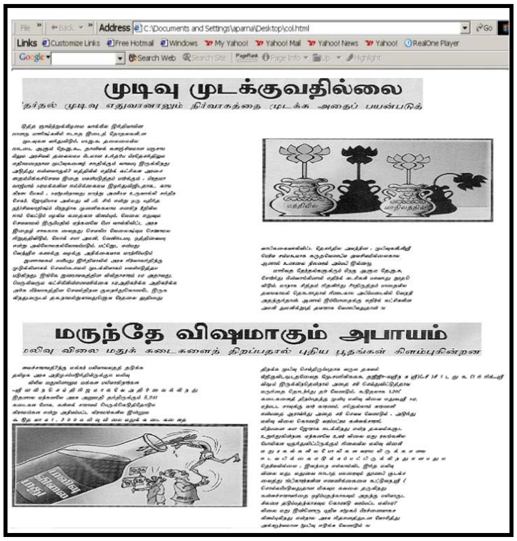
5. RECONSTRUCTION OF THE DOCUMENT IMAGE

5.1 Logical structure derivation

Derivation of logical structure of a document involves in finding the relationship among various blocks of the document and grouping document blocks into logical “units”, each unit representing an article or a story. The next step involves finding the reading order of the blocks within each unit. For finding the logical units we assume that, surrounding an article we have a sufficiently wide white space, which separates it from other articles. Within a logical unit, the read order starts at the top left corner block and goes from top to bottom and moves left to right covering all the blocks.

5.2 Reconstruction into HTML format

Finally the recognized text blocks, represented in a suitable symbolic code, and non-text blocks, represented as image files, are put together following the read order to reconstruct the original document in HTML format (Figure 5).



முடிவு முடக்குவதில்லை

தேர்தல் முடிவு எதுவானாலும் நிர்வாகத்தை முடக்க அதைப் பயன்படுத்

மருந்தே விஷமாகும் அபாயம்

மலிவு விலை மதுக் கடைகளைத் திறப்பதால் புதிய பூதங்கள் கிளம்புகின்றன

Figure 5: Document reconstructed in HTML format

6. 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.

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.

An important step in document image analysis is development of a suitable document model. The process of converting the physical structure of the magazine document into its logical structure is proposed. While reconstructing the logical structure, the read order is used in composing the document items. This helped to automatize the reconstruction process of converting into HTML format.

REFERENCES

1. Haykin, S. (1999) Neural networks: A Comprehensive foundation. Prentice Hall
2. 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).
3. N. Ostu, “A threshold selection method from gray scale Histograms”, *IEEE Trans on man Cybernet.*, pp 62-66, (1979).
4. T. Pavlidis and J. Zhou, “Page Segmentation and Classification”, *CVGIP* Vol. 54, No. 6, pp 484-496, November (1992).