A Complete Tamil Optical Character Recognition System

K.G. Aparna and A.G. Ramakrishnan

Biomedical Laboratory, Department of Electrical Engineering
Indian Institute of Science, Bangalore – 560 012
{prjocr,ramkiag}@ee.iisc.ernet.in

1 Introduction

Document Image processing and Optical Character Recognition (OCR) have been a frontline research area in the field of human-machine interface for the last few decades. Recognition of Indian language characters has been a topic of interest for quite some time. The earlier contributions were reported in [1] and [2]. A more recent work is reported in [3] and [9]. The need for efficient and robust algorithms and systems for recognition is being felt in India, especially in the post and telegraph department where OCR can assist the staff in sorting mail. Character recognition can also form a part in applications like intelligent scanning machines, text to speech converters, and automatic language-to-language translators.

Tamil is the official language of the southern state of Tamil Nadu in India, and also of Singapore, and is a major language in Sri Lanka, Malaysia and Mauritius. It is spoken by over 65 million people worldwide. The assumptions made in our work are that the document contains only printed text with no images and are uni-lingual.

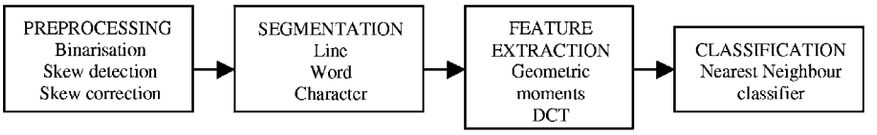


Fig. 1. Block diagram of our OCR System

2 Preprocessing

The block diagram of our OCR System is shown in Fig. 1. Preprocessing is the first step in OCR, which involves binarisation, skew detection [5] and skew correction. Binarisation is the process of converting the input gray scale image scanned with a resolution of 300 dpi into a binary image with foreground as white and background as black.

The skew introduced during the process of scanning is detected using the algorithm proposed by Kaushik et al [4], which is based on Hough transform and principal component analysis. An estimate of the skew angle is found to an accuracy of $\pm\ 0.06^0$. Such a high skew accuracy is needed as the first level of classification is based on the spatial occupancy of the characters as shown in Fig. 5. If skew is not properly detected, then characters will get misclassified in the first level itself.

While the skew detection is performed on the binarised document, correction, which involves rotating the image in the appropriate direction, is performed on a gray

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 53–57, 2002.

scale image to lessen the quantization effects [7], which is caused when a binary image is rotated. Bilinear interpolation is employed for this purpose. The output of skew correction is shown in Fig 2.

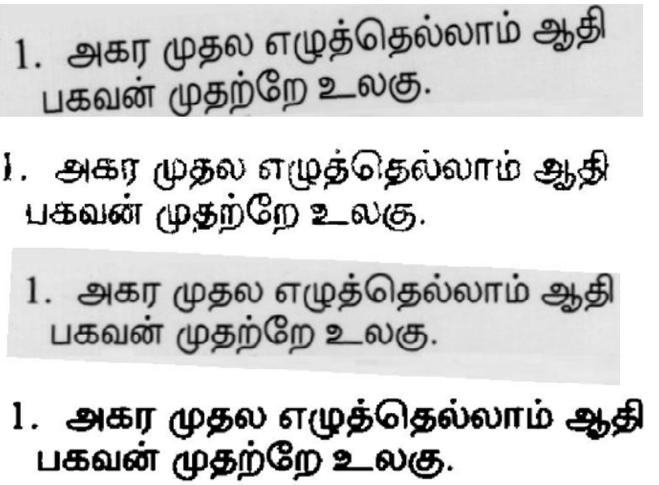


Fig. 2a. Original skewed image

Fig. 2b. Skew correction performed on binary image

Fig. 2c. Skew corrected gray scale image

Fig. 2d. Skew corrected binarised image

Fig. 2. Skew correction performed on Binary and Grayscale image

3 Segmentation

Segmentation is the process of extracting objects of interest from an image. The first step in segmentation is detecting lines. The subsequent steps are detecting the words in each line and the individual characters in each word, respectively.

Horizontal and vertical projection profiles are employed for line and word detection, respectively. Connected component analysis [6] is performed to extract the individual characters. The segmented characters are normalised to a predefined size and thinned before the recognition phase. Figures 3 and 4 show, respectively, the horizontal and vertical projections of a Tamil sentence.

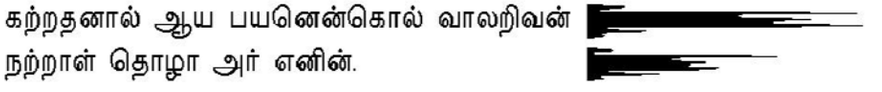


Fig. 3. Text lines with corresponding horizontal projection profiles

Fig. 4. Text lines with corresponding vertical projection profiles

4 Symbol Recognition

Tamil alphabet set contains 154 different symbols. This increases the recognition time and the complexity of the classifier, if single level classification is used. Hence, it is desirable to divide the characters into some clusters, so that the search space is reduced during recognition, which in turn results in lesser recognition time. Classification is based on spatial occupancy and on matras/extensions and recognition is based on orthonormal transform features.

5 Feature Extraction and Classification

5.1 Classification Based on Spatial Occupancy

This is the first level of clustering. The text lines of any Tamil text have three different segments as shown in Fig. 5. Depending upon the occupancy of these segments, the symbols are divided into one of the four different classes, defined as follows: Class 0 for symbols occupying segment 2, Class 1 for symbols occupying segments 1 & 2, Class 2 for symbols occupying 2 & 3 and Class 3 for symbols occupying all 3 segments.

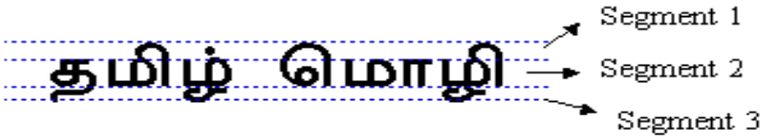


Fig. 5. The three distinct vertical segments of Tamil script.

5.2 Classification Based on Matras/Extensions

This level of classification is applied only to symbols of classes 1 and 2, which have upward matras and downward extensions. The classes are further divided into Groups, depending on the type of ascenders and descenders present in the character. This level of classification is feature based i.e. the feature vectors of the test symbol are compared with the feature vectors of the normalised training set. The features [10] used in this level are *second order geometric moments* and the classifier employed is the nearest neighbour classifier.

5.3 Recognition Based on Orthonormal Transform Features

In the third level, feature-based recognition is performed. For each of the groups, the symbol normalisation scheme is different. The dimensions of the feature vector are different for different groups, as their normalisation sizes are different. Truncated Discrete cosine transform (DCT) coefficients are used as features at this level of classification. DCT [6] is the most compact frequency-domain description. It is possible to reconstruct the image with a high degree of accuracy, even with very few

coefficients. This motivated us to use DCT of the image as a feature vector. Nearest neighbour classifier is used for the classification of the symbols.

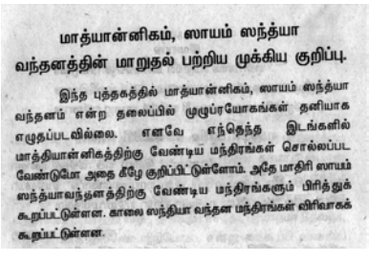
6 Training Set

In order to obtain good recognition accuracy, a vast database of training data exceeding 4000 samples was created. Each character has around 50 samples collected from various magazines, novels, etc. The database includes bold and italic characters along with few special symbols and numerals. Font sizes from 14 to 20 were handled while testing the system.

The training feature set contains the features obtained from normalised and thinned symbols and a label to identify the character. The features of the unknown symbol are compared with the features of the known symbols in the training set. The label of the training sample that closely matches with the test character is assigned to the latter. A symbol is declared unknown if its nearest neighbour is beyond a certain threshold distance.

7 Classification Results

The System is being tested on files taken from tamil magazines and novels which are scanned at 300 dpi (dots per inch).. Results on a set of 100 chosen samples are discussed below. About 40% of the samples were taken from the Training Set. These resulted in an accuracy of over 99%. The remaining samples (disjoint from the Training Set) resulted in an recognition accuracy of around 98%. Hence the average recognition accuracy stands at an appreciable 98%. The result obtained with one of the test documents is shown in Fig. 6. For this sample the total number of input samples were 351. There was one rejection (~ symbol) and two symbols were misclassified.


மாத்யான்னிகம், ஸாயம் ஸந்த்யா
வந்தனத்தின் மாறுதல் பற்றிய முக்கிய குறிப்பு.

இந்த புத்தகத்தில் மாத்யான்னிகம், ஸாயம் ஸந்த்யா வந்தனம் என்ற தலைப்பில் முழுப்ரயோகங்கள் தனியாக எழுதப்படவில்லை. எனவே எந்தெந்த இடங்களில் மாத்யான்னிகத்திற்கு வேண்டிய மந்திரங்கள் சொல்லப்பட வேண்டுமோ அதை கீழே குறிப்பிட்டுள்ளோம். அதே மாதிரி ஸாயம் ஸந்த்யாவந்தனத்திற்கு வேண்டிய மந்திரங்களும் பிரித்துக் கூறப்பட்டுள்ளன. காலை ஸந்தியா வந்தன மந்திரங்கள் விரிவாகக் கூறப்பட்டுள்ளன.

Fig. 6a. Original document

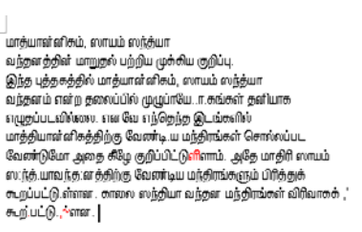

மாத்யான்விகம், ஸாயம் ஸந்த்யா
வந்தனத்தின் மாறுதல் பற்றிய முக்கிய குறிப்பு.
இந்த புத்தகத்தில் மாத்யான்விகம், ஸாயம் ஸந்த்யா
வந்தனம் என்ற தலைப்பில் முழுப்ரயோகங்கள் தனியாக
எழுதப்படவில்லை. எனவே எந்தெந்த இடங்களில்
மாத்யான்விகத்திற்கு வேண்டிய மந்திரங்கள் சொல்லப்பட
வேண்டுமோ அதை கீழே குறிப்பிட்டுள்ளோம். அதே மாதிரி ஸாயம்
ஸந்த்யாவந்தனத்திற்கு வேண்டிய மந்திரங்களும் பிரித்துக்
கூறப்பட்டுள்ளன. காலை ஸந்தியா வந்தன மந்திரங்கள் விரிவாகக்
கூறப்பட்டுள்ளன.

Fig. 6b. Recognised document

References

[1] Siromoney, G., Chandrashekaran, R., Chandrashekaran, M.: Computer recognition of printed Tamil characters, Vol. 10. Pattern Recognition, (1978) 243-247
[2] Sinha, R.M.K., Mahabala, H.: Computer recognition of printed Devnagari scripts, Vol. 9. IEEE trans. on Systems Man and Cybernetics, (1979) 435–441
[3] Pal, U., Choudhuri, B.B.: A Complete Printed Bangla OCR System, Vol. 31. Pattern Recognition, (1998)
[4] Kaushik Mahata, Ramakrishnan, A.G.: Precision Skew Detection through Principal Axis, Proc, Intern. Conf. on Multimedia Processing and Systems, Chennai, (2000)
[5] Chen, M., Ding, X.: A robust skew detection algorithm for grayscale document image, ICDAR, (1999) 617–620
[6] Gonzalez, R.C., Woods, R.E.: Digital Image Processing, Addison – Wesley Press, New York (1999)
[7] Dhanya, D., Ramakrishnan, A.G., Peeta Basa Pati.: Script Recognition in Bilingual Documents, Sadhana, (2002)
[8] Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis, John Wiley and Sons, New York (1973)
[9] Govindan, V.K., Shivaprasad, A.P.: Character recognition – a review, Pattern Recognition (1990) 671-683
[10] Trier, O., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition – a survey, Vol. 29. Pattern Recognition, (1996) 641–662