Simultaneous Recognition of Tamil and Roman Scripts

Dhanya D and A G Ramakrishnan

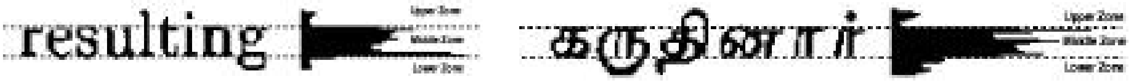
Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

INTRODUCTION

One of the important tasks in machine learning is the electronic reading of documents. All official documents, magazines and reports can be converted to electronic form using a high performance Optical Character Recogniser (OCR). In the Indian scenario, documents are often bilingual or multi-lingual in nature. English, being the link language in India, is used in most of the important official documents, reports, magazines and technical papers in addition to Tamil. Monolingual OCRs fail in such contexts and there is a need to extend the operation of current monolingual systems to bilingual ones. This paper describes one such system, which handles both Tamil and Roman scripts. Recognition of bilingual documents can be approached in two ways: (i) Recognition via script identification (ii) Bilingual approach.

RECOGNITION VIA SCRIPT IDENTIFICATION

In this approach, script recognition is first performed at the word level and this knowledge is used to identify the OCR to be employed. Individual OCRs have been developed for Tamil [1] as well as English and these could be used for further processing. Such an approach reduces the search space in the database and allows for the Roman and Tamil characters to be handled independently from each other.



(a) (b)

Fig. 1. The three zones of (a) an English word and (b) a Tamil word

In order to identify the script, the primary factor to be considered is the size of the textual block. Given the context of Indian scenario, where English words are inter-dispersed among Tamil words, it is prudent to identify the script at the word level. For script recognition, features are identified based on the following observations:

- Text lines of both Tamil and Roman scripts can be divided into three zones depending on the spatial occupancy of the characters : (i) Upper (ii) Middle and the (iii) Lower zones (see Fig. 1).
- All the upper case letters in Roman script extend into the upper zone while the lower case letters occupy the middle, lower and upper zones.
- Roman script has very few downward extensions (only for p, q, j and y), whereas the number of alphabets that extend downwards into the lower zone in Tamil is high.

- Roman script is dominated by vertical and slant strokes while Tamil script is dominated by horizontal and vertical strokes.
- The aspect ratio of Tamil characters is more as compared to the Roman set.

Based on the above observations, two sets of features have been explored for identifying the script of a word [2].

(i) The first set consists of spatial spread features, which quantifies the distribution of ascenders and descenders in a word. The ratios of the pixel densities in the lower and upper zones to that of the middle zone form the first two elements of this feature vector. The number of characters per unit width is less in Tamil as compared to English and this forms the third dimension of the feature vector.

(i) The second set of features makes use of the directional properties of various strokes in the two scripts. Each word is filtered by various directional filters and the energies of the filter responses form the features. Gabor filters with two frequencies and six different orientations are employed for this purpose, resulting in a 12-dimensional feature vector. Each filter has an angular bandwidth of 30°. English script has a higher response to 0° filter on account of the dominance of vertical strokes whereas Tamil script has a higher response to 90° filter due to the dominance of horizontal strokes.

The extracted feature vectors are classified using Nearest Neighbour and Support Vector Machine Classifiers. Once the script has been identified, then the corresponding OCR is used to recognise the character. The overall classification accuracy depends on the performance of the individual OCRs. The details of the performance of the above script identification schemes are discussed in the last Section.

BILINGUAL METHOD

In this approach, characters are handled in the same manner, irrespective of the script they belong to. In any classification problem, the feature dimension is very much dependent on the number of classes. As the number of classes increases, it is prudent to divide the classification problem and hence a hierarchical classification scheme is proposed.

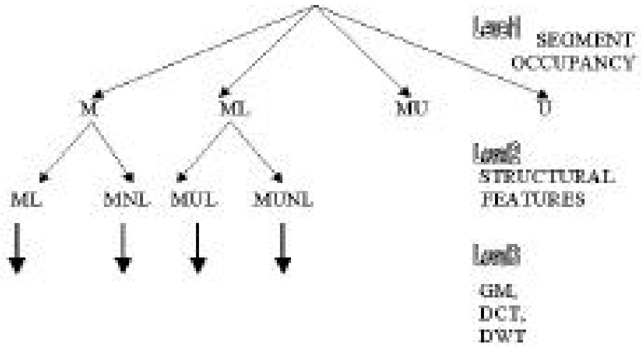


Fig. 2. Hierarchical feature extraction scheme

In this scheme, gross discriminations are made first, postponing the subtle ones to latter stages. In the first level, a character is classified as belonging to one of four groups depending on its spatial spread in the vertical direction. Thus characters are identified as occupying the Middle (M), Middle-Lower (ML), Middle-Upper (MU) or all three (A) zones. This division is based on the occupancy among the three segments (see Fig 2).

In the second level of classification, each character is normalized to a particular size depending on the zone it belongs to. The normalized characters are thinned and structural features such as the presence or absence of a loop (L) are extracted. This is performed for the groups M and MU only. A contour-tracing algorithm is used for this purpose.

In the final level, geometric moments and block DCT coefficients are explored as features with a view to the select the optimum set of features. In the case of geometric moments, each character is divided into sub-blocks of dimension 12x12. Second order geometric moments defined as

$$M_{mn} = \underset{x}{\_}\ \underset{y}{\_}\ f(x\ ,y)\ x^{p}\, y^{(2-q)} \qquad m, n = 0,1,2$$

where f(x,y) represents the input image, M_{mn} the (m,n)th order geometric moment, are extracted. These form the feature set. For discrete cosine transform based features, each character is subdivided into four blocks and DCT is performed on each block. DCT of an image I(x,y) is defined as

$$I'(x,y) = \overset{N-1}{\underset{i=0}{\_}}\ \ \overset{M-1}{\underset{j=0}{\_}}\ I(i,j)\cos((2i+1)x\text{ц}/N)\cos((2j+1)y\text{ц}/M)$$

where I'(x,y) is the transformed image. Since most of the energy is concentrated in the low frequency domain, only the low frequency coefficients are considered for classification. Both these methods undergo dimensionality reduction in order to avoid the peaking phenomenon. This is achieved via multiple discriminant analysis [3]. Classification is performed with nearest neighbour classifier, based on Euclidean distance. The hierarchical scheme for feature extraction is illustrated in Fig. 2.

SYSTEM DESCRIPTION

Figure 3 shows the overall flow diagram of the system. The input documents obtained from various magazines and journals are scanned at a resolution of 300 dpi. Binarization is the process of mapping a gray scale image into a binary image. For estimating the threshold required for binarization, the cdf of the image is computed. It is observed that the intensities above the mean value always correspond to the background. Based on this observation, these pixels are mapped to a particular value and cdf recomputed. Threshold value is given by the point at which the cdf crosses 0.5xMean . Proper selection of threshold takes care of denoising too, to some extent.

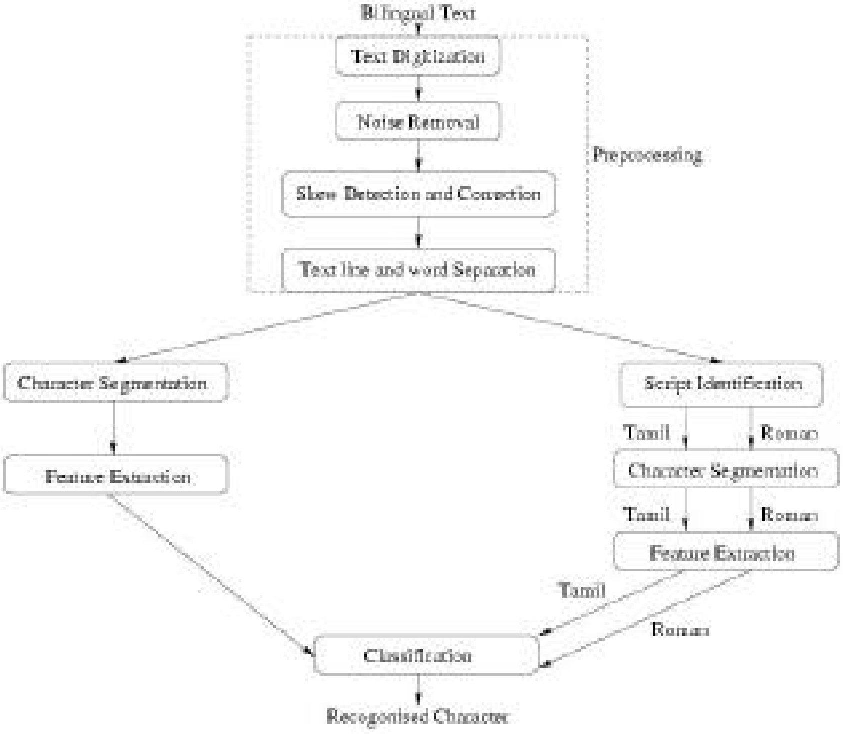


Fig 3 : Overall Flow diagram of the proposed system

Skew detection is based on the technique proposed in [4]. Correction is performed on the gray scale image in order to reduce quantization effects. Segmentation of characters is based on detection of valley points in the corresponding projection profiles. For script identification, features are extracted on a word level and classification is performed. Once the knowledge of the script is known, it is given to the corresponding monolingual OCR. For the combined database approach, feature extraction is performed on a character level.

RESULTS

Table 1 shows the recognition accuracies for the problem of script identification. The first set of features, though primitive, gives a reasonable performance. A good accuracy is obtained with directional filtering based features combined with SVM classifiers.

TABLE 1: RECOGNITION ACCURACIES (%) FOR SCRIPT IDENTIFICATION

	ACCURACIES WITH SPATIAL FEATURES			ACCURACIES WITH GABOR FEATURES		
	SVM	NN	K-NN	SVM	NN	K-NN
TAMIL	88.43	73.61	68.25	93.84	94.84	97.02
ENGLISH	87.76	71.23	84.72	98.21	88.88	84.92
TOTAL	88.09	72.42	76.49	96.03	91.86	90.97

Table 2 shows the results obtained for the combined character recognition approach. Table 3 shows the recognition accuracies based on DCT features. It can be seen that DCT based features give a better performance with reduced dimension.

TABLE 2: RECOGNITION ACCURACIES (%) WITH GEOMETRIC MOMENTS

CLASS	MNL	ML	MUNL	MUL	ML	MU	OVERALL
3^{RD} ORDER MOMENTS	96.09	96.06	92.52	93.69	92.98	93.27	94
REDUCED DIMENSION	97.39	96.21	94.36	94.32	96.22	95.01	95

TABLE 3: RECOGNITION ACCURACIES WITH DCT BASED FEATURES

CLASS	MNL	ML	MUNL	MUL	ML	MU	OVERALL
DCT BASED FEATURES	97.39	96.67	96.51	94.73	98.36	96.53	96
REDUCED DIMENSION	96.84	98.12	96.77	97.15	98.2	98	97.5

CONCLUSION

Two methods have been suggested for character recognition from bilingual text. The first one uses the knowledge of the script for identification. The performance depends on the efficiency of the individual OCRs. In the second approach, a hierarchical classification scheme has been suggested and two sets of features have been explored. It has been found that block-DCT based features, followed by a dimensionality reduction procedure give a very good performance.

BIBLOGRAPHY

1. K. Mahata and A. G. Ramakrishnan, "A complete OCR for printed Tamil text", Proc. Tamil Internet 2000, Singapore, July 22-24, 2000, pp. 165-170.
2. D. Dhanya and A. G. Ramakrishnan, "Script Indentification in Printed Bilingual Documents", in press, Sadhana, Special Issue on Document Analysis, October 2001.
3. R. O. Duda and P. E. Hart, Pattern Classifiacation and Scene Analysis, John Wiley and Sons, 1973.
4. Kaushik Mahata and A. G. Ramakrishnan, "Precision Skew Detection through Principal Axis", Proc. Intern. Conf. on Multimedia Processing and Systems, Chennai, Aug. 13-15, 2000, pp. 186-188.