Script Identification in Printed Bilingual Documents

D. Dhanya and A.G. Ramakrishnan

Department of Electrical Engineering,
Indian Institute of Science,
Bangalore 560 012, India
ramkiag@ee.iisc.ernet.in

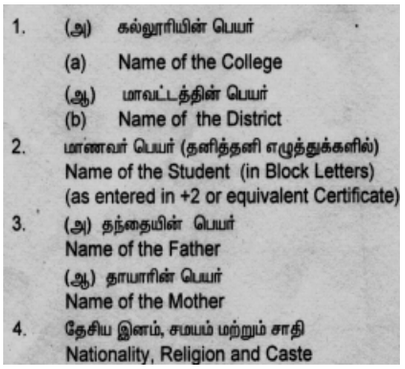
Abstract. Identification of script in multi-lingual documents is essential for many language dependent applications such as machine translation and optical character recognition. Techniques for script identification generally require large areas for operation so that sufficient information is available. Such assumption is nullified in Indian context, as there is an interspersion of words of two different scripts in most documents. In this paper, techniques to identify the script of a word are discussed. Two different approaches have been proposed and tested. The first method structures words into 3 distinct spatial zones and utilizes the information on the spatial spread of a word in upper and lower zones, together with the character density, in order to identify the script. The second technique analyzes the directional energy distribution of a word using Gabor filters with suitable frequencies and orientations. Words with various font styles and sizes have been used for the testing of the proposed algorithms and the results obtained are quite encouraging.

1 Introduction

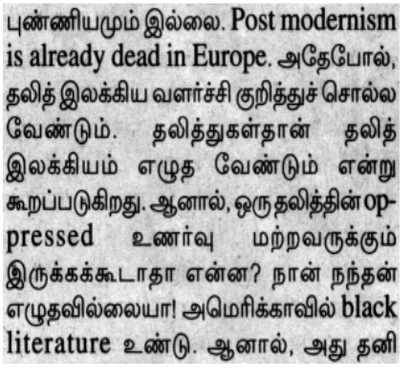
Multi-script documents are inevitable in countries housing a national language different from English. This effect is no less felt in India, where as many as 18 regional languages coexist. Many official documents, magazines and reports are bilingual in nature containing both regional language and English. Knowledge of the script is essential in many language dependent processes such as machine translation and OCR. The complexity of the problem of script identification depends on the disposition of the input documents. Recognition can be done on a block of text such as a paragraph, a line or a word. The features are to be selected depending on the size of input text blocks, to bring out the characteristics of the script. It is not advisable to work on individual characters, because one loses the whole advantage of script recognition, which is meant to reduce the search space for the OCR. Algorithms that work on text blocks of large size may or may not retain their performance when applied on a smaller block of text. The foremost deciding parameter for the algorithm to be used then is the size of the largest contiguous text block of any one of the scripts that one is always assured of being available in the given document. As shown in Fig. 1, in the Indian

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 13–24, 2002.

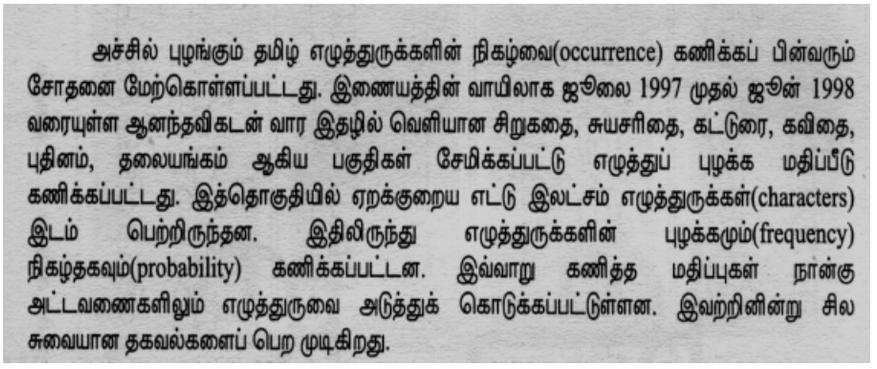
context, the bilingual documents contain single words of English interspersed in an otherwise Indian language text. The document 1 (a) is taken from a college application form; (b) from a weekly magazine and (c) from an International conference proceedings [1]. In order to be of general applicability then, script recognition needs to be performed at the word level.



(a)



(b)



(c)

Fig. 1. Typical bilingual documents (a) Official document (b) Magazine (c) Technical report

Among the work done in this area, Spitz et al. [2,3,4] have worked on textual paragraphs for recognizing Roman and Asian scripts. They have used spatial relationship of structural features of characters for differentiating Han and Latin based scripts. Asian scripts (Japanese, Korean and Chinese) are distinguished from Roman by a uniform vertical distribution of upward concavities. In the case of the above Asian scripts, the measure of optical density i.e. the number of ON-pixels per unit area is employed to distinguish one from the other. Hochberg

et al. [5] use cluster-based templates for script identification. They consider 13 different scripts including Devanagari, an Indian script. Their technique involves clustering of textual symbols (connected components) and creating a representative symbol or a template for each cluster. Identification is through the comparison of textual symbols of the test documents with the templates. This method necessitates a local approach in the sense that each connected component needs to be extracted for identifying the script. Wood et al. suggest a method based on Hough transform, morphological filtering and analysis of projection profile [6]. Though their work involves the global characteristics of the text, the results obtained are not encouraging.

Tan [7] has attempted a texture based approach to identify six different scripts - Roman, Persian, Chinese, Malayalam, Greek and Russian. The inputs for script recognition are textual blocks of size 128 x128 pixels, which, for the scanning resolution used by him, cover several lines of text. This method requires such image blocks containing text of single script. These blocks are filtered by 16 channel Gabor filters with an angular spacing of 11.25^o. The method has been tested for single fonts assuming font invariance within the same block. A recognition accuracy greater than 90% has been reported. However, the efficiency is reported to go down to 72% when multiple fonts are incorporated.

Tan et al. [8] have worked on three scripts - Latin, Chinese and Tamil. These scripts are used by the four official languages of Singapore. Their work is based on attributes like aspect ratio and distribution of upward concavities. The use of such primitive features necessitates long passages of input text for good performance. They report recognition accuracies above 94%.

Pal and Chaudhuri [9] have proposed a decision tree based method for recognizing the script of a line of text. They consider Roman, Bengali and Devanagari scripts. They have used projection profile besides statistical, topological and stroke based features. Initially, the Roman script is isolated from the rest by examining the presence of the headline , which connects the characters in a word. Devanagari is differentiated from Bangla by identifying the principal strokes [9]. In [10], they have extended their work to identification of the script from a given triplet. Here, they have dealt with many of the Indian scripts. Besides the headline, they have used some script dependent structural properties such as distribution of ascenders and descenders, position of vertical line in a text block, and the number of horizontal runs. Chaudhuri and Seth [11] have used the horizontal projection profile, Gabor transform and aspect ratio of connected components. They have handled Roman, Hindi, Telugu and Malayalam scripts. Their work involves identifying the connected components and convolving them with a six channel Gabor filter bank. The output is full-wave rectified and its standard deviation calculated. Their results vary from 85% for Hindi to 51% for Malayalam. Most of these works require large textual regions to achieve good performance. However, this necessity cannot be satisfied by most Indian documents, in which the script changes at the level of a word. Here, bilingual script recognition has been attempted to work at the word level. Each word is assumed to contain at least four patterns. Though quite a few number of English words

do not meet this requirement, our assumption is justified by the fact that the probability of finding such words, in a bilingual Tamil document, is quite low. In such a context, the above assumption guarantees high recognition accuracy.

2 Language Description

The spatial spread of the words formed by the scripts, as well as the orientation of the structural elements of the characters play a major role in our approach. A clear understanding of the properties of the associated scripts is essential for the design of an identifier system.

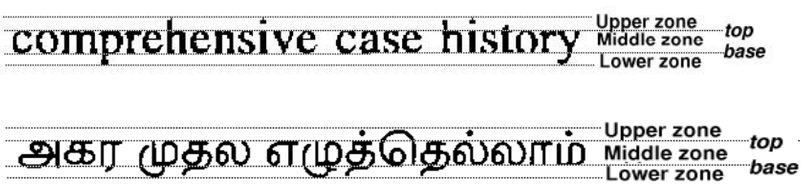


Fig. 2. Three distinct zones of (a) Roman script and (b) Tamil script

The various properties as analyzed are:

(1) Both Tamil and Roman characters (words) are structured into three distinct zones, viz. Upper, Middle and Lower, based on their occupancy in the vertical direction (2). For our discussion, we define the following terms: top line, base line, descenders and ascenders. We call the boundary that separates the upper and middle zones as the top line and the one that separates the middle and lower zones as the base line. The structures that extend into the lower and upper zones are called descenders and ascenders, respectively.
(2) Roman script has very few descenders (only in 'g', 'j', 'p', 'q' and 'y'), as compared to Tamil. The probability of lower zone occupancy is therefore less in Roman script. For example, in an analysis of 1000 words each of both scripts, it has been observed that 908 words of Tamil script (90%), and only 632 words (63%) of English have descenders.
(3) Roman alphabet contains more slant and vertical strokes as compared to Tamil, which has a dominance of horizontal and vertical strokes.
(4) The number of characters per unit area is generally less in Tamil.

3 Feature Extraction

Feature extraction aims at selecting features that maximize the distinction between the patterns. For the task at hand, those features that highlight the characteristic properties of the scripts have been chosen. As explained in Sec.2, the

relative distribution of ascenders and descenders in a word as well as the directional distribution of stroke elements of the alphabets differ for Tamil and Roman scripts. The spatial distribution of ascenders and descenders in a word can be quantified through the analysis of the projection profile of the word. The directional distribution of strokes can be extracted by looking at the energy distribution of the word in various directions.

3.1 Spatial Spread Features: Character Density and Zonal Pixel Concentration

It has been observed that the number of characters present per unit area of any Tamil word is generally less than that in English. Based on this observation, we define a feature, character density, as,

$$characterdensity = \frac{No.\ of\ Characters\ in\ a\ word}{Area\ of\ the\ bounding\ box} \ . \tag{1}$$

The analysis of the horizontal projection profile in the three zones of the words suggests use of zonal pixel concentration (ratio of number of ON-pixels in each zone to the total number of ON- pixels) as a feature for script recognition. Considering the top left most index as the origin, and denoting the horizontal profile (the row sum of the image) by P, the zone boundaries are defined as,

$$top = arg(max(P(y) - P(y-1))) \qquad \forall 0 <= y < H/2 \ . \tag{2}$$

$$base = arg(min(P(y) - P(y-1))) \qquad \forall H/2 >= y < H \ . \tag{3}$$

where H is the height of the word or line. Figs. 3(a) and 3(b) show the projection profiles of single English and Tamil words, respectively. The profiles show very sharp transitions at the zone boundaries, which are identified from the first difference of the profile as given by equations 2 and 3. If U, M and L represent the upper, middle and lower zones respectively, then

$$U = \{(x,y)|y < top\} \ . \tag{4}$$

$$M = \{(x,y)|top \leq y \leq base\} \ . \tag{5}$$

$$L = \{(x,y)|y > base\} \ . \tag{6}$$

where (x, y) are the image coordinates. Let $f(x, y)$ be the binary image, whose value is '1' for foreground pixels, and '0' for background pixels. Zonal pixel concentration is defined as,

$$PC_k = \frac{\sum_{(x,y)\epsilon k} f(x,y)}{\sum_{(x,y)} f(x,y)} \ . \tag{7}$$

where k is U, L.

The formation of the feature vector is as follows. The pixel concentrations in U and L zones are calculated and these form the first two elements of the feature

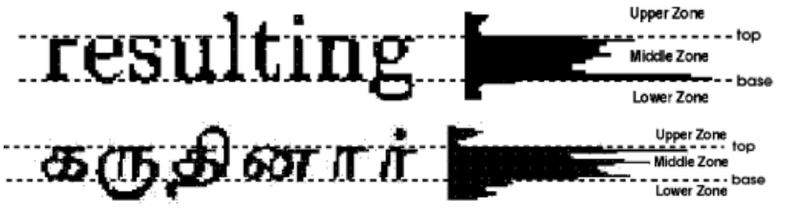


Fig. 3. The three distinct zones of (a) English and (b) Tamil words and their corresponding projection profiles.

vector. The character density forms the third dimension of the feature vector. Since only relative densities are used, there is no need for size normalization.

Figure 4 shows the scatter plot of the feature vectors for a typical bilingual Tamil and Roman document. There is clear distinction between the feature vectors of the two scripts. However, it can be seen that some vectors belonging to Tamil script fall near the cluster of English feature vectors. These are attributed to those sample words formed by characters having less downward extensions.

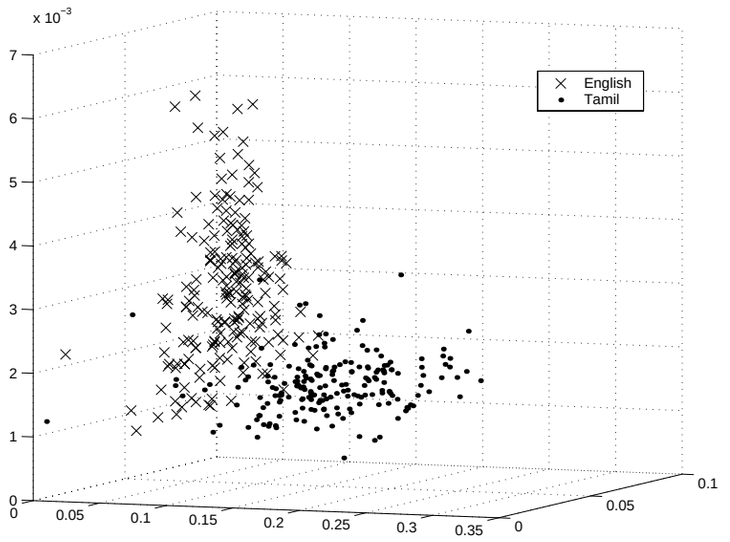


Fig. 4. Scatter plot of the spatial features : Upper zone pixel concentration Vs Lower zone concentration Vs No. of components per unit area

3.2 Directional Features: Gabor Filter Responses

The motivation for using directional features arose from the observation of the nature of strokes. The stroke information is effectively and inherently captured by the Human Visual System (HVS), the best-known pattern recognizer that identifies distinctive patterns through their orientation, repetition and complexity. Repetition is indicative of the frequency selectivity; orientation, of the directional sensitivity and complexity, of the type of pattern. Hence we attempted to have a feature extractor, which performs the same functions as HVS. Studies indicate that cells in primary visual cortex of human brain are tuned to specific frequencies with a bandwidth of one octave and orientations with an approximate bandwidth of 30^o each [12,13]. This type of organization in the brain, leading to a multi-resolution analysis, motivated us to use Gabor filters, which have been known to best model the HVS. These directional filters, with proper design parameters, are used to effectively capture the directional energy distribution of words.

A Gabor function is a Gaussian modulated sinusoid. A complex 2-D Gabor function with orientation θ and center frequency F is given by:

$$h(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\ exp\left\{-\frac{1}{2}\left[\frac{x^2}{\sigma_x^2}+\frac{y^2}{{\sigma_y}^2}\right]\right\} exp\{j2\pi F[xcos\theta + ysin\theta]\}\ . \quad (8)$$

The spatial spreads σ_x and σ_y of the Gaussian are given by:

$$\sigma_x = \frac{\sqrt{ln2}(2^{\Omega_F}+1)}{\sqrt{2}\pi F(2^{\Omega_F}-1)}\ . \quad (9)$$

$$\sigma_y = \frac{\sqrt{ln2}}{\sqrt{2}\pi F\tan(\Omega_\theta/2)}\ . \quad (10)$$

where Ω_F and Ω_θ are the frequency and angular bandwidths, respectively. Change of frequency and scaling of Gabor functions provide the parameters necessary to model the HVS. A filter bank, with both angular bandwidth and spacing set to 30^o, and the frequency spacing to one octave, closely models the HVS. With a circular Gaussian ($\sigma_x = \sigma_y$), we can obtain a variable spread (scale) that helps to capture information at various scales and orientations.

Figure 5 shows the filter bank designed to suit the purpose. Frequency spacing of one octave with two frequencies (0.25 and 0.50 cpi) is considered. For the formation of the feature vector, the word is thinned and filtered by the filter bank. Initially, two frequencies are specified and all possible directions with an angular bandwidth of 30^o, have been used. This leads to twelve feature coefficients. These are normalized by the norm of the feature vector. However, the analysis of the efficiency of individual feature element proved that not all coefficients are effective discriminators. Hence only a subset of these coefficients have been used as features.

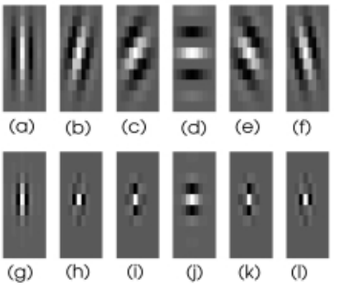


Fig. 5. Gabor filters: (a)-(f) F = 0.25 cpi and $\theta = 0^o$ to 150^o with angular spacing of 30^o; (g)-(l) F = 0.5 cpi and $\theta = 0^o$ to 150^o with angular spacing of 30^o

Feature selection for Directional Energy Coefficients. It is well known in Pattern Recognition literature that features should have small intraclass variance and a large interclass scatter. Several criterion functions have also been proposed in order to examine the clustering ability of the features. One such standard criterion function is the Fisher's ratio. This ratio is normally used in order to arrive at the transformation matrix for dimension reduction procedure. However, the same ratio can also be used for the purpose of feature selection.

For a two class problem, Fisher's ratio is as follows:

$$FR = \frac{var(X_1) + var(X_2)}{det(cov(X1, X2))} \ . \tag{11}$$

where X_i represents the feature vectors belonging to class i and $cov(.)$ represents the covariance. Figure 6 shows the ratio for all the 12 dimensions. It is apparent that features with small ratios are better discriminators. This is verified by comparing the performance of the complete 12-dimensional Gabor features with that of a subset of only 8 elements corresponding to lower ratio values.

4 Experiments and Results

The feature extraction techniques have been tested on a variety of documents obtained from various reports and magazines. The text-only input document is a gray scale image scanned with a resolution of 300 dpi. This is binarized using a two-stage process and deskewed to avoid errors due to tilt in text lines [14], [15]. Text lines and words are identified using the valleys in the profile in the corresponding direction. Segmented words are thinned and feature extraction is then performed on them. Thinning aids in a concise concentration of energy along particular directions.

The extracted features are classified using Support Vector Machines (SVM) [16], Nearest Neighbor (NN) and k-Nearest Neighbor (k-NN) classifiers. Euclidean metric is assumed. The value of k in the case of k-NN classifier is set at

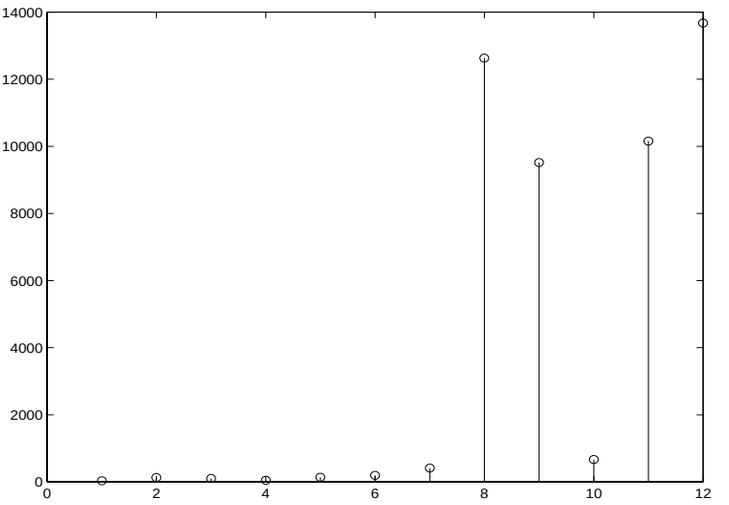


Fig. 6. Ratio for the twelve feature coefficients

30. We have used Gaussian kernel for the SVM classifier. The variance σ^2 for the Gaussian kernel is set to that of the reference data set. The results are tabulated in Table 1. The training and test patterns have 1008 samples each, consisting of equal number of Tamil and English words. Figure 7 shows some of the samples of various fonts used in the experiment.

Fig. 7. Sample words of various fonts used in the experiment.

Table 1. Results showing Recognition Accuracies (Tamil-English)

	% Accuracies with Spatial features			% Accuracies with 12 Gabor Responses			% Accuracies with 8 Gabor Responses		
	SVM	*NN*	*k-NN*	*SVM*	*NN*	*k-NN*	*SVM*	*NN*	*k-N*
Tamil	88.43	73.61	68.25	93.84	94.84	97.02	93.04	94.84	97.02
English	87.76	71.23	84.72	98.21	88.88	84.92	98.5	88.69	84.02
Total	88.09	72.42	76.49	96.03	91.86	90.97	95.77	91.76	90.52

Table 1 shows the results of script recognition using spatial spread and directional features with each of the classifiers. The lower efficiency of the method based on spatial spread features can be attributed to the presence of words with very few ascenders and descenders. The method based on the Gabor filters results in a superior performance, since it takes into consideration the general nature of the scripts rather than specific extensions. Thus, the presence or absence of a few strokes does not affect its performance. English script has a higher response to 0^o directional filter on account of the dominance of vertical strokes while Tamil script has a higher response to 90^o filter due to dominance of horizontal strokes. It is observed that the reduced set of Gabor features performs as well as the original set.

Among the works reported in the literature, Tan's approach [8] also uses Gabor filters. However, his work is based on large blocks of text. A recognition accuracy greater than 90% has been obtained using text containing a single font only. However, the efficiency has been reported to go down to 72% when multiple fonts are incorporated. Further, the reported results are based on a small test set of 10 samples each, for each script. On the other hand, our proposed approach works well with multiple fonts, achieving a good accuracy of above 90% and has been tested thoroughly on a set of 1008 samples each, for each script.

Pal and Chaudhuri [10] have used structural features, (principal strokes) and distribution of ascenders and descenders. Their work has given a good recognition accuracy of 97.7% for distinguishing among Tamil, Devanagari and Roman text lines. Chaudhury and Sheth's work [11], though uses Gabor filter based features, gives an accuracy of around 64% only. Better results are obtained (around 88%) using projection profile and height to width ratio. However, these methods operate at the line or paragraph level.

Our method works under the assumption that any word contains a minimum number of four connected components. The assumption is justified by the fact that the probability of occurrence of words with very few components is low. This assumption also eliminates symbols such as bullets and numerals. Difficulty is encountered while classifying Hindu-Arabic numerals since they are shared by both the scripts. Since most of the mono-script OCRs incorporate numerals also, this problem can be easily circumvented. Thus, irrespective of the script the numbers are classified into, they are taken care of by the respective OCRs.

The proposed method can be extended to other South Indian Dravidian languages as they too are quite distinct from Roman script. The algorithm was

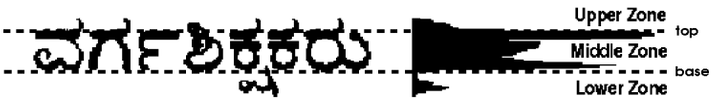


Fig. 8. Three distinct zones of Kannada and the corresponding projection profile

tested to isolate Kannada from Roman script. Figure 8 shows a typical word of Kannada script which also is structured into three zones. Also shown is its projection profile. The algorithms were tested on Kannada-Roman bilingual texts and the results are tabulated in Table 2. Kannada words have a uniform response to Gabor filters on account of their circular nature.

Table 2. Results showing Recognition Accuracies (Kannada-English)

	% Accuracies with Spatial features			% Accuracies with 12 Gabor Responses			% Accuracies with 8 Gabor Responses		
	SVM	*NN*	*k-NN*	*SVM*	*NN*	*k-NN*	*SVM*	*NN*	*k-NN*
Kannada	86.99	74.38	69.41	94.63	93.84	97.02	94.44	91.6	96.826
English	88.71	68.07	68.83	95.02	92.04	91.85	94.65	91.48	93.46
Total	87.85	71.22	69.12	94.83	92.94	94.44	94.53	94.43	92.84

5 Conclusion

Two different sets of features have been employed successfully for discriminating Tamil and English words. The first method uses the pixel concentration in the different zones and the average number of connected components per unit area in a word as features. Directional features, obtained as the responses of Gabor filters, are employed in the second approach. These features are observed to possess better discriminating capabilities than the spatial spread features. Experiments are conducted with documents covering a wide range of fonts. Accuracies as high as 96% have been obtained with SVM classifiers using directional energy features.

References

1. Proceedings of Tamil Internet (2000) 22-24 July, Singapore.
2. Spitz, A.L.: Determination of Script and Language Content of Document Images. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 235–245
3. Sibun, P., Spitz, A.L.: Natural Language Processing from Scanned Document Images. In: Proceedings of the Applied Natural Language Processing, Stuttgart (1994) 115–121

4. Nakayama, T., Spitz, A.L.: European Language Determination from Image. In: Proceedings of the International Conference on Document Analysis and Recognition, Japan (1993) 159–162
5. Hochberg, J., et al.: Automatic Script Identification from Images Using Cluster-Based Templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 176–181
6. Dang, L., et al.: Language Identification for Printed Text Independent of Segmentation. In: Proceedings of the International Conference on Image Processing. (1995) 428–431
7. Tan, C.L., et al.: Language Identification in Multi-lingual Documents. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 751–756
8. Tan, T.N.: Rotation Invariant Texture Features and their Use in Automatic Script Identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 751–756
9. Chaudhuri, B.B., Pal, U.: A complete Printed *bangla* OCR System. Pattern Recognition **31** (1998) 531–549
10. Chaudhuri, B.B., Pal, U.: Automatic Separation of Words in Multi-lingual Multi-script Indian Documents. In: Proceedings of the International Conference on Document Analysis and Recognition, Germany (1997) 576–579
11. Chaudhury, S., Sheth, R.: Trainable Script Identification Strategies for Indian languages. In: Proceedings of the International Conference on Document Analysis and Recognition, India (1999) 657–660
12. Hubel, D.H., Wiesel, T.N.: Receptive Fields and Functional Architecture in Two Non-striate Visual Areas 18 and 19 of the Cat. Journal of Neurophysiology **28** (1965) 229–289
13. Campbell, F.W., Kulikowski, J.J.: Orientational Selectivity of Human Visual System. Journal of Physiology **187** (1966) 437–445
14. Chen, Y.K., et al.: Skew Detection and Reconstruction Based on Maximization of Variance of Transition-Counts. Pattern Recognition **33** (2000) 195–208
15. Dhanya, D.: Bilingual OCR for Tamil and Roman Scripts. Master's thesis, Department of Electrical Engineering, Indian Institute of Science (2001)
16. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery **2** (1998) 955–974