Sādhanā Vol. 27, Part 1, February 2002, pp. 73–82. © Printed in India

Script identification in printed bilingual documents

D DHANYA, A G RAMAKRISHNAN* and PEETA BASA PATI

Biomedical Laboratory, Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, India
e-mail: {dhanya, ramkiag, pati}@ee.iisc.ernet.in

Abstract. Identification of the script of the text in multi-script documents is one of the important steps in the design of an OCR system for the analysis and recognition of the page. Much work has already been reported in this area relating to Roman, Arabic, Chinese, Korean and Japanese scripts. In the Indian context, though some results have been reported, the task is still at its infancy. In the work presented in this paper, a successful attempt has been made to identify the script, at the word level, in a bilingual document containing Roman and Tamil scripts. Two different approaches have been proposed and thoroughly tested. In the first method, words are divided into three distinct spatial zones. The spatial spread of a word in upper and lower zones, together with the character density, is used to identify the script. The second technique analyses 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 are quite encouraging.

Keywords. Script identification; printed bilingual documents; Tamil script.

1. Introduction

Most of the states in India have more than one language of communication. Thus, many official documents are multi-script in nature. Identification of the script is one of the challenging tasks facing a designer of an OCR system. Script identification makes the task of analysis and recognition of the text easier by suitably selecting the modalities of OCR. Quite a few results have already been reported in the literature, identifying the scripts in a multi-lingual and multi-script document dealing with Roman and other Oriental scripts such as Chinese, Korean and Japanese. A few attempts have already been made to isolate and identify the scripts of the texts in the case of multi-script Indian documents. Most of these attempts consider mono-script lines of text. In our work, we assume bilingual documents which require script recognition at word level.

Spitz and coworkers (Spitz & Nakayama 1993; Spitz & Sibun 1994; Spitz 1997) use the spatial relationship of structural features of characters for distinguishing between Han- and

*For correspondence

Latin-based scripts. Asian scripts (Japanese, Korean and Chinese) are differentiated from the Roman script by an 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* (1997) use cluster-based templates for script identification. They consider thirteen 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 comparison of textual symbols of test documents with those of the templates. However, the requirement of the extraction of connected components makes this feature a local one. Wood *et al* (1995) suggest a method based on the Hough transform, morphological filtering and analysis of projection profile. Though their work involves the global characteristics of the text, the results obtained are not encouraging.

Tan (1998) has suggested a method for identifying six different scripts using the texture-based approach. Textual blocks of 128×128 are taken and filtered by 16 channel Gabor filters with angular spacings of 11.25°. This method requires image blocks containing text of single script. The method has been tested for multiple fonts assuming font invariance within the same textual block. Roman, Persian, Chinese, Malayalam, Greek and Russian have been differentiated through this approach.

Pal & Chaudhuri (1997) have proposed a method based on a decision tree for recognizing the script of a line of text. They consider Roman, Bengali (Bangla) and Devanagari scripts. They have used the projection profile besides statistical, topological and stroke-based features. At the initial level, the Roman script is isolated from the other two by examining the presence of the *headline*[1] (*shirorekha*). Devanagari is differentiated from Bengali by identifying the principal strokes. They have extended their work (Pal & Choudhuri 1999) to the identification of the script from a given triplet. Here, they have dealt with almost all 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.

Chaudhury & Sheth (1999) have proposed techniques that use the horizontal projection profile, Gabor transform and the aspect ratio of connected components. They have handled Roman, Hindi, Telugu and Malayalam scripts.

Thus, all the reported studies accomplish script recognition either at the line level or at the paragraph level. To our knowledge, no one has yet successfully attempted to identify the script of a given word. Tan's (1998) work, having an approach similar to that of our proposed work, is applicable to mono-script blocks of text only. Also, considerable amount of textual regions are needed for discretion. Considering the nature of many of the documents in the Indian context (like application forms, examination question papers, technical reports, magazines and children's books), where the script could change at the word level, there is a need for a method to identify the script of a word. In such documents, it is unlikely to find enough textual blocks having the same font. Hence, there is a need for font-independent identification to be performed on a small region of text. In this paper, we make an attempt to separate the English and Tamil words present in a bilingual document containing various fonts from both the languages.

[1]Both Devanagari and Bengali (Bangla) script have a horizontal line at the top, known as *shirorekha*, which connects the characters in a word

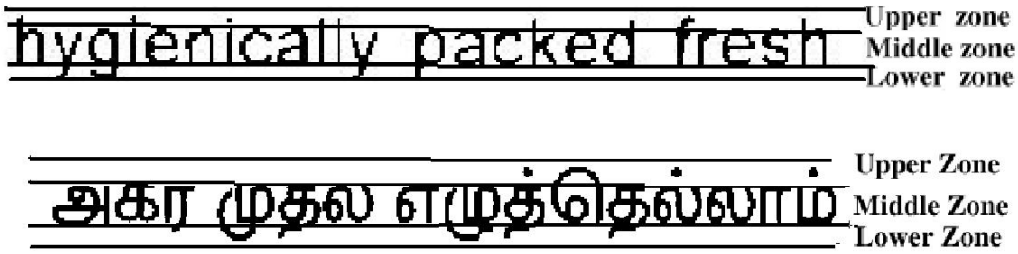

Figure 1. The three distinct zones of **(a)** Roman script and **(b)** Tamil script.

The identification is performed at the word level. Each word is assumed to contain at least four patterns. Though quite a few 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 with inter-dispersed English words, is quite low. In such a context, the above assumption guarantees high recognition accuracy.

2. Characteristics of the scripts

In our approach for the identification of the script of a word, the features that play a major role are the spatial spread of the words formed by the scripts and the direction of orientation of the structural elements of the characters in the word. So, a brief description of the properties of the associated scripts is in order, for the clear understanding that leads to the proper design of an identifier system. Some of the relevant properties are:

(1) The words of both the scripts, Roman and Tamil, can be divided into 3 distinct zones (see figure 1). They are: (i) upper zone, (ii) middle zone, and (iii) lower zone.
(2) The Roman script has 26 each of upper and lower case characters. In addition, there are some special symbols and numerals.
(3) While the capital letters of the Roman script occupy the middle and the upper zones, most of the lower case characters have a spatial spread that covers only the middle zone or the middle and the upper zones.
(4) Apart from some special symbols (such as ‘,’), only the characters ‘g’, ‘j’, ‘p’, ‘q’ and ‘y’ spread spatially to the lower zone. These being few in number, the probability of their occurrence is small.
(5) The structure of the Roman alphabet contains more vertical and slant strokes.
(6) The Tamil script has 12 vowels, 18 consonants and 6 special characters.
(7) Combinations of consonants with vowels give rise to new symbols or result in modified symbols. Hence a set of 262 symbols exists in the Tamil script.
(8) The modifier symbols occupy specific positions around the base characters. While the modifiers that get added on the left or the right side remain disjoint from the base character, the modifier symbols that are added either at the top or the bottom get connected to the base character and spread to the upper and the lower zones respectively.
(9) There is a dominance of horizontal and vertical strokes in the Tamil script.
(10) The *aspect ratio*[2] of the Tamil characters is, *generally*, more than the aspect ratio of the Roman characters.

[2]The ratio between the width and the height of the bounding box of a character

3. Feature extraction

Features are the representative measures of a signal, which distinguish it from other signals. Here we have tried to select features that maximize the distinction between Tamil and English words. The analysis of the horizontal projection profile of a word provides information on the spatial spread of the ascenders and descenders in the word. This distribution is script-dependent and the knowledge of the nature of spread aids in identifying the script. Another distinguishing feature is the energy distribution in the word in various directions, which could be different for different scripts. In this paper, we have proposed two methods for discriminating between Tamil and English words, based on the above two properties.

3.1 *Spatial spread features: Zonal pixel concentration and character density*

Analysis of the horizontal projection profiles of English and Tamil words reveals the following facts regarding the pixel concentration (i.e. the percentage of ON-pixels) in the three zones.

(i) Tamil words have more pixel concentration in the lower zone than English words.
(ii) The number of characters present per unit area in Tamil words is generally less than that in English words.

The above facts encourage us to employ them as features for distinguishing between Tamil and English words. Accordingly, the ratio of the number of ON-pixels in the lower zone to the total number of ON-pixels and the number of characters per unit area form the feature vector. Since the ratios of ON-pixels and the relative number of characters determine the final outcome, size normalization of words is not attempted.

Figures 2a and b show a word each of English and Tamil, and their corresponding projection profiles. At the zone boundaries, the profile shows a very sharp transition. The first difference of the profile gives a maximum at the junction of upper and middle zones, and a minimum at the junction between the middle and lower zones. It can be observed that the percentage of pixels that occupy the lower zone in the Tamil word is more than its English counterpart.

3.2 *Directional features: Gabor filter responses*

The orientations of the structural elements of Tamil and English scripts are exploited for differentiating one script from the other. A careful observation of the words of these scripts

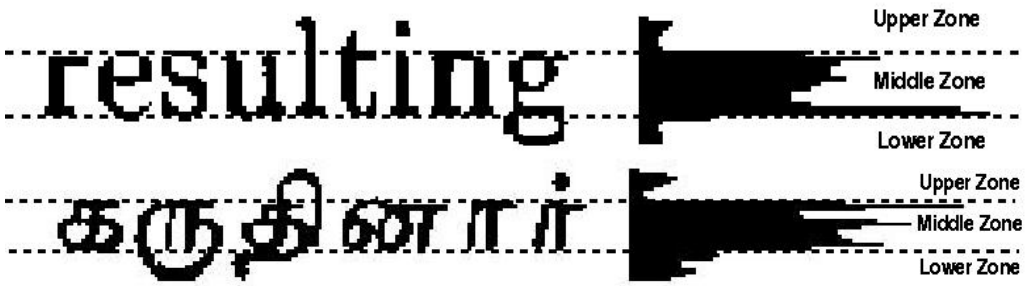

Figure 2. The three distinct zones of **(a)** an English word and **(b)** a Tamil word, and their corresponding projection profiles.

reveals the fact that Tamil script has more horizontal lines and strokes while English has more slant strokes. This motivates us to select a directional feature extractor, which can effectively capture the concentration of energies in various directions. Gabor filters are reported to accomplish this task efficiently (Clausi & Jernigan 2000). It is known that the human visual system (HVS) is sensitive to the spatial frequency as well as to specific orientations with an approximate angular bandwidth of 30° (Hubel & Wiesel 1965; Campbell & Kulikowski 1966). Gabor filters, when modelled to closely resemble the HVS, are capable of providing multi-resolution analysis and hence can be fruitfully used to extract directional features.

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

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

The spatial spreads σ_x and σ_y of the Gaussian, in the x and y directions, are given by:

$$\sigma_x = \sqrt{\ln 2}(2^{\Omega_F} + 1)/(\sqrt{2}\pi F(2^{\Omega_F} - 1)), \tag{2}$$

$$\sigma_y = \sqrt{\ln 2}/(\sqrt{2}\pi F \tan(\Omega_\theta/2)), \tag{3}$$

where Ω_F and Ω_θ are the frequency and the angular bandwidth, respectively. Change of frequency and the scaling of Gabor functions provide the necessary parameters to model the HVS. A filter bank, with both angular bandwidth and spacing set to 30°, 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), which is helpful in capturing information at various scales and orientations.

The test words are processed with filters designed for various orientations and frequencies. The angular spacing of 30° provides six different angles for each frequency (see figure 3). The frequency spacing is one octave and two such frequencies are considered (0.25 and 0.50

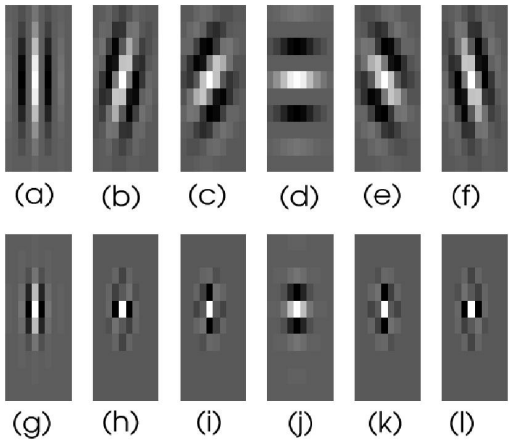

Figure 3. Gabor filters: **(a)-(f)** $F = 0.25$ cpi and $\theta = 0°$ to 150°, with angular spacing of 30°; **(g)-(l)** $F = 0.5$ cpi and $\theta = 0°$ to 150°, with angular spacing of 30°.

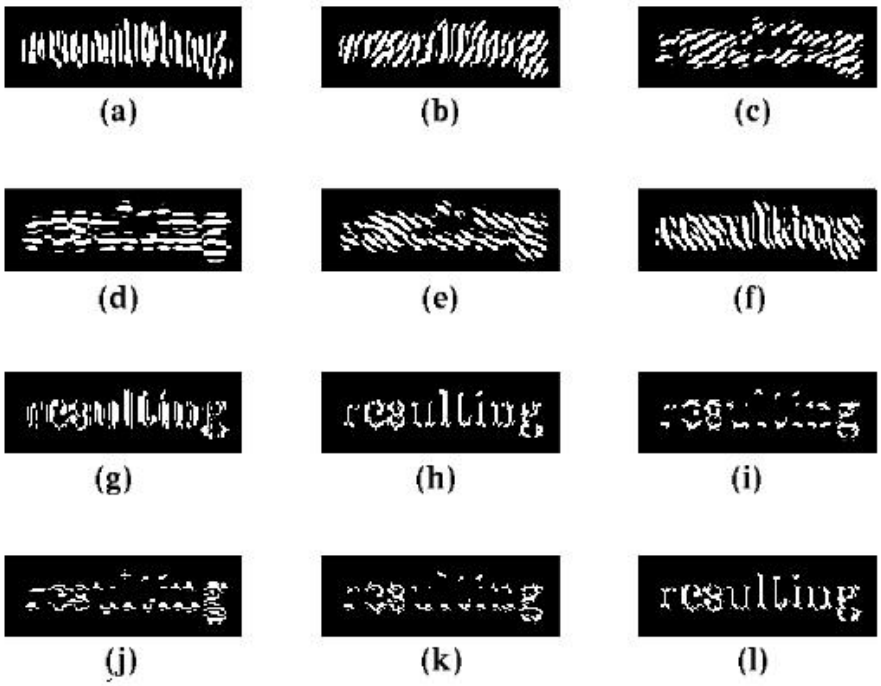

Figure 4. Gabor filter responses for the English word in figure 2a.

cpi). The filtered images contain information about the script at various scales and directions (see figures 4 and 5). The energies of the responses are good measures of the characteristics of the script. The filter responses for the two scripts under consideration are found to be distinct. The differences between the responses for English and Tamil words are clearly visible in the above mentioned figures. A feature vector of dimension twelve is constructed from the energy content of these responses.

4. System description

Figure 6 shows the general block diagram of the system. The input is a gray scale image obtained by scanning the document at a resolution of 300 dpi. The input document is assumed to contain only text and is thus free from graphics, figures, maps and tables. Binarization of the input document is performed using a global threshold. The pixel distribution of the input image is the sum of two Gaussian densities. The means represent the average foreground and background intensities. The threshold for binarization is chosen midway between the two mean positions in the distribution. The skew introduced during the process of scanning is detected using a two-stage process proposed by Mahata (2000) and Ramakrishnan & Mahata (2000). Initially, the input image is scaled down, run length smoothed and intermediate lines are located midway between the smoothed lines. Hough transform applied to the intermediate lines gives the approximate skew. The exact value of skew is detected in the next stage through principal component analysis. While skew detection is performed on the binarized document, correction, which involves rotating the image in the appropriate direction, is performed on the gray scale image to reduce the quantization effects. Bilinear interpolation is employed for this purpose.

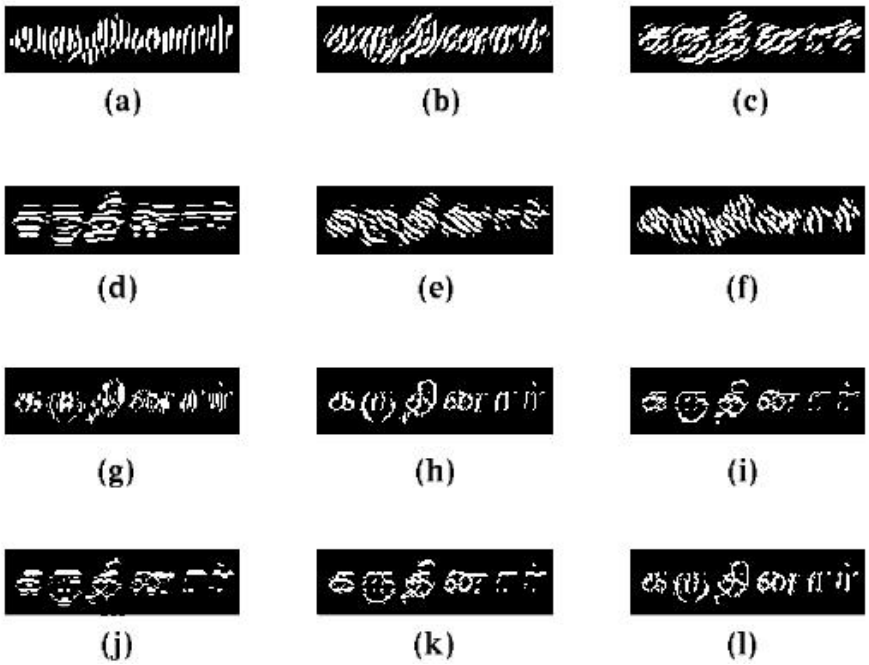

Figure 5. Gabor filter responses for the Tamil word in figure 2b.

To identify the textual blocks of interest, the skew corrected binary document is first segmented into lines, and then, into words. Line segmentation is performed by identifying the valley points in the projection profile taken along the rows of the image. Similarly, the valleys in the profile of the line taken in the vertical direction mark the word boundaries. These separated words are now the 'textual blocks' taken for feature extraction.

Two sets of features are extracted as described in § 3. In the first set, which contains the spacial spread features, the first two elements are formed by the ratios of the number of ON-pixels in the upper and lower zones with respect to the total ON-pixels of the word. Also the number of connected components per unit area is found out and is considered as the third feature element.

For the feature vector based on the directional energy concentration, the word is thinned and filtered by the twelve Gabor filters (see figure 3) with parameters described in § 3.2. The energy content of the filter responses form the feature elements.

The words are classified into one of the scripts using the extracted features employing various classifiers such as Support Vector Machines (SVM) (Burges 1998), nearest neighbour

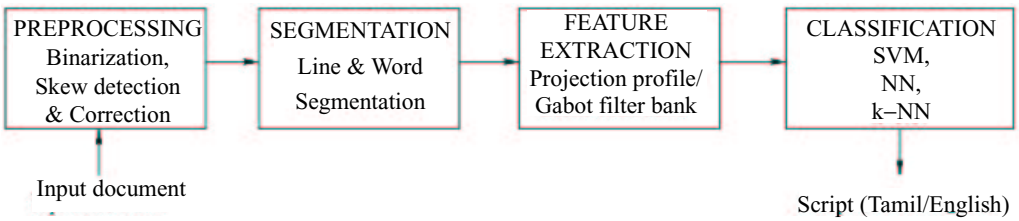

Figure 6. Block schematic of the script identification system.

(NN) and k-nearest neighbour (k-NN) classifiers. The software employed for SVM classifier is SVMTorch (Collobert & Bengio 2000). A Gaussian kernel with a variance σ^2 set to that of the reference data set is used. For the classifiers based on distance, Euclidean metric is assumed. For the k-NN classifier, various values of k are tried. It has been noticed that the performance remains the same for values of k ranging from 2% to 10% of the number of training samples. We have set the value at 30 (3%).

5. Results and discussion

Inputs were obtained from various magazines, newspapers and other such documents containing variable font styles and sizes. A scanning resolution of 300 dpi is employed for digitization of all the documents. The training and test patterns have 1008 samples each, consisting of equal number of Tamil and English words. Figure 7 shows an example of a typical bilingual document used in our work. One can clearly observe the inter-dispersion of isolated English words in the Tamil text. It can also be seen that the text block available for processing is very small. As compared to this, most of the reported studies in the literature work on larger blocks of text for their classification. Thus, difficulty arises in identifying the script of these words. Figure 8 shows some sample words of various fonts of both Tamil and Roman script used in the experiment.

Table 1 shows the results of script recognition using spatial spread and directional features with each of the classifiers. The first method, based on spatial spread features, though primitive, entails a fair performance. The lower efficiency can be attributed to the presence of words with very few ascenders and descenders. However, it is observed that the method based on the Gabor filter results in a superior performance. Since this method takes into consideration the general nature of the scripts rather than specific extensions, the presence or absence of a few strokes does not affect its performance. The English script, on account of the dominance of vertical strokes, gives a higher response to 0° filter, while the Tamil script has a higher response to 90° filter on account of the dominance of horizontal strokes.

Among the works reported in the literature, Tan's (1998) approach also uses features based on Gabor filters. However, his work is based on mono-script blocks of text. A recognition accuracy greater than 90% has been reported in his work, 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

அச்சில் புழங்கும் தமிழ் எழுத்துருக்களின் நிகழ்வை(occurrence) கணிக்கப் பின்வரும் சோதனை மேற்கொள்ளப்பட்டது. இணையத்தின் வாயிலாக ஜூலை 1997 முதல் ஜூன் 1998 வரையுள்ள ஆனந்தவிகடன் வார இதழில் வெளியான சிறுகதை, சுயசரிதை, கட்டுரை, கவிதை, புதினம், தலையங்கம் ஆகிய பகுதிகள் சேமிக்கப்பட்டு எழுத்துப் புழக்க மதிப்பீடு கணிக்கப்பட்டது. இத்தொகுதியில் ஏறக்குறைய எட்டு இலட்சம் எழுத்துருக்கள்(characters) இடம் பெற்றிருந்தன. இதிலிருந்து எழுத்துருக்களின் புழக்கமும்(frequency) நிகழ்தகவும்(probability) கணிக்கப்பட்டன. இவ்வாறு கணித்த மதிப்புகள் நான்கு அட்டவணைகளிலும் எழுத்துருவை அடுத்துக் கொடுக்கப்பட்டுள்ளன. இவற்றினின்று சில சுவையான தகவல்களைப் பெற முடிகிறது.

Figure 7. Example of a bilingual document containing both Tamil and Roman scripts.

Figure 8. Sample words of various fonts used in the experiment.

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 & Chaudhuri (1999) have used structural features for their task. Their work is based on identification of 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 scripts. However, such a recognition accuracy is obtained only at the line level.

Chaudhury & Sheth's work (1999), though using 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, both of these methods operate at the 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 Arabic numerals. However, this does not result in any practical problem, since the Tamil script also uses Arabic numerals only for representing numbers. 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.

Table 1. Results showing recognition accuracies.

	% Accuracies with spatial features			% Accuracies with Gabor filter responses		
Language	SVM	NN	*k*-NN	SVM	NN	*k*-NN
Tamil	86.70	73.61	68.25	93.84	94.84	97.02
English	88.49	71.23	84.72	98.21	88.88	84.92
Total	88.39	72.42	76.88	96.03	91.86	90.02

6. Conclusions

Two different sets of features have been employed successfully for discriminating between Tamil and English words. The first method takes the pixel concentration in the different zones and the average number of connected components per unit area in a word into consideration for forming the feature vector. Directional features are the responses of Gabor filters, each filter representing a particular orientation and frequency. The energy contents of these filtered responses are observed to possess good discriminating capabilities. Experiments have been conducted with documents covering a wide range of fonts, and encouraging results have been obtained.

References

Burges C J C 1998 A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* 2: 955-974

Campbell F W, Kulikowski J J 1966 Orientational selectivity of human visual system. *J. Physiol.* 187: 437–445

Chaudhury S, Sheth R 1999 Trainable script identification strategies for Indian languages. In *Proc. Int. Conf. on Document Analysis and Recognition* (IEEE Comput. Soc. Press) pp 657–660

Clausi D, Jernigan M E 2000 Designing Gabor filters for optimal texture separability. *Pattern Recogn.* 33: 1835–1849

Collobert R, Bengio S 2000 Support vector machines for large-scale regression problem. Technical Report, Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland

Hochberg P J, Kerns L, Thomas T 1997 Automatic script identification from images using cluster-based templates. *IEEE Trans. Pattern Anal. Machine Intell.* 19: 176–181

Hubel D H, Wiesel T N 1965 Receptive fields and functional architecture in two non-striate visual areas 18 and 19 of the cat. *J. Neurophysiol.* 28: 229–289

Mahata K 2000 *Optical character recognition for printed Tamil script*. Master's thesis, Department of Electrical Communication Engineering, Indian Institute of Science Bangalore

Pal U, Chaudhuri B B 1997 Automatic separation of words in multi-lingual multiscript Indian documents. In *Proc. Int. Conf. on Document Analysis and Recognition* (IEEE Comput. Soc. Press) pp 576–579

Pal U, Chaudhuri B B 1999 Script line separation from Indian multiscript document. In *Proc. Int. Conf. on Document Analysis and Recognition* (IEEE Comput. Soc. Press) pp 406–409

Ramakrishnan A G, Mahata K 2000 A complete OCR for Tamil printed text. In *Proc. Tamil Internet 2000*, Singapore, pp 165–170

Spitz A K 1997 Determination of script and language content of document images. *IEEE Trans. Pattern Anal. Machine Intell.* 19: 235–245

Spitz A L, Nakayama T 1993 European language determination from image. In *Proc. Int. Conf. on Document Analysis and Recognition* (IEEE Comput. Soc. Press) pp 159–162

Spitz A L, Sibun P 1994 Natural language processing from scanned document images. In *Proc. Applied Natural Language Processing*, Stuttgart, pp 115–121

Tan T N 1998 Rotation invariant texture features and their use in automatic script identification. *IEEE Trans. Pattern Anal. Machine Intell.* 20: 751–756

Wood S L, Yao X, Krishnamurthi K, Dang L 1995 Language identification for printed text independent of segmentation. In *Proc. Int. Conf. Image Processing*, pp 428–431