Script Identification in Printed Bilingual Documents

D. Dhanya and A. G. Ramakrishnan
Department of Electrical Engineering
Indian Institute of Science, Bangalore, 560 012
e-mail: ramkiag@ee.iisc.ernet.in

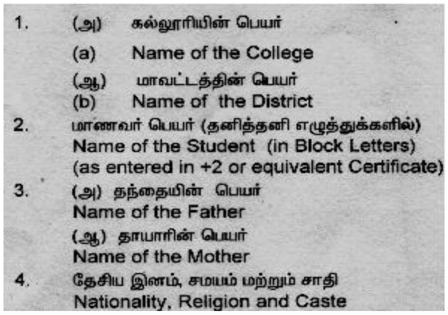
Abstract

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 contexts, 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 utilises 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

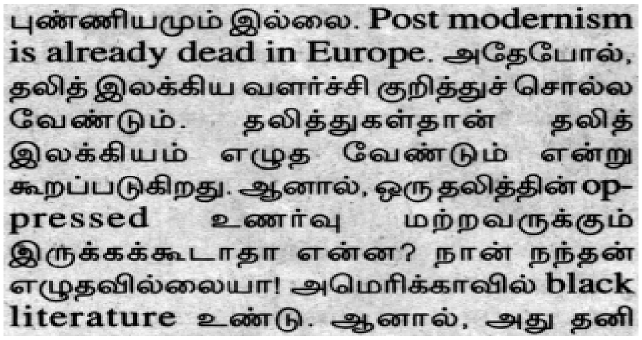
Multi-script documents are inevitable in countries housing a national language different from English. This effect is no less felt in India, where as many as 18 regional languages coexist. Many official documents, magazines and reports are bilingual in nature containing both regional language as well as 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 it can also be done on 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 (a), (b) and (c), in the Indian context, more often than not, the bilingual documents contain single words of English interspersed in an otherwise Indian language text. In order to be of general applicability then, script recognition needs to be performed at the word level.



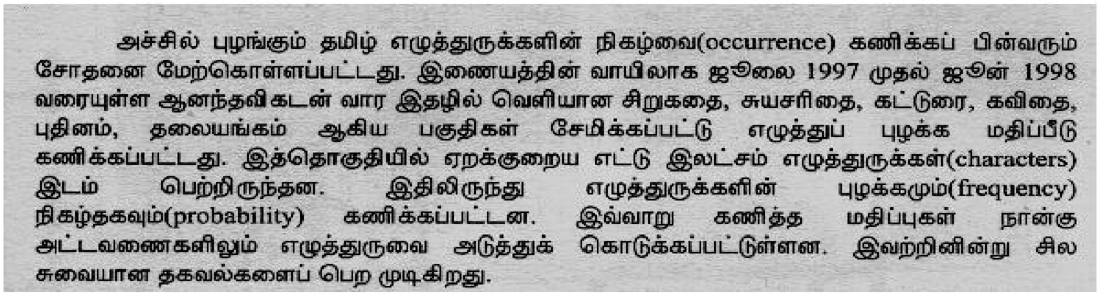
1. (அ) கல்லூரியின் பெயர்
(a) Name of the College
(ஆ) மாவட்டத்தின் பெயர்
(b) Name of the District
2. மாணவர் பெயர் (தனித்தனி எழுத்துக்களில்)
Name of the Student (in Block Letters)
(as entered in +2 or equivalent Certificate)
3. (அ) தந்தையின் பெயர்
Name of the Father
(ஆ) தாயாரின் பெயர்
Name of the Mother
4. தேசிய இனம், சமயம் மற்றும் சாதி
Nationality, Religion and Caste

(a)



புண்ணியமும் இல்லை. Post modernism is already dead in Europe. அதேபோல், தலித் இலக்கிய வளர்ச்சி குறித்துச் சொல்ல வேண்டும். தலித்துகள்தான் தலித் இலக்கியம் எழுத வேண்டும் என்று கூறப்படுகிறது. ஆனால், ஒரு தலித்தின் oppressed உணர்வு மற்றவருக்கும் இருக்கக்கூடாதா என்ன? நான் நந்தன் எழுதவில்லையா! அமெரிக்காவில் black literature உண்டு. ஆனால், அது தனி

(b)



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.

(c)

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

Among the work done in this area, Spitz *et al.* [1, 2, 3] 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.* [4] 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 [5]. Though their work involves the global characteristics of the text, the results obtained are not encouraging.

Tan [6] 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°. 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.* [7] 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 [8] have proposed a decision tree based method for recognising 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. At the initial level, the Roman script is isolated from the other two by examining the presence of the **headline**[1], which connects the characters in a word.

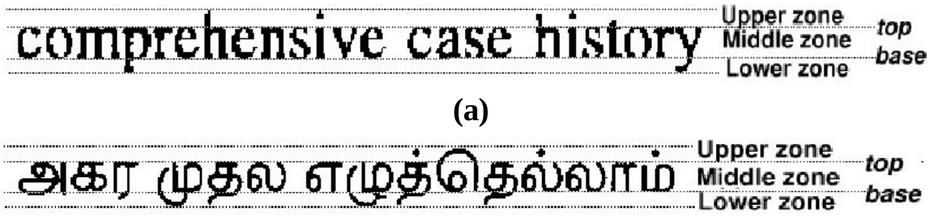
[1] Both Devanagari and Bangla script have a horizontal line known as ***shirorekha***.

Devanagari is differentiated from Bangla by identifying the principal strokes [8]. In [9], they have extended their work to 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. Chaudhuri and Seth [10] have proposed techniques using analysis of 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 85.24% for Hindi to 51.25% for Malayalam.

Most of these works require large textual regions to achieve good recognition accuracy. However, this necessity cannot be satisfied by most of the official Indian documents, technical journals, magazines, *etc.*, in which the script changes at the level of a word. In our work, bilingual script recognition techniques have been attempted that 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 with inter-dispersed English words, 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 direction of orientation of the structural elements of the characters play a major role in our approach for identification of the script of the word. So, a brief description of the properties of the associated scripts is essential for the clear understanding and design of an identifier system. The various properties as analyzed are:



(a)

(b)

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

1) Both Tamil and Roman characters (words) are structured into three distinct zones, *viz.* Upper, Middle and Lower, based on the occupancy of the characters in the vertical direction (Ref. Fig.2). For convenience in our discussion, we define the following terms: *top* line, *bottom* line, descenders and ascenders. We call the boundary that separates the top and middle zones as the *top* line, while the line that separates the middle and lower zones is called the *base* line. The structures that extend into the lower zone are called descenders, while those that extend into the upper zone are called ascenders.

2) Roman script has very few descenders (only in ‘g’, ‘j’, ‘p’, ‘q’ and ‘y’), while Tamil script contains many alphabets with descenders. The probability of lower zone occupancy is therefore less in Roman script than in Tamil. For example, in an analysis of 1000 words each of both scripts, it has been observed that 908 words of Tamil script have descenders (approximately 90%), while only 632 words of English have descenders (approximately 63%).

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 [2] present in Tamil words is generally less than that in English words.

3. Feature Extraction

Features are the representative measures of a signal, which distinguish it from other signals. Feature extraction aims at selecting features that maximize the distinction between the patterns. For the task at hand, those features that highlight the characteristic property 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. Hence, these attributes could be used as features for distinguishing between these two 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 words. Based on this observation, we define a feature, *character density,* as,

$$character\ density = \frac{No\ of\ characters\ in\ a\ word}{Area\ of\ bounding\ box\ of\ the\ word} \tag{1}$$

The analysis of the horizontal projection profile of the words in the three zones suggests consideration of zonal pixel concentration (the ratio of the 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,

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

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

[2] Characters per unit area = Number of characters in a word/Area of the bounding box of the word

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

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

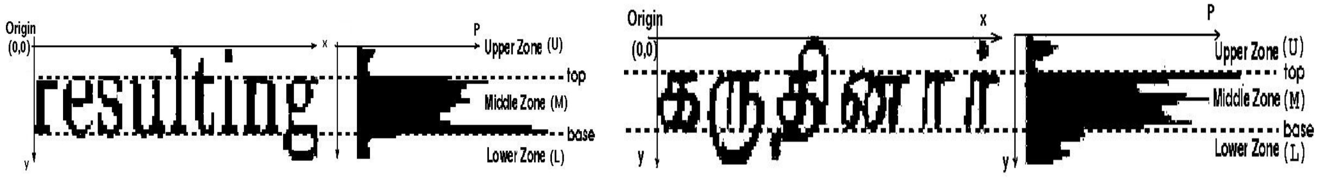
$$L = \{(x,y) \mid y > base\} \tag{5}$$

$$M = \{(x,y) \mid top \leq y \leq base\} \tag{6}$$

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,

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

where *k* is *U, L.*



(a) (b)

Fig. 3: Projection profiles. (a) An English word. (b) A Tamil word.

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 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 document containing both Tamil and Roman scripts. 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.

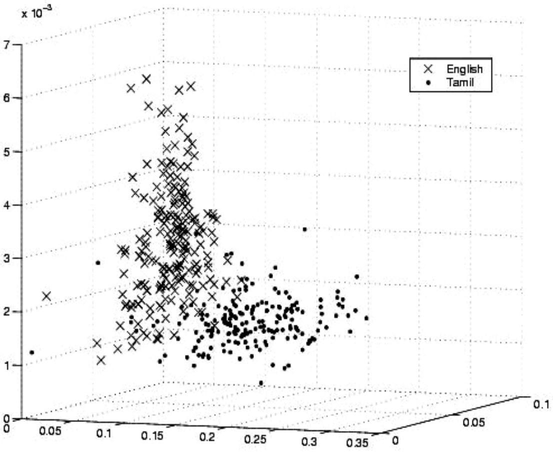


Fig 4: Scatter plot of the spatial features : Upper zone pixel concentration Vs Lower zone pixel concentration Vs Character density

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° each [11]. 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:

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

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

$$\sigma_x = \frac{\sqrt{\ln 2}\left(2^{\Omega_F}+1\right)}{\sqrt{2}\pi F\left(2^{\Omega_F}-1\right)} \qquad (9)$$

$$\sigma_y = \frac{\sqrt{\ln 2}}{\sqrt{2}\pi F \tan(\Omega_\theta / 2)} \tag{10}$$

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°, 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.

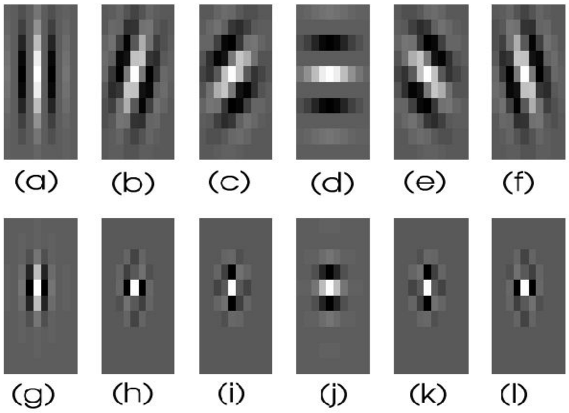


Fig. 5: Gabor filters: (a)-(f) F = 0.25 cpi and θ = 0° to 150°with angular spacing of 30°; (g)-(l) F = 0.5 cpi and θ =0° to 150° with angular spacing of 30°

Figure 5 shows the filter bank designed with two frequencies (0.25 and 0.50 cpi[3]) and a spacing of one octave. For the formation of the feature vector, the test word is thinned and filtered by the filter bank. With two frequencies and an angular bandwidth of 30°, twelve features are obtained. These are normalized to unit length. Figure 6 shows the average feature vectors for Tamil and Roman scripts.

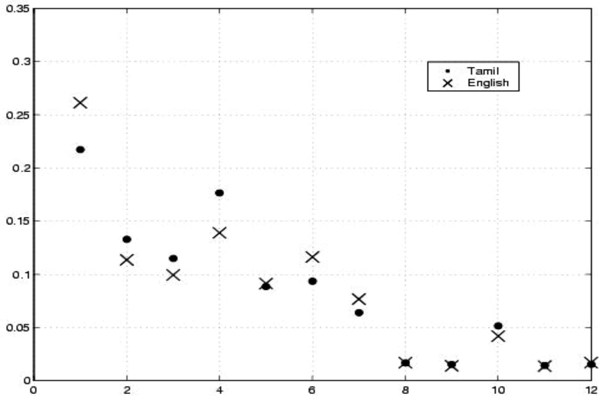


Fig. 6: Average energy responses for the twelve Gabor filters

[3] cpi: cycles per image width (height)

4. Experiment and results

The feature extraction techniques have been tested on variety of documents obtained from various reports and magazines. The input document is a gray scale image scanned with a resolution of 300 dpi. It is assumed to be free from graphics, figures or maps. The input document so obtained is binarized using the two-stage process discussed in [11]. The skew introduced during the process of scanning is detected and corrected using the method discussed in [12, 11]. 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) [13], 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 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 tables 1 & 2. 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.

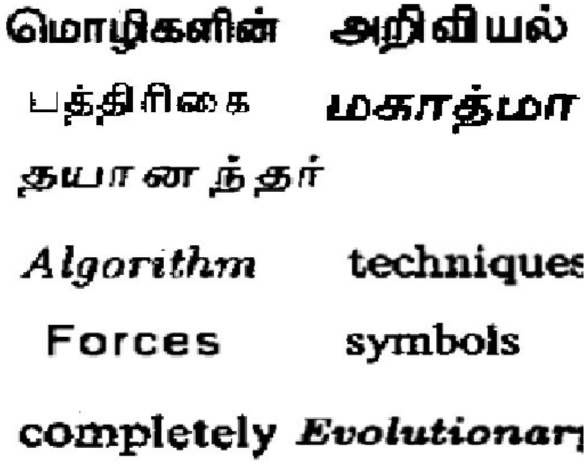

மொழிகளின்
அறிவியல்
பத்திரிகை
மகாத்மா
தயானந்தர்
Algorithm
techniques
Forces
symbols
completely
Evolutionary

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

TABLE 1: RESULTS SHOWING RECOGNITION ACCURACIES

	% Accuracies with spatial features			%Accuracies with Gabor filter responses		
	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

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, performs fairly well. 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 filters 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. English script has a higher response to 0° directional filter on account of the dominance of vertical strokes while Tamil script has a higher response to 90° filter due to dominance of horizontal strokes.

Among the works reported in the literature, Tan's approach [6] also uses features based on Gabor filters. However, his work is based on large 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 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 [9] 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 recognition accuracy is obtained only at the line level.

Chaudhury and Sheth's work [10], 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, 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.

The proposed method can also be extended to other South Indian languages, as they are quite distinctive from the Roman script. The algorithm was tested to isolate Kannada from Roman script. Figure 7 shows a Kannada word and the corresponding 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.

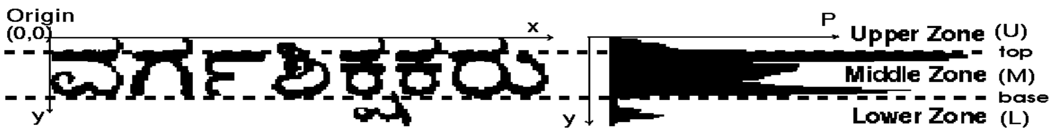


Fig. 8: Three distinct zones of Kannada and the corresponding projection profiles

TABLE 2: RESULTS SHOWING RECOGNITION ACCURACIES (KANNADA-ENGLISH)

	% Accuracies with spatial features			%Accuracies with Gabor filter responses		
	SVM	NN	k-NN	SVM	NN	k-NN
Kannada	86.99	74.38	69.41	94.63	93.84	97.02
English	88.71	68.07	68.83	95.02	92.04	91.85
Total	87.85	71.22	69.12	94.83	92.94	94.44

5. Conclusion

Two different features have been employed successfully for discriminating 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 formation of the feature vector. Directional features are the filter 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 are conducted with documents covering wide range of fonts and accuracies as high as 96% have been obtained with SVM classifiers for the directional energy features.

Bibliography

1) A. L. Spitz, "Determination of Script and Language Content of Document Images", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 235-245, 1997.

2) P. Sibun and A. L. Spitz, "Natural Language Processing from Scanned Document Images", *Proceedings of the Applied Natural Language Processing*, pp.115-121, Stuttgart, 1994.

3) T. Nakayama and A. L. Spitz, "European Language Determination from Image", *Proceedings of the International Conference on Document Analysis*, pp.159-162, Japan, 1993.

4) J. Hochberg *et al*, "Automatic Script Identification from Images Using Cluster-based Templates", *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp.176-181, 1997.

5) L. Dang *et al*, "Language Identification for Printed Text Independent of Segmentation", *Proceedings of the International Conference on Image Processing*, pp. 428-431, 1995.

6) T. N. Tan, "Rotation invariant texture features and their use in automatic script identification", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 20, no. 7, pp. 751-756, 1998.

7) C. L. Tan *et al*, “Language identification in Multilingual Documents”, *International Symposium on Intelligent Multimedia and Distance Education (ISIMADE'99)*, pp.59-64, Baden-Baden, Germany, 2-7 August 1999.

8) B. B. Chaudhuri and U. Pal, “A Complete Printed Bangla OCR System”, *Pattern Recognition*, vol. 31, no. 5, pp. 531-549, 1998.

9) B. B. Chaudhuri and U. Pal, “Automatic separation of words in multi-lingual multi-script Indian documents”, *Proceedings of the International Conference on Document Analysis*, pp. 576-579, Germany, 1997.

10) S. Chaudhury and R Sheth, “Trainable Script Identification Strategies for Indian languages”, *Proceedings of the International Conference on Document Analysis*, pp. 657--660, India, 1999.

11) D. Dhanya, “Bilingual OCR for Tamil and Roman scripts”, *Master’s thesis*, Indian Institute of Science, Bangalore, 2001.

12) Y. K. Chen *et al*, “Skew detection and reconstruction based on maximization of variance of transition-counts”, *Pattern Recognition*, vol. 33, pp 195-208, 2000.

13) C. J. C. Burges, “A tutorial on support vector machines for pattern recognition”, *Data Mining and Knowledge Discovery*, vol. 2, no 2, pp 955-974, 1998.