Optimal feature extraction for Bilingual OCR

D. Dhanya and A. G. Ramakrishnan

Department of Electrical Engineering,
Bangalore, India
Indian Institute of Science
e-mail: ramkiag@ee.iisc.ernet.in

Abstract

Feature extraction in bilingual OCR is handicapped by the increase in the number of classes or characters to be handled. This is evident in the case of Indian languages whose alphabet set is quite huge. It is expected that the complexity of the feature extraction process increases with the number of classes. Though the determination of the best set of features that could be used cannot be ascertained through any quantitative measures, the characteristics of some of the features can help decide on the feature extraction procedure. This paper describes a hierarchical feature extraction scheme for recognition of printed Tamil and Roman scripts. The scheme divides the alphabet set of both the scripts into subsets by the extraction of certain spatial and structural features. Three features *viz* geometric moments, DCT based features and Wavelet transform based features are extracted from the grouped symbols and a linear transformation is performed on them for the purpose efficient representation in the feature space. The transformation is obtained by the maximization of certain criterion functions. Three techniques : Principal component analysis, maximization of Fisher's ratio and maximization of divergence measure have been employed to estimate the transformation matrix. It has been observed that the proposed hierarchical scheme allows for easier handling of the alphabets and there is an appreciable rise in the recognition accuracy as a result of the transformation.

1 Introduction

Development of English as the universal language has resulted in the evolution of multi-script documents in many nations hosting a national language different from English. Borrowed words are usually followed by explanations in the language native to the reader. In a country like India, housing 18 official languages, the use of English along with the regional language in many official documents, reports and magazines has become mandatory. Conversion of such printed texts into editable format cannot be achieved through the use of a single monolingual OCR.

The problem of bilingual OCR can be viewed as an extension of the basic character set to include those of the second script. However, the increase in the number of symbols to be identified poses several problems such as the requirement of a large number of features for discrimination among all symbols which inturn calls for large number of training samples in order to overcome the 'peaking' phenomenon. Mixture of symbols from two different scripts might also result in the patterns having multi-modal distributions which cannot be efficiently represented by a single set of features. Hence it is prudential to resort to a hierarchical scheme having multiple stages of feature extraction. Such a scheme not only reduces the number of classes at each level, but also allows for independent handling of groups of patterns. One

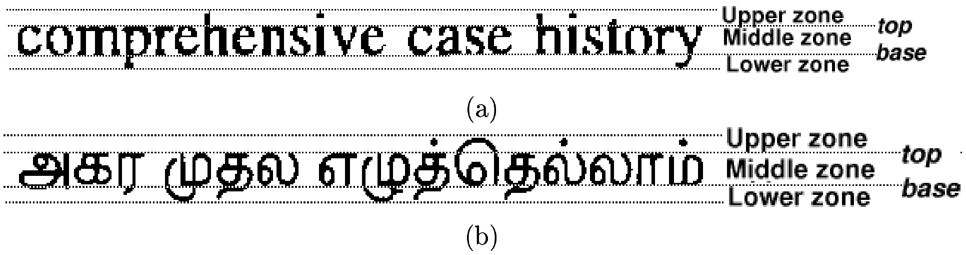


Figure 1: Three distinct zones of (a) Roman script and (b) Tamil script

such hierarchical scheme for classification of Tamil and Roman scripts is described in this paper. Section II describes the characteristics of the scripts that have been made use of for classification purposes. Section III describes the hierarchical scheme along with its different stages and section IV describes the experiments conducted and the results obtained using the scheme.

2 Characteristics of Tamil and Roman scripts

Tamil language belongs to the group of Dravidian languages and is the official state language of Tamil Nadu, a southern state of India, a national language in Singapore and a major language in Malaysia and Sri Lanka. Tamil script belongs to the group of Southern Indic scripts and is derived form Grantha script, a descendant of ancient Brahmi script. A brief description of the script follows:

1. Tamil alphabet set has 12 vowels, 18 native consonants and 5 borrowed consonants.

2. The twelve vowels combine with the consonants to form compound characters. These characters are modifications of the basic characters. A modification appears in the form of a matra or diacritic (consonant modifier) that either gets connected to the basic character or remains disjoint from it.

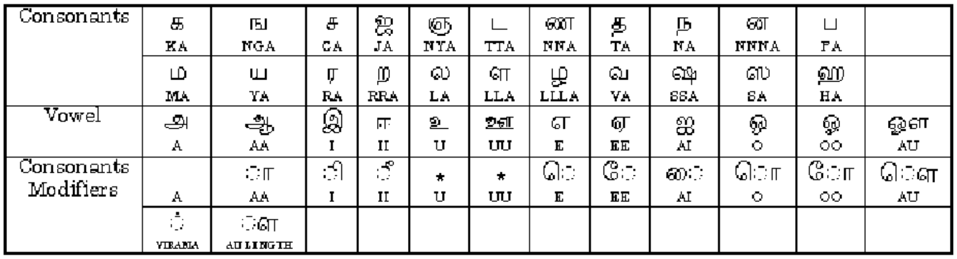
3. Consonant clusters, which are series of consonants without any intervening vowel, are represented in Tamil script by suppressing the inherent vowels except the last one in the series. The vowel suppresser is called as *pulli*; it appears as a dot placed as a superscript above the basic consonant.

4. In contemporary Tamil texts, vowels occur in their natural form only at the beginning of words; in any other location, they occur as modifiers or diacritics of the basic consonants.

5. The diacritics get added to the right, left, top or bottom of the consonants. Those added to the right and left are disconnected from the consonant they modify, whereas those added at the top and bottom are connected and change the shape of the consonant.

6. While all consonant-vowel combinations are derived as modifications of the basic set in Tamil script, no such concept exists in Roman script. All the vowels and consonants occur independently and there are no 'modifications' as in Tamil script. This makes the Roman script easier to handle.

7. 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 (1). For convenience in our discussion, we define the



Consonants	க KA	ங NGA	ச CA	ஜ JA	ஞ NYA	ட TTA	ண NNA	த TA	ந NA	ன NNNA	ப PA	
	ம MA	ய YA	ர RA	ற RRA	ல LA	ள LLA	ழ LLLA	வ VA	ஷ SSA	ஸ SA	ஹ HA	
Vowel	அ A	ஆ AA	இ I	ஈ II	உ U	ஊ UU	எ E	ஏ EE	ஐ AI	ஒ O	ஓ OO	ஔ AU
Consonants Modifiers	A	◌ா AA	◌ி I	◌ீ II	* U	* UU	ெ◌ E	ே◌ EE	ை◌ AI	ெ◌ா O	ே◌ா OO	ெ◌ள AU
	◌் VIRAMA	◌ௗ AU LENGTH										

Figure 2: Basic alphabet set of Tamil script with corresponding diacritics

following terms: *top* line, *bottom* line. 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.

8. Though Tamil has native numerals, they are hardly used in contemporary texts; only the Hindu-Arabic numeral set is being used.

Figure 2 shows the basic alphabet set of Tamil characters along with the modifiers or matras. One can see that some of the vowel modifiers appear as separate symbols and these are dealt as separate patterns for identification. The * in the figure indicate that the matras change the entire shape of the characters differently for different consonants. In such cases, each modified consonant is considered a separate class for identification. In the rest of the discussion, the term “symbol" is used to denote any pattern taken for classification: a character, a disjoint matra or a numeral.

3 Three-level hierarchical scheme

The primary factors to be considered in a hierarchical scheme are the features to be used at each level and the corresponding decision rule. In the proposed scheme the symbols are grouped initially into different subsets depending on the spatial distribution of symbols and presence or absence of a loop. Features such as geometric moments, discrete cosine transform coefficients and wavelet coefficients are then extracted and each of them transformed into a subspace of reduced dimension for optimum representation.

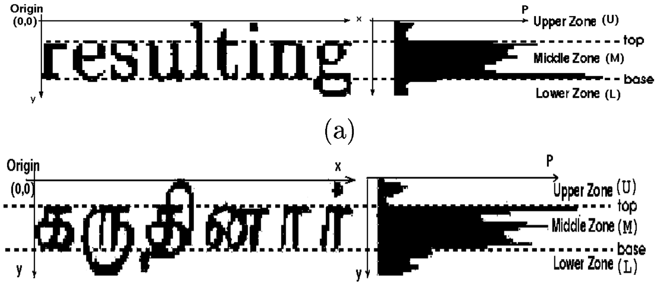
3.1 Primary grouping

In the first level of classification, the zonal occupancy of the characters is used to divide the observation space into four major groups as follows.

Group1: All three zones
Group2: Middle and Lower zones
Group3: Middle and Upper zones
Group4: Middle zone

The following observations are made after the primary grouping.

→ Group 1 consists solely of Tamil characters, since Roman script does not have any alphabet that occupies all three zones.



(a)

(b)

Figure 3: Projection profiles of English and Tamil words

- → Group 2 has a majority of Tamil symbols consisting both basic characters and modified ones; only a few alphabets from Roman, namely, p, q, y, j and g come under this group.
- → All upper case letters in Roman alphabets and all Hindu-Arabic numerals are members of group 3. In Tamil script, some vowel modified symbols and certain basic characters come under this group. This set has the largest cardinality.
- → Many basic characters in Tamil and a few vowel modified consonants form the elements of group 4. Many of the lower case letters in English are also elements of this set.

Feature extraction involves identification of the bounding box of the character or symbol and locating its position with reference to the zone boundaries. Considering the top left most index as the origin, and denoting the projection profile by P, the zone boundaries are defined as,

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

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

where H is the height of the word or line. Fig. 3 shows 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 1 and 2. The boundary points are marked at the corresponding extrema points.

3.2 Subgrouping

Traditionally, structural features have exploited for the purpose of character recognition. Structural features need to be consistent across different fonts and their variations. Features such as presence or absence of loops, strokes and their distributions have been explored. For handwritten digits, presence of loops is a very important feature [1]. In the proposed scheme, this feature is used for the purpose of subgrouping. A loop is a closed curve or a hole and its presence is detected on a thinned contour by means of a contour tracing algorithm. Subgrouping is performed on Groups 3 & 4 thereby dividing them into two subsets: one consisting of symbols with loops and other without them. However, the presence of noise often distorts certain symbols which have a very small hole in them. On such characters, grouping is not performed at all.

3.3 Tertiary features

An efficient representation can be described as the ability of the features to form disjoint clusters in the feature space such that their intra-class variance is small while the inter-class variance is large. Though there is no quantitative measure to determine the suitability of a particular feature, the of the features can provide some insight into their suitability. Though the direct application of the 'suitable' features can ensure fair performance, performance improvement could be obtained by transforming these features into a different space where symbols are better represented. The transformation can also help reduce the dimension of the feature vectors thereby reducing the 'curse of dimensionality problem'. The extraction of an optimal set of features with minimum dimension is the essential part any pattern recognition problem. In this paper, three features have been studied with respect to their ability for compact representation in the feature space: geometric moments, Discrete Cosine Transform (DCT) based features and Discrete Wavelet Transform (DWT) based features. Considerable improvement in the performance in terms of reduced dimension and increased efficiency has been obtained through linear transformations of these features.

3.3.1 Geometric moments

The use of moments as a tool for character recognition has been widely explored [2],[3] and [4]. Such features capture the global information of the character. Moments such as geometric moments, Zernike moments and Legendre moments are some of the important features used for character recognition. Moments are projections of the symbol onto different polynomials. Their representation efficiency is described in terms of their susceptibility to noise and affine transforms. The simplest of them is the geometric moment (GM), in which the polynomial of order $p+q$, onto which the pattern $f(x,y)$ of dimension M x N with coordinates x and y is projected, is given by $x^p y^q$. GM of order $p+q$ is defined by

$$M_{pq} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} x^p y^q f(x,y) \tag{3}$$

The main disadvantage of geometric moments is the non-orthogonal nature of the monomials which leads to redundancy in representation. The parameters to be chosen in using geometric moments as features are (i) the order of moments (ii) area of operation on the image. Though it is well known that these moments represent the global properties of the image, little is known about what the higher order moments indicate. As the order of moments increases, the features become sensitive to each image pixel. Very low order features provide a fair description of the pattern, but are not powerful discriminators when the number of classes is quite high. Selection of the area of operation on the input pattern involves a trade-off between discrimination capability and noise sensitivity. Operating on the whole pattern results in loss of details while having a very small area makes the features too sensitive to small perturbations.

In our work, upto third order moments have been considered for classification. Localization is better achieved by dividing the symbol image into blocks of size 12x12 pixels and extracting features from each of the blocks.

3.3.2 Discrete cosine transform

Transform based features have found widespread use because they are easy to compute and have good reconstruction properties. DCT is a very efficient transform with regard to decorrelation and energy compaction properties. It is asymptotically equivalent to the Karhunen-Loeve transform (KLT) which is the optimal transform. The DCT of an image f of dimension MxN is given by

$$F^{II}(u,v) = \alpha(u)\alpha(v)\sum_{y=0}^{N-1}\sum_{x=0}^{M-1} f(x,y)\cos\left(\frac{(2x+1)u\Pi}{2M}\right)\cos\left(\frac{(2y+1)v\Pi}{2N}\right) \tag{4}$$

where

$$\alpha(u) = \begin{cases} \frac{1}{P} & u = 0 \\ \frac{2}{P} & u > 0 \end{cases} \tag{5}$$

and P is the size of the image in the corresponding spatial direction. The subscript II denotes that the DCT is of type-II. In this work, each size-normalizedsymbol is divided into four sub-blocks and DCT is taken on each sub-block. DCT being a separable transform, the above operation is executed via one-dimensional DCT. Owing to the nature of the input patterns, most of the signal energy (information) is concentrated in few low frequency coefficients. Thus, only the low frequency coefficients of DCT are considered for recognition. In our work, we have considered only $H/6$ x $W/6$ low frequency coefficients where $H$and W are the dimensions of the sub-blocks.

3.3.3 Wavelet based features

The lack of localization property in most of the transform based techniques including DCT leads to ambiguous results for all pairs of similar looking characters or symbols. In the recent times, a lot of research work has been done on wavelet transforms and their localization properties. Some of the properties of wavelet transforms, which are responsible for their popularity are:

(i) Time-frequency localization
(ii) Efficient representation in the case of orthogonal basis functions.

In general, if $\Psi_{s,n}$ represents the wavelet basis with a dilation scale parameter s and translation parameter n, then the wavelet transform of a signal $f(x)$at scale s and position n is given by,

$$Wf(s,n) = f * \Psi_{s,n} \tag{6}$$

where $*$ denotes the convolution operation. Discrete wavelet transform (DWT) is implemented through a filter bank as shown in Fig. 4. Decomposition is performed by low and high pass filters operating in parallel and decimation of the filtered signal. Binary images are efficiently represented using Haar wavelets. For Haar wavelet transform, the low and high pass filters h and g are given by

$$h = \frac{1}{\sqrt{2}}\left(\delta(n) + \delta(n+1)\right) \tag{7}$$

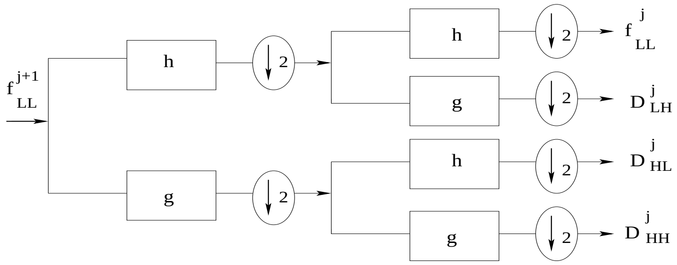


Figure 4: Filter bank implementation of wavelet transform

$$g = \frac{1}{\sqrt{2}}\left(\delta(n) - \delta(n+1)\right) \tag{8}$$

A two level decomposition is performed using the above filters. Since Haar wavelet has a separable kernel, the transform is implemented by operating on the rows first, followed by columns. The coefficients corresponding to the low pass filtered image in the second level are used to form the feature vector.

3.4 Feature transformation

In general, as the dimension of the feature vector increases, the efficiency increases initially and then starts to decrease. This phenomenon called 'peaking phenomenon' is a major handicap in feature extraction process. Also, as the dimension of a feature vector increases, the number of training samples required also increases. Hence, dimensionality reduction is an essential. Linear transformations are used to project the feature vector in a higher d-dimensional space onto a lower m-dimensional space such that they are better represented in the transformed space. Thus if x represents a feature vector in the original domain, y represents the transformed vector, the transformation is defined by,

$$y = W^T x \tag{9}$$

where W is the transformation matrix. The aim is to find the matrix W which produces a better representation in the transformed domain. This is brought about by maximizing certain criterion functions. Three transformation techniques have been employed for the purpose : *principal component analysis* (PCA), maximization of *Fisher's ratio* and maximization of *divergence measure*. In order to have a common platform for comparing the efficiency of these methods, m is always set to the number of symbols in each group irrespective of the transformation used.

3.4.1 Principal Component Analysis (PCA)

PCA demands that the projection vectors in W maximize the variance in the projected data and give uncorrelated distributions. The projection vectors that maximize the variance on the projected features are the eigenvectors corresponding to m largest eigenvalues of the covariance matrix, S, of the original feature set.

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x - m_x)(x - m_x)^T \tag{10}$$

where x is the feature vector, n the sample size and m_x is the global mean of all feature vectors. The transformation matrix W is formed by m eigenvectors corresponding to m largest eigenvalues where m is the required dimension.

3.4.2 Fisher's ratio

Fisher's ratio is one of the important criterion functions and is defined as the ratio of total between-class scatter to within-class scatter. Maximization of this criterion function ensures maximum inter-class variability and small intra-class variability. The scatter matrices are defined as follows : The between-class scatter matrix S_B is given by

$$S_B = \sum_{i=1}^{C} (m_i - m_x)(m_i - m_x)^T \tag{11}$$

where m_x is the global mean feature vector across all classes and m_i is the mean vector of i^{th} class. The within-class scatter matrix S_W is given by

$$S_W = \sum_{i=1}^{C} S_i \tag{12}$$

where S_i is the within class scatter matrix for i^{th}class given by,

$$S_i = \sum_{x} (x - m_i)(x - m_i)^T \tag{13}$$

with the summation taken over all feature vectors x belonging to class i. If J represents the criterion function, (Fisher's ratio), then

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{14}$$

The required transformation matrix W is the one that maximizes the above criterion function. W is solved as a generalized eigenvalue problem

$$S_B w_i = \lambda_i S_W w_i \tag{15}$$

where λ_i is an eigenvalue and w_i is the eigenvector.

3.4.3 Divergence measure

In [5]a criterion function that uses a weighted average divergence measure is proposed for the purpose of speech identification. This method takes into account the difficulty associated with certain pairs of characters. The criterion function is the divergence measure and using the same notation as above, the criterion function that is maximized is given by

$$J = tr\left[\left(W^T S_W W\right)^{-1}\left(W^T M W\right)\right] \tag{16}$$

where M is given by

$$M = \sum_{i=1}^{C} \sum_{j=1}^{C} P_i P_j \left(m_i - m_j\right) \left(m_i - m_j\right)^T \tag{17}$$

and P_i is the priori probability of class i. The transformation matrix W that maximizes the above function is found by solving the generalized eigenvalue problem:

$$M w_i = \lambda_i S_W w_i \tag{18}$$

Using the above three methods, the transformation matrix is found

3.5 Results and discussion

In order to analyse the performance of each of the features, the algorithms have been tested on samples collected from various magazines, reports and books. Documents are scanned at 300dpi and binarized using the two stage method described in [6]. The skew introduced during the process of scanning is estimated and corrected. Textual lines and words are segmented by determining the valley points in the projection profiles and symbols are isolated using connected component analysis.

The primary and secondary levels of classification are based on spatial occupancy and presence of structural features and do not require training samples. Initially, samples are grouped depending on the zone they occupy. To accommodate for varying thickness of the symbol on account of font differences, a 2 pixel margin is kept while detecting zone boundaries. However, it is assumed that all symbols within a line are of the same font size.

Symbols are then passed through the hierarchical feature extractor. Each symbol is normalized in size and thinned. The normalization size is dependent on the group to which the symbol belongs. Symbols belonging to groups 1, 2 and 3 are normalized to size 48x48 while those belonging to group 4 are normalized to 36x36. Further grouping is done on these thinned characters depending on the presence or absence of a loop structure. Final level features such as, geometric moments, DCT and wavelet coefficients are then extracted from these sets of samples.

The training file contains 150 to 200 samples of each of the frequent prototypes. The infrequent symbols are represented by the bootstrapped samples[7]. The transformation matrix is computed for each of the feature sets, and dimensionality reduction is performed for the entire training set. For testing purposes, samples are similarly collected from various books, magazines, reports *etc* with a large font variability. Test set consists of 6000 symbols containing samples from all of the 183 classes. These samples are obtained from as many as 20 scanned pages, each containing a minimum of 300 characters. Many of the frequently occurring characters are represented by as many as 40 samples, while some of the infrequent characters are represented by 20 samples or less. In order to compare the performance of the each of the feature extraction methods, the algorithms have been tested on the same set of test samples that are pre-grouped, normalized and thinned. The recognition accuracy of each of the sets S1 to S6 has been found using the nearest neighbour classifier and the overall efficiency, which is the percentage of symbols correctly classified in all the sets, is calculated. The flow diagram of the procedure is explained in Fig. 5.

The results are tabulated in tables 1, 2 and 3. From the tables, the following can be observed :

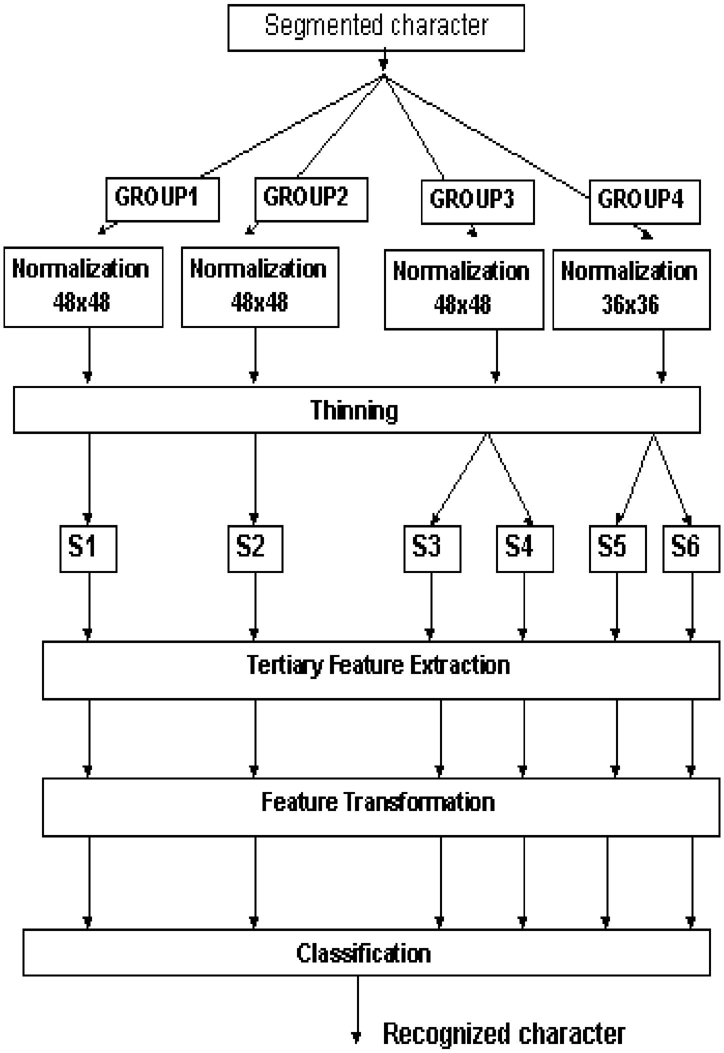


Figure 5: Flow diagram of the recognition scheme

Table 1: Recognition Accuracies With Geometric Moments
(PCA: Principal Component Analysis; FR: Fischer's Ratio; DM: Divergence Measure)

Class	S1	S2	S3	S4	S5	S6	S7
3rd Order Moments	93.27	92.98	93.69	92.98	96.06	96.09	94
PCA	89.58	92.98	93.37	93.18	95.30	95.35	93.64
FR	95.01	96.22	94.32	94.36	96.21	97.39	95
DM	88.5	95.91	94.84	95.28	97.27	95.16	94.5

Table 2: Recognition Accuracies With DCT Based Features

Class	S1	S2	S3	S4	S5	S6	S7
DCT	95.69	96.00	92.58	95.84	96.96	94.42	95
PCA	97.05	97.69	93.11	95.84	96.66	95.53	96.29
FR	98.41	97.92	95.48	96.49	97.27	94.79	96.73
DM	98.41	97.77	95.37	96.75	97.72	94.79	96.80

Table 3: Recognition Accuracies With DWT Based Features

Class	S1	S2	S3	S4	S5	S6	S7
DWT	96.29	93.86	92.75	92.60	94.36	92.75	94
PCA	96.73	94.7	92.22	92.73	94.65	94.6	94.33
FR	97.16	96.85	95.31	93.77	92.91	92.75	95
DM	94.98	94.16	94.78	92.09	95.08	91.63	93.79

Geometric moments

1. Geometric moments achieve an overall recognition rate of 94% without any dimensionality reduction.

2. The overall accuracy increases when the dimension of the features is reduced using Fisher's ratio and divergence measure, while PCA retains the performance. These results reiterate the fact that there exists some amount of correlation and redundancy in the GM based features.

3. A good recognition rate of 95 to 97% is obtained for the individual sets S5 and S6 and the minimum rate is obtained for set S1. Since S5 contains a large number of commonly occurring classes, this result is encouraging. Also, S1 contains many symbols of low occurrence probability. Hence, the performance in general, for a scanned document, will not be much affected because of the low efficiency of S1.

4. For certain sets, the method based on divergence maximization gives a better performance. Since this method tries to optimize the separation between each pair of classes, the performance improves to some extent where the number of similar patterns is more.

DCT based features

1. They give a good performance of above 95% without dimensionality reduction.

2. All methods of dimension reduction give good performance improvement.

3. For individual sets, DCT shows a good performance. The performance for sets S3 and S6 is lower compared to the others. However, the recognition rates for all the other sets are higher than those of the GM features and in general outperform the corresponding results of geometric moments.

DWT based features

1. Wavelet based features show a fair performance with an overall performance rate of approximately 94%.

2. Among the feature dimension reduction methods, PCA gives results comparable to that of the original set without apparent increase in the efficiency. Fisher's method gives an improved performance of 95%. Similarly, the method based on divergence measure gives a performance comparable to those of the original and PCA based methods.

3. One of the reasons for ineffective performance of the dimension reduction techniques could be the following. The features in wavelet domain still retain the apparent shape of the character, with the values being real. There is no clustering of the coefficients, which is very much needed for dimensionality reduction. Hence, there is some loss of information when the dimension is reduced.

4. The overall efficiency for wavelet based features is comparable to that of GM features and is less than that of DCT.

3.6 Conclusions

Effective representation is an important aspect in feature extraction as evidenced by the results. The hierarchical approach taken has reduced the number of classes to be handled into manageable limits. Of the three features extracted, DCT gives a very good performance and further improvement has been obtained by the transformation of the extracted features into a different space. Of the transformation methods, Fisher's method performs consistently well. Principal component analysis, though in some cases does not improve the efficiency appreciably, does not degrade it. Hence, the advantage of dimensionality reduction still remains. The method based on maximization of divergence measure, also does improve the performance and these methods when incorporated in combination with DCT would certainly ensure an good performance OCR system.

References

[1] A. Sinha, "Improved Recognition module for the Identification of Handwritten Digits," *Master's thesis*, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1999.

[2] C. H. Teh and R. T. Chin, "On Image Analysis by method of moments," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 10, no. 4, pp. 496-513, 1993.

[3] A. Khotanzad and Y. H. Hong, "Rotation Invariant Image Representation using Features Selected via a Systematic Method," *Pattern Recognition*, vol. 23, no. 10, pp-1089-1101, 1990.

[4] R. R. Bailey, "Orthogonal Moment Features for use with Parametric and Non-Parametric Classifiers," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 18, no. 4, pp. 389-399, 1996.

[5] P.C.Loizou and A.S. Spanias, "Improved Speech Recognition Using a Subspace Projection Approach," *IEEE Transaction on Speech and Audio Processing*, vol. 7, no. 3, 1999.

[6] D. Dhanya, "Bilingual OCR for Tamil and Roman scripts," *Master's thesis*, Department of Electrical Engineering, Indian Institute of Science, 2001.

[7] Y. Hamamoto *et al.,* "A Bootstrap Technique for Nearest Neighbour Classifier Design," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 19, no. 1, pp. 73-79, 1993.