An Approach for Structural Feature Extraction for Distorted Tamil Character Recognition

Nirase Fathima Abubacker
Lecturer
School of Information Technology,
City University College of Sci &Tech.,
KualaLumpur, Malaysia.

Indra Gandhi Raman
Associate Professor
Department of Computer Applications,
G.K.M College of Engineering and Technology,
Tamil Nadu, India

ABSTRACT

Feature extraction is an important task for designing an OCR for recognizing degraded documents. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance in character recognition systems. Shape inconsistency among characters of the same structure is sometimes quite large because of the poor resources and environmental impact on the document images. Therefore, it is necessary to select features which can adapt to the shape variations irrespective of the distortions. Hence, in this paper, selection of appropriate standard structural features is taken as the primary task for various distortion types that are considered to recognize the Tamil distorted characters

Keywords

Character Normalization, Distorted Character Recognition, Structural Features.

1. INTRODUCTION

Feature extraction is a method of automatic pattern recognition in which recognition is achieved by making measurements on the patterns to be recognized, and then deriving features from these measurements. These strategy used for recognition can be broadly classified into structural, statistical and hybrid. Structural techniques use some qualitative measurements as features. Structural features are the features that are physically a part of the structure of the character, such as straight lines, arcs, circles, intersections etc. The features used for recognition are the positions of vertical lines, horizontal lines and branching in a character. Statistical techniques use some quantitative measurement. Selection of a feature extraction method is probably the single most important factor in achieving high recognition performance of OCR systems. However, the other steps in the system also need to be optimized to obtain the best possible performance and these steps are not independent. The choice of feature extraction method limits or dictates the nature and output of the preprocessing step.

2. REVIEW OF LITERATURE

Structural features can represent various global and local properties of characters with high tolerance to distortions and style variations. These features describe a pattern in terms of its topology and geometry by giving its global and local properties. Many different types of features have been identified by Suen *et al.* [1] in the literature that may be used for character and numeral. Two main categories of those features are Structural (topological) and Global (statistical). Trier *et al.* [2] summarized a good survey on feature extraction method for character recognition. The author mentioned that different types of features can be extracted depending on the representation forms of characters, which can be grouped as grayscale images, Binary images, character contour and character skeletons. Heutte *et al.* [3] says that some of the main structural features include features like number and intersections between the character and straight lines, number of vertical and horizontal lines, holes position, end points, presence of loops, number of loops, number of intersections and junctions. These features are generally hand crafted by various authors for the kind of pattern to be classified. Olszewski [4] designed a structural recognition approach for extracting morphological features and performing classification without relying on domain knowledge. This system employs a statistical classification technique on structural features is a natural solution. Lee and Gomes [5] have used the structural features for handwritten numeral recognition such as number of central, left and right cavities, location of each central cavity, the crossing sequences, the number of intersections with the principal and secondary axes and the pixel distribution. Kam-Fai and Dit-Yan [6] proposed a syntactic (structural) approach for the analysis of on-line handwritten mathematical expressions. Amin [7] has used seven types of structural features such as number of sub words, number of peaks of each word, number of loops of each peak, number and position of complimentary characters, the height and width of each peak for recognition of printed Arabic text. In a classic paper Kahan *et al.* [8] have developed a structural feature set for recognition of printed text of any font and size. The feature set includes number of holes, location of holes, concavities in the skeletal structure, crossings of strokes, endpoints in the vertical direction and bounding box of the character.

Penman [9] the word recognizer in this is based on feature extraction, in which the loops, lines and distinguishing characteristics of the scanned data are extracted and analyzed to determine the letters. The sequence of recognized letters is compared against words in a dictionary to aid accuracy. Rocha and Pavlidis [10] have proposed a method for the recognition of multi-font printed characters using the following structural features: convex arcs and strokes, singular points and their relationships. Pal and Anirban [11] used the topological features for recognition of printed Urdu script, which are used at each non-terminal node, should be robust and maximally divides the character in a node into two groups to get an optimal tree. Sometimes character came to same node because of same topological properties. Leedham and Pervouchine [12] have used global features like handwriting size, word spacing, line spacing, arrangement of words, margin patterns, baseline patterns, line quality, spelling and grammar etc. for recognition of handwritten text. According to Anisimovich [13] the structural pattern is matched against a character image by establishing correspondence between structural elements and parts of the image which satisfy all spatial relations. The developed matching procedure can successfully find this correspondence on broken and distorted images.

3. TOPOLOGICAL / STRUCTURAL FEATURES

Structural features capture shape information of the characters. These features should be chosen keeping in mind that the shape variations should affect feature set minimally. It was not an easy step to decide which structural features technique should be chosen to extract the structural features from distorted characters of Tamil script as there were large shape variations in characters of the same class. The structural features set have the following advantages over common characteristics.

a) The structural features are font independent.
b) These features are less sensitive to character size and very much tolerant to noise.
c) It can compensate heavy variations in input data and certain kind of shape distortions.
d) The features representing different characters have a very low probability to coincide.
e) It is perceptive over parameters and minute details of the process.
f) The result and performance of a statistical method, depends heavily on the parameters, features set used and the training set.

We have used the following features of Tamil characters for constructing feature vectors.

3.1 Presence of Loop (FS1)

A small circle like sub-symbol is seen in all parts of the characters. This feature is present if the character contains a closed loop as present in the character 'ண'. Expect in 'ஃ' the closed loop present in remaining characters is noticed as one of the parts of the character. This feature is sensitive to noise, were a broken loop may resemble some other characters. We have used a tolerance level to deal with broken loop. As such, there are totally 55 characters (vowels, consonants and vowel consonants) containing this feature. This feature varies from 0 to 3. For example, characters 'ட', 'ர' and 'ற' does not contain any loop, so the feature value is 0. Similarly characters like 'அ', 'ே' and 'ள' contain feature value as 2 and 'ண' hold the feature value as 3. Features have Boolean value 0 if character does not contain loop and value 1 if the character has got at least one loop. Figure 1 shows the Tamil character having loops.

Figure Error! No text of specified style in document.**: Characters having looping structure**

3.2 Presence of Lower Arc (FS2)

An arc like concavity sub-symbol is seen almost at the lower part (lower zone) of the character and sometimes slightly extended up to upper part. As such, there are totally 24 characters (vowels, consonants and vowel consonants) containing this feature. The feature is true if an arc is present at lower junction or else it is false. For example, characters 'ஞ', 'ஒ' and 'ழ' contains lower arc and the Boolean value of this type is 1. Figure 2. shows the Tamil characters having lower arc.

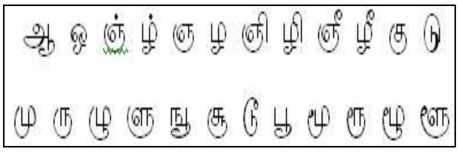

Figure 2: Characters having Lowe arc

3.3 Presence of Upper Arc (FS3)

An arc like convexity sub-symbol is seen almost at the upper part (upper zone) of the character and sometimes slightly extended up to base line. All characters of these types are called as "Kurill" letters. As such, there are totally 20 characters (vowels, consonants and vowel consonants) containing this feature. This feature is true if an arc is present at upper junction or else false. For example, characters 'இ', 'கி' and 'டு' contains upper arc and the Boolean value of this type is 1 which is otherwise a 0. Figure 3. shows Tamil characters having upper arc characters.

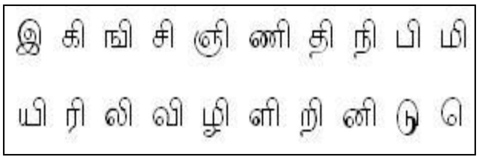

Figure 3: Characters having Upper arc

3.4 Presence of Slat Line (FS4)

Features of type "/" are always seen at the end part of the characters. Continuous decrease in white pixel count indicates a slant. The position of the horizontal and vertical lines in the character can be defining the character to a considerable extent. Hough Transform defines the slant line in the image in terms of parametric co-ordinates θ and ρ lines with angles $-45° \le \theta \le 45°$ considered as vertical lines, while others are considered as horizontal lines. The ρ gives position of the line, in terms of perpendicular distance from the origin. The region in which a particular line falls is determined and used as a slant line. As such, there are totally 16 characters (vowels, consonants and vowel consonants) containing this feature. Characters 'ஏ', 'ர', 'த', 'ந' and 'ற' with combination of matra belong to this feature category. They have Boolean value either zero or one. The feature is true if a slant line is present at the end of the character or else false. Figure 4. shows the Tamil characters with slant line.

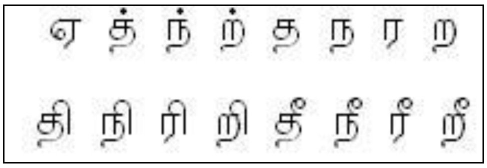

Figure 4: Characters having Slant Line

3.5 Presence of Lower Arc along with loop (FS5)

Loop along with the concave like arc is presented only at the lower part of characters. There are only three characters with these sub-symbols that end at lower zone namely, 'ஒ', 'யூ' and 'ஆ', remaining characters of this type ends with side bar (middle zone). As such, there are totally 19 characters (vowels, consonants and vowel consonants) containing this feature.

These features have Boolean value either zero or one. The feature is true if an arc along with the loop is present at the end of the character or else it is false. Figure 5 shows the Tamil characters having lower arc.

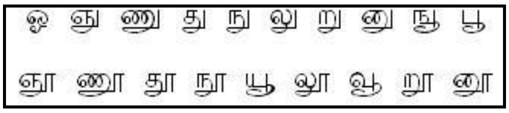

Figure 5: Characters contains Lowe arc with Loop

3.6 Presence of Upper Arc along with loop (FS6)

Loop with the convexity sub-symbol along with arc “∂” is presented only at the upper part of characters. All characters of these types are called as “Nedeil - நெடில்” letters. As such, there are totally 18 characters (vowels, consonants and vowel consonants) containing this feature. Majority of 11 characters ‘கீ’, ‘ஙீ’, ‘சீ’, ‘ணீ’, ‘பீ’, ‘மீ’, ‘யீ’, ‘லீ’, ‘வீ’, ‘ளீ’ and ‘னீ’ of this type occupy upper-middle zone and remaining 7 characters ‘ஞீ’, ‘தீ’, ‘நீ’, ‘ரீ’, ‘ழீ’, ‘றீ’ and ‘டு’ of this type occupies all-zones. These features have Boolean value either zero or one. The feature is true if an arc along with the loop is present at the upper part of the character or else it is false. Figure 6 shows Tamil character having upper arc.

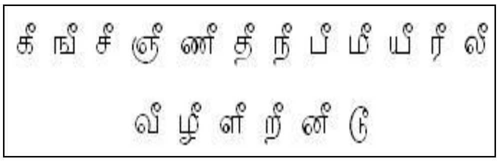

Figure 6: Characters contains Upper arc with Loop

3.7 Presence of Dot (FS7)

Features of these characters are always called as “Pully – ● ”. Expect “ஈ” remaining characters of these type have dot only at upper part of the characters. As such, there are total 19 characters (vowels-consonants) contains this feature. Majority of 13 characters ‘க்’, ‘ங்’, ‘ச்’, ‘ட்’, ‘ண்’, ‘ப்’, ‘ம்’, ‘ய்’, ‘ர்’, ‘ல்’, ‘வ்’, ‘ள்’ and ‘ன்’ of this type occupy upper-middle zone, 5 characters ‘ஞ்’, ‘த்’, ‘ந்’, ‘ழ்’ and ‘ற்’ occupy all zone and remaining 1 characters ‘ஈ’ occupies only middle zone (vowel). These features have Boolean value either zero or one. The feature is true if a loop is present or else it is false. Figure 7 shows Tamil character having “Pully” characters.

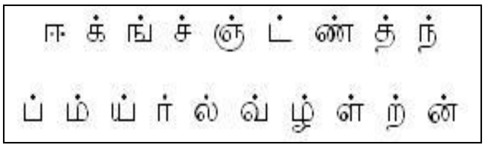

Figure 7: Characters with Dot

3.8 Presence of Side Bar (FS8)

A vertical line, of approximately the height of the full character is present on the rightmost side of the sub-symbols. As such, there are total 39 characters (vowels-consonants) containing this feature. Characters of this type almost occupy all upper, middle and middle zones. These features have Boolean value either zero or one. The feature is true if a side bar is presented at rightmost or else it is false. Figure 8. shows the Tamil character having ‘Pully’ characters.

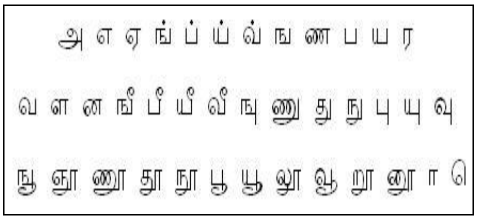

Figure 8: Characters with Side Bar

3.9 Number of Junctions with the Baseline (FS9)

It is observed that all characters merge at baseline at one or more than one point. This feature value is very important to find recognition accuracy whenever the characters are distorted especially for broken characters. Letter ‘ே’ holds the least value (one) as the junction value with baseline. Expect ‘ே’ remaining almost all Tamil characters hold higher value than other features. This feature is true if the number of junction is one or more than one. As seen in all the Figures from Figure 1 to Figure 8, all Tamil character have this feature.

3.10 Aspect Ratio (FS10)

Aspect ratio is obtained by dividing the height of the character by width of the character. We have divided the entire set of characters and the result obtained is used in recognition of the distorted characters especially for touching, broken and heavily printed characters. Distorted character aspect ratio is always less than its original value and in case of a touching character it is always higher than the original. For our calculation purpose we have rewritten the value of aspect as 1 when even it is less than 0.5, similarly 0.5 < and > 2 = 2, greater than 2 = 3. Even an extremely distorted character contains minimum aspect ratio.

3.11 Entry Point Directional (FS11- FS16)

Various entry points are used on left, right, top and bottom of character matrix profiles. The count of entry points is calculated by adding all the entry points in left, right, top and bottom of character matrix. Left and right entry points can further be classified into upper and lower portions. Based on this, we have sub divided the entry point direction as follows.

- ➢ Left Upper Entry Point (F11)
- ➢ Left Lower Entry Point (F12)
- ➢ Right Upper Entry Point (F13)
- ➢ Right Lower Entry Point (F14)
- ➢ Top Entry Point (F15)
- ➢ Bottom Entry Point (F16)

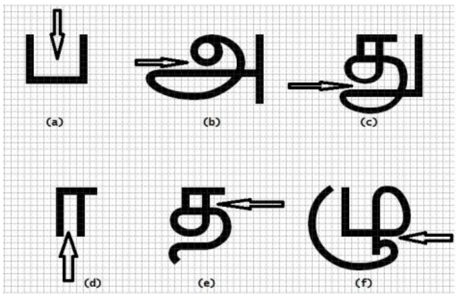

Figure 9: Sample characters with six directional entry points

An entry point is found if no object pixel occurs in the respective direction up to certain threshold (simulation_threshold). The threshold value defines the depth of entry points. To find out the entry points in particular direction, we first find the continuous projections for that particular direction by counting the first run of consecutive white pixels. An entry point is said to occur if the value of continuous projection is greater than threshold (simulation_threshold). The different entry points are shown in Figure 9. However, maximum number of entry points can be 4. Except 'இ', 'ட', 'ஜ', 'கி' and 'ஸ்ரீ' all other characters have at least one entry point. There are various feature extraction methods already reported in the literature, based on this we have tried our best to select features which are well suited for accurate recognition of different distorted types.

4. PERFORMANCE ANALYSIS OF STRUCTURAL FEATURES

In previous sections we have categorized various structural features of Tamil scripts. These categorizations were formulated with an aim of reproducing the actual characters from distorted documents based on zone wise sub-symbols that form the characters.

4.1 Feature Value for Middle Zone Characters

As per as feature value of middle zone characters that come under FS1 to FS10 for all Tamil characters is concerned, it satisfies only five structural features viz., FS1, FS7, FS8, FS9 and FS10. Middle zone covers almost 25% of the FS1, 0.18% of FS7 and 18% of FS8. This feature will be very useful in recognition touching and heavily printed characters. When broken characters are involved, the results might be misleading.

4.2 Feature Value for Upper-Middle Zone Characters

As per as feature value of upper-middle zone characters that come under FS1 to FS10 for all Tamil characters is concerned, it satisfies only seven structural features viz., FS1, FS3, FS6, FS7, FS8, FS9 and FS10. Upper-Middle zone covers almost 38.18% of the FS1, 65% of FS3, 61.11% of FS6, 68.42% of FS7, 23.08% for FS8 and 18% of FS8. This feature will be very useful in recognition touching and heavily printed characters. When broken characters are involved, the results might be misleading.

4.3 Feature Value for Middle-Lower Zone Characters

As per as feature value of middle-lower zone characters that come under FS1 to FS10 for all Tamil characters is concerned, it satisfies only five structural features viz., FS1, FS2, FS4, FS5, FS8, FS9 and FS10. Middle-lower zone covers almost 40% of the FS1, 67% of FS2, 31.25% of FS4, 100% of FS5 and 43.59% FS8. One special feature to note in this zone is that, this is the only zone satisfying structural features of FS5. This feature will be very useful in recognition touching and heavily printed characters. When broken characters are involved, the results might be misleading.

4.4 Feature Value for All Three Zone Characters

Similarly for all three zone characters that come under FS1 to FS10 for all Tamil characters is concerned, this holds the maximum number of structural features when compared with other zones viz., FS1, FS2, FS3, FS4, FS6, FS7, FS9 and FS10. All three zones have the least FS1 percentage of less than 1, 33.34% of FS2, 35% of FS3, 68.75% of FS4, 38.89% of FS6 and 26.32% FS7. Aspect ratio is always greater than 1 in this zone. During character segmentation, FS10 helps in separating the characters perfectly. This feature will be very useful in recognition touching and heavily printed characters. When broken characters are involved, the results might be misleading but would eventually lead to another valid character.

5. PERFORMANCE ANALYSIS OF INDIVIDUAL STRUCTURAL FEATURE OVER OTHER STRUCTURAL FEATURES

No single character will have all the 17 structural features. This is evident from the table 1. If FS2 is true then FS4 and FS5 is false, i.e., if lower arc is present then there won't be any lower arc & loop and upper arc & loop. If FS3 is true then FS5, FS6, FS7 and FS8 is false, i.e., if upper arc is present then there won't be any lower arc & loop, upper arc & loop, dot and side bar. If FS4 is true then FS2 and FS5 is false, i.e., if slant line is present then there won't be any lower arc and lower arc and loop. If FS5 is true then FS2, FS3, FS4, FS6 and FS7 is false, i.e., if lower arc and loop is present then there won't be any upper, lower arc, slant line, upper arc & loop and dot. If FS6 is true then FS3, FS5 and FS8 is false, i.e., if upper arc and loop is present then there won't be any upper arc, upper arc & loop and side bar. If FS7 is true then FS3, FS5 and FS6 is false, i.e., if dot is present then there won't be any upper arc, lower arc & loop and upper arc & loop.

Table Error! No text of specified style in document.**: Analysis of Individual Structural Features over other Features**

Structural Features	**Loop (FS1)**	**Lower Arc (FS2)**	**Upper Arc (FS3)**	**Slant Line (FS4)**	**Lower Arc & Loop (FS5)**	**Upper Arc & Loop (FS6)**	**Dot (Pully) (FS7)**	**Side Bar (FS8)**
FS1		1	1	1	1	1	1	1
FS2	1		1	0	0	1	1	1
FS3	1	1		1	0	0	0	0
FS4	1	0	1		0	1	1	1
FS5	1	0	0	0		0	0	1
FS6	1	1	0	1	0		1	0
FS7	1	1	0	1	0	0		1
FS8	1	1	1	1	1	1	1	

Table 2: Selection of different features used for different Zones

Sub-Symbols	Features Used
Only Middle Zone	FS1, FS7 to FS17
Upper & Middle	FS1, FS3, FS6 to FS17
Middle & Lower	FS1, FS2, FS4 to FS6, FS8 to FS17
All Zone	FS1 to FS4, FS6, FS7, FS9 to FS17

6. CONCLUSION

Based on all the above analysis, selection of different features used for different zones is listed out in table 2. Structural features of FS1 to FS17 are satisfied in all the zones. Other features selections, structurally differ for different zones. Thus by selecting appropriate structural features for different zones of distorted Tamil characters, we can recognize the characters to better extent.

7. REFERENCES

[1] C. Y. Suen, M. Berthod and S. Mori, "Automatic recognition of hand printed characters- the state of the art", Proceedings of the IEEE, Vol. 68(4), pp. 469-487, 1980.

[2] O.D.Trier, A.K.Jain and T.Taxt, "Feature Extraction methods of Character Recognition: - a survey", Pattern Recognition, Vol.29 (4), PP. 641-662, 1996.

[3] L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier and C. Olivier, "A structural/statistical feature based vector for handwritten character recognition", Pattern Recognition Letters, Vol. 19(7), pp. 629-641, 1998.

[4] Robert T. Olszewski, "Generalized Feature Extraction for Structural Pattern Recognition in TimeSeries Data", Ph.D. thesis, University- Pittsburgh, 2001.

[5] L. L. Lee and N. R. Gomes, "Disconnected handwritten numeral image recognition", in the Proceedings of 4th ICDAR, pp. 467-470, 1997.

[6] Kam-Fai Chan and Dit-Yan Yeung, "An effecient syntactic approach to structural analysis of on-line handwritten mathematical expressions", Pattern Recognition, Vol. 33, pp. 375-384, 2000.

[7] A. Amin, "Recognition of printed Arabic text based on global features and decision tree learning techniques", Pattern Recognition, Vol. 33, pp. 1309-323, 2000.

[8] S. Kahan, T. Pavlidis and H. S. Baird, "On the recognition of printed characters of any font and size", IEEE Transaction, .Pattern Analalysis . Mach. Intell, Vol. 9, pp. 274-288, 1987.

[9] CEDAR. "Penman: Handwritten Text Recognition Project Description". Available Thein from http://www.cedar.buffalo.edu/Penman/description.htm

[10] J. Rocha and T. Pavlidis, "A shape analysis model with applications to a character recognition system", IEEE Transactions on PAMI, Vol. 16(4), pp. 393-404, 1994.

[11] U.Pal and Anirban Sarkar, "Recognition of Printed Urdu Script", Proceedings of 7th Int. Conf., on Document Analysis and Recognition, pp.1183-1187, 2003.

[12] G. Leedham and V. Pervouchine, "Validating the use of handwriting as a biometric and its forensic analysis", in the Proceedings of International Workshop on Document Analysis (IWDA), India, pp. 175-192, 2005.

[13] Anisimovich, K. Rybkin, V. Shamis, A. Tereshchenko, V. "Using combination of structural, feature and raster classifiers for recognition of handprinted characters",., Proceedings of the Fourth International Conference on Document analysis and Recognition, Aug1997/ICDAR. 1997.620638.