DOI: 10.21917/ijivp.2021.0346

RECOGNIZING TAMIL PALM-LEAF MANUSCRIPT CHARACTERS USING HYBRIDIZED HUMAN PERCEPTION BASED FEATURES

Paramasivam Muthan Eswaran, Dinesh Manib and Sabeenian Royappan Savarimuthu
Department of Department of Electronics and Communication Engineering, Sona College of Technology, India

Abstract

In this paper, we present a zoning strategy for recognizing manuscript character images, based on human structural perception of characters. The deficiencies in a uniform zoning approach are filled by drawing significant information using the proposed method. The obtained feature set when applied on a SVM classifier, substantially improves the recognition rate for character images having structural variation at significant regions of characters. As a initiative, we have formulated the Tamil Palm-Leaf Character dataset. Preliminary results show that the incorporation of this hybridized zoning approach has improved the symbol recognition rate to 9.06% (from 81.07% to 90.13%). The average rejection rate has been nullified using this generic non-symmetrical zoning for the proposed dataset.

Keywords:

Manuscript Character Recognition, Visual perception, Triangular Zoning, Shape Based Zoning, Significant Zone Slicing

1. INTRODUCTION

Centuries before Egyptians invented 'Papyrus', knowledge was shared by inscribing on tree barks, skin hides, rocks and leaves [1]. Information present on manuscripts vary from traditional medicines, land documents, astrology, astronomy and many more. Palm leaves due to its enormous availability and capability to withstand rigorous conditioning was the most widely used manuscript material. Despite such conditionings [2, 3], the organic nature of palm leaves restricted its life to an average of 300 - 400 years.

A few samples of highly degraded palm leaf manuscripts are shown in the Fig.1. The traditional mundane of preserving information from decaying manuscripts was to copy contents to a fresh set of conditioned leaves. With the invention of printing press, the art of scribing has been lost today. The advancements in imaging technology has intervened heavily in culture preservation, particularly for photographing manuscripts. International Organizations [4], Governmental [5] [6] and Non-Governmental organizations [7]-[10] have started utilizing this technological intervention extensively for preserving information on palm-leaf manuscripts.

Manuscript Character Recognition from images has been a challenge to computer scientists for more than a decade. Few literatures have reported on OCR systems for recognition of ancient Latin and Greek scripts. Markus Diem et al. [11] have made an OCR system to recognize printed Latin text from ancient manuscripts. Barbuti and Caldarola [12] have invented an apparatus to recognize text in images of an ancient printed books and manuscripts.

The character ligatures of ancient manuscript Greek characters was investigated and a fast, efficient recognition technique was proposed by Gatos [13]. Lanna script, an obsolete script of Thailand, was recognized by Thammano and Pravesjit [14] using self-organizing maps for dividing the image to several clusters and classified using particle swarm algorithm.

For the past few decades, many literatures have reported in the areas of offline handwritten character recognition of numeric, English, Chinese, Indian and Arabic scripts. On the input perception, one might categorize manuscript images for the manuscript character recognition, while scanned document character images for the offline character recognition. A few artifacts that discriminate manuscript and offline character recognition have been elaborated in Section 2.2

Fig.1. Tamil Palm-Leaf Manuscripts degraded due to termites and dynamic climatic conditions

1.1 FEATURES AND CLASSIFIER

The extensive survey carried out by Divined Due Trier, Jain AK [15] and Texts gives a summary on various features used for character recognition. Feature extraction approaches are either directly applied on the binary image (HOG [16], R-HOG [17], CNN [18]), or to the character's simplified boundary (thinned image). This paper has utilized thinned images of characters.

Govindan and Shivaprasad [19] have broadly classified features for character recognition into two major types viz. structural and statistical. Heutte et al. [20] introduced the possibility of combining both these complementary features for better character recognition. A contour-based feature was extracted by [21] for multi-orientated and multi-size texts present in newspapers, magazines and maps.

Dynamic time warping approach implemented on English character by Rath and Manmatha [22] was tried on Tamil Characters by Niels and Vuurpijl [23]. Prasanth and et.al [24] extended the elastic matching technique with a few local features for Tamil and Telugu characters. The presence of larger character set, along with contrasting characteristics and degrees of complexity, bottlenecks the replication of existing methods used for English as such to Indian scripts. This has enabled development of numerous novel approaches.

A support vector machine Soman et al. [25] is a supervised pattern classifier for a two-class recognition. Multi-class recognitions can be approached by combining several binary SVMs with a one-versus-one (OVO) technique. The performance of any SVM classifier largely depends on the effective selection of the feature sets. The features obtained by the above approaches

can be applied to an appropriate statistical classifier, which sorts each character symbol to its respective classes. .

The choice of a good set of features along with an optimal classifier plays a key role in obtaining better recognition with minimal mis-prediction and rejection rate. Shanthi and Duraiswamy [26] carried out an initiatory in recognizing handwritten Tamil Characters using support vector machines.

Summarizing the handwriting recognition approaches, it can be classified into three basic categories: (a) direct analysis of binary pixels in an image (b) structural analysis of the image, such investigating bends, end points, intersections, loops, measures of concavity and distance information (c) applying global mathematical transformations.

Despite the success of existing methods, there is still room for improvement. One major source of error is due to similar character symbols which differ by only a very small amount, e.g. க (/ka/), சு (/su/) and ச (/sa/), ய (/ya/), ப (/pa/) and ம (ma), etc. Such dwarfish variations become soared up in handwritten characters due to varying writing styles on a paper. In case of manuscript characters these variations would still be higher, leading to confusion in the recognition process. This paper has focused on extracting features, centric to such marginal variations.

1.2 ZONING

Zoning refers to dividing an image to smaller regions. Invariable of features used, zoning methods have supported in extracting local characteristics, thereby improving the recognition rate. For an image *C*, an increased vector size is obtained by zoning the image to *M* sub-images ($M>1|\in R$) and in turn extracting features from each sub-image. Impedovo and Prilo [27] have presented an extensive survey on various zoning methods proposed and broadly classified them into static and dynamic approaches.

Literatures falling under static zoning have used uniform grids of size ($u \times v$), overlaid on the character image. The authors in [28] have indicated a loss of continuity in the feature vector even for a small variations in the character, for an intensified zoning approach. Rules for filling or removing vector spaces involve tedious tuning. Very minimal literatures have experimented character recognition under shape zoning. This paper has proposed two non-uniform zoning approaches, viz. Triangular Zoning (shape based) and Significant Zone Slicing (slice based) with a focus to extract features from marginally varying character structure of manuscript Tamil Script.

The rest of the article is organized as follows. Section 2 presents a detailed description of the proposed dataset utilized in this work along with a summary of Tamil script structure. A comprehensive analysis of various challenges for recognition manuscript Tamil characters is available in Section 2.2. The proposed zoning strategies are described in Section 3 with adequate details. Section 4 presents the performance of proposed strategies. In section 5, we summarize our work.

2. PROPOSED DATASET

The recognition of Tamil manuscript characters has not been much explored by the pattern recognition community and hence there has been no standard dataset available. This section shall detail the various steps carried out by the authors in developing the proposed dataset, which has been used for evaluating the proposed method. A brief description of Tamil script and its structural attributes have been described in this section, thereby emphasizing the morphological complexity involved in the recognition process.

2.1 TAMIL SCRIPT

Tamil with its classical status [29], has 12 vowels (V), 18 consonants (C), one special character (S) and 216 compound characters (CC) which are formed by combining consonants and vowels. The prettiness of Tamil scripts constitutes its closed and open loops, intersection points and end points. While most language scripts have only one symbol representing each vowel, Tamil language has one character in the vowel set represented with two isolated symbols (ஔ/au/), making the language distinct from other languages. The method in [30] have detailed on various properties of Tamil Script.

Identical to Indian scripts (Devanagari, Oriya, Punjabi, etc.), certain members of the CC set of Tamil script also possess upper, middle and lower zones (Connected Compound Characters-CCC). Fig.2 shows a compound character categorized over upper, middle and lower zones. Literatures [31] – [33] have focused on categorizing characters by analyzing these longitudinal regions, thereby classifying compound characters. A hybridized latitudinal and longitudinal region analysis will effectively target curves present in the consonant, which is very much essential for improvising recognition rate of manuscript characters.

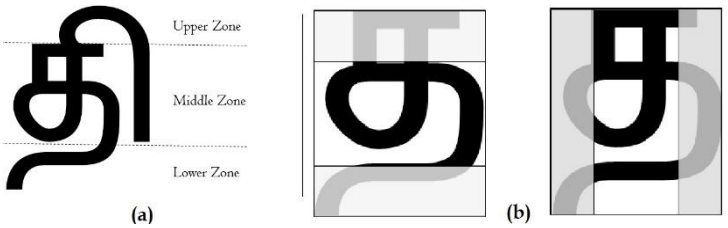

Fig.2. (a) Upper, Lower and Middle Zones in a Compound Tamil Symbol (b) Proposed middle zone centric zoning for character symbols

A few members in set CC which have the consonant either preceded or followed by unique symbols (Y). In definite cases, consonant members are preceded as-well-as followed by these symbols. The possible precedes and follows of consonants are: ◌ெ , ◌ே, ◌ொ, ◌ோ, ◌ௌ and ◌ை. (The dotted circle indicates its replacement with elements of C set). This paper has termed such characters as Multi-Symbol Characters (MSC). A complete character set in printed form is presented in Appendix. The authors would recommend the reader to look on the same for further understanding.

2.2 CHALLENGES IN MANUSCRIPT CHARACTERS

Contrasting to the printed form of scripts, the handwritten forms fluctuate in their structural representation. Palm leaf being a supple material is endangered of being damaged during scribing (writing) process. This constraint entangled the scriber (writer) to alter writing fashion for every character, thereby aggregating the

complexity to modern day recognition systems. The other constraints to be addressed for obtaining an efficient recognition rate are:

- Marginally low contrast between fore-ground and back-ground.
- Uneven back-ground color, thereby challenging the gray-scale conversion algorithms.
- Touching/overlapping in closed loops of character structures, thereby creating confusion amongst character symbols.

A sample palm-leaf manuscript exhibiting the above listed confronts is shown in Fig.3

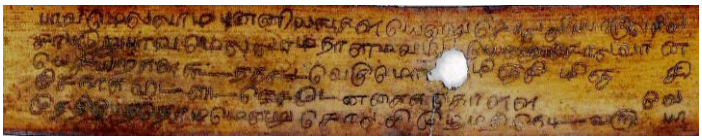

Fig.3. Poorly imaged palm-leaf manuscript exhibiting the listed confronts

A group of identical symbols extracted from a set of palm-leaf manuscripts written by a scriber would explicitly show a discrepancy in the character structure. This paper has predominantly tried to recognize such with-class structural differences in manuscript Tamil character symbols. In Fig.4, a few samples of isolated manuscript character வ (/va/) have been given. These samples depict the degree of challenge involved in manuscript character recognition.

Fig.4. Samples from Dataset for Tamil Character வ (/va/).(Samples have been normalized to equal size for better visual perception)

2.3 DATA COLLECTION

The dataset used in this paper has been developed utilizing palm-leaf manuscripts from 'Agathiyar Vaithiyam' [34], a treatise on traditional medicine (Siddha) written by a saint Agasthiyar, centuries ago. Around 50 manuscript images were randomly selected. The selection criteria was based on better visibility of the scribed characters with minimal degradation of the leaf. Expert advice was sought to ensure that the chosen images were inscribed by the same scriber.

Isolated character symbols (20,000 images) were clipped and in-turn binned in 60 different classes. In case of multi-symbol characters (MSC), each symbol was considered as a separate class. An investigation on the total number of images obtained under each class was carried upon. The constraint of manuscript getting damaged due to intense curved characters has restricted the usage of many characters. For example, a compound character ழ (/za), considered as the retroflex approximant of Tamil, contributed only 37 numbers in our analysis. A few other members in CCC character set that have been minimally utilized on palm-leaf manuscripts are :(/tii/),(/Nuu/),(/thuu/),(/Nuu/), (/lii/),(/Luu/) and(/sri/).

2.4 DATASET SIZE

Around 8000 isolated manuscript Tamil character symbols from CC and Y set were identified. Ultimate care has been taken such that, there has been no drastic variation of degradation within a particular image. Our experiments started with applying geometric features on a M = 9(3×3) uniform grid. A group of 20 symbols were chosen viz., ள, ன, ை, ெ, ே, க, , ச, ண, ட, ய, ப, ம, ர, ா, ல, வ, த, ந and உ. The choice for these selected character symbols was due to its predominant usage in palm-leaf manuscripts (nominal structural complexity), thereby fairly increasing the investigation and training samples.

The confusion matrix was analyzed and character symbols ெ, ே,உ,ண and ட exhibited a higher recognition rate of close to 89%. The structure of these symbols do not represent any structural similarity with other symbols. An abstract of the confusion matrix is shown in Fig.5. In this paper, marginally varying character symbols have been termed as Confusion Groups (CG). Out of the 20 character symbols used, 15 were grouped to form six CG sets.

The Table.1 shows the confusion group sets, along with the number of images under each class in the proposed dataset. The first three groups CG1, CG2 and CG3 have a combination of three symbols, while the later (CG4, CG5 and CG6) have a two symbol confusion. Though the structural variation of symbols in printed form is very meagre, the variations on a manuscript relatively highly.

	ள	ன	ை
ள	53.33	2	1.6
ன	42.22	91	21
ை	2.96	2.6	58

	க	சு	ச
க	89	18	14
சு	0.5	55	0.3
ச	6.3	7	80

	ய	ப	ம
ய	72	27	3.6
ப	14	68	0.3
ம	3.7	1.5	90

	ல	வ
ல	80	36
வ	12	58

	த	ந
த	90	20
ந	1.6	66

	ர	ா
ர	68	1.1
ா	23	67

Fig.5. Synopsized Confusion Matrix suggesting possible Confusion Groups (Values indicate percentage)

Table.1. Confusion Groups of various classes along with number of images in training and test dataset

Confusion Group	Character Symbol	Character Set	# Training Dataset	# Test Dataset
CG1	ள ன ை	CC CC Y	25 119 26	135 608 127
CG2	க சு ச	CC CC CC	141 19 78	742 98 374
CG3	ய ப ம	CC CC CC	100 56 144	573 343 777

CG4	ல வ	CC CC	82 54	442 325
CG5	த ந	CC CC	140 43	752 229
CG6	ர ெ◌ர	CC Y	39 97	202 521
		Total	1163	6248

3. PROPOSED METHOD

Jean-Luc Chevillard [35] had proposed as how neural networks can be used for character recognition of symbols extracted from Palm-leaf manuscripts. Since then, very few scientist [36, 37] have taken up this specific area of application research. To the best of our knowledge, there has been no experiment conducted to recognize characters from images of Tamil palm-leaf manuscripts and this paper has made a maiden approach. An important contribution of this work is to provide an effective perceptual zoning for manuscript character recognition.

3.1 PERCEPTUAL ZONING MECHANISM

People categorize physical entities into different classes by effectively sorting out its surroundings. This is done using basic elements of perception based on Gestalt Theory [38]. These perceptions can be properly mapped to build features and we call these 'Visual Element Features'. In case of character identification, humans converge on Visual Element Features, such as significant Concavities/Convexities present in characters for effective recognition.

Invariant of the script, in the recent years, recognition of characters is much focused on character symbols which differ marginally [39]. We examined how Visual Element Features under a uniform grid based zoning gets comprehensible for structurally similar Tamil manuscript character symbols. The region's leading such confusions were dissected, based on which two approaches have been proposed in this paper. The proposed non-uniform and shape based zoning does not use any complex and extensive algorithms to design the zoning.

We were able to determine confusions occurring at certain concrete regions for characters in the CG set. Fig.6 demonstrates samples of manuscript symbols க (/ka/), ச (/sa/), த (/tha/) and ந (/nha/). It is important to realize that these are only samples demonstrating the structural variations and those indicated are not homogeneous. Nevertheless, the region identified were found to remain consistent for members in the six CG sets.

Fig.6. Sample Manuscript Character Skeletons (க (/ka/), ச (/sa/), த (/tha/) and ந (/nha/)) having marginal variation in their structure have been indicated in red circles.

Enciphering of curves present in structures would be more complex and hence the last two parts mentioned above can be used for the same. Unlike English characters, having minimal intersection points, Tamil characters have fair number of intersections. Approaches [40] - [57] have focused on increasing the number of sub-images M, thereby intensifying the feature vector space with an attempt to achieve minimal error rate. To have a better separability amongst the extracted features, a 3×3 grid has been used (Zones M1 − M9).

With the proposed dataset having marginal fluctuations in character pattern, an intensified zoning approach makes it infeasible to capture all vital feature information. Visual inspection of confusions between character symbols indicate that they share common structures and are meagerly different in some critical parts. As an example, we observe that the symbols ல (/la/) and வ (/va/) differ primarily in the middle region. The confusion pair க (/ka/), (/su/) and ச (/sa/) present structural differences at the end points. We experimented on various samples of manuscript characters and identified as how the feature vector obtained by uniform zoning misses such marginal structural dissimilarities.

For the reader to infer, we present the geometric features extracted on a confusion pair க (/ka/ and ச (/sa/)) using a 3×3 uniform zone. Fig.7 shows the superimposition of a uniform 3×3 grid on the both these characters. The basic perception on the structure of க (/ka/ and ச (/sa/)) is the presence of ending curve in the symbol (க (/ka/)). A uniform 3×3 zoning flushes out this marginal vital curve indicated, thereby leading to a misclassification. In the Fig.7, identical pixel values have been mapped for the sake of analysis and the critical curve region has been clearly focused. The presence and absence of concavity in the confusing symbols has been apparently missed thereby leading to misclassification.

In order to extract the curves and intersection points present in the Tamil script, we have proposed a perceptual shape based zoning and a significant slicing approach.

The regions of a character that claim, marginal structural variations were identified and hence, windows centered on such regions have been proposed.

- Diagonals leading to Triangular Zoning
- Longitudinal and Latitudinal slicing of Middle Zones

3.1.1 Triangular Zoning:

The set of features extracted by this approach are focused to discriminate the marginal concavity and convexity of manuscript characters. The skeleton of character image is fitted using bounding box. Each corner of the box is labeled as (x_0,y_0),(x_1,y_1),(x_2,y_2) and (x_3,y_3). Diagonal vectors are calculated using Euclidean distance measure between points (x_0,y_0),(x_2,y_2) and $(x_1,y_1)(x_2,y_2)$. With the computed diagonal as the hypotenuse, right angled triangles are constructed to mask the character. Fig.8 gives a better picture of how triangular zoning is carried out. The proposed approach has provided a better visualization of perceptual information contained in the character symbols of CG set.

The Fig.9 demonstrates the effective capture of minor concavity between manuscript character symbols for the confusion sample discussed in the previous section.

3.1.2 Significant Zone Slicing:

The idea behind observing the middle region of character symbols is to give more emphasis to confusion zones. The appropriate measure for region to be analyzed was identified using a printed character symbol. Fig.10 shows the regions in sample characters that vitally contribute towards the structural

variations. Unmasked regions depicted in the image are termed as Significant Zones in this paper. The discriminative feature between ன and ன is the vertical line in ன and the closed curve in ன. Likewise, the closed loop in த differentiates it from ந. Due to space constraints, we have restricted in showing the discriminative zones for two confusions pairs viz. (ன and ன) and (த and ந). Nevertheless, this analysis is true for all the members in the CG set. A measure of concavity / convexity on sub-images obtained by both these proposed methods show a vital discrimination between confusing characters.

Unlike the printed character symbols, the size of manuscript characters varied over case-by-case. A universal mask cannot be overlaid on the character symbols. We conducted a number of investigations on manuscript samples to identify the exact measurement for masking significant regions. On perceptual assessment, it was identified that the significant regions contributed almost half the size of character symbol. An effective slicing approach aimed at the significant region of characters was identified as shown in Fig.11.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed recognition process involves the following stages:

Step 1: Preprocessing

Step 2: Hybridized Feature Extraction using

a. Uniform Zoning (3×3)

b. Triangular Zoning

c. Significant Zone Slicing

Step 3: Classification using a SVM classifier

4.1. Preprocessing Steps

We have used the classical weighted grayscale conversion, rgb2gray command of MATLAB for the sake of simplicity. An analysis of various binarization algorithms on palm-leaf manuscripts was carried out by [58] [59] and [60]. We have utilized the widely exploited Global Thresholding method proposed by Otsu [61].

The proposed approach emphasizes on the structural characteristics of symbols and hence the thinned version of the binary image without normalization (resizing the image size) has been used. Fig.12 illustrates binarized samples from the proposed dataset constituting structurally distorted character symbols. On each row, the left most character is the character in printed form, and the rest are sample images extracted from the proposed dataset.

4.1 HYBRIDIZED FEATURE VECTOR

The recognition of characters from images of palm-leaf manuscripts has not been much attractive in the image processing community. The main motive of this work is to target the minor structural disparities in curves and loops of Tamil Manuscript Characters. Features of a character can be termed as vital information that helps computers to categorize them based on human perception. Dana Ballard and Chris Brown [62] have outlined Euler Number of a binary image as the number of connected components in the image minus number of 'holes' in the image. Connected components have been chosen both on 4 and 8 connectivity basis. These features along with normalized skeleton area were evaluated on a global and localized ground.

Blumenstein and et al. [63] have proposed a 3×2 gridded direction features and hence compared it with the transition of pixels in vertical and horizontal direction from background to foreground. The structure of character symbols comprise of straight lines, curves and in certain cases a period on the character (mostly in Indic scripts). Considering only the case of lines and curves, the features have been categorized into four main parts. viz., Vertical Line (v), Horizontal Line (h), Right Diagonal Line (r) and Left Diagonal Line (l). The number of elements falling under each of the above mentioned category are identified on a localized basis and normalized to the zone. The length of each element is also normalized on a zonal basis.

The Fig.13 shows structural features that can be extracted from the skeleton of Tamil Palm-Leaf Manuscript character க (/ka/) on a uniform grid 3×3. This constitutes a feature vector set of size 81.

In our experimental implementation, the directional geometry have been extracted from 4 triangular zones (36 feature set) and 2 significant zones (18 feature set). These feature vectors have been amended with the 81 features thereby forming hybridized approaches, viz., Method 1: 3×3 Uniform Grid + 4 Triangular Zone Features; Method 2: 3×3 Uniform Grid + Significant Zone Slicing; Method 3: 3×3 Uniform Grid + 4 Triangular Zone Features + Significant Zone Slicing.

4.2 CLASSIFICATION USING SVM CLASSIFIER

The augmented feature vectors obtained by the three methods are fed to SVM, with parameters $C = 3$ and a quadratic polynomial kernel. A five-fold cross validation was performed for the kernel and parameters to be optimally set. To have some basis for comparison, we have used was $Z = 9$ (3×3 Grid) and $Z = 6$ (3×2 Grid). The average recognition rates of both these methods varied marginally by 0.12% (Refer Table.2).

Table.2. Performance comparison of Average Recognition Rate of Proposed Feature Set

Method	Number of Features	Avg. Recognition %
Blumenstein	57	81.06
3Method I ×3 Grid	84 97	81.18 83.08
Method II	111	85.48
Method III	125	**90.13**

An average recognition rate of 90.5% is achieved on the proposed dataset using the Method 3 hybridized feature set. Fig.14 and Fig.15 show the average recognition rates for the 15 classes grouped in their respective CG set. On a general perspective the proposed hybridization in Method 3 has outmatched the other hybridization approaches.

4.3 ANALYZING MIS-CLASSIFICATIONS

In general, the misclassification group can be categorized into three: (1) character symbols identifiable through visual perception

to the correct class, but has been misclassified by the approach to another class, (2) character symbols that are not identifiable through human perception (due to heavy degradation, over binarization and many other hurdles) and (3) character symbols wrongly binned in a different class.

We made a careful investigation on the 608 misclassified samples (out of 6248 test samples) to identify the exact misclassification. Utmost care was taken during the dataset creation and hence sample in the last category never appeared. The first category contributed about 150 images (one fourth), while the second category constituted almost three fourth (458 samples) of the misclassified set. Excluding the unrecognizable category, the theoretical maximum achievable recognition rate is about 97.40%.

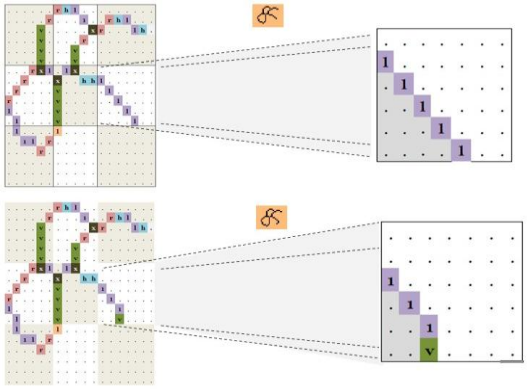

Fig.7. Ambiguity of concaveness demonstrated in the uniform grid. Zone M6 has been detailed to show the concavity acquired for characters க (/ka/ and ச (/sa/)

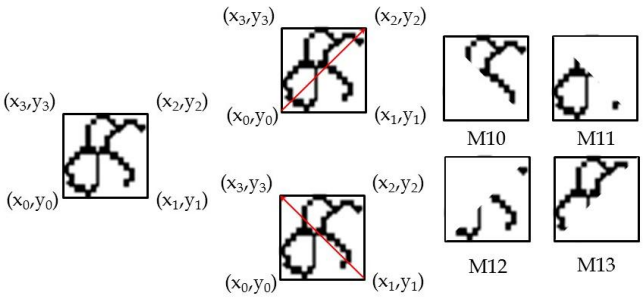

Fig.8. Triangular Zoning (a) Bounding Box and fixing co-ordinates to corners (b) Identifying diagonals (c) Masking upper and lower regions of diagonals alternatively

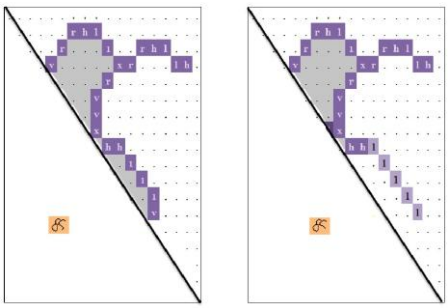

Fig.9. Triangular Zoning discriminating character symbols க (/ka/ and ச (/sa/) based on concaveness of the end points

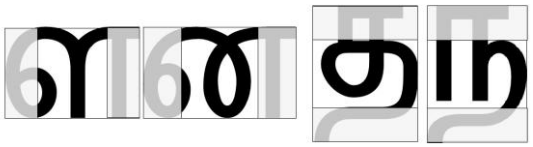

Fig.10. Significant Zone Slicing on a Printed Tamil Character

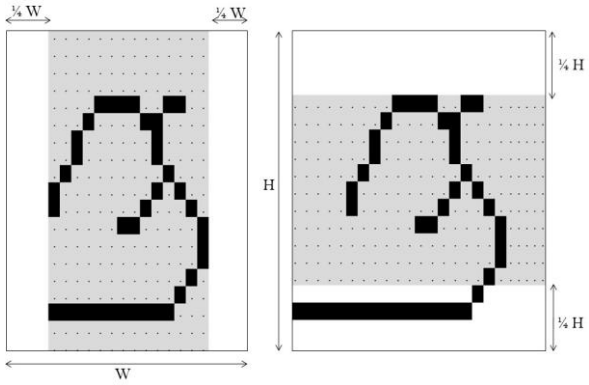

Fig.11. Significant Zone Slicing of a Binary Manuscript Character Symbol ந (/nha/))

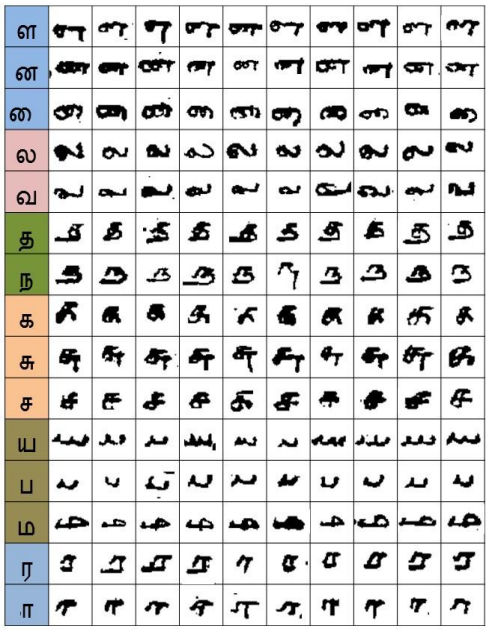

Fig.12. Binarized Character Samples from the dataset utilized for experimentation

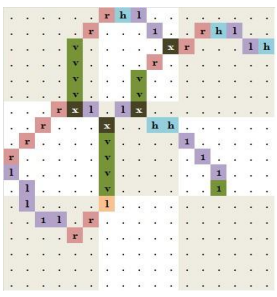

Fig.13. 3×3 Uniform Zone Directional Features on a sample manuscript character க (/ka/) [v. Vertical Line, h. Horizontal Line, r. Right Diagonal and l. Left Diagonal]

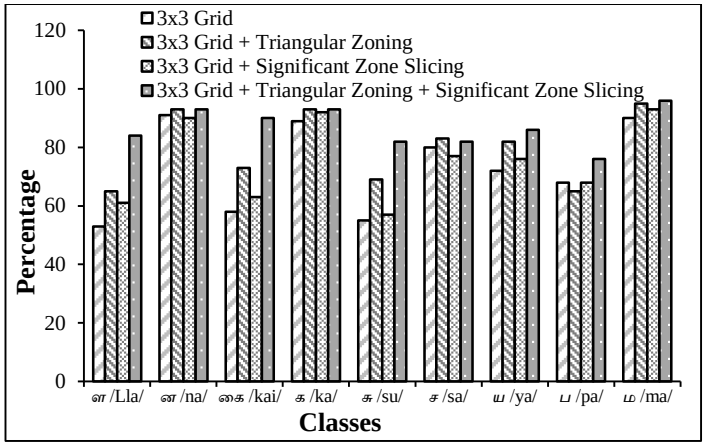

Fig.14. Recognition Rate of Individual Classes using different methods for Cluster Groups CG1, CG2 and CG3

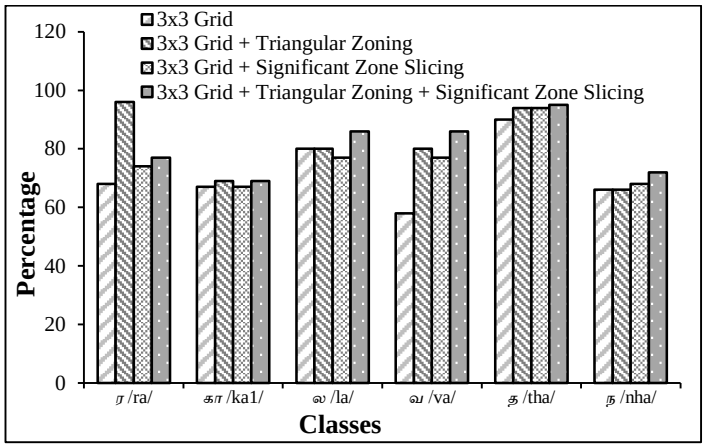

Fig.15. Recognition Rate of Individual Classes using different methods for Cluster Groups CG4, CG5 and CG6

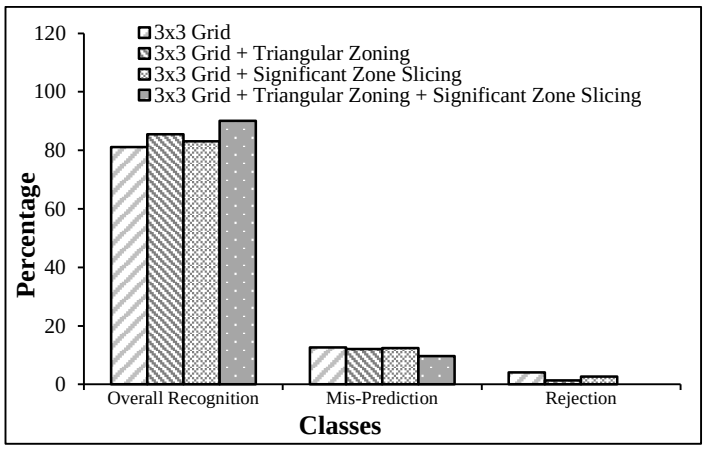

Fig.16. Comparison of Average Recognition, Mis-prediction and Rejection rates

Our main objective was to explore, as how far we can go ahead in decreasing the mis-prediction rate of the proposed system, using simple zoning approaches and geometries of characters. The Fig.16 shows a comparison of average recognition, mis-prediction and rejection rates for the proposed methods with existing approaches. The hybridization has explicitly reduced the rejection rate to zero. To the best of our knowledge, this work is novel towards recognizing characters present in Tamil palm-leaf manuscripts.

5. CONCLUSION

This paper has presented character recognition approach for manuscript characters present on a palm-leaf manuscripts. The proposed approach is focused on improving the mis-prediction of characters due to marginal structural variations. Multiple SVMs with a quadratic kernel have been employed to classify the global and local feature set. The zoning approaches proposed in this paper have been studied on a basis of human perception. While triangular zoning, a shape based zoning, focuses on marginal concavity and convexity of characters, the significant zone slicing helps in featuring small and distinct information present in middle zones. A dataset was formed by grouping an elite set of character symbols from palm-leaf manuscripts. The mis-prediction rate of existing uniform grid approaches was studied. Confusion groups were identified from the results, thereby paving way for improving the recognition rate.

REFERENCES

[1] D.B. Diskalkar, "*Materials used for Indian Epigraphical Records*", Bhandarkar Oriental Research Institute, 1979.

[2] G. John Samuel, "*Uthirum Malargal*", The Institute of Asian Studies, 1994

[3] G.John Samuel, "*Kumari Muthal Warsaw Varai*", The Institute of Asian Studies, 1994.

[4] UNESCO, "Memory of the World Programme", Available at http://www.unesco.org/new/en/communication-and-information/flagship-project-activities/memory-of-the-world/homepage/, Accessed at 2020.

[5] National Mission for Manuscript-The Republic of India, Available at http://www.namami.org/, Accessed at 2020.

[6] Tamil Virtual Academy, Available at http://www.tamilvu.org/library/suvadi/html/index.htm, Accessed at 2020.

[7] The Institute of Asian Studies, Available at www.instituteofasianstudies.com, Accessed at 2020.

[8] Project Madurai, Available at http://www.projectmadurai.org/pmworks.html, Accessed at 2020.

[9] Chinmaya International Foundation (CIF), Available at http://www.chinfo.org, Accessed at 2020.

[10] Tara Prakashana, Available at http://www.taraprakashana.org/, Accessed at 2020.

[11] M. Diem and R. Sablatnig, "Recognizing Characters of Ancient Manuscripts", *Proceedings of Computer Vision and Image Analysis of Art*, pp.1-13, 2010.

[12] N. Barbuti and T. Caldarola, "An Innovative Character Recognition for Ancient Book and Archival Materials: A Segmentation and Self-learning Based Approach", *Proceedings of International Conference on Digital Libraries and Archives*, pp. 261-270, 2012.

[13] K. Pratikakis, I. Petridis, S. Konidaris and S.J. Perantonis, "An Efficient Segmentation-Free Approach to Assist Old Greek Handwritten Manuscript OCR", *Pattern Analysis and Applications*, Vol. 8, No. 4, pp. 305-320, 2006.

[14] Arit Thammano and Sakkayaphop Pravesjit, "Recognition of Archaic Lanna Handwritten Manuscripts using a Hybrid Bio-Inspired Algorithm", *Memetic Computing*, Vol. 7, No. 1, pp. 1-17, 2015.

[15] Oivind Due Trier and Anil K. Jain and Torfinn Taxt, "Feature Extraction Methods for Character Recognition- A Survey", *Pattern Recognition*, Vol. 29, No. 4, pp. 641-662, 1996.

[16] S. Iamsa-At and P. Horata, "Handwritten Character Recognition using Histograms of Oriented Gradient Features in Deep Learning of Artificial Neural Network", *Proceedings of International Conference on IT Convergence and Security*, pp. 1-5, 2013.

[17] Hari Vasudevan, Abhijit R. Joshi, Narendra M. Shekokar and Parshuram M. Kamble, "Handwritten Marathi Character Recognition using R-HOG Feature", *Proceedings of International Conference on Advanced Computing Technologies and Applications*, pp. 266-274, 2015.

[18] K. Mehrotra, S. Jetley and S. Belhe, "Unconstrained Handwritten Devanagari Character Recognition using Convolutional Neural Networks", *Proceedings of International Conference on Multilingual OCR*, pp. 1-5, 2015.

[19] V.K. Govindan and A.P. Shivaprasad, "Character Recognition: A Review", *Pattern Recognition*, Vol. 23, No. 7, pp. 671-683, 1990.

[20] J.V. Moreau, Y. Lecourtier and C. Olivier, "A Structural Statistical Feature Based Vector for Handwritten Character Recognition", *Pattern Recognition Letters*, Vol. 19, No. 7, pp. 629-641, 1998.

[21] P.P. Roy, U. Pal, J. Llados and M. Delalandre, "Multi-Oriented and Multi-Sized Touching Character Segmentation using Dynamic Programming", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 11-15, 2009.

[22] T.M. Rath and R. Manmatha, "Word Image Matching using Dynamic Time Warping", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 521-527, 2003.

[23] R. Niels and L. Vuurpijl, "Dynamic Time Warping Applied to Tamil Character Recognition", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 730-734, 2005.

[24] L. Prasanth, V. Jagadeesh Babu, R. Raghunath Sharma and G.V. Prabhakara Rao, "Elastic Matching of Online Handwritten Tamil and Telugu Scripts using Local Features", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 1028-1032, 2007.

[25] K.P. Soman, R. Loganathan and V. Ajay, "*Machine Learning with SVM and Other Kernel Methods*", PHI Learning Publisher, 2009.

[26] N. Shanthi and K. Duraiswamy, "A Novel SVM-Based Handwritten Tamil Character Recognition System", *Pattern Analysis and Applications*, Vol. 13, No. 2, pp. 173-180, 2010.

[27] D. Impedovo and G. Pirlo, "Zoning Methods for Handwritten Character Recognition: A Survey", *Pattern Recognition*, Vol. 47, No. 3, pp. 969-981, 2014.

[28] Jun Cao, M. Ahmadi and M. Shridhar, "Recognition of Handwritten Numerals with Multiple Feature and Multistage Classifier", *Pattern Recognition*, Vol. 28, No. 2, pp. 153-160, 1995.

[29] The Hindu, "World Classical Tamil Conference - A Perspective", Available at http://www.thehindu.com/opinion/op-ed/world-classical-tamil-conference-a-perspective/article444941.ece, Accessed at 2012.

[30] S. Sundaram and A.G. Ramakrishnan, "Performance Enhancement of Online Handwritten Tamil Symbol Recognition with Re-Evaluation Techniques", *Pattern Analysis and Applications*, Vol. 17, No. 3, pp. 587-609, 2014.

[31] N. K. Garg, L. Kaur and M. Jndal, "Recognition of Offline Handwritten Hindi Text using Middle Zone of the Words", *Proceedings of International Conference on Computer and Information Science*, pp 325-328, 2015.

[32] A.K. Bhunia, A. Das, P.P. Roy and U. Pal, "A Comparative Study of Features for Handwritten Bangla Text Recognition", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 636-640, 2015.

[33] Partha Pratim Roy, Ayan Kumar Bhunia, Ayan Das, Prasenjit Dey and Umapada Pal, "HMM-based Indic handwritten Word Recognition using Zone Segmentation", *Pattern Recognition*, Vol. 47, No. 1, pp. 1-24, 2016.

[34] Digital Manuscript Gallery, Available at http://www.bdu.ac.in/suvadi/1.1/, Accessed at 2020.

[35] Jean Luc Chevillard, "A Proposal for the Digital Encoding of Palm-Leaf Tamil Manuscripts", *Proceedings of International Conference on Tamil Internet*, pp. 109-121, 2003.

[36] T.R. Vijaya Lakshmi, Panyam Narahari Sastry and T.V. Rajinikanth, "A Novel 3D Approach to Recognize Telugu Palm Leaf Text", *International Journal Engineering Science and Technology*, Vol. 12, No. 3, pp. 34-45, 2016.

[37] Narahari Sastry Panyam and N.V. Koteswara Rao, "Modeling of Palm Leaf Character Recognition System using Transform based Techniques", *Pattern Recognition Letters*, Vol. 84, No. 2, pp. 29-34, 2016.

[38] Randolph Blake and Robert Sekuler, "*Perception*", McGraw-Hill Higher Education, 2005.

[39] K.C. Leung and C.H. Leung, "Recognition of Handwritten Chinese Characters by Critical Region Analysis", *Pattern Recognition*, Vol. 43, No. 3, pp 949-961, 2010.

[40] M. Blumenstein, B. Verma and H. Basli, "A Novel Feature Extraction Technique for Recognition of Segmented Handwritten Characters", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 122-134, 2003.

[41] C.Y. Suen, J. Guo and Z.C. Li, "Analysis and Recognition of Alphanumeric Handprints by Parts", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 24, No. 4, pp. 614-631, 1994.

[42] F. Bortolozzi and C.Y. Suen, "Segmentation and Recognition of Handwritten Dates: An HMM-MLP Hybrid Approach", *Document Analysis and Recognition*, Vol. 6, No. 4, pp. 248-262, 2003.

[43] L.S. Oliveira, R. Sabourin, F. Bortolozzi and C.Y. Suen, "Automatic Recognition of Handwritten Numerical Strings: A Recognition and Verification Strategy", *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 11, pp. 1438-1454, 2002.

[44] Glenn Baptista and K.M. Kulkarni, "A High Accuracy Algorithm for Recognition of Handwritten Numerals", *Pattern Recognition*, Vol. 21, No. 4, pp. 287-291, 1988.

[45] A.L. Koerich and P.R. Kalva, "Unconstrained Handwritten Character Recognition using Metaclasses of Characters", *Proceedings of IEEE International Conference on Image Processing*, pp. 542-545, 2005.

[46] B. Verma, J. Lu, M. Ghosh and R. Ghosh, "A Feature Extraction Technique for Online Handwriting Recognition", *Proceedings of IEEE International Joint Conference on Neural Networks*, pp. 1337-1341, 2004.

[47] S. Singh and M. Hewitt, "Cursive Digit and Character Recognition in CEDAR Database", *Proceedings of IEEE International Conference on Image Processing*, pp. 569-572, 2000.

[48] G. Pirlo and D. Impedovo, "Adaptive Membership Functions for Handwritten Character Recognition by Voronoi-Based Image Zoning", *IEEE Transactions on Image Processing*, Vol. 21, No. 9, pp. 3827-3837, 2012.

[49] Sung Hyuk Cha, C.C. Tappert and S.N. Srihari, "Optimizing Binary Feature Vector Similarity Measure using Genetic Algorithm and Handwritten Character Recognition", *Proceedings of IEEE International Conference on Document Analysis and Recognition*, pp. 662-665, 2003.

[50] Atul Negi, Shanker, K. Nikhil and Chandra Kanth Chereddi, "Localization, Extraction and Recognition of Text in Telugu Document Images", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 1-12, 2003.

[51] F. Kimura and M. Shridhar, "Handwritten Numerical Recognition based on Multiple Algorithms", *Pattern Recognition*, Vol. 24, No. 10, pp 969-983, 1991.

[52] Francesco Camastra and Alessandro Vinciarelli, "Combining Neural Gas and Learning Vector Quantization for Cursive Character Recognition", *Neurocomputing*, Vol. 51, pp. 147-159, 2003.

[53] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai and L. Lam, "Computer Recognition of Unconstrained Handwritten Numerals", *Proceedings of the IEEE*, Vol. 80, No. 7, pp. 1162-1180, 1992.

[54] P. Vanaja Ranjan and V.N. Manjunath Aradhya, "Isolated Handwritten Kannada and Tamil Numeral Recognition: A Novel Approach", *Proceedings of International Conference on Emerging Trends in Engineering and Technology*, pp. 1192-1195, 2008.

[55] S.V. Rajashekararadhya and P.V. Ranjan, "Support Vector Machine based Handwritten Numeral Recognition of Kannada Script", *Proceedings of IEEE International Conference on Advance Computing*, pp. 381-386, 2009.

[56] S.V. Rajashekararadhya and P.V. Ranjan, "Zone Based Hybrid Feature Extraction Algorithm for Handwritten Numeral Recognition of South Indian Scripts", *Proceedings of IEEE International Conference on Contemporary Computing*, pp. 138-148, 2009.

[57] S.V. Rajashekararadhya and P.V. Ranjan, "The Zone-Based Projection Distance Feature Extraction Method for Handwritten Numeral/Mixed Numerals Recognition of Indian Scripts", *Proceedings of IEEE International Conference on Frontiers in Handwriting Recognition*, pp. 617-622, 2010.

[58] Chun Che Fung and R. Chamchong, "A Review of Evaluation of Optimal Binarization Technique for Character Segmentation in Historical Manuscripts", *Proceedings of 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 236-240, 2010.

[59] Chun Cheand Wong and Kok Wai, "Comparing Binarisation Techniques for the Processing of Ancient Manuscripts", *Proceedings of International Conference on Cultural Computing*, pp. 55-64, 2010.

[60] R.S. Sabeenian, M.E. Paramasivam and P.M. Dinesh, "Appraisal of Localized Binarization Methods on Tamil Palm-leaf Manuscripts", *Proceedings of IEEE International Conference on Wireless Communications, Signal Processing and Networking*, pp. 1-13, 2016.

[61] N Otsu, "A Threshold Selection Method from Gray-Level Histograms", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 9, No. 1, pp. 62-66, 1979.

[62] Dana Ballard and Chris Brown, "*Computer Vision*", Prentice Hall, 1982.

[63] M.Blumenstein and B. Verma, "A Novel Feature Extraction Technique for Recognition of Segmented Handwritten Characters", *Proceedings of International Conference on Document Analysis and Recognition*, pp. 1-8, 2003.