International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Classification of Handwritten Tamil Characters using Variable Length Puzzle Pieces

Ashlin Deepa R N, Rajeswara Rao R

Abstract*: — *Offline handwritten character recognition system has been a challenge for Indian scripts, especially for South Indian languages. Huge number of characters of local languages including alphabets, consonants and composite characters make the recognition system more complicated. A good recognition system for subset of Tamil script, a famous South Indian script, is proposed in this work. Variable length feature vector is extracted from the thinned character image. This extracted feature is given to a novel simple classification algorithm which works based on probability. A subset of Tamil script, 20 character classes, is considered for experiment. The samples were taken from HP Labs dataset for Tamil language and a recognition accuracy of 88.15% has been produced.

Keywords *: offline; Handwritten character; Recognition; classification; feature extraction*

I. INTRODUCTION

Handwritten character recognition system is a branch of pattern recognition which recognizes electronic form of handwritten character images. India, being multilingual country, has given the freedom for its citizens to use their own local languages in their day-to-day applications. According to senses 2017, only 12 percentage of people in India, are comfortable in using English [1] casually. Paper based form-filling, railway reservation, postal address, signature verification etc. are the main real time applications in which pen-based data entry is being practiced. People residing in rural areas prefer using their local languages in everyday life. Huge number of people are involved in generating digitized version of the data which is collected on paper. If proper Handwritten Character Recognition (HCR) system is developed for local languages in India, it would help people in rural areas to explore many digital applications in their daily routine.

II. RELATED WORK

Tamil, a famous South Indian script, is the native language of huge population in India especially in Tamil Nadu. Significant number of research has been done for HCR systems of few languages like Chinese, Arabic, Bangla, Devanagari etc., [2-8].

Revised Manuscript Received on October 30, 2019.
***** Correspondence Author
Ashlin Deepa R N*, Department of Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technology Hyderabad, India. Email:deepa.ashlin@gmail.com
Rajeswara Rao R, Department of Computer Science and Engineering, JNTUK University College of Engineering, Vizianagaram Andra Pradesh, India. Email: raob4u@yahoo.com

But, very few literature is available for the Tamil script. Due to the great similarity between inter-class character symbols, the handwritten character recognition system for Tamil language fails to show desired
performance [9-12]. Feature extraction plays a major role in any pattern recognition problems [13]. Most of the approaches in literature used fixed length feature vector [9-12, 14-16]. The strokes are extracted from thinned character image and number of matching segments and the distance between the image and frame at 16 directions has been considered as Normalized Feature Vector (NFV). This NFV is given for classification and an accuracy of 87.45% is produced for 20 character classes. Here, the classification is done on fixed length feature vector [15].

As handwritten characters are in varying shape, the number of features generated for any character image varies according to the size and structure of image contour. In many cases, fixed length feature vectors are generated by clubbing variable length features. In our previous approaches, variable length feature vectors are considered and better recognition accuracy is produced. As the standard classifiers work on fixed length vectors, novel classifiers which can process with variable length features and can classify the character images without compromising on recognition accuracy are presented [17-19].

In this paper, character image is represented as variable length feature vector in a sequence of 3 characters as an element of a set. Thus, the features are enclosed in a set. The classification of character image is done through simple matching and decision is taken using probability.

III. METHODOLOGY

The architecture of the proposed offline HCR system is shown in Fig. 1. The character images are collected from HP Labs dataset for Tamil script. The character image is binarised and used for further processing. The image is thinned, segmented and made into a polygonal shape according to the procedure in [15, 16]. The segments in polygon take a length of 6 pixel points in our experiment. Each segment of the polygon need to be labelled to form primitive string. The segmented polygon of Tamil character 'KA' is given in Fig. 2.

Four basic features form pattern in our proposed method. The pattern primitives need to be identified based on the Algorithm 1, in Section 3. The four primitives include horizontal stroke (-), vertical stroke (|), right slanted stroke (\) and left slanted stroke (/).The sequence of four primitives generated according to the Algorithm 1, form feature vector.

As handwritten character image shows massive variability in shape, the primitives developed from those images will be of variable in length. The variable length feature vector generated for the character image 'KA' in Fig. 2 is,
hvlhrvhvhvh
which is a combination of character strings.

Algorithm1

The following algorithm is meant for Primitive stroke identification and labeling. Four basic features are used for labeling the polygonal image. The features are given below.

Feature	Label	Feature	Label
Horizontal stroke -	*h*	Right slant stroke \	*r*
Left slant stroke /	*l*	Vertical stroke |	*v*

The membership functions μ_h and μ_v for the feature to be horizontal and vertical strokes are given as

$$\mu_h = 1 - |m_x| \quad \text{for } |m_x| < 1 \qquad (1)$$
$$= 0 \quad \text{for } |m_x| > 1$$
$$\mu_v = 1 - |1/m_x| \quad \text{for } |m_x| > 1 \qquad (2)$$
$$= 0 \quad \text{for } |m_x| < 1$$
$$\mu_{ob} = 1 - |(\theta - 45)/45| \quad \text{for } 0 < |m_x| < \infty \qquad (3)$$

Where m_x is the slope of the feature. The Membership function for the feature to be oblique i.e., either left slant or right slant is expressed as where $\theta = \tan^{-1} m_x$. The sign of m_x determines whether the feature is left slant or right slant.

Algorithm**:** Labelling Algorithm

1. The slope m_x of the feature is obtained by least square error method. μ_h, μ_v and μ_{ob} are determined using Eqs. (1)-(3).
2. The maximum of μ_h, μ_v and μ_{ob} is computed.
3. If μ_h is found maximum then the feature is labeled as horizontal stroke *h*. If μ_v is found maximum, then the feature is labeled as vertical stroke *v*. Else if μ_{ob} is found maximum, then the feature is labeled as right slant *r* for positive m_x and is left slant *l* for negative m_x.

A. Formation of Multiset of Puzzle pieces (MPP)

The primitive string of characters produced through the Algorithm 1 is divided to form Multiset of puzzle pieces (MPP). Each element of the multiset is the sequence of 3 primitive strokes in the order of generation of primitive string. This series of primitive strokes represents the direction of writing of the character image partially. The formation of MPP consists of following steps.

Step 1: The primitive string is padded with anchors ($) at both ends.

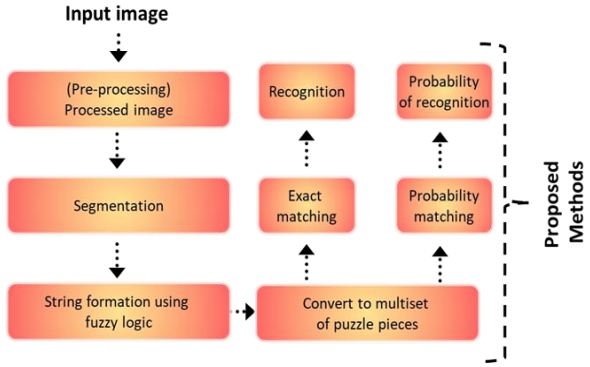

Fig. 1 Architecture of MPP method

Fig. 2 Segmented polygon of the Tamil character 'KA'

Step 2: A mask of 3 consecutive one's, {111}, is placed over the substring of the primitive string and a dot product is applied to generate one puzzle piece.

Step 3: This process in step 1, is repeated for all characters in the primitive string beginning at each character and thus producing multiset of puzzle pieces. Puzzle pieces produced for the character image 'KA' in Fig. 2, is given below.

{$hv,hvl,vlh,lhr,hrv,rvh,vhv,hvh,vhv,hvh,vh$}

The multiset of puzzle pieces are generated for all the images in the training set and stored in the database. The number of puzzle pieces generated for each character image varies as the primitive string is variable in length according to the size and shape of the character image. These puzzle pieces undergo classification procedure.

B. Classification

Two classification approaches are defined in our work.

1) Exact matching
2) Probability matching

Exact matching:

The input character image is classified only if the puzzle pieces are available in the database exactly in the same order and in the same number for any character image in the training set. For example, the puzzle pieces for the 'KA' will be classified only if the puzzle pieces {$hv, hvl,vlh,lhr, hrv,rvh,vhv,hvh, vhv,hvh, vh$} are available in the database in the same order and in the same number.

Probability Matching:

The input character image generates probability of matching with the images in the training set. The procedure for probability matching is as follows.

Step 1: Generate puzzle pieces for training set and input image.

Step 2: Find the probability matching for the input image with each entry of the database using (4).

$$prob(i) = \frac{Number\ of\ puzzle\ pieces\ of\ test\ image\ matching\ with\ ith\ training\ image}{Number\ of\ puzzle\ pieces\ of\ ith\ training\ image} \qquad (4)$$

Step 3: The class of the training image with highest value in *Step 2*, and is above .7, is considered as the class of the input image.

Table 1 shows an example of small dataset, generation of multiset of puzzle pieces and probability matching for input character image. Twenty classes are used for experiment but, only seven classes and one entry per class is shown in the table for convenience.

C. Result and discussion

In this paper, 20 character classes of Tamil script are considered for experiment. The only standard dataset available for Tamil language is given by HP Labs [17] which is used for the experiment.

Our proposed probability method produces good recognition accuracy of 88.15%. The strings are traced in clockwise direction and are concatenated serially. The size of the character image is adjusted to a smaller frame size of 20x20 which helps in generating least possible number of significant information from the character image. The exact matching method produced less recognition accuracy of 78.4% because a small mismatch in the multiset of puzzle pieces leads to misclassification of input image.

The result in NFV [15] has been used for comparison. In NFV, the same set of character classes were used and the experiment was done on 2500 manually collected samples. Though the primitive string generated from each character image is variable in length, the count of each type of strokes matching between training and test image and the distance between frame and the image at 16 directions, used for classification. This makes the features fixed in length during classification. The accuracy produced using NFV is 87.15% for the same set of 20 classes used in our experiment. In the proposed probability method, the use of ordered strokes in the form of puzzle pieces increases the recognition accuracy over count of strokes in NFV from 87.45% to 88.15%.

1) Parameter Selection

The main parameters used in our proposed method are size of the frame, length of the mask and length of a primitive segment. The optimal values for these parameters were decided after repeated runs of the experiment for different training and test samples.

Table 1. Puzzle pieces generated for seven handwritten characters and the probability matching of input character

Character	Labelled string in database	Multiset of Puzzle pieces	Input	Probability Matching Count (MP/Input)
KA	hvlhrvh rvhvh	{$hv,hvl,vlh,lhr,hrv,rvh,vhr,hrv,rvh,vhv,hvh,vh$}	$hvlhrvhvh vh$ {$hv,hvl,vlh,lhr,hrv,rvh,vhv,hvh,vhv,hvh,vh$}	**8/11**
THU	hvlhrlv hrvlhrh v	{$hv,hvl,vlh,lhr,hrl,rlv,lvh,vhr,hrv,rvl,vlh,lhr,hrh,rhv,hv$}		5/11
CHA	hvlrvhv h	{$hv,hvl,vlr,lrv,rvh,vhv,hvh,vh$}		6/11
RA	vlhrvlh rvlhv	{$hv,hvl,vlr,lrv,rvh,vhv,hvh,vh$}		6/11
TA	Vh	{$vh}		0/11
MA	vhvhlvr	{$vh,vhv,hvh,vhl,hlv,lvr,vr$}		2/11
YA	vrhvhv	{vr,vrh,rhv,hvh,vhv,hv}		2/11

A frame size of 20x20 is used because increase in size results in disconnection of contour of the character image during thinning process. Reducing the frame size further resulted in loss of significant information. The mask {111} of length 3 runs over each character of the primitive string to produce a puzzle piece. If the length of the mask is decreased, inter-class similarity between certain classes like 'TA' and 'PA' increases. The number of pixels used to determine segment length is considered as 6. Increase in segment length decreases the dissimilarity between classes like 'KA' and 'CHA', and 'TA' and 'PA'.

Figure 3(a-c), shows the recognition accuracy of 20 character classes for varying values of parameters discussed above. The efficiency of the probability method is shown in Table 2 through a comparison with exact matching and NFV. All character classes used for the experiment is also listed in Table 2.

IV. CONCLUSION

In this paper, a handwritten character recognition system for 20 classes of Tamil script was proposed using multiset of puzzle pieces as feature vector and probability based classification method. The variable length of the feature vector used for the classification retains all vital information on the shape of the character image. The formation of multiset of puzzle pieces as a sequence of 3 primitive elements preserves the shape of the character image partially. This adds novelty to our approach which produced an accuracy of 88.15% for 20 character classes. When the same approach is applied to all 156 character classes of Tamil script in HP Labs dataset, a recognition accuracy of 68% was obtained. As inter-class similarity is a major challenge in handwritten characters, the recognition accuracy is at risk when more number of classes are included in classification.

Table 2. Recognition accuracy (%) of probability matching, exact matching and NFV [15] for tamil characters

Character Classes	NFV [15]	Exact matching	Probability Matching
KA	88	72	91
THU	76	69	80
CHA	94	69	92
RA	83	72	86
MA	94	80	93
TA	94	90	95
YA	94	82	94
AH	88	75	86
LA	83	81	85
GNA	86	82	86
RAE	96	92	94
MU	88	70	90
RU	94	81	96
NAA	78	72	81
NEE	84	80	86
THEE	78	75	81
THA	85	82	80
ZHA	86	80	88
MEE	88	79	88
LY	90	85	91
Average	**87.45**	**78.4**	**88.15**

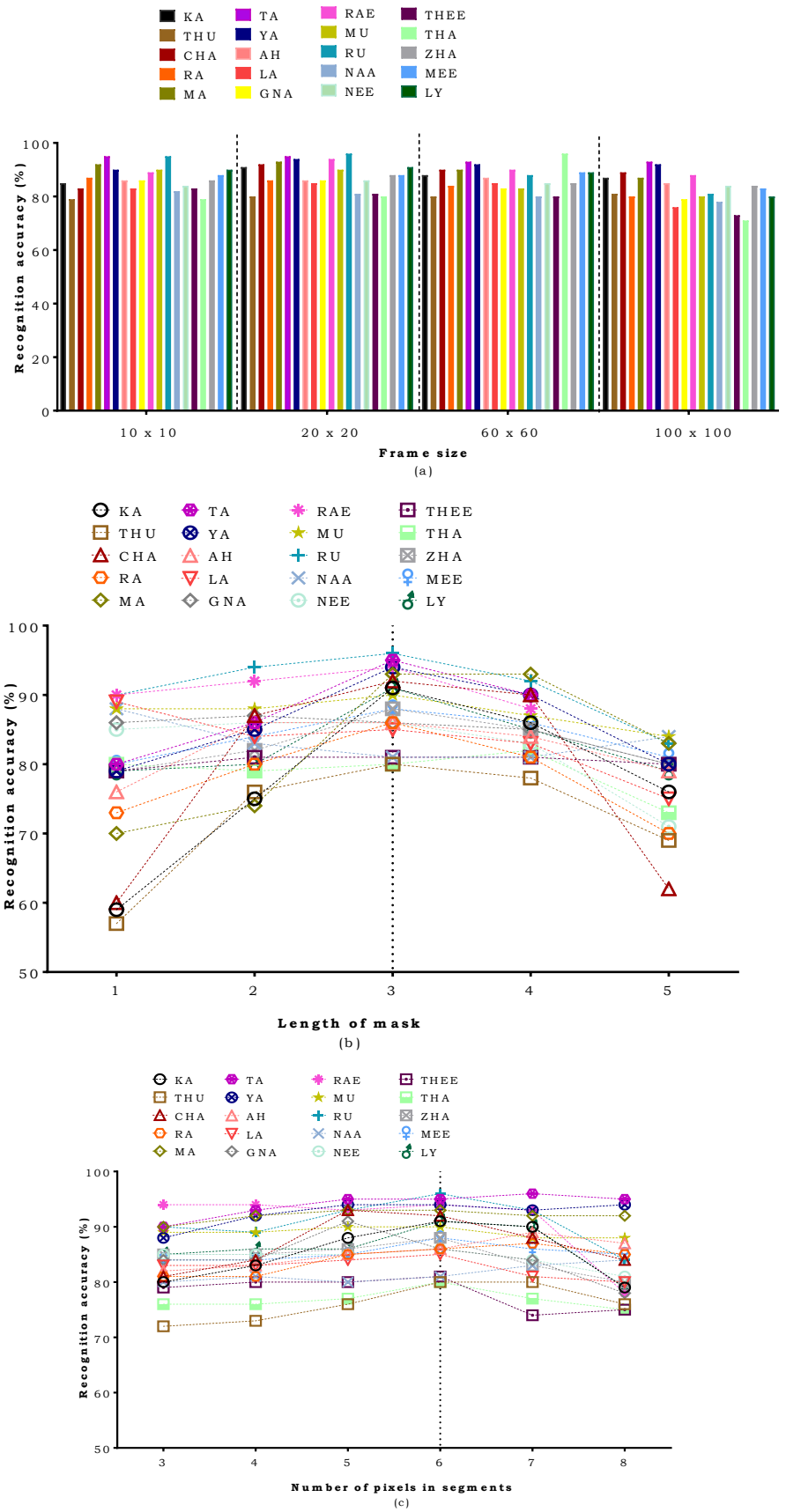

Fig. 1 (a-3). The recognition accuracy of 20 character classes for varying values of parameters. Optimal values for frame size, length of mask and the number of pixels used in segments formation with recognition accuracy.

REFERENCES

1. "India will gain 100% literacy in next 5 years: Javadekar", Hindustan Times, 5 August 2017.
2. Xuefeng Xiao, Lianwen Jin, Yafeng Yang, Weixin Yang, Tianhai Chang, "Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition", Pattern Recognition, Volume 72, December 2017, pp 72-81
3. Xiwen Qu, Weiqiang Wang, Ke Lu, Jianshe Zhou, "Data augmentation and directional feature maps extraction for in-air handwritten Chinese character recognition based on convolutional neural network" , Pattern Recognition Letters, Volume 111, 1 August 2018, pp. 9-15
4. Xiwen Qu, Weiqiang Wang, Ke Lu, Jianshe Zhou , "In-air handwritten Chinese character recognition with locality-sensitive sparse representation toward optimized prototype classifier" ,Pattern Recognition, Volume 78, June 2018, pp. 267-276
5. Ihab Khoury, Adrià Giménez, Alfons Juan, Jesús Andrés-Ferrer, " Window repositioning for printed Arabic recognition", Pattern Recognition Letters, Volume 51, 1 January 2015, pp. 86-93
6. Adarsh Trivedi, Siddhant Srivastava, Apoorva Mishra, Anupam Shukla, Ritu Tiwari, "Hybrid evolutionary approach for Devanagari handwritten numeral recognition using Convolutional Neural Network", Procedia Computer Science, Volume 125, 2018, pp. 525-532
7. Soumen Bag, Gaurav Harit, Partha Bhowmick, "Recognition of Bangla compound characters using structural decomposition", Pattern Recognition, Volume 47, Issue 3, March 2014, pp. 1187-1201
8. Saikat Roy, Nibaran Das, Mahantapas Kundu, Mita Nasipuri, "Handwritten isolated Bangla compound character recognition: A new benchmark using a novel deep learning approach", Pattern Recognition Letters, Volume 90, 15 April 2017, pp. 15-21
9. Shanthi N, Duraiswami K. "A Novel SVM-based handwritten Tamil character recognition system", Springer; Pattern Analysis and Application, Volume13(2), (2010), pp:173-80.
10. Vijayaraghavan, Prashanth, Misha Sra, "Handwritten Tamil Recognition using a Convolutional Neural Network", NEML Poster 2015.
11. Ashlin Deepa, R.N, Rajeswara Rao R. "An efficient offline Tamil handwritten character recognition system using zernike moments and diagonal-based features", International Journal of Applied Engineering Research, Volume 11, Number 4 (2016) pp:2607-2610.
12. Ashlin Deepa RN, Rajeswara Rao R, "An Eigen characters method for Recognition of Handwritten Tamil Character Recognition", Proceedings of the First International Conference on Intelligent Computing and Communication, Advances in Intelligent Systems and Computing 458, DOI 10.1007/978-981-10-2035-3_51, Springer 2017.
13. Ashlin Deepa R.N and Rajeswara Rao R," Feature Extraction Techniques for Recognition of Malayalam Handwritten Characters: Review", International conference on Emerging Trends in Engineering and Technology, Vol. 3(1), 2014, pp. 481– 485.
14. Bhattacharya U, Ghosh SK, Parui SK, "A two stage recognition scheme for handwritten Tamil characters", Proceedings of the ninth international conference on document analysis and recognition (ICDAR 2007). IEEE Computer Society, Washington, DC, pp 511-515.
15. R.M Suresh, "Printed and handwritten Tamil character recognition using fuzzy approach", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008, Volume I 2008, pp:19-21.
16. R.M. Suresh, S.Arumugam, "Fuzzy technique based recognition of handwritten characters", Im-age and Vision computing, , January 2006, pp 1-10.
17. Isolated Handwritten Tamil Character Dataset, hpltamil-iso-char http://www.hpl.hp.com/india/research/penhw/resources/tamil-iso-char .html

AUTHORS PROFILE

Ashlin Deepa R.N, with 10 years of teaching experience, is working as Assistant Professor in Computer Science and Engineering Department at Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad. She is a research scholar of JNTU Hyderabad under the guidance of Dr.Ramisetty Rajeswara Rao, Professor and HOD of CSE department at JNTUK, University College of Engineering, Vizianagaram, Andhra Pradesh, India. Her area of interest includes Image Processing and Pattern Recognition. She has 3 international conferences and 4 international journals to her credit.

Dr. Ramisetty Rajeswara Rao, is presently working as Professor & HOD in the Department of Computer Science & Engineering (CSE), JNTUK-UCEV, Vizianagaram. Dr. Rao completed his B.Tech in CSE in the year 1999 from V.R.Siddartha Engineering College, Vijayawada. M.Tech in CSE from JNTUH- Hyderabad in the year 2003, PhD in CSE from JNTUH-Hyderabad in the year 2010. Prior to joining in JNTUK-UCEV, he worked as Professor and HOD in Mahatma Gandhi Institute of Technology (MGIT), Hyderabad. He authored one monograph titled with Automatic Text Independent Speaker Recognition using Source Feature (Lap LABERT Publishing GmbH Co. KG, Germany) in the year 2012 and one Text-Book with titled Cloud Computing and virtualization (BSP Publications) in the year 2014. Two students have received Ph. D under his guidance from JNTUH-Hyderabad. To his credit he had published papers in ACM, ELSEVIER, SPRINGER and other reputed journals. He authored 76 Journals, presented papers in 35 conferences and chaired 23 session chairs and gave 23 invited talks in various reputed colleges in Andhra Pradesh and Telengana. He received VIDYA RATAN award from T.E.H.E.G, New Delhi for the year 2011. He is an academic advisor to National Cyber Safety and Security Standards (NCSSS). He is a Member of CSI and Sr. Member of IEEE. His Areas of Interest are Speech Processing, Pattern Recognition, NLP and Cloud Computing.