A MULTISET APPROACH FOR RECOGNITION OF HANDWRITTEN CHARACTERS USING PUZZLE PIECES

Ashlin Deepa R.N[1] R.Rajeswara Rao[2] Y.Vijayalata[3]

[1]Assistant Professor, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad,Andhra Pradesh, India
`deepa.ashlin@gmail.com`
[2]Professor, Dept of CSE,
Mahatma Gandhi Institute of Technology, Hyderabad, Andhra Pradesh, India
`raob4u@yahoo.com`
[3]Professor, Gokaraju Rangaraju Institute of Engineering and Technology,
Hyderabad,Andhra Pradesh, India
`vijaya@ieee.org`

ABSTRACT

Image pattern matching is one of the most widely used techniques in character recognition. A pattern is the description of an object. Although a large number of papers have been reported on handwritten character   recognition, it is still a very challenging problem. Recognition of characters is based on pattern classification. In this paper a hybrid approach is proposed using fuzzy technique and multiset comparison to recognize handwritten characters. Fuzzy technique is used to measure the variations in the features of handwritten numerals to form strings of character. Features are selected from the preprocessed image and the feature is labeled using fuzzy logic. Strings of a character are formed from these labeled primitives. Each string is divided into a multiset of puzzle pieces by repeatedly applying a mask which is comprised of all ones. The string is padded with anchors at the beginning and end. The mask is placed over the string, beginning at a certain character, and reading off all characters corresponding to 1 bit in the mask, thus producing one puzzle piece. To divide a string into multiset of puzzle pieces, the mask is applied at all shifts in the string starting at each character. Thus for each prototype character, we associate a multiset of puzzle pieces, which form the database. The   multiset of puzzle pieces of an unknown character is constructed and compared with the same in the database, leading to the identification of the unknown character.

KEYWORDS

Fuzzy technique; Thinning; Segmentation; Labeling; Multiset; Puzzle pieces

1. INTRODUCTION

A character recognition system is a pattern recognition system which receives scanned preprocessed data as input and identifies the name of the character by comparing it with the prototype characters already in the memory. PR techniques are an important component of intelligent systems and are used for Character recognition and Handwriting analysis. In recent years pattern recognition scientists are working with a vision to build a machine with new technology and methodology, which can   identify, understand and process pictorial information for human use with more and more efficiency. Pattern recognition system can be defined as the classification of input

data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant   details. Our proposed approach handwritten character recognition through multiset of puzzle pieces has been designed to address the following issues:

- It avoids sorting of the prototype characters in the database.

- It eliminates the small variations in different images of the same character.

- It provides a probability of similarity for each input handwritten character with the proto-type character

The input characters in the document are digitized by the optical scanner. The resulting characters are then fed into a preprocessor for smoothing, elimination of noise, edge detection and thinning etc. This preprocessed image is segmented into primitives. The membership values for the feature are measured and the feature is labeled. Strings of a character are formed from these labeled primitives. Each string is divided into a multiset of puzzle pieces. These puzzle pieces are compared with the puzzle pieces of prototype character in the database and handwritten character is classified under a specified prototype class.

2. RELATED WORK

Different methods for automatic pattern recognition are motivated by the way in which pattern classes are characterized and defined.

In Heuristic approach [3] the characterization of a pattern class is by making use of template matching and feature matching concepts. The disadvantage of this approach is that the recognition rate of the character is based on the size of frame and according to the random placement of the   character in the frame. Handwritten character recognition algorithm based on artificial immune (HCRA) is proposed by Yuefeng Chen, Chunlin Liang, Lingxi Peng, Xiuyu Zhong, Xiuyu Zhong[14].

In Deterministic Approach [7] the feature selection can be computed from a signal or an image. When the pattern classes under considerations do not share identical patterns, it is always possible to determine a set of decision functions using some attributes which will correctly classify all the patterns of a training set. In many classes the specified percentage is not achievable within the complexity constraints placed on the classifier. Under these conditions, either the percentage of acceptable misclassifications must be raised, or the complexity of the classifier increased, or an alternative method of classification chosen.

The Statistical Approach [1][9] is based on Baye's classification rule and is an optimum   classifier when the probability density function of each pattern population and the probability of each pattern class are known. But the poor convergence rates of statistical classifiers tend to overshadow their potential for optimum performance.

In Linguistic (Syntactic or Structural) methods [8][4][6] Characterizing of patterns by primitive elements (Sub patterns) and their relationships suggests automatic pattern recognition by the linguistic or syntactic or structural approach, making use of common-property concept. The disadvantage of this approach is the structural information about the inter connectives in complex patterns cannot be   handled very well. For handwritten character images, inconsistent shape variations are often found in different samples of the character images of the same category allowing distinct descriptions for   categorically different patterns. Distribution based algorithm is proposed for handwritten character recognition[3]. The primary difficulty in handwriting recognition lies in the variety of writing styles by individuals at different times and by different individuals. Therefore the hybrid fuzzy approach combined with syntactic techniques for character recognition is proposed [10] .Since 1980 a large amount of work has been carried out in character recog-

nition. However, there is still a long way to go in order to reach the ultimate goal of machine simulation of fluent human reading, especially for   unconstrained on-line and off-line handwriting.

3. PROBLEM DEFINITION

The primary difficulty in handwriting recognition lies in the variety of writing styles by individuals at different times and by different individuals. Hand-written characters differ not only from person to person but also according to the state of mood of person where the intra-class variations are infinite and cannot strictly obey the mathematical constraints set by the formal language theory. Also when the primitive string is compared with the prototype string, the string matching algorithm was simple binary search and the strings are to be ordered before searching. Hence sorting of the strings is essential in that recognition approach and is time consuming. Taking these problems into account we proposed an approach which recognizes the handwritten characters through multiset of puzzle pieces.

The input handwritten characters are preprocessed and the segmented primitives are labeled to form strings using fuzzy logic[11] since the handwritten characters cause vagueness. These strings are processed by multiset of puzzle pieces.

4. IMPLEMENTATION

The idea behind our approach to handwritten character recognition is to is to divide each character string into a multiset of "puzzle pieces."Various steps in Recognition:

Step 1: Grey-level scanning

Step 2: Preprocessing (Noise Removal and Smoothing, Binarization,Edge detection, Thinning)

Step 3: Feature Extraction (Segmentation, Identification, Labeling using fuzzy technique   and string formation.)

Step 4: Construction of multiset of puzzle pieces.

Step 5:Comparing with the prototype multiset for handwritten character recognition. The steps are shown in the Figure 1.

4.1 Grey-level scanning and preprocessing

i. The input characters in the document are digitized by the optical scanner. Grey level scanning is done at an appropriate resolution, typically 300 to 1000 dots per inch.
ii. Noise removal and smoothing is performed using ROM filter approach for filling and cleaning of the image.
iii. Histogram based threshold approach (Otus binarization) is used to convert the gray image into two tone image i.e, binary image.
iv. The edges of the two tone image are obtained using Sobel edge detection procedure. The edge detected image is undergone thinning process by the procedure described by Ramesh in [10].

v. Using the thinned character image, the pattern primitives are segmented [2], identified, and labeled for further recognition using fuzzy function[11][13]. The following algorithm is meant for Primitive identification and labeling. The four basic features are used for labeling the polygonal image. The features are given below.

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

$$\mu_h = 1 - |m_x| \qquad \text{for } |m_x| < 1 \qquad (1)$$

$$= 0 \qquad \text{for } |m_x| > 1$$

$$\mu_v = 1 - |1/m_x| \qquad \text{for } |m_x| > 1 \qquad (2)$$

$$= 0 \qquad \text{for } |m_x| < 1$$

$$\mu_{ob} = 1 - |(\theta - 45)/45| \qquad \text{for } 0 < |m_x| < \infty \qquad (3)$$
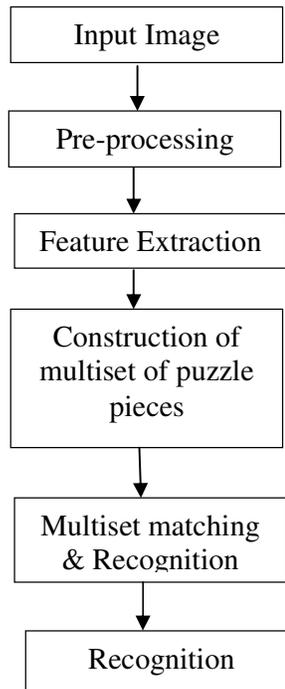
Figure 1. Various steps in Recognition

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**:** Labeling 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. (4.1)-(4.3).

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   b   c
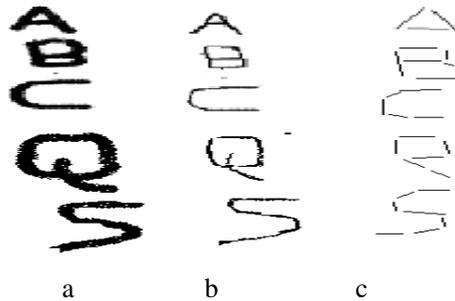
Figure 2. (a) Handwritten character, (b) Preprocessed image, (c) Segmented Image

4.2 Recognition of handwritten characters through multiset of puzzle pieces

The string representing the input character is the input of this recognition approach. The process is as follows:

Divide each string into a multiset of "puzzle pieces." Puzzle pieces are constructed out of proximate characters of the original string by repeatedly applying a mask. A mask is a binary array and applying a mask to a string involves computing a dot product of the mask with a substring of the string. Mask is comprised of all ones. The mask is applied by placing it over the string, beginning at a certain character, and reading off all characters corresponding to 1 bit in the mask, thus producing one puzzle piece. To divide a string into pieces, the mask is applied at all shifts in the string starting at each character.As an example, consider the character string **hvhrvhv** which is the labeled output using the fuzzy logic for the handwritten character **'B'.** Under the mask 111, which has length lm=3 the string is processed for recognition. The string is padded with anchors at the beginning and end to get \$ hvhrvhv \$, which is subdivided into the following multiset of puzzle pieces:

{\$hv, hvh, vhr, hrv, rvh, vhv, hv\$}

The multiset of puzzle pieces for the prototype characters is stored in the database through the same procedure as a standard for classification of unknown character string. The classification is done in two ways.

1.Exact matching   2. Probability matching

4.2.1 Exact matching

Using the exact matching the unknown character string is classified under a prototype class, if and only if the multiset of puzzle pieces obtained from the unknown character string matches exactly with any one of the entries in the database. For example the multiset {\$hv, hvh, vhr, hrv,

rvh, vhv, hv$} is obtained from the unknown character string $hvhrvhv$. If the database for the prototype characters contains an entry of the same multiset in the same order then the unknown handwritten character is classified under the class represented by that entry. Thus the handwritten character is recognized as 'B'.

4.2.2 Probability matching

In probability matching the input character string is classified with some probability. The procedure of probability matching is as follows:

Step 1: Get the input string.

Step 2: Apply the mask and construct the multiset of puzzle pieces.

Step 3: Count the entries in the multiset of puzzle pieces of unknown character.

Step 4: Construct the database of prototype character strings (multiset of puzzle pieces of each character string) and count the number of puzzle pieces in the multiset of each character string and keep it in the array "Total no of pieces". The prototype database contains the multiset of n prototype characters.

Step 5: Initialize two arrays "MatchingPieces" and "ProbabilityMatching" where the size of each array is equal to the number of entries in the prototypes database.

Step 6: For all puzzle pieces in the multiset of unknown character

    For all entries in the prototype database

      For all puzzle pieces in each entry of the prototype database

     If any single piece in the unknown character set matches   with any one of the piece in the single entry of prototype database then

        Increment the corresponding entry in the "Matchingpieces" array.

 Else

    Move to the next single piece of unknown character set.

End If End for,   End for, End for

 Step 7: For all entries in the prototype database

Find the probability matching by using the   following formula

$$\text{ProbabilityMatching}_i = \frac{\text{MatchingPieces}_i}{\text{Total no of pieces}_i} \quad \text{where } i = 1 \text{ to } n$$

 End for

Step 8: Initialize MAX=ProbabilityMatching[1]

Step 9: For all entries in the ProbabilityMatching

    If ProbabilityMatching [i+1]> MAX

        MAX= ProbabilityMatching [i+1]

        Priority=i

End If

End For

Step 10: Retrieve the class name corresponds to ith entry in the database and place it in variable "Known".

Step 11: Print "The unknown character is classified to the class known with probability MAX".

The algorithm is applied to the input string and the classification is done with some probability. The unknown character is classified under the prototype class which has the maximum value in the Probability Matching array.

Table 1. Strings of characters and the percentage of recognition using exact matching and   probability matching

	String obtained using fuzzy logic	% of Recognition by Exact matching	% of Recognition by Probability matching		String obtained using fuzzy logic	% of Recognition by Exact matching	% of Recognition by Probability matching
A	llhrr	90.0	92.0	N	Vrv	91.0	92.0
B	hvhrvhv	91.0	93.0	O	hlvrhlvr	92.0	93.0
C	hlvrh	86.0	90.0	P	Hvhvv	90.0	92.76
D	hrvlhv	89.0	91.0	Q	Hvrrhv	89.0	90.0
E	hvhvh	91.0	91.21	R	hvhrrvv	87.0	89.0
F	hvhv	90.0	90.5	S	hlvrvlh	88.0	90.0
G	hlvrhvh	87.0	89.2	T	Hhv	89.0	91.0
H	Hvhvv	92.0	93.0	U	Vrhlv	90.0	92.0
I	hhvhh	89.0	90.0	V	Rl	92.0	93.0
J	hhvl	88.0	91.0	W	Rlrl	90.0	91.0
K	vrlv	90.0	90.42	X	Rrll	91.0	92.0
L	vh	91.0	92.3	Y	Rlv	89.0	92.0
M	vrlv	92.0	92.6	Z	Hlh	92.0	95.0

5. RESULT AND CONCLUSION

Unconstrained handwritten characters are considered as an image whose size and position are considered as variant. The concatenation among the pattern was considered as in serial and the strings obtained are by considering the trace in clockwise direction. The frame for writing characters is converted into 20X20 binary matrix. Our algorithm for recognition of handwritten characters through multiset of puzzle pieces is analyzed and the result is shown in the Table 1. Standard English alphabets have been used for the recognition purpose. The Comparison of both methods is done and is shown in Figure 3. Probability Matching is found better than Exact matching.
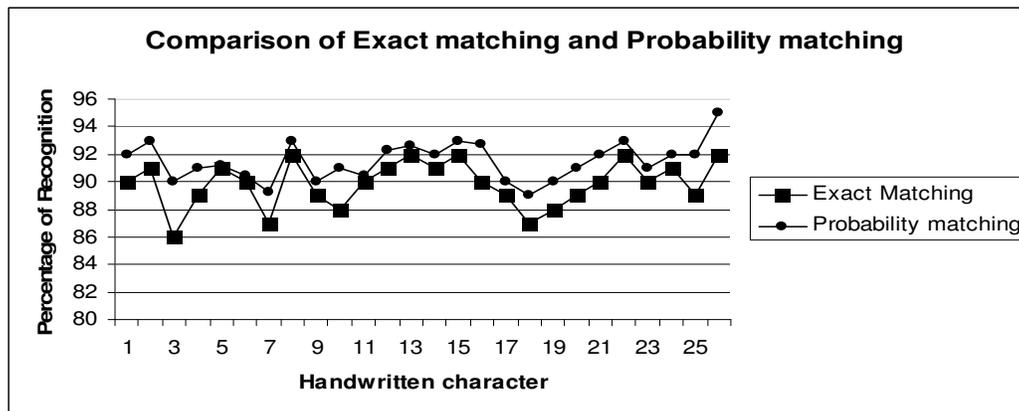
Figure 3. Comparison of Probability Matching and Exact matching, In X axis 1-A, 2-B …, 25-Y, 26-Z

6. REFERENCES

[1]   Alessandro L. Koerich, "Unconstrained handwritten Character Recognition Using Different Classification Strategies, " IEEE International Conference on Image Processing, Volume: 2, on page(s): II- 542-5, Sep-2005.

[2]   Amer Dawoud and Mohamed S. Kamel, "Iterative Multimodal Subimage Binarization for Handwritten Character Segmentation," IEEE Transactions On Image Processing, Vol.13 No.9, September 2004.

[3]   Chiu HP, DC Tseng, A novel stroke-based feature extraction for handwritten Chinese character recognition, Patt.Rec.32, pp 1947-1959, 1999.

[4]   E. Kavallieratou, N.Fakotakis and G.Kokkinakis, "Handwritten Character Recognition based on Structural Characteristics," Proceedings. 16th International Conference on Pattern Recognition, 139-142 vol.3, 2002.

[5]   Huiqin Lin, Wennuan Ou, Tonglin Zhu," The Research of Algorithm for Handwritten Character Recognition in CorrectingAssignment System", 2011 Sixth International Conference on Image and Graphics.

[6]   Jinhai Cai and Zhi-Qiang Liu, "Integration of Structural and Statistical Information for Unconstrained Handwritten Numeral Recognition," IEEE Transactions on Pattern Analysis And Machine Intelligence, VOL. 21, NO. 3 MARCH 1999.

[7]   Ling Zhang   , Dorothea Blostein, Richard Zanibbi, "Using Fuzzy Logic to Analyze Superscript and Subscript Relations in Handwritten Mathematical Expressions," Proceedings. Eighth International Conference on Document Analysis and Recognition, 972- 976 Vol. 2, Sep-2005.

[8]   Madasu Hanmandlu, Mohd. Hafizudin Mohd.Yusof and Vamsi Krishna Madasu" Fuzzy based approach to the recognition of Multi-Font Numerals," 2nd National Conference on Document Analysis and Recognition (NCDAR-20\03), 11-12 July, 2003.

[9]   McLachlan, G.J., Discriminant Analysis and Statistical Pattern Recognition, Wiley, 1995.

[10] R.M. Suresh, S.Arumugam, "Fuzzy technique based recognition of handwritten characters", Im-age and Vision computing pp 1-10, January 2006.

[11] R.M. Suresh and S.Arumugam," Fuzzy Context Free Grammar for Handwritten Numeral Recogni-tion," IT for new Generation-Proceedings of Annual Convention 90th Computer Society of India, pages 411-419, 1998.

[12] R.M. Suresh, S.Arumugam, L. Ganesh," Handwritten Characters Recognition using Finite Fuzzy Au-tomata," Information technology for computer Innovations, pages 62-73, 1999.

[13] R.M. Suresh ,S.Arumugam,L. Ganesh" Fuzzy Approach to Recognize Handwritten Tamil Characters, " Information conference on Computational Intelligence and Multimedia Ap-plications, IEEE Com-puter Society , pages 459-463,1999

[14] Yuefeng Chen, Chunlin Liang, Lingxi Peng, Xiuyu Zhong, Xiuyu Zhong," A Handwritten Charac-ter Recognition Algorithm based on Artificial Immune", 2010 International Conference on Com-puter Application and System Modeling (ICCASM 2010), 2010 IEEE V12-273.

Authors

Ashlin Deepa R.N, with 3 years of teaching experience, is working as Assistant Pro-fessor in Computer Science and Engineering Department at Gokaraju Rangaraju Insti-tute of Engineering and Technology, Hyderabad. She is a research scholar of JNTU Hyderabad under the guidance of Dr.R.Rajeshwara Rao, Professor and Head of CSE department at Mahatma Gandhi Institute of Technology, Hyderabad. Her area of inter-est includes Speech Processing and Character Recognition. She has 3 national confe-rences to her credit.

R.Rajeswara Rao, with 12 years of teaching experence, presently working as Professor and Head of CSE department at Mahatm a Gandhi Institute of Technology, Hyderabad. did graduation from VR Siddhartha Engineering College, Vijayawada, M .Tech Com-puter Science and Engineering in JNTU College of Engineering , Hyderabad during 2001-2003. Obtained Ph. D in Speech Processing from JNTU Hyderabad in the year 2010 under the guidance of Prof.V. Kamakshi Prasad, Professor of CSE , JNTU Hyde-rabad. To his credit he has 12 national and international journals and 8 national and international conferences.

Y. Vijayalata, with more than 15 years teaching experience, is working as Professor in Computer Science and Engineering Department in Gokaraju Rangaraju Institute of Engineering and Technology at Hyderabad. Her area of interest includes Algorithm Analysis and Pattern Recognition. She is the chairperson for IEEE WIE-AG, Hydera-bad section. She has many national and international journals to her credit.