SUPPORT VECTOR MACHINE BASED GUJARATI NUMERAL RECOGNITION

Mamta Maloo
SNJB's KBJ College Of Engineering, Chandwad, Nashik (M.S.), India mamtaji_61079@rediffmail.com

K.V. Kale
Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Mararthwada University, Aurangabad (M.S.), India
kvkale91@gmail.com

Abstract— **In this paper we propose the Support Vector Machine (SVM) based recognition scheme towards the recognition of Gujarati handwritten numerals. The preprocessing is done considering morphological operations. For computing the features, each isolated numeral is segmented into blocks. These blocks create base for four sets of features. Then we derived affine invariant moments as features. The features obtained from these blocks are fed to SVM classifier. We obtained the recognition rate of 91% approximately.**

Keywords- handwritten character recognition, affine invariant moments, support vector machine, Gujarati script

I. INTRODUCTION

Now days, each computer-user wants the computer to have user-friendly interactions. Also due to use of computers in all aspects of human life, it is now always being desired that computers should recognize native languages to help common man to perform his daily tasks. From this front, it poses a problem of pattern recognition to recognize individuals handwriting from different other handwritten scripts.

To solve the problem of pattern classification, a powerful classifier called Support Vector Machine (SVM) classifier is used. It is having good generalization and convergence property. SVM finds its applications in various fronts of human interactions like face recognition, text recognition, bioinformatics, etc. SVM is basically a linear classifier but because of its kernel functions, SVM can be used as non-linear classifier making it work on data of high dimensions

In this paper we propose the SVM based recognition scheme towards the recognition of Gujarati handwritten numerals. Section I describes the introduction of the paper. Section II gives the details of Gujarati Numerals. Detailed literature survey is done in Section III. Section IV describes the process of data acquisition and preprocessing done for the system. Section V elaborates the feature extraction method used in the paper. SVM classifier along with the results obtained is described in Section VI. Finally Section VII concludes the paper.

II. PROPERTIES OF GUJARATI NUMERALS

Numerals belonging to Gujarati script are shown in figure 1.

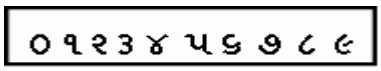
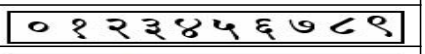
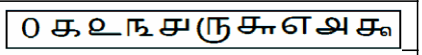
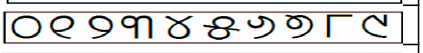
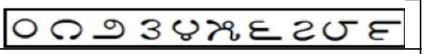
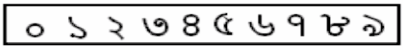


Figure 1 Numerals of Gujarati Script

Numeral 0 resembles almost all zeros on all the Indian regional scripts but there is wide variety in the rest of the digits. This variety can be viewed in table 1. All the numerals are in the sequence 0,1,2,3,4,5,6,7,8,9. It could be seen that numeral 8 in Gurumukhi is 180° rotated shape of that for numeral 8 in Gujarati. It can also be seen that numeral 1 in Gujarati is numeral 7 in Bengali. Numeral 4 in Gujarati, Devanagari, Gurumukhi, Kannada, Oriya and Telugu is same. Numeral 5 in Gujarati and Gurumukhi resemble each other. There is controversy in numeral 3 from Bengali and numeral 7 from Gujarati and Devanagari. Due to these similarities and dissimilarities it motivated us to continue with the recognition of Gujarati numerals.

TABLE 1NUMERALS OF DIFFERENT SCRIPTS

Sr.no	Script	Numerals
1	Devanagari	
2	Kannada	
3	Gurumukhi	
4	Tamil	
5	Oriya	
6	Telugu	
7	Bengali	

III. LITERATURE SURVEY

Sameer Antani [2] in 1999 has given the primitive effort to Gujarati printed text. For classification the author has used two classifiers, K-NN classifier and minimum hamming distance classifier. The best recognition rate was for 1-NN for 600 dimensional binary feature space i.e. 67% 1-NN in regular moment space gave 48% while minimum distance classifier had the recognition rate of 39%. The Euclidean minimum distance classifier recognized only 41.33%.

Dholakia [6] attempted to use wavelet features, GRNN classifier and KNN classifier on the printed Gujarati text producing 97.59 and 96.71 as their respective recognition rates. In 2005, Jignesh Dholakia et. al [7] have presented an algorithm to identify various zones used for Gujarati printed text. In the algorithm they have proposed the use of horizontal and vertical profiles.

Desai [3] reported that for feature extraction four profile vectors are used as an abstracted feature of identification of digit. Five more patterns for each digit are created in both clockwise and anticlockwise directions with the difference of 2degrees each up to 10°. A feed forward back propagation neural network is used for Gujarati numeral classification with success rate for standard fonts as 71.82%, for handwritten training sets as 91.0% while for testing sets as a score of 81.5% was recorded.

Another effort contributed for Gujarati script was by Shah & Sharma [8]. They used template matching and Fringe distance classifier as distance measure. By this effort, for connected component recognition rate was 78.34%. For upper modifier recognition rate was 50% where as for lower modifier it was 77.55% and for punctuation marks it was 29.6%. Cumulative for overall it was 72.3%.

Mahmud et al [12] has reported accuracy of 98% for recognizing Bengali isolated characters and 96% for continuous characters using chain code as feature vector and using a feed forward neural network as classifier. Using nearest neighbor classifier and string connectivity as feature vector, Ray & Chatterjee [13] developed a recognition system for Bengali characters. Combining template and feature matching approach, Chaudhari & Pal [14], reported 99.10% recognition accuracy for printed Bengali characters.

Lehal and Singh presented an OCR system for printed Gurumukhi script [16]. The skew angle is determined by calculating horizontal and vertical projections at different angles at fixed interval in the range [0° to 90°].A recognition rate of 96.6% at a processing speed of 175 characters/second was reported. Lehal & Singh [15] also developed a post processor for Gurmukhi.

Sinha & Mahabala [17] presented a syntactic pattern analysis system with an embedded picture language for the recognition of handwritten and machine printed Devanagari characters. Problems that arise in developing OCR systems for noisy images are addressed in the work by Parvati Iyer et al [18]. Character recognition rate of only 55% is reported. The authors also trained a feed-forward back propagation neural network, with a single hidden layer. Character recognition rate of 76% is reported with the neural network approach. Veena [18] Performance of 93% accuracy at character level is reported. Pal & Chaudhari [19] reported a complete OCR system for printed Devanagari. A structural feature-based tree classifier recognizes modified and basic characters, while compound characters are recognized by hybrid approach combined with

structural and run based template features. The method reports about 96% accuracy.

Pujari et al [27] proposed a recognizer that relies on wavelet multi-resolution analysis for capturing the distinctive characteristics of Telugu script .The performance across fonts and sizes is reported as varying from 93% to 95%. An OCR for Telugu is reported by Negi, et al [25]. Raw OCR accuracy with no post processing is reported as 92%. Performance across fonts varied from 97.3% for Hemalatha font to 70.1% for the newspaper font. Non- linear normalization to improve performance was used by Negi et al, [26] by selectively scaling regions of low curvature in the glyphs.

Jawahar et al [22] proposed a Bilingual OCR for Hindi-Telugu documents. It is based on Principal Component Analysis followed by support vector classification. An overall accuracy of approximately 96.7% is reported. Anuradha Srinivas, et al [23] developed a Telugu optical character recognition system for a single font. A 2-stage classifier with first stage identifies the group number of the test character, and a minimum-distance classifier at the second stage identifies the character. Recognition accuracy of 93.2% is reported.

Mohanty & Behera [22] described a complete OCR development system for Oriya script. Ashwin & Sastry [20] developed a font and size-independent OCR for Kannada. Classification based on the Support Vector Machines is adopted.

IV. DATA ACQUISITION AND PREPROCESSING

A. Data Acquisition

As there was no database available for Gujarati Script, we have created the database by taking the samples of handwritten imprints on that datasheets created to take the samples. Then these samples were scanned using HP 2400 Scanjet scanner at the resolution of 300dpi. The samples were withdrawn at random from various writers belonging to different profession, age, sex and education levels and were using different ink for preparing the samples on datasheets. Ten samples were taken from eight persons. The database was created manually.

B. Preprocessing

Before extracting the features there is need to preprocess the image which comprises of removing any noise, skew, or apply some morphological methods to enhance the image. Initially there can be variation in the size of sample as it depends on the variety of writers. To apply any unbiased algorithm we require that all images should be normalized to same size. For this, in the paper, we have resized the initially captured image into a normalized image of size 40x40 using nearest neighbor interpolation technique. This resized image is then binarized with the optimum threshold of 0.2 binarization level so that the image is converted into two tones i.e. 0 for background and 1 for foreground. Due to binarization there is a chance that the image is broken down or may lose some of the information. So the image is then dilated using the structuring element “diamond” with radius size 1. Finally the image is skeletonized to one pixel thin. This image is finally ready for extracting features.

V. FEATURE EXTRACTION

Each image has some features and finally to make a comparison and recognition we require using the features that describe each numeral separately and distinctly. Before computing the features of the image, the image is subjected for image division technique where each image is divided into a matrix of boxes each of size 10X10, 8x8 and 5x5. Thus each image is converted into a matrix of size 4x4, 5x5 and 8x8. These image matrices create a base for four separate feature sets namely;

- Feature Set1 (FS1) considering the image as whole, no division is applied.
- Feature Set2 (FS2) considering the image of matrix size 4x4
- Feature Set3 (FS3) considering the image of matrix size 5x5
- Feature Set4 (FS4) considering the image of matrix size 8x8

To fulfill this purpose we have derived the affine invariant moments by means of the theory of algebraic invariants [16, 17].

The AMIs is invariant under general affine transformation

$$\left.\begin{aligned} u &= a_0 + a_1 x + a_2 y \\ v &= b_0 + b_1 x + b_2 y \end{aligned}\right\} \text{.....(1)}$$

where, (x, y) and (u, v) are coordinates in the image plan before and after the transformation ,

respectively. The basic affine invariant moments are given below:

$$
\begin{aligned}
I_1 &= (\mu_{20}\mu_{02} - \mu_{11}^2)/\mu_{00}^4 \\
I_2 &= (\mu_{30}^2\mu_{03}^2 - 6\mu_{30}\mu_{21}\mu_{12}\mu_{03} + 4\mu_{30}\mu_{12}^3 + 4\mu_{03}\mu_{21}^3 - 3\mu_{21}^2\mu_{12}^2)/\mu_{00}^{10} \\
I_3 &= (\mu_{20}(\mu_{21}\mu_{03} - \mu_{12}^2) - \mu_{11}(\mu_{30}\mu_{03} - \mu_{21}\mu_{12}) + \mu_{02}(\mu_{30}\mu_{12} - \mu_{21}^2))/\mu_{00}^7 \\
I_4 &= (\mu_{20}^3\mu_{03}^2 - 6\mu_{20}^2\mu_{11}\mu_{12}\mu_{03} - 6\mu_{20}^2\mu_{02}\mu_{21}\mu_{03} + 9\mu_{20}^2\mu_{02}\mu_{12} + 12\mu_{20}\mu_{11}^2\mu_{21}\mu_{03} \\
&\quad + 6\mu_{20}\mu_{11}\mu_{02}\mu_{30}\mu_{03} - 18\mu_{20}\mu_{11}\mu_{02}\mu_{21}\mu_{12} - 8\mu_{11}^3\mu_{30}\mu_{03} - 6\mu_{20}\mu_{02}^2\mu_{30}\mu_{12} + \\
&\quad 9\mu_{20}\mu_{02}^2\mu_{21} + 12\mu_{11}^2\mu_{02}\mu_{30}\mu_{12} - 6\mu_{11}\mu_{02}^2\mu_{30}\mu_{21} + \mu_{02}^3\mu_{30}^2)/\mu_{00}^{11}
\end{aligned}
\quad \} \ldots\ldots (2)
$$

VI. CLASSIFICATION

Support vector machine (SVM) is the state-of-the-art classifiers derived from statistical learning theory. It is a supervised learning algorithm with better generalized properties and with limited number of training patterns [9]. SVM is introduced for classifying linearly separable classes of objects. SVM resolves the classification problems by separating the data into two categories by using an n dimensional hyper plane. SVM determines the hyper plane that maximizes the margin between classes. For any particular set of two classes of objects, an SVM finds the unique hyper plane having the maximum margin. SVM represents the classified outputs as support vectors that determine the maximum margin hyper plane. This maximum margin solution enable SVM to outperform compared to other nonlinear classifiers, particularly in noisy environments. In this paper, we have focused on noisy numerals.

A unique property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence it is also known as maximum margin classifier [10]. A support vector machine for pattern classification is built by mapping the input pattern x into a high-dimensional feature vector v using a non linear transformation f(x), and by constructing an optimal hyper plane in the feature space. A function called 'kernel' is used to map the data from input space to feature space. In this paper we have used linear kernel function for SVM. The Recognition rate indicates the ratio of correctly recognized numerals to total number of samples.

$$
Recognition\ rate = \frac{Correct\ Recognition}{Total\ Number\ of\ Samples} * 100
$$

This recognition rate is summarized in table 2

TABLE 2 COMPARISON OF RECOGNITION RATE

Gujarati Numerals	Number of samples are 800			
	FS1	FS2	FS3	FS4
0	97.00	95.25	94.75	96.25
1	89.00	86.50	87.75	86.25
2	89.25	87.75	89.50	86.25
3	88.75	87.50	88.50	87.00
4	90.75	89.00	91.50	91.50
5	89.00	88.50	87.50	89.50
6	89.25	87.25	88.25	90.50
7	88.00	88.50	86.25	86.25
8	96.75	94.50	96.50	94.00
9	87.75	88.50	90.75	89.00
Average	**90.55**	**89.33**	**90.13**	**89.65**

From the experiment conducted it has been found that numeral '0' has the highest recognition rate i.e. 97% for FS1whereas it is minimum for FS3with 94.75%. It can be also noticed that for numerals 0,1,3,8 the recognition rates are highest in FS1 as compared to FS2, FS3 and FS4. For FS3 numerals 2, 4 and 9 have shown maximum results. For FS4 numerals 4, 5, and 6 show to have results that are highest among the four feature sets. Only for numeral 7
FS2 shows the maximum results i.e. 88.50%. Besides of this variation it has been seen that FS1 gives overall maximum results with recognition rate 90.55%. It has also been seen that one can get remarkable results using FS3.

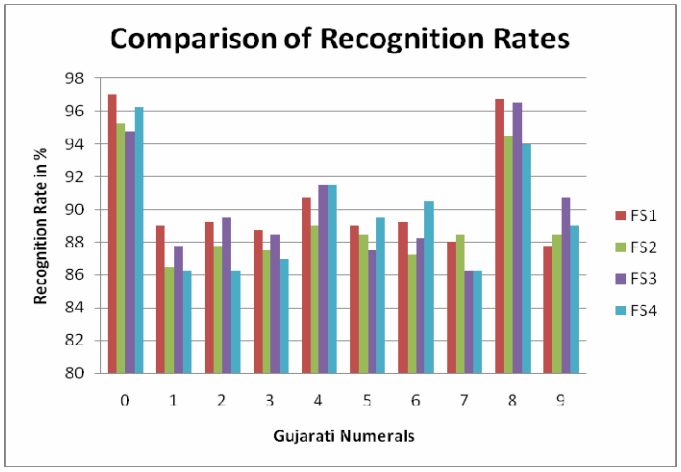


Figure 1 Comparison of Recognition Rates

VII. CONCLUSION

This paper attempts to apply a technique based on affine invariant moments for feature extraction. A database of characters has been required to extract the features, which form a template (GENERATED Trained Database). The paper has dealt with feature-based recognition of handwritten characters by giving the added approach of morphological dilation and skeletonization. Overall recognition rate for this approach is 90.55%. It has shown promising results for numerals 0, 4 and 8. Due to the division of image, the recognition rate for numerals 2, 4, 5, 6, 7 and 9 shows the enhanced graph but it has not shown good results for other numerals. As compared with [3, 29] the results are good. In future scope we tend to enhance results for rest of the numerals by doing modifications in the current system.

REFERENCES

[1] Rafael C. Gonzalez, Richard E. Woods, "Digital Image Processing", Second Edition, Pearson Education.
[2] Antani S. and Agnihotri L.: (1999) Gujarati Character Recognition 5th ICDAR, pp 418-422
[3] Desai A. A.:(2010). Gujarati handwritten numeral optical character reorganization through neural network, Pattern Recognition, Vol. 43, pp 2582-2589
[4] van den Boomgard, R, and R. van Balen(1992) Methods for Fast Morphological Image Transforms Using Bitmapped Images, Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing, Vol. 54, Number 3, pp. 254-258.
[5] Otsu, N.:(1979) A Threshold Selection Method from Gray-Level Histograms, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66.
[6] Dholakia J., Yajnik A., Negi A.(2007) Wavelet Feature Based Confusion Character Sets for Gujarati Script, ICCIMA, p366-371
[7] Dholakia J., Negi A., S. Rama Mohan (2005) Zone Identification in the Printed Gujarati Text, Proc. of 8th ICDAR, p272-276
[8] S K Shah and A Sharma Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching IE(I) Journal−ET Vol.86 pgs. 44-49 2006
[9] Ganapathiraju, 1. Hamaker, and 1. Picone, "Hybrid SVMlHMM architectures for speech recognition," in Proc. ICSLP, Beijing, 2000
[10] Chandra Sekhar, W.F.Lee, K. Takeda and F. Itakura Acoustic Modeling of Sub word Units Using Support Vector Machines, Workshop on Spoken Language Processing, TIFR, Mumbai, India
[11] Cristianini, N., and Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, First Edition (Cambridge: Cambridge University Press).
[12] Jalal Uddin Mahmud, Mohammed Feroz Raihan and Chowdhury Mofizur Rahman 2003 A Complete OCR System for Continuous Bengali Characters. Conference on Convergent Technologies for Asia-Pacific Region (TENCON) Volume 4, Issue , 15-17 Page(s): 1372 – 1376
[13] Ray K, and Chatterjee B 1984 Design of a nearest neighbor classifier system for Bengali character recognition. Journal of . Inst. Electronics. Telecom. Eng. 30 pgs.226–229.
[14] Chaudhuri, B.B. and Pal, U. 1998 A complete printed Bangla OCR system. Pattern Recognition, Vol. 31, 1998, pp 531-549
[15] Lehal G S and Chandan singh, 2002 A post-processor for Gurmukhi OCR Saadhana Vol. 27, Part 1, February, pp.99–111

[16] Lehal G S, Singh C 2000 A Gurmukhi script recognition system. Proc. 15th Int. Conf. on Pattern Recognition (Los Alamitos, CAIEEE Comput. Soc.) vol. 2, pp 557–560
[17] Sinha R.K, Mahabala 1979 Machine recognition of Devnagari script. IEEE Trans. Systems Man Cybern. Pgs.435–441.
[18] Veena Bansal 1999 Integrating knowledge sources in Devnagari text recognition. Ph.D. Thesis, IIT Kanpur,
[19] Chaudhuri B B, PalU1997 An OCR system to read two Indian language scripts: Bangla and Devnagari (Hindi). Proc. Fourth Int. Conf. on Document Analysis and Recognition (Los Alamitos, CA: IEEE Comput. Soc.) pp 1011–1016
[20] Ashwin T. V. and Sastry P. S.: (2002) A font and size-independent OCR system for printed Kannada documents using support vector machines Sadhana 27(1): 35-58.
[21] Jan Flusser,Tomas Suk: (1994) Affine Moment Invariants: a new tool for Character Recognition, Pattern Recognition Letters, Vol.15: 433-436
[22] Jawahar C. V., M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran 2003. A Bilingual OCR for Hindi-Telugu Documents and its Applications.International Conference on Document Analysis and Recognition.
[23] AnuradhaSrinivas, Arun Agarwal, C.R.Rao 2007 Telugu Character Recognition. Proc. Of International conference on systemics, cybernetics, and informatics, Hyderabad, pgs.654-659..
[24] Mohanty S., H.K.Behera, 2004 A Complete OCR Development System for Oriya Script. Proceeding of symposium on Indian Morphology, phonology and Language Engineering, IIT Kharagpur.
[25] Negi Atul, Chakravarthy Bhagvati and.Krishna B 2001 An OCR system for Telugu. Proc. Of 6th Int. Conf. on Document Analysis and Recognition IEEE Comp. Soc. Press, USA,. Pgs. 1110-1114.
[26] Negi Atul, Chakravarthy Bhagvati, and V.V.Suresh Kumar. 2002 Non-linear Normalization to Improve Telugu OCR Proc. of Indo-European Conf. on Multilingual Communication Technologies, pgs 45-57,Tata McGraw Hill Book Co., New Delhi,
[27] Pujari Arun K , C Dhanunjaya Naidu & B C Jinaga 2002 An Adaptive Character Recognizer for Telugu Scripts using Multiresolution Analysis and Associative Memory. ICVGIP, Ahmedabad.
[28] Esa rahtu, M. Salo, J. Heikkila , Jan Flusser: Generalization affine moment invariants for object recognition
[29] M. J. Baheti, R. J. Ramteke and K. V. Kale, Gujarati Handwritten Numeral Recognition, Proc. of International Conference on Cognition and Recognition Mandya 2008 pp 51-56

AUTHORS PROFILE

Mamta Maloo has received M.Sc. Degree in Computer Science from Sant Gadge Baba Amravati University in 2003. Since 2007, she has been a Ph.D. student of Dr. K. V. Kale. Her current research interests include pattern recognition, image analysis and document processing. Presently she is working as Leturer at SNJB's College of Engineering, Nashik.

Dr. Karbhari Vishwanath Kale, M.Sc., MCA (Engg & Tech), Ph.D., FIETE., Professor and Head Dept. of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada, University, Aurangabad, INDIA Completed M.Sc Physics (Electronics) in 1987, B.Ed in 1989, MCA (Engg. And Tech) in 1995 and Ph.D on Superionics in 1997. Fifteen students awarded Ph. D. in Computer Science. Ten students have working for Ph.D. under the guidance. One student awarded and two are working for M.Phil. under the guidance