I J M I

International Journal of Machine Intelligence
ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 3, 2011, pp-160-163
Available online at http://www.bioinfo.in/contents.php?id=31

COMPARISON OF CLASSIFIERS FOR GUJARATI NUMERAL RECOGNITION

BAHETI M. J.[1]*, KALE K.V.[2], JADHAV M.E.[3]
[1]Department of Computer Engineering, SNJB's College of Engineering, Chandwad, Nashik, India
[2]Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India.
[3]Institute of Management Studies & Information Technology, Aurangabad, India
*Corresponding Author: Email- mamta.baheti610@gmail.com

Received: September 29, 2011; Accepted: November 03, 2011

Abstract- Handwritten script has been the most acknowledged method of collecting, storing and transmitting information all the way through the centuries. In this paper an attempt is made to compare the offline handwritten character recognition system for the isolated Gujarati numerals. For feature extraction affine invariant moments based model is used. We are using KNN classifier and PCA (to reduce dimensions of feature space) and used Euclidean similarity measure to classify the numerals. KNN classifier yielded 90 % as recognition rate whereas PCA scored recognition rate of 84%. The comparison of KNN and PCA is made and it can be seen that KNN classifier has shown better results as compared to PCA classifier.

Key words - Affine Moment Invariant, PCA, KNN.

Introduction

Handwritten script has been the most acknowledged method of collecting, storing and transmitting information all the way through the centuries. This is useful for making digital copies of handwritten documents, and also in many automated processing tasks, such as automatic mail sorting or cheque processing. In automated mail sorting, letters are directed to the correct location by recognition of the handwritten address. Similarly, cheque processing involves recognizing the words making up the cheque amount.

In India approximately twenty-two official languages are used across various regions of the country [1-4]. Each language has the variety representing its peculiarity as well as shares some of the similarities with other languages. Gujarati is spoken and used as official language in Gujarat. Gujarati language is written by using Gujarati script which in turn is derived from Devanagari. Fig. (**1**) shows the numerals belonging to Gujarati language.

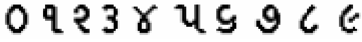

Fig. 1Gujarati numerals

Gujarati numerals show some same appearances like other numerals in Devanagari. Numerals like 0, 2, 3, 4, 7 and 8 are same as that in Devanagari but numeral 1 has. This paper deals with the comparison of classifiers for recognition of Gujarati handwritten numeral.

This paper is organized in following sections; Section 2 describes brief literature survey done for Indian languages recognition. Section 3 details the steps taken for preprocessing. Section 4 describes algorithm which we have used to implement the paper. Section 5 elaborates the feature extraction done. Section 6 describes PCA (Principal Component Analysis) and KNN (K-Nearest Neighbor) based numeral recognition. Section 7 details the conclusion of work done.

Literature Survey

Pieces of work were found for the recognition of Indian scripts [5, 6]. Pal & Chaudhuri made a wide-ranging appraise for recognition of Indian scripts [6]. A modified quadratic classifier based scheme had been proposed for recognition of off-line handwritten numerals of Devnagari, Bangla, Telugu, Oriya, Kannada and Tamil scripts. The bounding box of a numeral was segmented into blocks and the directional features were computed in each of the blocks. These blocks were then down sampled by a Gaussian filter and the features obtained from the down sampled blocks were fed to a modified quadratic classifier for recognition. After using five-fold cross validation technique accuracy was 99.56%, 98.99%, 99.37%, 98.40%, 98.71% and 98.51% for Devanagari, Bangla, Telugu, Oriya, Kannada, and Tamil scripts, respectively [7]. In the work [8] it had been reported that four sets of feature namely directional opening, directional closing, direction erosion and k-curvature features for writer recognition on Telugu handwritten documents. Each of the features was extracted from the words after dividing them into a number of cells and then subjected to a nearest neighbor classifier for writer recognition. Rahiman [9] proposed an algorithm where characters were grouped in to different classes based on their HLH intensity patterns and produced an accuracy of 88%. Anuradha developed a Telugu optical character recognition system for a single font. A 2-stage classifier with first stage identified the group number of the test character, and a minimum-distance classifier at the second stage identified the character. Recognition accuracy of 93.2% was reported [10]. Jawahar [11] proposed a Bilingual OCR for

Hindi-Telugu documents. It was based on Principal Component Analysis followed by support vector classification. An overall accuracy of approximately 96.7% was reported. Chaudhari [12] used preprocessing techniques like skew correction, line segmentation, zone detection, word and character segmentation and then the combination of stroke and run-number based and water reservoir based features were used as classifiers. They achieved 96.3% of accuracy. The features of Oriya OCR developed at the Indian Statistical Institute, Kolkata were similar to the Bangla OCR developed by the same team Chaudhuri [12, 13]. On average, the system reported an accuracy of about 96.3%.

Gujarati Handwritten Database

For the identification of Gujarati handwritten numerals the scheme required to have database but at present there is no standardized database available for Gujarati handwritten numerals, the database has been formed. Data was collected from people of different age groups, belonging to different profession, illiterate but knowledge of writing Gujarati irrespective of gender. From such diversity of group, one sample of each digit from 80 persons was collected on a specially designed datasheet as shown in Fig. (2) for data collection.

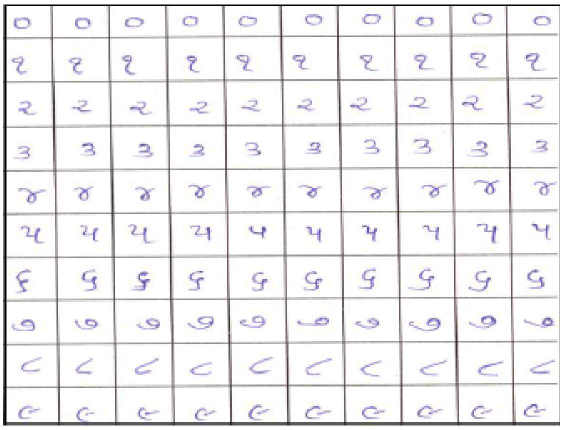

Fig. 2 Sample sheet for database collection

The algorithm (as shown in Fig. (**3**)) identifies the handwritten Gujarati numerals based on iterative approach. It identifies more than one numeral. The overall recognition rate if finally achieved by ratio of sum of the correctly recognized numerals by total numerals for KNN as well as PCA.

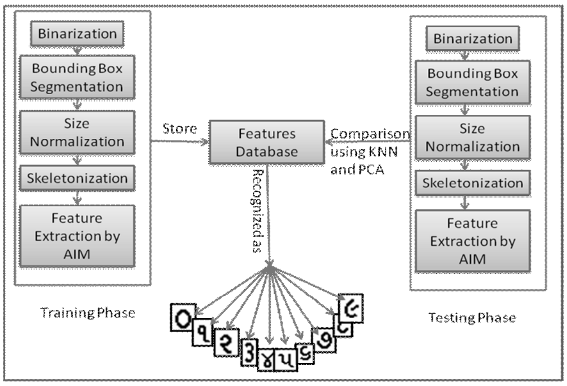

Fig. 3 Algorithm employed for Gujarati numerals recognition

Affine Moment Invariant

Now the affine invariant moments [14, 15] are derived for each of the numeral image as follows. The AMIs is invariant under general affine transformation

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

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:

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

Recognition

For the derived feature set from affine invariant moments, we apply the KNN and PCA classifier.

K-Nearest Neighbor Classifier

In pattern recognition [16], the k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor [17].

Principal Component Analysis

Principal Components Analysis (PCA) is a multivariate procedure, which rotates the data such that maximum variability's are projected onto the axes [18]. The main use of PCA is to reduce the dimensionality of a data set while retaining as much information as is possible. It computes a compact and optimal description of the data set. Data points are vectors in a multidimensional space. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on [19-21]. PCA is theoretically the optimum transform for a given data in least square terms. The Principal Component Analysis module in proposal system generates a set of data, which can be used as features in building feature vector section. Fig. (**4**) shows a co-ordinate system (X_1, X_2). Choose a basis vector such that these vector points in the direction of max variance of the data, say (Y_1, Y_2), and can be expressed as

$$\left.\begin{aligned} Y_1 &= X_1 \cos\theta - X_2 \sin\theta \\ Y_2 &= X_1 \sin\theta + X_2 \cos\theta \end{aligned}\right\}\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

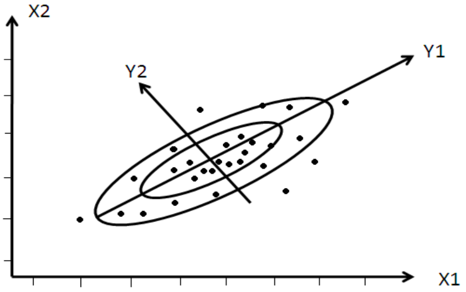

Fig. 4 Ellipse Distribution with PCA

We have used Euclidean distance as the similarity measure for comparing the testing and training samples.

Results and Discussion

After applying the KNN and PCA [27] classifier we have tabulated the results in table 1.

Table 1 Comparison of Recognition rate in % for affine invariant moments for PCA and KNN classifiers

Numerals	**PCA**	**K-NN**
Zero	85	95
One	79	93
Two	80	88
Three	86	90
Four	90	93
Five	80	92
Six	80	80
Seven	82	88
Eight	91	95
Nine	88	90
Average	**84.1**	**90.04**

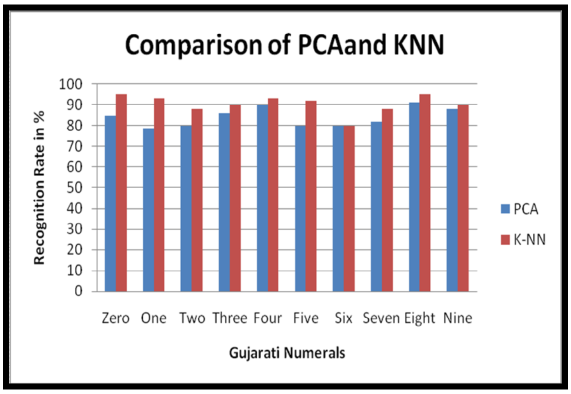

Fig. 5 Comparison of classifiers for affine invariant moments

By using KNN classifier we have found that for numeral 0 reported the recognition rate as 95%, whereas for numeral 1 as 93%. For numerals 2, 6 and 7 recognition rates is found less as compared to others as 88%, 80% and 88% respectively. Numerals 3 and 9 have shown same results as 90%. Numerals 0 and 8 reported to have maximum recognition rate 95% among all these recognition rates. The overall recognition rate is 90.04%. It showed good results for numerals 0, 4 and 8.

By using PCA classifier we have found that for numeral 0 reported the recognition rate as 85%, whereas for numeral 1 as 79%. For numerals 2, 5 and 6 recognition rates is found as 80% respectively. Numerals 3, 4, 7 and 9 have shown some better results as compared to 2, 5 and 6 as 86%, 90%, 82% and 88% respectively. Numeral 8 reported to have maximum recognition rate 91% among all these recognition rates. The overall recognition rate was 84.1%. It showed good results for numerals 4 and 8.

Conclusion

As compared to overall recognition rate from table 1 KNN have shown recognition rate of 90.04% where as PCA shows 84.1%. One can observe from Fig. (**5**) that KNN proves to be better classifier than PCA classifier for affine invariant moments as feature extraction technique. The results were compared with [22-25] and were found to be better because we have applied our algorithm on noisy numerals. In future we will try to improve the recognition by doing the modification in the current system.

References

[1] *http//ccat.sas.upenn.edu Accessed 15 June 2007*

[2] *http//languages.iloveindia.com Accessed 15 June 2007*

[3] *http//india.mapsofindia.com Accessed 15 June 2007*

[4] *http//en.wikipedia.org Accessed 15 June 2007*

[5] Chaudhuri B. B. (2007) *In Digital Document Processing: Major Directions and Recent Advances Springer-Verlag London Ltd,* 99–119.

[6] Internet Archive. at http://www.archive.org.

[7] Pal U., Wakabayashi T., Sharma N., Kimura F., (2007) *International Conference on Document Analysis and Recognition, Published by IEEE Computer Society in IEEE Xplore USA,* 2:749 – 753.

[8] Purkait P., Kumar R., Chanda, B. (2010) *International Conference on Frontiers in Handwriting Recognition*, *Published by IEEE Computer Society in IEEE Xplore USA,* 658 – 663.

[9] Rahiman M.A., Shajan A., Elizabeth A., Divya M.K., Kumar G.M. and Rajasree M.S. (2010) *International Conference on* Machine Learning and Computing, *Published by IEEE Computer Society in IEEE Xplore USA,* 147 – 151.

[10] Srinivas B. A., Agarwal A. and Rao C.R. (2007) *International Conference On Systemics, Cybernetics, And Informatics*, Hyderabad, 654-659.

[11] Jawahar C. V., Pavan M. N., Kumar S. S., Ravi Kiran S. (2003) *International Conference on Document Analysis and Recognition, Published by IEEE Computer Society in IEEE Xplore USA,*1, 408-412.

[12] Chaudhari B.B., Pal U. and Mitra M. (2002) *Sadhana*, 27(1), 23-34.

[13] *Journal of Language Technology, Vishwabharat@tdil*, July2003.

[14] Rahtu E.,Salo M.,Heikkila J. and Flusser J. (2006) *18th International Conference on Pattern Recognition, Published by IEEE Computer Society in IEEE Xplore USA,634-637.*

[15] Thomas Suk and Jan Flusser (2002) *Object recognition supported by user interaction for service robots Published by IEEE Computer Society*, 339-342.

[16] Duda R. O., Hart P. E. and Stork D.A., *Pattern Recognition, Second Edition, Wiley Student Edition.*

[17] Hall P., Park B.U., Samworth R.J. (2008) *Annals of Statistics,* 36 (5), 2135-2152.

[18] Ramteke R. J., Borkar P. D. and Mehrotra S. C. (2005) *International Conference on Cognition and Recognition*, 482-489.

[19] Smith L. I. (2002), http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf.

[20] Shlens J. A. (2005) http://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition _jp.pdf.

[21] Jolliffe I. T. (2002), *Springer Series in Statistics, 2nd ed., Springer.*

[22] Desai A. A. (2010) *Pattern Recognition* 43, 2582-2589.

[23] Desai A. A. (2010) *International Conference on Image Processing, Computer Vision, and Pattern Recognition*, 733-739.

[24] Prasad, J.R., Kulkarni U.V. and Prasad R.S. (2009) *International Conference on Emerging Trends in Engineering and Technology, Published by IEEE Computer Society in IEEE Xplore USA,* 263 – 268.

[25] Prasad J.R., Kulkarni U.V. and Prasad R.S. (2009) *3rd International Conference on Anti-counterfeiting, Security, and Identification in Communication*, 611 – 615.

[26] Baheti M.J., Mane A.V., Hannan M.S. and Kale K.V. (2011) *CiiT International Journal of Digital Image Processing,* 3(11), 709-715.