Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

KNN and PCA Approach for Gujarati Script

Mamta Baheti

Hislop College, Nagpur, Maharashtra, India

Researchers are very interested in the essential pattern recognition used for offline handwriting recognition. Any handwritten material may be converted into system-editable textual data by identifying concealed forms and understanding the typescripts included within the papers. Gujarati is one of India's 22 officially recognized languages. OCR in Gujarati has a number of problems, making it challenging to spot worldwide invariant forms and irregularities in handwritten Gujarati script. Another significant problem with handwritten Gujarati script is the shortage of an important comparison dataset. This problem was discovered, and a review was made. We created a dataset and treated it with invariant moments and used Gaussian and KNN Classifier for classification. KNN gave better results as compared to Gaussian.

Keywords: KNN, Guassian, Invariant Moments, Gujarati.

1. Introduction

For scripts like Roman, Chinese, Japanese, and others, OCR systems have been around for a while and are already fairly accurate. However, there isn't a commercial OCR solution for Indian characters yet [1][2]. The proliferation of information and communication technologies in India has increased the need for OCR tools for Indian scripts, but it is more challenging to create it due to the more complex character shapes.

India's defining characteristic is the diversity of its cultures, castes, and languages. The country is divided into many linguistic areas. In addition to Hindi, Devanagari, Gujarati, Kannada, Gurumukhi, Oriya, Malayalam, Telugu, Urdu, and Tamil are regarded as fundamental scripts. The Devanagari language family includes Gujarati, which is spoken by 65 million people in Gujarat. In addition to those who reside in Gujarat, Gujarati is a language that is spoken all over the world. There is not a concept of cursive writing in Gujarati script, which is written from left to right. The Gujarati script has several characters that have a similar appearance to one another, which makes identification difficult.

2. Related Work:

For the first time, Sinha and Mahabala [3] and Sethi [4] worked with hand-printed characters

and typed Devanagari script, respectively. For the purpose of recognising Devanagari letters written by hand and printed by machines, they provided a syntactic pattern analysis system with an embedded image language. Using structural characteristics and neural networks for classification, Devanagari text recognition in the Sanskrit manuscript 'Saddharmapundarika' [5] was successfully completed with an accuracy of 98.09%. In [6], Pal & Chaudhuri tried to OCR two scripts: Bangla and Devanagari. For printed Devanagari, Pal & Chaudhuri [7] claimed a comprehensive OCR system. Modified and elementary typescripts are recognized using a structural feature-based tree classifier, whilst compound characters are recognised by a hybrid technique combining structural and run-based template features. The approach claimed an accuracy of 96%.

A study of several structural strategies for feature extraction in OCR of various scripts was provided by Veena [8]. On a database of 20 writers, Connel [9] reported an online Devanagari script recognition effort with 86.5% accuracy. In terms of performance on individual characters, Sinha & Bansal [10] scored 93%. K-nearest neighbour (KNN) and neural network classifiers have been used in attempts to recognise printed Devanagari characters [11].

Bhattacharya & Chaudhari [12] gathered a database for Devanagari numerals from postal addresses and job application forms, while Kompalli [13] detailed database assessment methods. Parvati Iyer's work [14] deals with issues that emerge while creating OCR algorithms for noisy photos. Only a 55% character recognition rate was recorded.

Recent years have seen a number of studies on recognizing handwritten Gujarati characters. For Gujarati character recognition, these researches have mostly investigated various feature abstraction procedures and classifiers. In one such attempt, the researchers in the reported work [15] classified handwritten Gujarati numerals employing a multi-layered FFNN classifier and features based on projection profile achieving an accuracy of 81.66%. Radial histogram-based features and a Euclidean Distance classifier were coupled by the authors [16] to achieve an accuracy of 26.86%. Using moment-based characteristics that are affine invariant and an SVM classifier, authors in the reported work [17] were able to recognise handwritten Gujarati numerals with a 91% accuracy rate. The researcher in [18] employed structural characteristics and used classifier as decision tree which achieved 88.78% accuracy. Using structural decomposition-based features and SVM classifier, the authors in [19] were able to recognise handwritten Gujarati characters with an accuracy of 99.48%. In their study of affine invariant moment-based features utilising KNN and PCA-based classifiers, the authors [20] achieved accuracy rates of 90% and 84%, respectively. Authors in [21] introduced a zoning-based feature extraction method that, when combined with naive Bayes and neural network classifiers, produced handwritten Gujarati character recognition accuracy of 95.92%.

In the reported work [22], authors introduced features based on the prediction algorithm and SVM grid and achieved an accuracy of 98.93%. The reported work [23] employed low-level stroke characteristics with KNN and SVM classifiers and obtained accuracy on handwritten Gujarati and Hindi of 98.46% and 98.65%, respectively. In the research by the authors [24], an LSTM model was created for Gujarati character recognition using a dataset of 58,000 pictures. They managed to obtain an accuracy of about 97%. Using a KNN classifier in conjunction with moment-based and centroid distance-based features, authors [25] obtained a recognition accuracy of 63.1%. In the reported work [26] used a weighted KNN classifier with

a pattern descriptor and Gabor phase XNOR pattern-based features to achieve an 86.33% identification accuracy for handwritten Gujarati characters. In another reported work [27] cross correlation-based features with naive Bayes and support vector machine classifiers led to recognition accuracy of 53.12%, 68.53%, and 66.43%. In a collection of 10,000 pictures, the authors [28] recognition of handwritten Gujarati characters using CNN and MLP yielded success rates of 97.21% and 64.48%, respectively. Using transfer learning and CNNs, the authors [29] concentrated on handwritten Gujarati character recognition. They had adjusted the relevant hyperparameters through arduous and thorough testing. According to the authors, VGG16 and DenseNet both offered accuracy that was equivalent to MobileNet's, which was above 97%.

3. Experimental Methods

Feature-based recognition of printed and written characters based on their position, size, orientation, inclination, and other variables is the goal of ongoing research. This could be handled by extracting unique features. We use Hu's [30] moment invariants, taking into account the independence of the underlying variables.. A set of seven invariant moments can be derived from out of which six moments are absolute orthogonal invariants (and one skew orthogonal invariants) [31-33]

Algorithm based on Invariant Moments using Gaussian distribution function

1. Image is taken from database
2. Resize it to 40x40
3. Complement the image
4. Image is binarized
5. Dilate the binarized image with structuring elements ‘line’ and ‘diamond’
6. Thin the image with various iterations (1,2,3,4,5,Inf)
7. Apply Invariant Moments Approach
8. Use Gaussian membership function as classifier
9. Recognition rate is computed

Algorithm based on Invariant Moments using K-Nearest Neighbor

1. Input image is fetched from database
2. Image is resized to 40x40
3. Image is complemented
4. Image is binarized
5. Dilate the binarized image with structuring elements ‘line’ and ‘diamond’
6. Thin the image with various iterations (1,2,3,4,5,Inf)

7. Apply Invariant Moments Approach
8. Use K-Nearest Neighbor as classifier
9. Compute the recognition rate on the basis of misclassified and classified numerals

Data Collection:

Considering various aspects of handwritten characters ten different forms have been filled by the different writers. Data was collected from people of different age groups from 15 to 60 years, belonging to profession like student, professor, housewife, illiterate but knowledge of writing Gujarati, lawyer, warden, etc. irrespective of gender. In addition also, the data was collected from other possible places where Gujarati is practiced in speaking and writing. Different samples of each number were taken on a specially designed and an aimlessly designed datasheet from 200 persons. While creating database different pen and ink was used. The data sheet was then scanned using a HP 2000 flatbed scanner with resolution of 200 dpi and HP 2400 flatbed scanner with resolution of 300dpi. Numeric database comprised of numerals 0 to 9. They were isolated manually from some sheets and from remaining sheets an algorithm was prepared to store the numerals in appropriate folders formed rendering to the writers. Each writer wrote 102 samples of each numeral. Likewise data was collected from 200 writers. The isolated digits were applied with Invariant moments feature extraction technique and then these features were given as input to Gaussian membership function and KNN classifiers.

The results were either as properly recognized or misrecognized numeral. The confusion matrix with correctly recognized and misrecognized numerals is formed.

Invariant Moments Approach using Gaussian Classifier

Even though the input picture was noisy, the number 0 is correctly recognized in 93.75% of cases, whereas it was incorrectly identified as 5 and 7 in 3.75% and 2.5% of cases, respectively. The numbers 6 and 3 are misclassified as 2, whereas the remaining numbers are correctly classified as other numbers. 15% of the time, the number 1 is assigned the wrong value, 4, instead. The numbers 1, 4, 5, and 8 are interchanged. There is a 92.5% recorded recognition rate for the number 8. Being a two-part sign, the number 9 is frequently mistaken for the number 7. The overall recognition rate is around 72%. Even though the overall findings are less encouraging, the numbers 0, 4, and 8 have shown positive outcomes when compared to Desai's results. [39]

Invariant Moments Approach using K- Nearest Neighbor as classifier

The numbers 0 and 4 had a 93% recognition rate, while the number 1 had a 90% recognition rate. The same identification rate for the numbers 2 and 5 was determined to be 88%. It was discovered that the identification rate for the numbers 3 and 9 was 86%. Numbers 6 and 7 had lower recognition rates than the rest, at 80% and 82%, respectively. In all of these identification rates, the number 8 was reported to have the highest recognition rate at 95%. Overall, 88.1% of numerals were recognized. Results for the numbers 0, 4, and 8 were satisfactory.

Comparison of Gaussian and KNN classifiers with features extraction approach for invariant moments

Numerals	Gaussian	K-NN
0	93.75	93
1	72.5	90
2	51.25	88
3	48.75	86
4	90	93
5	66.25	88
6	53.75	80
7	66.25	82
8	92.5	95
9	76.25	86
Average	71.25%	88.1

The recognition rates of the numbers 1, 2, 3, 5, 6, and 7 were lower with Gaussian than with KNN. The Gaussian performs at its best when the numbers are 0. The recognition rates reveal that KNN has performed well across the board for all numbers. For zero, KNN has demonstrated results that are 20% better than Gaussian. Additionally, KNN has demonstrated a recognition rate of 88.1% when compared to Gaussian's 71.25% for the total recognition rate. When using KNN as a feature extraction approach for invariant moments, one can see that it performs better as a classifier than Gaussian. A typical OCR system for the Indian language contains several classes and high-dimensional feature vectors. Variability of characters is also very high at each occurrence. So for these constraints we have procured good results

4. Conclusion

In this paper, we have proposed an algorithm for recognition of handwritten numerals in Gujarati. The system is based on Invariants Moments. The Gaussian Distribution Function, and K-Nearest Neighbor Classifier have been adopted for the classification of extracted features. In the present work, we emphasized on Handwritten Gujarati Numerals. As there was no standardized database available for Handwritten Gujarati numerals, a database was prepared by collecting the data samples from people belonging to various domains.

It was found that it was possible to enhance performance of system if a character is divided in a systematic manner and features of each divided part are used in recognition system. KNN (88.1%) proved to be better classifier than Gaussian (72%) for invariant moments as feature extraction technique.

References

1. Pal U and Chaudhuri B B (2004) Indian script character recognition: a survey. Pattern Recognition. 37(9): 1887–1899
2. Bag S and Harit G (2013) A survey on optical character recognition for Bangla and

Devanagari scripts. Sadhana. 38(1): 133–168
3. R. M. K. Sinha and H. N. Mahabala, (1979) Machine Recognition of Devnagari Script. IEEE Transactions on Systems, Man and Cybernetics, Pattern Recognition, Vol. 9(8), pp. 435 - 441.
4. I.K. Sethi, Machine (1977) Recognition of Online Handwritten Devnagari Characters. Pattern Recognition, Vol. 9, pp. 69 - 75.
5. K. Keeni, T-Nishino, H. Shimodaria and Y. Tan, (1996) Recognition of Devnagari characters using Neural Networks. IEICE Transactions on Information and Systems, Vol. E79-D, No.5, pp. 523 -528.
6. Chaudhuri B.B., Pal U., (1997) An OCR system to read two Indian Language Scripts: Bangla and Devnagari(Hindi). ICDAR, pp. 1011 - 1015.
7. Pal U and Chaudhuri B B (1997) Printed Devnagari script OCR system. Vivek 10:12-24
8. Veena Bansal (1999) Integrating knowledge sources in Devnagari text recognition PhD Thesis, IIT Kanpur.
9. Connel S. D., Sinha R.M.K., Jain A. K., (2000) Recognition of Un-constrained On-Line Devnagari Characters. Proceedings of ICPR, pp. 2368 - 2371.
10. Sinha R.M.K. and Bansal V., (2001) A complete OCR for Printed Hindi Text in Devnagari Script. Proceedings of ICDAR, pp. 800 - 804.
11. Kompalli S., Setlur S., Govindaraju V. and Vemulapati R., (2003) Creation of data resources and design of an evaluation test bed for Devnagari script recognition. Proceedings of the thirteenth International Workshop Research Issues on Data Engineering: Multilingual Information Management, pp. 55 – 61.
12. Parvati Iyer, et.al (2005) Optical Character Recognition System for Noisy Images in Devnagari Script UDL Workshop on Optical Character Recognition with Workflow and Document Summarization
13. U. Bhattacharya and B. B. Chaudhuri, (2005) Databases for Research on Recognition of Handwritten Characters of Indian Scripts. Proceedings of Eighth International Conference on Document Analysis and Recognition, pp. 789 - 793.
14. Desai A. A. (2010) Gujarati handwritten numeral optical character reorganization through neural network. Pattern recognition. 43(7): 2582–2589
15. Shah L, et. al (2014) Handwritten character recognition using radial histogram. International Journal of Research in Advent Technology (4). 24-28
16. Maloo M and Kale K V (2011) Support vector machine based Gujarati numeral recognition. International Journal on Computer Science and Engineering. 3(7): 2595–2600
17. Thaker H and Kumbharana C (2014) Analysis of structural features and classification of Gujarati consonant for offline character recognition. International Journal of Scientific and Research Publications. 4(8): 1–5
18. Sharma A K, et al (2019) Handwritten Gujarati Character Recognition Using Struc- tural Decomposition Technique. Pattern Recognition and Image Analysis, 29(2), 325–338
19. Baheti M. J., Kale K. V.and Jadhav M.E., (2011) Comparison of classifiers for Gujarati numeral recognition. International Journal of Machine Intelligence, 3(3), 160–163
20. Sharma A K, Adhyaru D M, Zaveri T H and Thakkar P B (2015) Comparative analysis of zoning based methods for Gujarati handwritten numeral recognition. Nirma University International Conference on Engineering pp. 1–5.
21. Nagar R, Mitra S K 2015 Feature extraction based on stroke orientation estimation technique for handwritten numeral. In: 2015 eighth international conference on advances in pattern recognition (ICAPR) (pp. 1–6).
22. Goswami M M and Mitra S K 2015 Offline handwritten Gujarati numeral recognition using low-level strokes. Inter- national Journal of Applied Pattern Recognition. 2(4): 353–379
23. Rajyagor B and Rakholia R (2021) Isolated Gujarati Hand- written Character Recognition

(HCR) using Deep Learning (LSTM). International Conference on Electrical, Computer and Communication Technologies (ICECCT) pp. 1–6.

24. Patel C and Desai A 2013 Gujarati handwritten character recognition using hybrid method based on binary tree- classifier and k-nearest neighbour. International Journal of Engineering Research & Technology (IJERT). 2(6): 2337–2345
25. Prasad J R and Kulkarni U 2015 Gujrati character recognition using weighted k-NN and mean v2 distance measure. International Journal of Machine Learning and Cybernetics. 6(1): 69–82
26. Sharma A K, Adhyaru D M and Zaveri T H 2018 A novel cross correlation-based approach for handwritten Gujarati character recognition. In: Proceedings of First International Conference on Smart System, Innovations and Computing (pp. 505–513). Springer, Singapore
27. Pareek J, Singhania D, Kumari R R and Purohit S 2020 Gujarati handwritten character recognition from text images. Procedia Computer Science. 171: 514–523
28. Krishn Limbachiya, Ankit Sharma* , Priyank Thakkar and Dipak Adhyaru Identification of handwritten Gujarati alphanumeric script by integrating transfer learning and convolutional neural networks Sådhanå (2022)47:102
29. M. K. Hu, "Visual Pattern Recognition by Moment Invariants", IRE Trans. On Information Theory, Vol. IT-8, pp. 179-187, 1961.
30. Trier O D, Jain A K, Taxt T(1996) Feature extraction methods for character recognition- a survey. Pattern Recognition 29:641-662.
31. Dhandra BV, Hangarge M, Hegadi R, Malemath V S (2006) Handwritten Script Identification Using Fuzzy K-Nearest Neighbor. Proc. of IEEE 1st ICSIP II: 587-591.
32. Andrews M J (1999) Moment Representation of blobs in 2-D Intensity Images, http//wwwgweepnet/~rocko/Moment/ Accessed 15 June 2007
33. Hanmandlu M and Murthy OV R, (2005) Fuzzy Model Based Recognition of Handwritten Hindi Numerals. Proc ICCR, p 490-496.
34. R. O. Duda, P. E. Hart and D.A. Stork, "Pattern Recognition" Second Edition, Wiley Student Edition.
35. Hall P., Park B.U., Samworth R.J. "Choice of neighbor order in nearest neighbor classification", Annals of Statistics vol 36 (5): 2135-2152 2008.
36. http://en.wikipedia.org/wiki/k_nearest_neighbour_algorithm
37. D. G. Terrell; D. W. Scott (1992). "Variable kernel density estimation". Annals of Statistics 20 (3): 1236–1264. doi:10.1214/aos/1176348768.
38. Dholakia J, Negi A and Mohan S R 2009 Progress in Gujarati document processing and character recognition. In: Guide to OCR for Indic Scripts (pp. 73–95). Springer, London