

Offline Handwritten Character Recognition for Gujarati Language

Sumit G. Trivedi¹ Arun Nandurbarkar²

¹PG Student ²Associate Professor

^{1,2}Department of Electronics & Communication Engineering

^{1,2}L.D. College of Engineering Ahmedabad, Gujarat, India

Abstract— This paper deals with the problem of recognizing handwritten Gujarati characters from scanned image. First, pre-processing steps are applied on input scanned image. i.e. RGB to binary conversion, removal of noise and different morphological operation to find skeleton of an image and normalization of each character image to 30x30 sizes. Second, Features are extracted from the normalized 30x30 image for each character. Here we have extracted gradient feature and also proposed a new combination of features which is combination of correlation function based feature, invariant moments and projection profiles. Third, these extracted features are supplied as an i/p to classifiers. i.e. SVM, k-NN and Naïve Bayes. The performance of different classifiers measured using 10-fold cross validation.

Key words: OCR, HCR, Gujarati Language, Feature Extraction, SVM, k-NN, Naïve Bayes

I. INTRODUCTION

OCR is a system which automatically retrieve, analyse and save information. It is the translation of handwritten or printed text into machine editable text which is also called as digitization. It is field of research in pattern recognition. Basically OCR is machine replication of human function and it also has various commercial application such as postal automation, bank cheque processing, automatic data entry application etc. All OCR system can be examined into two categories: (i) Systems classified according to data acquisition techniques (online and offline). (ii) Systems classified according to the text type (printed and handwritten). Various techniques has been proposed towards offline handwritten character recognition by many authors for different languages i.e. English, Chinese, Roman and Arabic script. Lots of work also has been proposed for Indian languages as well. Comparatively very less work has been reported for Gujarati language. The text type (printed and handwritten). Various techniques has been proposed towards offline handwritten character recognition by many authors the text type (printed and handwritten). Various techniques has been proposed towards offline handwritten character recognition by many authors for different languages i.e. English, Chinese, Roman and Arabic script. Lots of work also has been proposed for Indian languages as well. Comparatively very less work has been reported for Gujarati language.

Gujarati language has its origin from Sanskrit, like most other Indian writing system viz. Marathi, Devanagari, Gurmukhi. Gujarati language is largely used in the state Gujarat in western India. It contain total 34 consonants, 12 vowels and some special character as well. Our dataset contains 150 samples for each Gujarati character [1].

II. PREVIOUS WORK

Many authors have published their research work on OCR and contributed in this field, but very less work found on HCR especially for Gujarati language.

Swital J. macwan [1] took different features form transform domain, statistical domain, structural domain and Geometrical domain and used support vector machine as a classifier to classify Gujarati consonants and obtained 96.65% accuracy.

Archana N. Vyas [2] took feature as modified chain code, DFT and DCT and also used there different classifiers k-NN, SVM and ANN and obtained highest accuracy 93.60% for Gujarati Numerals.

Ankit K. Sharma [3] took zoning based features and classified them using naïve Bayes classifier and ANN and got maximum accuracy 95.92% for Gujarati Numeral.

Jayashree R. Prasad [4] took template as a feature and classified them using Neural Network and got accuracy 71.66%.

Apurva A. Desai [5] took four different profiles horizontal, vertical, positive and Negative slant used as template and classified them using multi-layered feed forward Neural Network and achieved 82% of accuracy

III. PROPOSED METHODOLOGY

Our proposed work presents a complete handwritten character recognizer. The system can be split into three stages (as shown in figure 1): a) pre-processing, b) proposed feature extraction scheme, and c) SVM, k-NN and Naïve Bayes -based training and classification. In the following we will describe each of these in detail.

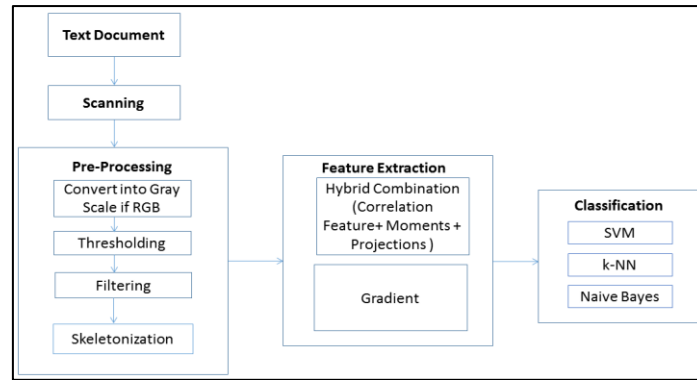


Fig. 1: Proposed Diagram for OCR

IV. IMAGE ACQUISITION AND PRE-PROCESSING

Data samples of handwritten Gujarati characters from different writers on plain white papers. The writers were allowed to write the characters without any constraint on writing style or equipment. The written documents were then stored in the PC as gray-scale images. [1]

A. Thresholding

This is used to convert gray-scale image to binary image. A threshold is defined for this purpose; the pixels above this threshold are set to white and those below the threshold are set to black.

B. Filtering

This is used to remove the noise present in the binary image. To remove small black specks in the image and the black shade appearing at the edges, we can use different filters for different noise. The thresholding and filtering steps often resulted in some broken characters. To re-join the broken characters, we performed image dilation operation on the filtered images.

C. Boundary Tracing

This step identifies the connected components i.e. characters in the filtered images and stores them in an array. To find the connected components, our algorithm starts by traversing the rows of filtered image. If in any row the algorithm finds a foreground pixel, it marks that pixel then it picks and marks all the neighbours of found pixel in different search directions till all the pixels of the potential character have been traversed and marked. If in any row the algorithm does not find any foreground pixel, it will continue its search in the next row.

D. Skeletonization

This step reduces the thickness of character image to one-pixel by removing the pixels on the boundary of the character but without breaking the character. Features will be extracted from the character skeletons.

V. FEATURE EXTRACTION

Feature Extraction is the process of collecting different and very useful information of an object or a group of objects, so based on that collected information, we can classify new unknown objects by matching it. Feature is the robust representation of the raw data.

A. Correlation Function based Features

We extract correlation function based features [6] from the skeletonized characters. The normalized character images are of size 30X30. We divide each character into different segments, each of size 9X9. We define four elementary shapes, each of the size same as that of segments, i.e. 9X9. The four elementary shapes chosen are horizontal line, vertical line, positive slant line and negative slant line.

To compute the correlation, we need to define the correlation points in the 30X30 window of the normalized character samples.

Horizontal axis of normalized 30X30 character image:

$$m = [5; 10; 15; 20; 25]$$

Vertical axis of normalized 30X30 character image:

$$n = [5; 10; 15; 20; 25]$$

Combining the axes in above equation, we obtain following correlation points around which we take the different segments of each character:

$$\begin{aligned}
 mn = & [5\ 5; 5\ 10; 5\ 15; 5\ 20; 5\ 25; \\
 & 10\ 5; 10\ 10; 10\ 15; 10\ 20; 10\ 25; \\
 & 15\ 5; 15\ 10; 15\ 15; 15\ 20; 15\ 25; \\
 & 20\ 5; 20\ 10; 20\ 15; 20\ 20; 20\ 25; \\
 & 25\ 5; 25\ 10; 25\ 15; 25\ 20; 25\ 25]
 \end{aligned}$$

We have computed normalized correlation value of different segments with the elementary shapes by Pearson's correlation function [7] as shown below:

$$h(m, n) = \frac{\sum e_i(x, y) * ch(x, y)}{\sqrt{\sum |e_i(x, y)|^2} \sqrt{\sum ch(x, y)^2}}$$

This gives total 100 features, 25 for each elementary shape.

B. Invariant Moment

The invariant moments [8] describes the rate of change in a shape's area and measure pixel distribution around the centre of character image. These moments are invariant to position, size and orientation of the character. So, here we compute seven invariant moments for each character which gives 7 dimensional feature vector.

C. Projection Profile

Projection profile is an accumulation of black pixels along rows or columns of an image. The discriminating power of horizontal and vertical projection profiles make them well suitable for the recognition of a complex language like Gujarati. We have extracted both vertical and horizontal projection profiles by counting the pixels column wise and row wise respectively which together forms a 20 dimension feature vector.

D. Gradient Feature

Gradient method is a search based edge detection method. To find the gradients, we have applied Sobel, Robert and Prewitt operator to calculate the horizontal gradient (gx) and vertical gradient (gy) components. Gradient directions are considered as real value that has range from [0 360). So we took gradient direction as a feature vector but obtained very less accuracy [1]. So to increase the accuracy we equally quantized direction range into a smaller no of integer values. There are 9 integer values used representing gradient scope: [0 40], [41 80], [81 120]... [321 360) and K= 1, 2, 3... 9 and we got 900 (30x30 image) feature vector of thinned image.

VI. CLASSIFICATION TECHNIQUES

Classification or Recognition process is for decision making, like this new character fit in which class or looks like. It means, in the phase of classification characters are identified and assign labelling. Performance of the classification depends on good feature extraction and selection. Various classification techniques are available and they all are ultimately based on image processing and artificial intelligence.

A. Support Vector Machine

Support vector machines (SVMs) solve binary classification problems. The SVM (binary classifier) is applied to multiclass character recognition problem by using one-versus-all strategy. SVM is supervised learning technique. SVM not only has a rigorous theoretical foundation, but also performs classification very accurately, especially for high dimensional data. We have used libSVM [9] tool for classification.

B. K Nearest Neighbour

K- Nearest Neighbour is non-parametric recognition method. It is a supervised learning algorithm. No training is required. It estimates the posteriori probability from frequency of nearest neighbours of unknown pattern. i.e. it stores the data as a template and whenever new data comes it finds the minimum distance (Euclidean distance) between the unknown data and the training sample's templates. Then after it calculates the majority of nearest neighbours and assign that class label to unknown sample. The value of k, interprets the number of neighbours on which calculation is based. We have considered k=1.

C. Naïve Bayes Classifier

When the input dimensionality is high, the naive Bayes classifier technique [3] is used which is based on Bayesian theorem. Though naïve Bayes is simple yet it can outperform most sophisticated classification methods. The below equation is represents Naive Bayes rule.

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Based on the observation of evidences (E), the output of event (H) or hypothesis can be predicted. This is the basic idea of Bayes' rule. (1) A priori probability of H or P (H): Probability of an event before the evidence is observed. (2) A posterior probability of H or P(H | E) : Probability of an event after the evidence is observed.

VII. RESULT ANALYSIS

The overall dataset considered in this research are 5100 (34*150) [1]. Results obtained after application of all the mentioned methods of feature extraction analysed using 10 fold cross validation using three different classifiers are shown in below tables. Performance of each of three classifiers is measured with the help of obtained accuracies of the classifiers.

Feature	Feature Vector	Dataset	Classifier	Accuracy
Hybrid Combination (Correlation Feature+ Invariant Moment + Projection)	1*127	5100	SVM	89.82 +/- 1.49%
			Naïve Bayes	95.33 +/- 1.20%
			k-NN	97.27 +/- 0.46%

Table 1: Results of Hybrid Combination of Feature

Feature	Feature Vector	Dataset	Classifier	Accuracy
Gradient Feature (Prewitt)	1*900	5100	SVM	97.92 +/- 0.78 %
			Naïve Bayes	85.90 +/- 0.75 %
			k-NN	91.49 +/- 1.26 %

Table 2: Results of Gradient Feature

VIII. CONCLUSION

We have studied several techniques for optical character recognition for different languages. We have extracted the features using four different techniques. We have proposed a new combination of features namely correlation features, moments and projection profiles and also extracted the gradient features for the consonants. Then we have applied these features to three different classifiers namely SVM, k-NN and Naïve Bayes classifier for recognition of consonants. Then we have measured the accuracies of different classifiers using 10 fold cross validation and we have got highest accuracies 97.92% (SVM) and 97.27% (k-NN) using Gradient and Hybrid Combination of feature respectively.

REFERENCES

- [1] Swital J. Macwan, Archana N. Vyas, "classification of offline gujarati handwritten character", International Conference On Advances in Computing, Communication and informatics (ICACCI), IEEE-2015, p:1535-1541.
- [2] Archana N. Vyas, Mukesh M. Goswami, "Classification of Gujarati handwritten Gujarati numeral", International Conference on Advances in Computing, Communication and informatics (ICACCI) IEEE-2015, p: 1231-1237.
- [3] Ankit K. Sharma, Dipak M. Adhyaru , Tanish H. Zaveri , Priyank B Thakkar, "Comparative analysis of zoning based methods for Gujarati numeral recognition.", 5th Nirma University International Conference On engineering(NUiCONE), IEEE-2015, p:1-5.
- [4] Prasad, Jayashree R., U. V. Kulkarni, and Rajesh S. Prasad. "Offline handwritten character recognition of Gujrati script using pattern matching." Anti-counterfeiting, Security, and Identification in Communication, 2009. ASID 2009. 3rd International Conference on. IEEE, 2009.
- [5] Desai, Apurva A. "Gujarati handwritten numeral optical character reorganization through neural network." Pattern recognition 43.7 (2010): 2582-2589.
- [6] Muhammad Naeem Ayyaz, Imran Javed and Waqar Mahmood, "Handwritten Character Recognition Using Multiclass SVM Classification with Hybrid Feature Extraction", PJEAS, VOL.10, JANUARY 2012, p:57-67.
- [7] A. Miranda Neto, L. Rittner, N. Leite, D. E. Zampieri, R. Lotufo and A. Mendeleck, "Pearson's correlation coefficient for discarding redundant information in real time autonomous navigation system", 16th IEEE International Conference on Control Applications, Part of IEEE Multi-conference on Systems and Control, 1-3 October 2007, Singapore.
- [8] M.K. Hu, "Visual pattern recognition by moment invariants", IRE Transactions on Information Theory, pp. 179-187, 1962.
- [9] Chang, Chih-Chung; Lin, Chih-Jen, "LIBSVM: A library for support vector machines".
- [10] D. K. Patel, T. Sam, Manaj Kumar Singh, "English character recognition using learning rule & Euclidean distance metric", International Conference on signal processing and communication (ICSC), IEEE-2013, p:207-2
- [11] R. Jayadevan, Satish R. Kolhe, Pradeep M. Patil, and Umapada Pal, "Offline Recognition of Devanagari Script: A Survey"- IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 41, NO. 6, November 2011
- [12] Mark Nixon & Alberto Aguado "Feature Extraction & Image Processing" 2nd edition.
- [13] Rafael C. Gonzalez and Richard E. Woods, "Digital Image Processing". 3rd edition.
- [14] Dhanashree Joshi, Sarika Pansare, "Combination of multiple image features along with KNN classifier for classification of Marathi Barakhadi.", International Conference on computing communication, control & automation, IEEE-2015, p: 607-610.