ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

Gujarati Character Identification: A Survey

Mitul Modi[1], Fedrik Macwan[2], Ravindra Prajapati[3]

PG Scholar, Faculty of Tech & Engineering, M S University, Baroda, India [1]

PG Scholar, Faculty of Tech & Engineering, M S University, Baroda, India [2]

PG Scholar, Faculty of Tech & Engineering, M S University, Baroda, India [3]

Abstract: English Character Recognition techniques have been studied extensively in the last two decades and it gain unbelievable high progress and success ratio. But for regional languages these are still emerging and their success ratio is very poor. In Gujarat, there are thousands of people who can speak, write and understand only Gujarati language. Rapid growing computation may increase Indian CR methodology. Today the whole world is digitized. And heavy demand of digital documentation in any field like postal services, publishing house, automation, data entry, text entry and communication technology. Gujarati is mother tongue of Gujarat, spoken by thousands of people. There is less development in this area due to complexity in script. In this paper, we are presents different technique through which GOCR (Gujarati Optical Character Recognition) is possible.

Keywords: Gujarati Character Recognition, Segmentation, Feature Extraction, Image Classification, OCR

I. INTRODUCTION

The traditional way of entering data into a computer is through the keyboard. However, this is not always the best nor the most efficient solution. In many cases automatic identification may be an alternative. Various technologies for automatic identification exist, and they cover needs for different areas of application. Like

- Speech recognition
- Radio frequency
- Magnetic strips
- Optical Mask Reader
- Optical Character Recognition

Optical Character Recognition (OCR) is a process of converting printed or handwritten scanned documents into ASCII characters so that a computer can easily recognize. In other words, automatic text recognition using OCR is the process of converting an image of textual documents into its digital textual equivalent. The advantage is that the textual material can be edited, which otherwise is not possible in scanned documents in which these are image files.

II. FEATURE OF GUJARATI LANGUAGE

The Gujarati script was adapted from the Devanagari script. As other Indian languages the character set of Gujarati comprises of 36 consonants, 12 vowel and 6 signs, 12 dependent vowel signs, 10 digits. The consonants can be combined with the vowels and can form compound characters shown in Fig-1.

A word can be formed by combining the basic character(s), which may by combine with vowel(s). Basically Gujarati text can be divided into three parallel lines i.e. Cap line, Middle line, and Base line as shown in Fig. 2 [1]. In between base line and cap line, consonants and independent vowels are written (Middle zone). The line below the base line, used for writing dependent (lower) vowels (Lower Zone). The line above the mean line, used for writing dependent (upper) vowels (Upper zone).

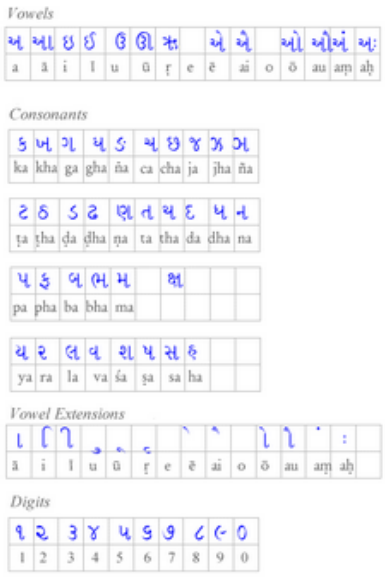


Fig.-1

Seminal and comprehensive work in Devnagari OCR is already carried out by R.M.K. Sinha and V. Bansal, [2-8]. A general Review of Statistical Pattern Recognition can also be found in [9-12]. These can be taken as good starting point to reach the recent studies in various types and applications of the Gujarati OCR problem.

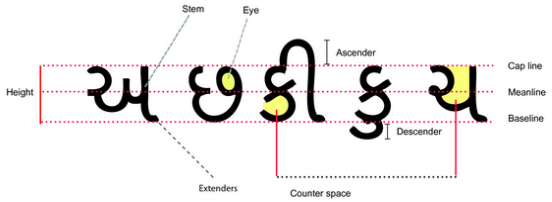
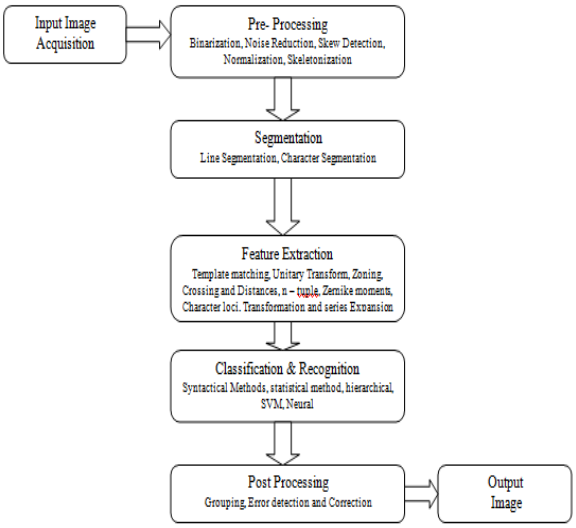


Fig. -2

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

III. ARCHITECTURAL VIEW OF GOCR



Above block diagram shows steps of Gujarati optical Character Recognition.

IV. PRE- PROCESSING

Pre-processing converts the image into a form suitable for subsequent processing and feature extraction. Data captured by optical scanning and stored in a file called pixels. These pixels may have values: OFF (0) or ON (1) for binary images, 0– 255 for gray-scale images, and 3 channels of 0–255 colour values for colour images. This further analysed to get useful information. Such processing includes the following.

A. Thresholding / Binarization

Thresholding can be used to convert captured image into binary image. There are two methods to achieve thresholding i.e. global thresholding and adaptive thresholding. In global A threshold is said to be global if the number of misclassified pixels is minimum, Histogram is bimodal (object and background), Ground truth is known OR the histograms of the object and the background are known. Problem with global thresholding is that changes in illumination across the scene may cause some parts to be brighter (in the light) and some parts darker (in shadow). With adaptive thresholding, such uneven illumination by determining thresholds locally. That is, instead of having a single global threshold, we allow the threshold itself to smoothly vary across the image.

B. Noise Elimination

Digital images are prone to a variety of types of noise. Noise is the result of errors in the image acquisition process. Due to that while obtaining pixel values from real scene that do not reflect the true intensities. There are many ways through which we can introduced noise into an image, depending on how the image is created. How the image is scanned from a photograph made on film. Noise can also be the result of damage to the film, or be introduced by the scanner itself. The distortion including local variations, rounding of corners, dilation and erosion, is also a problem. Median filter is a process that replaces the value of the pixel by the median of gray level.

C. Skew detection and Correction

While scanning the image, if paper/ source document is not aligned properly, it may course component to be tilted. There are various algorithms which can be used for skew correction. Some of the techniques [13] are as follows: projection profile technique, Linear Regression analysis, Fourier Transform based method, nearest neighbour chain, Edge based connected component approach, interline cross-correlation, Entropy based methods.

D. Size Normalization

Normalization is applied to obtain character of uniform size, slant & rotation. To be able to correct rotation, the angle of rotation must be found for rotated pages and text variant of Hough transform are commonly used for skew.

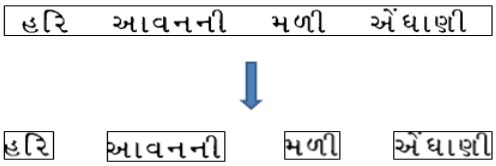
E. Skeletonization / Thinning

Skeletonization is a morphological operation that is used to remove selected foreground pixels from binary images [24]. Skeletonization extracts the shape information of the characters. Skeletonization is also called Thinning refers to the process of reducing the width of a line from many pixels to just single pixel. This process can remove irregularities in letters and in turn, makes the recognition algorithm simpler because they only have to operate on a character stroke, which is only one pixel wide. It also reduces the memory space required for storing the information about the input characters and also reduces the processing time too. The final stage in pre-processing is Skeletonization. Image Thinning extracts a skeleton of the image without loss of the topological properties [14]. The Skeletonization algorithm consists of both boundary pixel analysis and connectivity analysis.

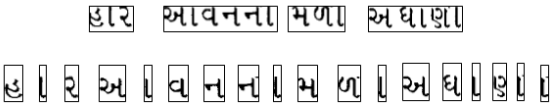
V. SEGMENTATION

The most basic step in OCR is to segment the input image into individual glyphs. This step separates out sentences from text and subsequently words and letters from sentences.

- Word Segmentation:



- Character Segmentation:



There are various different methods is used for segmentation. Like

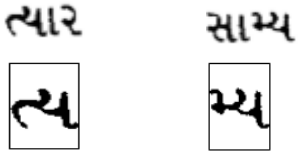
ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

- Region-growing algorithm. It starts at the first encountered text pixel and grows the character by looking for the presence of background pixels until some threshold level is reached.

- MLP and CHP [26] [27] based algorithms for joint character segmentation

e.g.



VI. Feature Extraction

After segmentation process, each character is processed through a feature extraction routine where the most descriptive features are extracted and used in training and testing. Following are the techniques for feature extraction.

A. Template Matching

Template characters are stored in the database and the entire input character is compared to every template in the database. The closest one is chosen based on some similarity measure like the mean squared distance:

$$W = \sum_{i=1}^{N} (Z(a_i, b_i) - T_j(a_i, b_i))^2$$

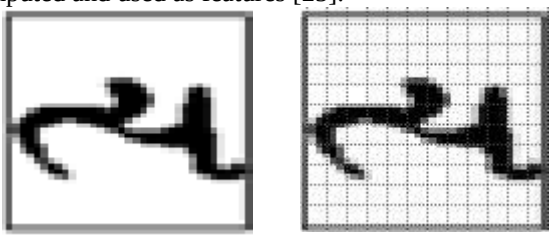
Where, Z is the input character and Tj is the jth template in the database. The limitations of this method are apparent.

B. Unitary Transforms

if A is the matrix form of the image transform, the $A^{-1} = A^{-T}$. Where A^{-} is the adjoint matrix to A and the superscripted T stands for the transpose. In most cases, these transforms can be represented by a series of orthogonal basis functions. The classic example is that of the Fourier Transform. Other transforms include the KL and Hadamard transforms.

C. Zoning

The rectangle circumscribing the character is divided into several overlapping, or none overlapping, regions and the densities of black points within these regions are computed and used as features [25].



D. Crossings and distances.

In the crossing technique features are found from the number of times the character shape is crossed by vectors along certain directions. This technique is often used by commercial systems because it can be performed at high speed and requires low complexity. When using the distance technique certain lengths along the vectors crossing the character shape are measured. For instance the length of the vectors within the boundary of the character.

E. n – tuples

The relative joint occurrence of black and white points (foreground and background) in certain specified orderings, are used as features. The word 'moment' here refers to the some of the characteristics that can be calculated from the images. There are moments of different orders that are used in pattern recognition as they are in statistics and elsewhere. The regular moments of order (p + q) for a given image of M pixels Z are given by

$$m_{pq} = \sum_{i=1} MZ(a_i, b_i)(a_i)^p\ (b_i)^q$$

F. Zernike moments

Moment descriptors have been studied for image recognition and computer vision since the 1960s. Teague [15] first introduced the use of Zernike moments to overcome the shortcomings of information redundancy present in the popular geometric moments. Zernike moments are a class of orthogonal moments which are rotation invariant and can be easily constructed to an arbitrary order. Khotanzad and Hong [16] proved that Zernike moments are effective for the optical character recognition (OCR).

G. Characteristic loci

For each point in the background of the character, vertical and horizontal vectors are generated. The numbers of times the line segments describing the character are intersected by these vectors are used as features.

H. Transformations and series expansions

These techniques help to reduce the dimensionality of the feature vector and the extracted features can be made invariant to global deformations like translation and rotation. The transformations used may be Fourier, Walsh, Haar, Hadamard, Karhunen-Loeve, Hough, principal axis transform etc.

VII. Classification & Recognition

The Extracted features are given as the input to the Classification process. A bag-of-key point extracted from the feature extraction approaches are used for classification.

A. Syntactic methods.

Measures of similarity based on relationships between structural components may be formulated by using grammatical concepts. The idea is that each class has its own grammar defining the composition of the character. A grammar may be represented as strings or trees, and the structural components extracted from an unknown character is matched against the grammars of each class.

B. Statistical methods

Statistical classifiers are rooted in the Bayes decision rule,and can be divided into parametric ones and non-

ISSN (Online) 2321 – 2004
ISSN (Print) 2321 – 5526

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN ELECTRICAL, ELECTRONICS, INSTRUMENTATION AND CONTROL ENGINEERING
Vol. 2, Issue 2, February 2014

parametric ones [17] [18]. Non-parametric methods, such as Parzen window and k-NN rule, are not practical for real-time applications since all training samples are stored and compared.

- *Non-parametric Recognition*

The finest known method of non-parametric categorization is the Nearest Neighbor (NN) and is widely used in CR. An incoming pattern is classified using the cluster, whose center is the minimum distance from the pattern over all the clusters. It does not involve a priori information about the data [19].

- *Parametric Recognition*

Since prior information is available about the characters in the training data, it is possible to obtain a parametric model for each character [20]. Once the consideration of the model, which is based on some probabilities, is obtained, the characters are classify according to some decision rules such as Baye's method or maximum Likelihood. Paper [21] presented by Manish Mangal and Manu Pratap shows that novel character recognition system. By using the virtual reconfigurable architecture-based evolvable hardware, a series of recognition systems are evolved. To improve the recognition accuracy of the proposed systems, a statistical pattern recognition-inspired methodology is introduced. The performance of the proposed method is evaluated on the recognition of characters with different levels of noise. The experimental results show that the proposed statistical pattern recognition-based scheme significantly outperforms the traditional approach in terms of character recognition accuracy. For 1-bit noise, the recognition accuracy is increased from 84.8% to 96.7%. Paper [22] presented by Sandhya Arora shows that handwritten Kannada and English Character recognition system based on spatial features is presented. Directional spatial features via stroke length, stroke density and the number of stokes are employed as potential & relevant features to characterize the handwritten Kannada numerals/vowels and English uppercase alphabets. KNN classifier is used to classify the characters based on these features with four fold cross validation. The proposed system achieves the recognition accuracy as 96.2%, 90.1% and 91.04% for handwritten Kannada numerals, vowels and English uppercase alphabets respectively.

C. Nearest Neighbour Classifier and Weighted Euclidean Distance

Nearest neighbour classifier is used on Zernike moment features with a simple weighted Euclidean distance (WED). For each test sample, the classification is based on the distance between this sample and each class. The feature vector is in a d-dimensional space and the computed mean and standard deviation feature vectors for class i are μ(i), α(i), where i = 1. . .M and M is the number of classes. For each test sample x_R d , the distance between this sample and each class is computed using the following formula

$$d^{(i)}(x) = \sum_{k=1}^{d} \left| \frac{x_k - \mu_k{}^{(i)}}{\alpha_k{}^{(i)}} \right|$$

D. Hierarchical Classification

Kanji and south-east Asian scripts have a large number of symbols. Hence, one stage discrimination does not generally suffice. In this approach, two-stage classification (coarse and fine) is used. The aim of coarse classification is to cluster similar-looking characters into groups and then perform fine classification to extract the right class.

- *City Block Distance with Deviation (CBDD)*

Let ν = (ν1,ν2,...,νn) be an n-dimensional input vector and μ = (μ1, μ2, . . ., μn) be the standard vector of a category. The CBDD is defined as

$$d_{CBDD}(v) = \sum_{j=i}^{n} max\{0, |v_j - \mu_j| - \theta\, s_j\}$$

where sj denotes the standard deviation of jth element and θ is a constant.

- *Asymmetric Mahalanobis Distance:*

For each cluster, the correct class is obtained by finding the minimum asymmetric Mahalanobis distance from the templates in that cluster. The function is given by

$$d_{AMD}(v) = \sum_{j=1}^{n} \frac{1}{\hat{\sigma}_j^2 + b} (v - \hat{\mu}\ \emptyset_j)^2$$

where b is the bias, ˆμ is the quasi-mean vector of the samples of class m, φj is the eigenvector of covariance matrix of this category, and ˆσj is the quasi-variance. In case of a tie, N-nearest neighbor is used, with N = 3.

E. Neural Network

Neural network consist, feed forward and feedback (recurrent) networks. The most common neural networks used in the OCR systems are the multilayer perception (MLP) of the feed forward networks and the Kohonen's Self Organizing Map (SOM) of the feedback networks.

F. Support Vector Machines (SVM)

Support vector machines (SVM), when applied to text classification provide high accuracy, but poor recall. One means of customizing SVMs to improve recall, is to adjust the threshold associated with an SVM. Shanahan and Roma described an automatic process for adjusting the thresholds of generic SVM [23] with better results. SVMs have achieved excellent recognition results in various pattern recognition applications.

VIII. POST - PROCESSING

Once character is recognised then the final stage is post processing in which grouping and error identification and correction processes is carried out.

A. Grouping

The process which performing association of symbols into strings, is commonly known to as grouping. The grouping of the symbols into strings is based on the symbols' location in the document. Symbols that are found to be sufficiently close are grouped together. For fixed pitch fonts the process of grouping is easy as the place of each character is known. For typeset characters the distance between characters are variable. The real problems arise for handwritten characters or when the text is skewed or slanted.

B. Error identification and Correction

After groping of Characters we have to identify errors and correct them. One of the approaches is the use of dictionaries, which has proven to be the most efficient method for error identification and correction. Given a word, in which an error may be present, the word is looked up in the dictionary. If the word is not present in the dictionary, an error has been initialized, and may be corrected by changing the word into the most similar word.

IX. CONCLUSION

In a typical OCR systems input characters are digitized by an optical scanner. In this paper we presented different techniques in each stage. First character is located and then segmented, and the resulting character image is fed in to a pre-processor for noise reduction and normalization. Certain characteristics are the extracted from the character for classification. The feature extraction is critical and many different techniques exist, each having its strengths and weaknesses. After classification the identified characters are grouped to reconstruct the original symbol strings, and context may then be applied to detect and correct errors.

REFERENCES

[1] S.Rama Mohan, Jignesh Dholakia, Atul Negi. "Zone identification in the printed Gujarati Text" *Processing of the 2005 Eight International Conference on Document Analysis & Recognition (ICDAR'05)*

[2] R.M.K. Sinha and Veena Bansal, "On Automating Trainer For Construction of Prototypes for Devnagari Text Recognition", *Technical Report TRCS*-95-232, I.I.T. Kanpur, India.

[3] R.M.K. Sinha and V. Bansal, "On Devnagari Document Processing", *Int. Conf. on Systems*, Man and Cybernetics, Vancouver, Canada, 1995.

[4] R.M.K. Sinha and Veena Bansal, "On Integrating Diverse Knowledge Sources in Optical Reading of Devnagari Script".

[5] R.M.K.Sinha., "Rule Based Contextual Post-processing for Devnagari Text Recognition", *Pattern Recognition*, Vol. 20, No. 5, pp. 475-485, 1987.

[6] R.M.K.Sinha, "On Partitioning a Dictionary for Visual Text Recognition", *Pattern Recognition*, Vol 23, No. 5, pp 497-500, 1989.

[7] Veena Bansal and R.M.K. Sinha, "On Automating Generation of Description and Recognition of Devnagari Script using Strokes", *Technical Report TRCS*-96-241, I.I.T. Kanpur, India.

[8] R. M. K. Sinha, "A Journey from Indian Scripts Processing to Indian Language Processing", *IEEE Annals of the History of Computing*, pp8-31, Jan–Mar 2009.

[9] R.G. Casey, D. R. Furgson, "Intelligent Forms Processing", *IBM System Journal*, Vol. 29, No. 3, 1990.

[10] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp- 4-37, January 2000.

[11] George Negi, "Twenty years of Document analysis in PAMI", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 1, pp- 38-62, January 2000.

[12] U. Pal, B. B. Chaudhuri, "Indian Script Character recognition: A survey", *Pattern Recognition*, vol. 37, pp. 1887-1899, 2004

[13] Praveen Krishnan, "Preprocessing Algorithms for Tamil Optical Character Recognition", *CEN, Amrita vishwavidyapeetham*, Master's Thesis, Page no. 18, 40, 45 and 48, July 2011

[14] Anil K. Jain, "Orivind Due Trier and Torfinn Taxt ―Feature Extraction Methods For Character Methods- A survey", *Pattern Recognition*, Vol. 29. No 4, PP 641-662, 1996.

[15] Teague, M.: "Image analysis via the general theory of moments". *Journal of the Optical Society of America* 70(8), 920–930 (1979)

[16] Wang Jin, Tang Bin-bin, piao Chang-hao, Lei Gai-hui "Statistical method-based evolvable character recognition system" *Key Lab. of Network control & Intell. Instrum., Chongqing Univ. of Posts & Commun., Chongqing, China.*

[17] K. Fukunaga, "Introduction to Statistical Pattern Recognition", *2nd edition, Academic Press*, 1990.

[18] Rajiv Kumar Nath, Mayuri Rastogi, "Improving Various Off-line Techniques used for Handwritten Character Recognition: a Review," *International Journal of Computer Applications* (0975 – 8887) Volume 49– No.18, July 2012.

[19] S. O. Belkasim, M. Shridhar, M. Ahmadi, "Pattern Recognition with Moment Invariants: A comparative Survey", *Pattern Recognition*, vol.24, no.12, pp.1117-1138, 1991.

[20] A.L Knoll, Experiments with "Characteristics Loci" for Recognition of Hand printed characters.

[21] Manish Mangal, Manu Pratap Singh,"Handwritten English Vowels Recognition Using Hybrid Evolutionary Feed-Forward Neural Network".

[22] Sandhya Arora et al., "Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition", *IJCSI International Journal of Computer Science* Issues, Vol. 7, Issue 3, May 2010.

[23] Vikas J Dongre Vijay H Mankar "A Review of Research on Devnagari Character Recognition" *International Journal of Computer Applications* (0975 8887) Volume 12– No.2, November 2010

[24] B.Indira, M.Shalini, M.V. Ramana Murthy, Mahaboob Sharief Shaik. "Classification and Recognition of Printed Hindi Characters Using Artificial Neural Networks" *MECS I.J. Image, Graphics and Signal Processing, 2012*, 6, 15-21

[25] Vijay Laxmi Sahu, Babita Kubde. "Offline Handwritten Character Recognition Techniques using Neural Network: A Review" *International Journal of Science and Research* (IJSR), India Online ISSN: 2319-7064 Volume 2 Issue 1, January 2013

[26] Bhattacharya, U., Chaudhuri, B.B., Parui, S.K. "An MLP-based texture segmentation technique which does not require a feature set." *Proceedings of the 13th International Conference on Pattern Recognition, 1996., (Volume:2) Print ISBN: 0-8186-7282-X*

[27] Patel C. & Desai A. , "Segmentation of Text Lines into Words for Gujarati Handwritten Text", *International Conference on Signal & Image processing, 130–134.*