RESEARCH ARTICLE OPEN ACCESS

Gujarati Ocr: Compound Character Recognition Using Zernike Moment Feature Extractor

Honey Patel
Computer Science & Engineering Department, ITM UNIVERSE, and Vadodara

ABSTRACT
Today most of people are e-reader. Gujarati documents are mostly in hardcopy; very less e-material is available. For converting those hardcopies into editable text form we need digitization. Optical Character Recognition is a technique for converting scanned documents into digitized text. So also for Gujarati language, optical character recognition system is required. For the recognition of printed characters of Indian languages like Gujarati, Marathi and many more; optical character recognition is required. Efforts of many researchers are going on for developing efficient OCR systems for Indian languages, especially for Gujarati, a popular language state Gujarat. There is very less material available for Gujarati script. Devnagari language is the base of Gujarati language. Only difference is shirorekha. Since 1999, efforts are going on for Gujarati character recognition but still that much accuracy cannot be obtained as compared to other Indic languages.
Keywords —Gujarati Character Recognition, Pre-processing, Segmentation, Feature Extraction, OCR

I. INTRODUCTION

Now a day, everything is going in the form of digital. Most of users are becoming e-users then all documents which are going to delivered them should be in digitized format. In recent years, the huge usage of internet had turned people to more usage to electronic documents compared to physical documents. The creation of electronic documents facilitate to easy communication and storage of documents. Thus, developing computer algorithms to identify the characters in the text is the principal task of OCR. Optical character recognition belongs to the family of techniques performing automatic identification.

The overwhelming volume of paper-based data in corporations and offices challenges their ability to manage documents and records. Computers, working faster and more efficiently than human operators, can be used to perform many of the tasks required for efficient document and content management. Computers understand alphanumeric characters as ASCII code typed on a keyboard where each character or letter represents a recognizable code [1].

However, computers cannot distinguish characters and words from scanned images of paper documents. Therefore, where alphanumeric information must be retrieved from scanned images such as commercial or government documents, tax returns, passport applications and .credit card applications, characters must first be converted to their ASCII equivalents before they can be recognized as readable text. Optical character recognition system (OCR) allows us to convert a document into electronic text, which we can edit and search etc. OCR is the machine replication of human reading and has been the subject of intensive research for more than five decades [1].

A. Introduction to Gujarati Script

Gujarati script is derived from Devnagari script. It has very rich set of consonants and vowels. It has 34 consonants as shown in Fig. 1 and 12 vowels as in Table I.
By combining consonants and vowels, compound characters are created. Compound characters are also known as conjuncts. Conjuncts are shown in Table II. These conjuncts make Gujarati script more complex compared to other Latin language. Gujarati script is written from left to right. The basic difference between Devnagari script and Gujarati script is absence on shirorekha [2].

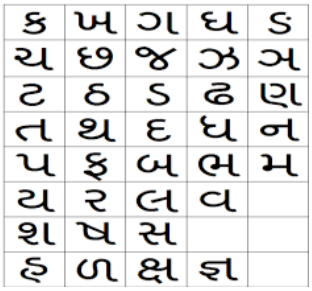

Fig. 1 Gujarati Consonants

TABLE I
GUJARATI VOWELS

અ	આ	ઇ	ઈ	ઉ	ઊ	એ	ઐ	ઓ	ઔ	અં	અઃ

TABLE III
EXAMPLE OF GUJARATI CONJUNCTS

વ્ય	ત્ય	જ્ય	ર્ય
બ્હ	ભ્ય	ચ્હ	ફ્ર
દ્ધ	હ્ય	ન્ન	શ્વ

B. Flow of Optical Character Recognition

Fig. 2 shows the flow of character recognition system [1].

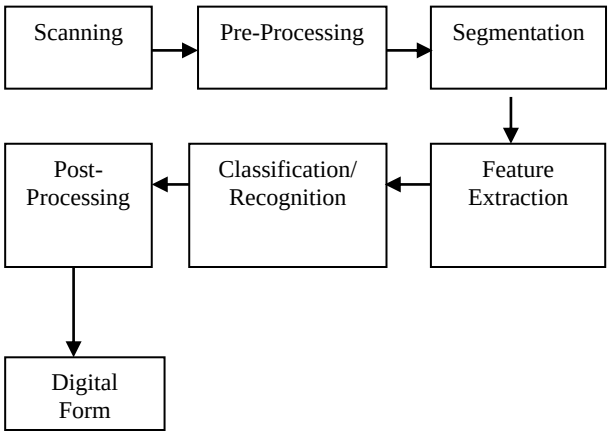

Fig. 2 Flow of OCR

II. PROPOSED APPROACH

A. Pre-Processing

Pre-Processing includes Binarization, Noise removal, Skew detection and correction. We have assumed that our scanned images are skew-free.

1) **Binarization:** Binarization converts the gray scale image into binary image. From the reported work [11], there are mainly two approaches for converting image into binary image. Binarization separates the foreground and background information. This separation of text from background is a prerequisite to subsequent operations such as segmentation.

- Global algorithm
- Local Algorithm

Examples of global algorithm are Otsu's method, Histogram peaks, and K-mean clustering based method. Examples of local algorithm are Niblack's method, Sauvola's method,Bernsen's Method. The most commonly used method is Otsu's method. We have used Otsu's method.

2) **Noise removal:** Due to printer, scanner, print quality, age of the document, the documents which are scanned may have noise. So, it is pre-requisite to filter this noise before processing the image. Commonly used approach is to process the image through a low-pass filter and use it for later processing. Median filter is best suitable for salt and pepper kind of noise. Maximum filter is used for removing negative outlier noise. Minimum filter is used for removing dark values in image [12]. We have used median filter for removing noise.

3) **Segmentation:** An image which is passed from the pre-processing phase is ready for the segmentation process. Commonly used segmentation algorithms are Projection profile, Run-length smearing, Hough transforms [15]. Segmentation contains three steps:

- Line Segmentation-Projection and Hough-based methods are suitable for clearly separated lines [15]. Projection-based methods can cope with few overlapping or touching components. Projection-profiles are commonly used for printed document segmentation. The global horizontal projection is used to compute the sum of pixels in each row and construct corresponding histogram Find the rows which are not contain white pixel using histogram. Based on the peak/valley points of the histogram, individual lines are separated. As shown in Fig-8, each binarized line is separated.
- Word Segmentation- For separate words from segmented line, line is scanned vertically. Number of black pixels in each column is calculated to construct column histogram. Vertical histogram is used for word segmentation [16]. Using the vertical Histogram, find the points from which the word starts and ends. Fig-11 shows the words segmented from Fig-8.
- Character Segmentation- For segmenting characters, vertical histogram of segmented word is created. Using the histogram, find the points from which the character starts and ends [16]. In a word, spaces between characters are the separators between the characters. ગ, ધ, ક, ત are the major cause for over segmentation resulting into inaccuracy. It is thus necessary to combine the over segmented characters at letter stage to form a single character for recognition as shown in Fig-13. Solution for this problem is as below:
  - Do character separation
  - Find the width of current character and previous character
  - If it's 3/4th of the previous character width then we merge both the character and it became one as character.

4) **Feature Extraction:** In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. Feature extraction methods are based on 2 types of features [17]:

- Global or Statistical
- Structural or topological features

5) **Feature Extraction using Zernike Moments:** Zernike moments are a class of orthogonal moments. The orthogonal property of Zernike polynomials enables the contribution of each moment to be unique and independent of information in an image. Fig. 3 shows the complete flow of calculation of Zernike moments [19].

The translation normalization is achieved by moving the image center to the image centroid. The scale normalization is achieved by set the image's 0th order regular moment to a predetermined value. To compute the Zernike moments of a digital image, the range of the image should be mapped to the unit circle first with its origin at the image's center. The pixels falling outside the unit circle are discarded in the computation process. Zernike polynomials take the unit disk x2 + y2 =1 as their computation domain [20].

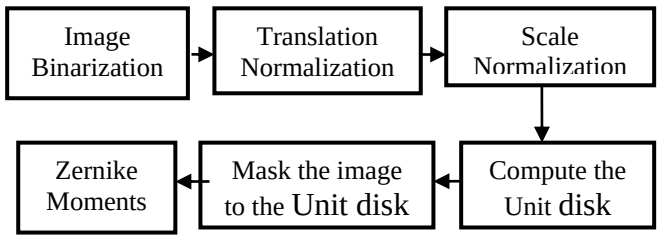

Fig. 3 Block diagram of computing Zernike moments

In my implementation of Zernike moments, I use binary images with spatial resolution of 128×128. All of these binary images are normalized into a unit circle with fixed radius of 64 pixels. Following steps are necessary to extract features of any character image using Zernike moments.

Step 1: First of all converts gray-scale image into the binary numeral image

Step 2: To map over a unit disc image be convert into polar coordinate

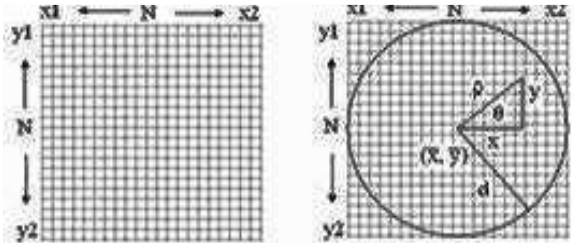

Fig. 4 (a)NxN pixels (b)Unit Circle Mapped image bitmap onto NxN pixel size image

In the above Fig. 4(a) the center of the image and disk is same. Where x1x2 are X-axis dimensions and y1y2 is Y-axis dimensions of the pixel rectangle. $(\bar{x},\bar{y})$ is the center of the unit disk, ρ is polar value and θ is polar angle. Now the image is mapped into polar co-ordinates and onto unit circle as:

Compute the distance d in Fig. 4(b) above as

$$d = \sqrt{(x2-\bar{x})^2\ \tfrac{1}{2} + (y2-\bar{y})^2\ \tfrac{1}{2}} \quad (1)$$

Compute the distance vector ρ and angle θ for any (x,y) pixel in $f(x,y)$ polar coordinates as

$$\rho = \sqrt{(x-\bar{x})\ + (y-\bar{y})/d} \quad (2)$$

$$\theta = \tan^{-1}\left[\frac{x-\bar{x}}{y-\bar{y}}\right] \quad (3)$$

This step maps pixel coordinate (x1, x2) to (-1, +1) and (y1, y2) to (-1, +1) in polar. In this way almost all the pixels in image bound box as above are inside unit circle except some foreground pixels.

Step 3: Fruits Zernike introduced a set of complex polynomials $\{V_{nm}(x, y)\}$ which form a complete orthogonal set over the unit disk of x2 + y2 <=1 in polar coordinates. The form of the polynomials is:

$$V_{nm}(x,y) = V_{nm}(\rho,\theta) = R_{nm}(\rho) = e^{jm\theta} \quad (4)$$

Where $j=\sqrt{-1}$, $\theta=\tan^{-1} y/x$, ρ is the length of the vector from the origin to the pixel (x, y); θ is the angle between the vector ρ and x axis in counter clockwise direction. $R_{nm}(\rho)$ is Radial polynomial defined as [20]:

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2}(-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!}\rho^{n-2s} \quad (5)$$

Where n>=0, |m|<=n, n-|m|=even. When the image is mapped onto unit disc, take desired value of order of moment, i.e. n and compute real and imaginary parts of the Zernike moment using Radial polynomials.

Step 4: Then compute Zernike moment of order n and repetition m for function $f(x,y)$ is defined as:

$$A_{nm}=\frac{n+1}{\pi}\sum_x\sum_y f(x,y)V^*_{nm}(x,y), x^2+y^2 \le 1 \quad (6)$$

Where n>=0, |m|<=n and * is the complex conjugate operator [21].

6) **Classification using K-NN classifier:** The classification stage is the decision making part of a recognition system and it uses the features extracted in the feature extraction stage. Classification stage in an OCR process assigns labels to character images based on the features extracted and the relationships among the features. K-NN is an instance based classification algorithm. To classify a new character, the system finds the k -nearest neighbors among the training datasets, and uses the categories of the k- nearest neighbors to weight the category candidates. Euclidean distance is considered as distance function between train data and test data. Euclidean distance between two n dimensional vectors a and b can be calculated as [21]:

$$\Delta_{Euclidean}(x,y) = \sqrt{(\sum_{i=1\ldots n}(a_i-b_i)^2)} \quad (7)$$

III. EXPERIMENTAL RESULTS

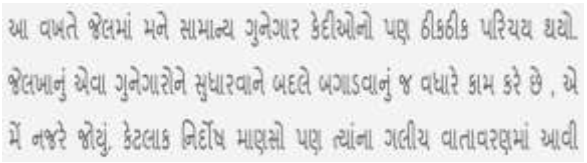
આ વખતે જેલમાં મને સામાન્ય ગુનેગાર કેદીઓનો પણ ઠીકઠીક પરિચય થયો.
જેલખાનું એવા ગુનેગારોને સુધારવાને બદલે બગાડવાનું જ વધારે કામ કરે છે , એ
મેં નજરે જોયું. કેટલાક નિર્દોષ માણસો પણ ત્યાંના ગલીચ વાતાવરણમાં આવી

Fig. 5 Original Image

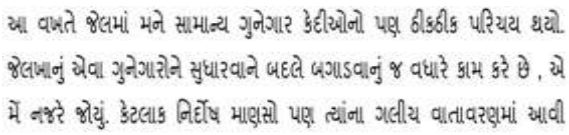

Fig. 6 Gray-Scale Image

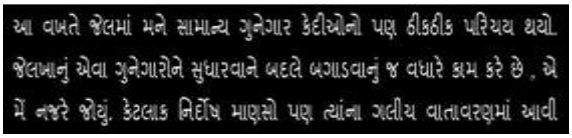

Fig. 7 Binary Image

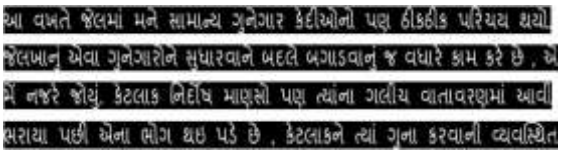

Fig. 8 Segmented Lines

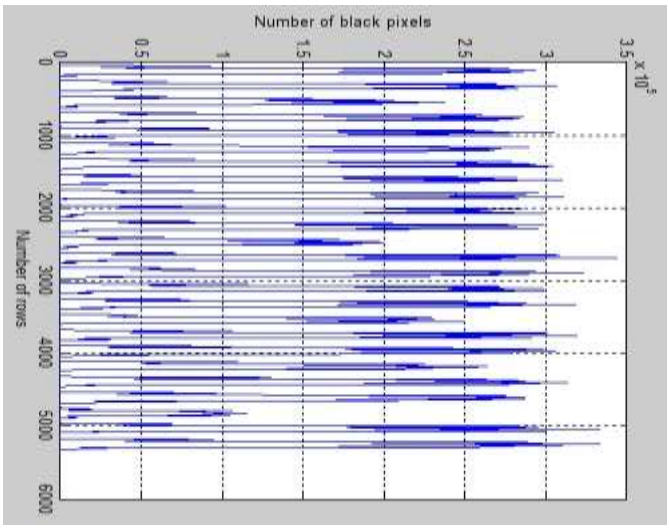

Fig. 9 Horizontal histogram for line segmentation

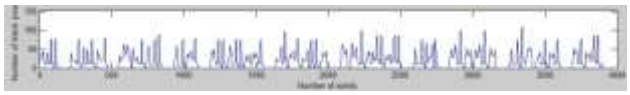
Fig. 10 Vertical histogram for word segmentation


Fig. 11 Segmented Words

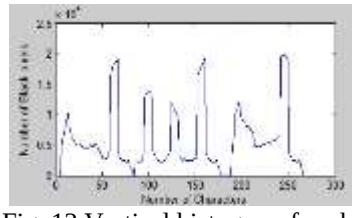
Fig. 12 Vertical histogram for character segmentation

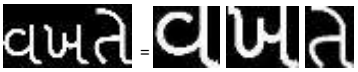
Fig. 13 Segmented Characters

TABLE III
RESULT OF LINE SEGMENTATION

Name	Total No. of Lines	Correctly segment -ed Lines	Incorrectly Segmen -ted Lines	Accu racy (%)
gocr-1	25	25	0	100%
gocr-2	24	24	0	100%
gocr-3	25	25	0	100%
gocr-4	25	25	0	100%
gocr-5	25	25	0	100%
gocr-6	25	25	0	100%
gocr-7	25	25	0	100%
gocr-8	24	24	0	100%
gocr-9	24	24	0	100%
gocr-10	25	25	0	100%
			Average Accuracy	**100%**

TABLE IV
RESULT OF CHARACTER SEGMENTATION

Names	Total No. of chara -cters	CorrectlSegment-ed Characters	Incorrect Segmented Characters	Accu Racy (%)
gocr-1	758	694	64	91.55
gocr-2	709	631	78	88.99
gocr-3	720	645	75	89.58
gocr-4	727	659	68	90.64
gocr-5	648	602	46	92.90
gocr-6	766	694	72	90.60
gocr-7	751	686	65	91.34
gocr-8	761	698	63	91.72
gocr-9	715	655	60	91.60
gocr-10	751	674	77	89.74
			Average Accuracy	**90.866**

TABLE V
CONJUNCTS USED FOR CLASSIFICATION

ચ્ચ	દ્ધ	ધ્ય	ગ્ન	ગ્ર	હ્ન	હ્ય	ક્સ	ક્ટ	ક્ત
લ્પ	લ્ય	મ્પ	ન્ડ	ન્મ	ન્ન	ણ્ય	ન્ય	પ્ર	ર્ય
ર્ડ	ર્ગ	ર્મ	ર્ણ	ર્શ	ર્ષ	ર્ટ	ર્ત	ર્થ	ર્વ
ર્ય	શ્ચ	શ્ર	ષ્ટ	શ્ચ	સ્મ	સ્પ	સ્ત	સ્થ	સ્વ
થ્થ	થ્ય	ત્ર	ત્સ	ત્ય	વ્ર	વ્ય	ક્ષ	દ્ર	ન્શ

Fig. 14 Original image

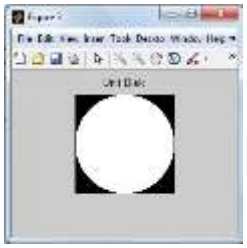

Fig. 15 Unit disk

Fig. 16 Mapping of image Reconstruction On Unit Disk

Fig. 17 of Character

From the 50 conjunct characters 2 characters are misclassified. The characters which are misclassified are ધ્ય and ર્ષ. From a basic consonant, a misclassified basic consonant is ઘ. It is classified as ય. As ઘ and ય are very similar. There is a chance of misclassification. In case of ધ્ય same reason is there. For ર્ષ. the character in train set is cropped at one end so its feature doesn't match with test set. The class label which is assigned to classified character contains combination of Unicode.

TABLE VI
UNICODE FOR CONSONANTS

x=	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+ 0A8x		ઁ	ં	ઃ		અ	આ	ઇ	ઈ	ઉ	ઊ	ઋ	ઌ	એ		ઐ
U+ 0A9x	ઑ	ઓ		ઔ	ક	ખ	ગ	ઘ	ઙ	ચ	છ	જ	ઝ	ઞ	ટ	
U+ 0AAx	ઠ	ડ	ઢ	ણ	ત	થ	દ	ધ	ન		પ	ફ s	બ	ભ	મ	ય
U+ 0ABx	ર		લ	ળ		વ	શ	ષ	સ	હ			઼	ઽ	ા	િ
U+ 0ACx	ી	ુ	ૂ	ૃ	ૄ	ૅ		ે	ૈ	ો		ો	ૌ	્		
U+ 0ADx	ૐ															
U+ 0AEx	ૠ	ૡ	ૢ	ૣ			૦	૧	૨	૩	૪	૫	૬	૭	૮	૯
U+ 0AFx		ૹ														

Fig. 18 Character in editable text

IV. CONCLUSIONS

From this experimental results, I have concluded that feature extraction of Gujarati language can be done using Zernike moments.

ACKNOWLEDGMENTS

I would like to thank my institute ITM UNIVERSE, Vadodara for giving me this opportunity to work in the great environment.

I would also like to thank my husband, parents, my friends and well wishers who helped me to come out of woods whenever any difficulty was encountered. It was only due to their support, motivation and encouragement that I could steer through the project on an honest course to the splendour of success.

REFERENCES

[1] Sameer Antani and Lalitha Agnihotri, “Gujarati Character Recognition, Document Analysis and Recognition" in Proc. ICDAR '99 1999

[2] Jignesh Dholakia, S. Rama Mohan And Atul Negi ,“Zone Identification in the Printed Gujarati Text”, Document Analysis and Recognition, 2005. pp272 - 276 Vol. 1

[3] Jignesh Dholakia, Archit Yajnik and Atul Negi, “Wavelet Feature Based Confusion Character Sets for Gujarati Script”, International Conference on Computational Intelligence and Multimedia Applications ,2007

[4] Jayashree R.Prasad, Dr. U.V.Kulkarni, Rajesh S. Prasad,“Template matching algorithm for Gujarati Character Recognition”, Second International Conference on Emerging Trends in Engineering and Technology,2009

[5] Mrs. Chhaya Patel, Mr. Apurva Desai, “Segmentation of Text Lines into Words for Gujarati Handwritten Text”, Second International Conference on Emerging Trends in Engineering and Technology,2010

[6] Mandar Chaudhary, Gitam Shikkenawis, Suman K. Mitra, Mukesh Goswami, “Similar looking Gujarati printed character recognition using Locality Preserving Projection and Artificial Neural Networks”, Third International Conference on Emerging Applications of Information Technology (EAIT),2012

[7] Jayashree Rajesh Prasad, Uday Kulkarni, "Gujarati character recognition using adaptive neuro fuzzy classifier with fuzzy hedges", Springer,2014

[8] Avani R. Vasant, Sandeep R. Vasant, Dr. G.R.Kulkarni, "Performance Evaluation of Different Image Sizes for Recognizing Offline Handwritten Gujarati Digits using Neural Network Approach", International Conference on Communication Systems and Network Technologies,2012

[9] Himanshukumar U. Prajapati, Prof. S. Ramamohan, Ms Sonam Chauhan "Removal of Graphics from Text-document and Segmentation of Gujarati Documents Using Connected Component Theory", IOSR Journal of Engineering,2012

[10] K.V.Kale, S.V.Chavan, M.M.Kazi, Y.S.Rode, "Handwritten Devanagari Compound Character Recognition using Legendre Moment an Artificial Neural Network Approach", International Symposium on Computational and Business Intelligence,2013

[11] Ntogas, Nikolaos,Ventzas, Dimitrios,"A BINARIZATION ALGORITHM FOR HISTORICAL MANUSCRIPTS",12th WSEAS International Conference on COMMUNICATIONS, Heraklion, Greece, July 23-25, 2008

[12] Yasser Alginahi, "Preprocessing Techniques in Character Recognition"

[13] Deepak Kumar, Dalwinder Singh, "A Review of Scanned Document Skew Detection and Correction Techniques", IJCST Vol. 3, Issue 2, April - June 2012

[14] Rodolfo P. dos Santos, Gabriela S. Clemente, Tsang Ing Ren and George D.C. Calvalcanti,"Text Line Segmentation Based on Morphology and Histogram Projection",10th International Conference on Document Analysis and Recognition,2009

[15] Laurence Likforman-Sulem, Abderrazak Zahour, Bruno Taconet, "Text Line Segmentation of Historical Documents: a Survey", International Journal on Document Analysis and Recognition, Springer, 2006

[16] Vijay Kumar, Pankaj K. Sengar, "Segmentation of Printed Text in Devanagari Script and Gurmukhi Script", International Journal of Computer Applications (0975 – 8887) Volume 3 – No.8, June 2010

[17] Rohit Verma , Dr Jahid Ali, " A-Survey of Feature Extraction and Classification Techniques in OCR Systems ",International Journal of Computer Applications & Information Technology Vol. I, Issue III, November 2012

[18]Youssef Bassil, Mohammad Alwani ," OCR Post-Processing Error Correction Algorithm Using Google's Online Spelling Suggestion ",Journal of Emerging Trends in Computing and Information Sciences ,VOL. 3, NO. 1, January 2012

[19] P. Bhaskara Rao,D.Vara Prasad,Ch.Pavan Kumar,"Feature Extraction Using Zernike Moments",International Journal of Latest Trends in Engineering and Technology

[20] "Classification of Printed Gujarati Characters using SOM based K-Nearest Neighbor Classifier"

[21] Chhaya Patel,Apurva Desai, "Gujarati Handwritten Character Recognition Using Hybrid Method Based OnBinary Tree-Classifier And K-Nearest Neighbour",International Journal of Engineering Research & Technology,Vol. 2 Issue 6, June – 2013

[22] Apurva A. Desai, "Segmentation of Characters from old Typewritten Documents using Radon Transform", International Journal of Computer Applications (0975 – 8887) Volume 37– No.9,January 2012