Structural Feature Extraction to recognize some of the Offline Isolated Handwritten Gujarati Characters using Decision Tree Classifier

Hetal R. Thaker
Atmiya Institute of Technology & science,
Kalawad Road,
Research Scholar, Saurashtra University,
Rajkot – Gujarat, INDIA

C. K. Kumbharana, PhD
Head - Dept. of Computer Science,
Saurashtra University,
Rajkot – Gujarat, INDIA

ABSTRACT

Large amount of information is prevailing on paper and in an era of digital technology it requires it to store this information in electronic format. Using scanner this information can be digitized. Later any modification in terms of add, editing, removing and searching to it requires a technique or methodology which will identify text from image and convert into ASCII or Unicode. This paper presents recognition of offline handwritten character for Gujarati script using structural features. For proposed experimental work five characters of Gujarati script are considered**.** Decision tree classifier is proposed for classification. Various phases of character recognition are implemented such as collecting handwritten input, digitization, preprocessing, extracting structural features and recognition. By proposed work authors are able to achieve 88.78% of average success rate for defined five characters.

General Terms

Pattern Recognition, Character Recognition.

Keywords

Offline handwritten character recognition, Gujarati handwritten Character recognition, Structural feature extraction.

1. INTRODUCTION

Use of Pen and paper to store information is convenient media for many people even in an era of digital technology. Some historical documents are also available on paper. There are two ways to convert these papers and documents into digital format. One is manually keying which will be a time consuming process. Another better approach will be to scan the document with device called scanner which converts and saves document into image format. It requires some special application to carry out further operations which will identify character from image and will convert it into Unicode or ASCII format so as operations like adding, deleting and updating text, searching text from image can be carried out. Recognizing handwritten text from image is called offline handwritten character recognition and is a challenging area of research due to diversities involved in the field. There are two perspective for diversities (1) writer wise (2) Language script wise. Writing pattern will vary from writer to writer, pen and paper used for writing. For language various characteristics that includes no. of alphabets, cursive or non-cursive, alphabet modifiers, direction, curves etc. From research carried out so far it's observed that character recognition technology has matured for English and Arabic script. Proposed paper focuses on recognition of handwritten characters written in Gujarati script. Characters will be identified based on structural features. This paper is organized into different sections as previous work, characteristic of Gujarati script, and various stages of character recognition process such as dataset, preprocessing, feature extraction and character identification, result analysis.

2. PREVIOUS WORK

Extensive work is observed for recognition for English and Arabic script. Work can also be traced for south Indian script: Telugu, Tamil, Kannada etc. Few research work can be traced for recognizing Gujarati characters, yet it is an open area of research to achieve higher accuracy.

Chhaya Patel et.al. [1] have proposed character recognition for Gujarati script. Binary Tree classifier and K-nearest neighbor is used for classification. 200 handwritten sample characters for each characters were digitized at 200-300 dpi to store in e-format. Preprocessing where contrast adjustment, intensity adjustment, 2D adaptive wiener filter to remove noise, crop boundary, binarization and normalization was proposed. Based on structural features such as vertical line, No of objects in character, no. of objects in upper half , lower-half, right-half and left-half various groups are proposed. Authors have stated 63.1% of overall accuracy for character recognition.

Jayshree R. Prasad et. al. [2]have proposed pattern matching approach for offline handwritten character of Gujarati script. Character sets are divided into 6 sets. Overall recognition efficiency for proposed task cited is 71.66%.

Lipi shah et. al. [3]have proposed Radial Histogram approach for recognizing handwritten character of Gujarati script. Structural characteristics were identified for feature extraction. For this task 72 directions at 5 degree interval radial histogram calculation made to obtain 72 unique feature vector. Euclidean distance classifier was used for classification.

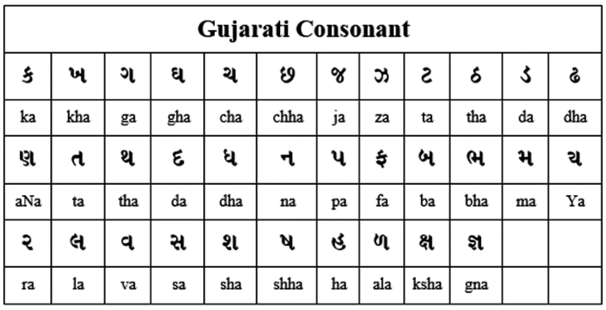
Kamal Moro et.al. [4] has reported that there is no standard database available for Gujarati and hence developed a database collecting handwritten characters from large number of writers and scanned at 300 dpi and have binarized and skeletonized images. For feature extraction horizontal and vertical and two diagonal profiles used and classified using neural network in a task of recognizing Gujarati handwritten numerical optical character.

Prasad J. R. et. Al. [5] have proposed a preprocessing approach in which they have used median filter to remove salt and pepper noise from the scanned images stored in png file format and have applied thinning to reduce character to minimum one pixel thickness, template matching for Gujarati character recognition. Various steps for template matching involves classification of templates, correlation analysis and calculating cross correlation coefficient which is repeated for every position and values were saved. Average overall recognition rate of 71.66 % is reported in an attempt. [8].

3. CHARACTERISTICS OF GUJARATI SCRIPT

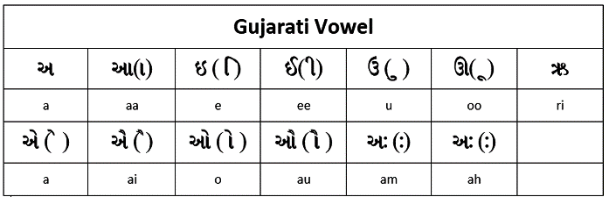
India is a versatile country where so many regional languages are spoken across various part of India. In India more than 20 official languages are there, Bengali, Malayalam, Hindi, English, Guajarati, Tamil, Kannada, Urdu etc [6]. Gujarati language is very popular language and it is an official language of the Gujarat State of India. More than 50 million people speak Gujarati language [7].

Gujarati-script used to write the Gujarati language. The Gujarati alphabet utilizes overall 75 distinct legitimate and recognized shapes, which mainly includes 59 Characters and 16 diacritics. Fifty-nine characters are divided into 36 consonants (34 Singular and 2 Compound (not lexically though)) means ornamented sounds, 13 vowels (pure sounds), and 10 numerical digits [8] [9] [10]. Sixteen diacritics are divided into 13 vowel and 3 other characters.



Gujarati Consonant											
ક	ખ	ગ	ઘ	ચ	છ	જ	ઝ	ટ	ઠ	ડ	ઢ
ka	kha	ga	gha	cha	chha	ja	za	ta	tha	da	dha
ણ	ત	થ	દ	ધ	ન	પ	ફ	બ	ભ	મ	ચ
aNa	ta	tha	da	dha	na	pa	fa	ba	bha	ma	Ya
ર	લ	વ	સ	શ	ષ	હ	ળ	ક્ષ	જ્ઞ		
ra	la	va	sa	sha	shha	ha	ala	ksha	gna		

Fig 1 : Gujarati Consonants



Gujarati Vowel						
અ	આ(ા)	ઇ (િ)	ઈ(ી)	ઉ (ુ)	ઊ(ૂ)	ઋ
a	aa	e	ee	u	oo	ri
એ (ે)	ઐ (ૈ)	ઓ (ો)	ઔ (ૌ)	અઃ (:)	અઃ (:)	
a	ai	o	au	am	ah	

Fig 2 : Gujarati Vowels

Five characters of Gujarati script ‘aNa’, ‘Ga’, ‘Sha’, ‘La’ and ‘Ja’ are considered for proposed experimental work.

4. DATASET OF HANDWRITTEN CHARACTERS

For proposed work sample handwritten characters were collected from various writers on a predefined datasheet. 150 samples of each five characters were collected. Digitization is carried out using Brother DCP – 7030 scanner at a resolution of 300 dpi. To extract isolated character from image program proposed in [11] is used.

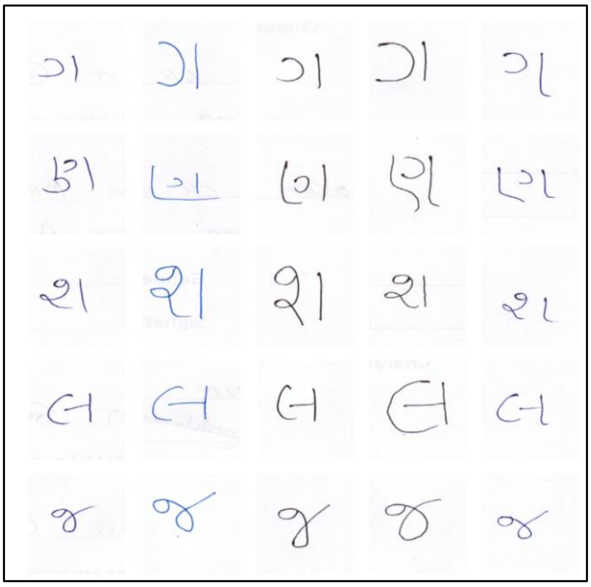


Fig 3 : Sample Handwritten Gujarati Characters

5. PREPROCESSING

RGB image Fig 2(a) is converted to Grayscale Fig 2(b) where hue and saturation is eliminated and luminance is retained. Contrast limited adaptive histogram equalization is used to adjust contrast of the image. 2-D Median filtering is used to remove salt and pepper noise Fig 2(c) which operates on small region rather than entire image. To convert grayscale image to binary image Fig 2(d) threshold is determined based on Otsu’s method which selects threshold in a way to minimize intra-class variance of black and white pixel. Morphological close operation is used to complete line strokes in image. To remove small objects and spurious pixels morphological open operation and spur is performed respectively. Boundary is clipped to acquire region of interest, Fig 2(e). On output image thinning is applied Fig 2(f) to reduce lines to single pixel. Boundaries are defined to obtain region of interest from image.

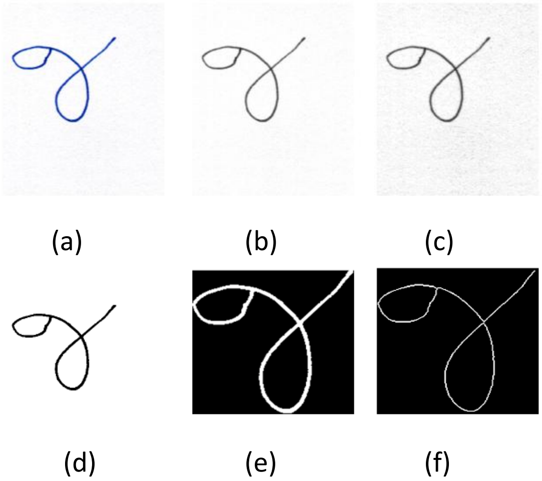


Fig 4 : Preprocessing steps on Gujarati Character 'ja'

6. FEATURE SELECTION AND EXTRACTION

Feature extraction is the process of collecting distinguishable information of an object or a group of objects so that on the basis of this information we can classify objects with different features. I.S. Oh [16] has defined that feature extraction and selection is a process of extracting the most representative information from the raw data.

Proposed methods uses structural features to uniquely classify character. Structural features have many advantages such as it can distinguish character irrespective of its size and writer's style. Structural feature describes how character is made up using basic or complex geometrical structure and shapes. These features can be simple such as vertical, horizontal and diagonal line or it can be complex such as curvature. For this paper three features are examined i.e. Connected or disconnected component, Number of End Point and Number of close loop.

6.1 Connected and Disconnected Components

How many separate elements combines together to make a character. If character comprises of single element implies character is connected.

→

1	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	1	0
1	0	0	0	0	1	0	1	0
1	0	0	0	0	1	0	1	0
1	0	0	0	0	1	0	1	0
0	1	0	0	1	0	0	1	0
0	0	1	0	0	0	0	1	0
0	0	0	1	0	0	0	0	1
0	0	0	0	1	0	0	0	1

Fig 5 : Sample handwritten character 'na' having three components and its binary representation

Next examination required is to find out how many separate elements exist in a character incase character is disconnected.

1	0	0	0	0	0	0	3	0
1	0	0	0	2	0	0	3	0
1	0	0	0	0	2	0	3	0
1	0	0	0	0	2	0	3	0
1	0	0	0	0	2	0	3	0
0	1	0	0	2	0	0	3	0
0	0	1	0	0	0	0	3	0
0	0	0	1	0	0	0	0	3
0	0	0	0	1	0	0	0	3

Fig 6 : Labeling separate component of character

To find out whether character is connected or not, connected components are labeled as shown in fig (6). If label number exceeds 1 which indicates that character is connected else it is disconnected. In case character is disconnected then last digit used to give label indicates number of separate components within that character. Here in example Gujarati Character 'aNa' is a disconnected component having 3 separate sub components.

6.2 Number of End Point

End point defines beginning and ending mark of structure elements. Every pixels of thinned binary image are examined for value of its eight neighbor pixels as shown in fig (7). To find out whether current pixel is end point its eight neighbors are examined for their on status and if out of eight only one neighboring pixel is on then current point is defined as end point. For Gujarati character 'aNa' such examination is demonstrated in fig (8) which has six end points.

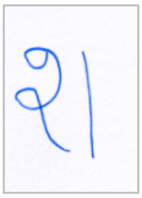
i-1, j-1	i-1, j	i-1 , j+1
i, j-1	i,j	i, j+1
i+1, j-1	i+1, j	i+1, j+1

Fig 7 : Eight neighbors of pixel i,j

1	0	0	0	0	0	0	1	0
2	0	0	0	1	0	0	2	0
2	0	0	0	0	2	0	2	0
2	0	0	0	0	2	0	2	0
2	0	0	0	0	2	0	2	0
0	2	0	0	1	0	0	2	0
0	0	2	0	0	0	0	2	0
0	0	0	2	0	0	0	0	2
0	0	0	0	1	0	0	0	1

Figure 8 End points of Character 'na'

6.3 Number of Close Loop



→

0	0	1	0	0	0	0	0	1
0	1	0	1	0	0	0	0	1
0	1	1	1	0	0	0	0	1
0	0	0	1	0	0	0	0	1
0	0	1	1	0	0	0	0	1
0	1	0	1	0	0	0	0	1
0	1	1	0	1	0	0	0	1
0	0	0	0	0	1	0	0	1
0	0	0	0	0	0	1	0	1

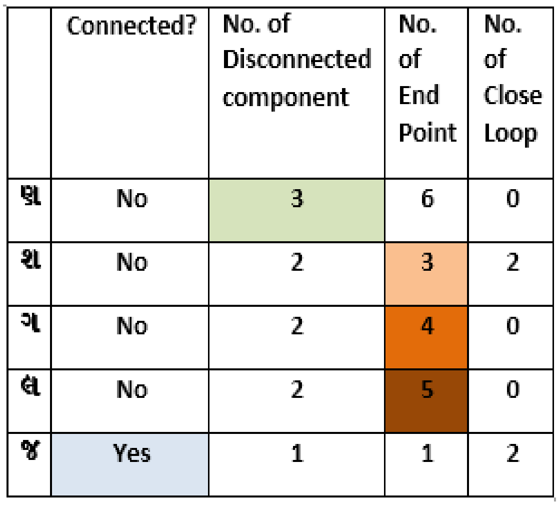
Fig 9 : Gujarati Character 'sa' having two close loops and its binary representaiton

Close loop defines close region, as represented in fig (9), Gujarati character 'Sha' is having two close region that is it contains two loops.

6.4 Feature Table

Based on structure analysis for Gujarati characters ‘aNa’, ‘Sha’, ‘Ga’, ‘La’, ‘Ja’ following decision table is derived. (Table 1). Tracing column by column of table unique cell is highlighted with separate color, which helps in identifying character uniquely. While extracting features from given input image if it is connected then based on following table it can be concluded that it will be ‘Ja’. If character is disconnected having components three will separate ‘aNa’ from rest three characters. Further no. of disconnected components are same for ‘Ga’, ‘La’ and ‘Ja’ but they can be differentiated based on another feature i.e. no. of end points as shown in third column of below table.

Table 1. Structural Feature Table for Gujarati Characters



	Connected?	No. of Disconnected component	No. of End Point	No. of Close Loop
ણ	No	3	6	0
શ	No	2	3	2
ગ	No	2	4	0
લ	No	2	5	0
જ	Yes	1	1	2

7. DECISION TREE CLASSIFIER

Decision tree classifier is one of the classification method used for recognizing character. Proposed tree is designed on the basis of structural features extracted and table presented above for five characters as presented in the fig (10). Five Gujarati characters forms a base of tree and that defines root node. Leaf node of tree are individual characters. This tree can also be defined as decision tree as at every step based on answer of key question set of possibility reduces. Progressing this way one can reach to leaf node which yields individual characters. Another alternative approach can be used is decision table where it requires to extract all the mentioned feature and based on it character will be recognized. Comparing two approaches i.e. tree and table for classification of character tree yields faster result in a case when distance of that character to root node is less. In another words time and effort required for identifying character depends on how far character is from root node. Decision table approach is better if character needs to be recognized is farthest from root as every time decision requires and based on answer alternate branch will be followed if feature is not extracted properly due to noisy image or shape in handwritten character is not properly recognized result obtained will be incorrect.

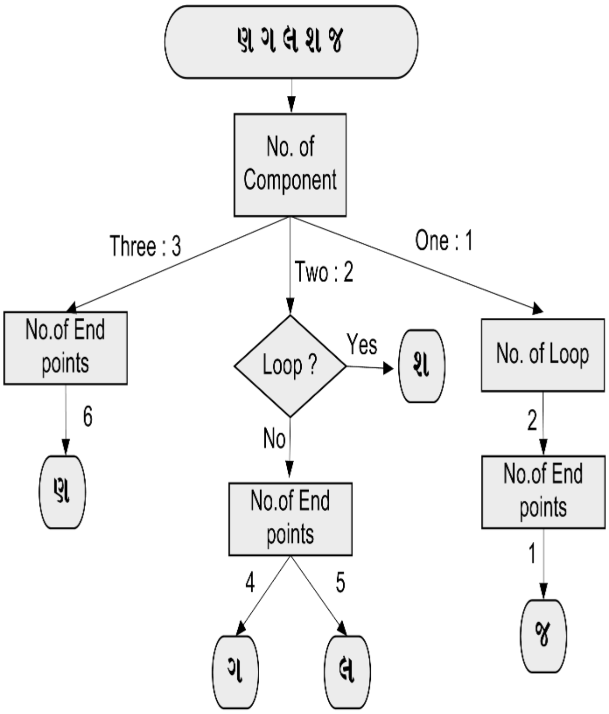
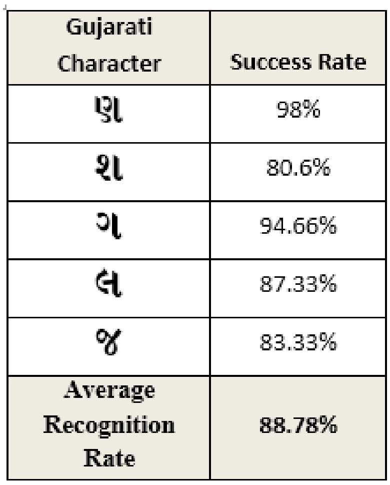


Fig 10 : Decision Tree for Classification of Characters

8. RESULT ANALYSIS

Proposed methodology of structural feature extraction and classification based on decision tree is applied on 750 handwritten character where 150 samples of each character, result obtained is shown in below table (Table 2). Success Ratio for ‘aNa’ is 98%, ‘Sha’ is 80.6%, ‘Ga’ is 94.66%, ‘La’ is 87.33% and ‘Ja’ is 83.33%.

Table 2 Success Ratio for recognition of isolated offline handwritten Gujarati Characters



Gujarati Character	Success Rate
ણ	98%
શ	80.6%
ગ	94.66%
લ	87.33%
જ	83.33%
Average Recognition Rate	**88.78%**

Table 3 shows Confusion matrix Where NR represents not recognized. Given an input image how many times given input image is misclassified with other characters as well as it also represents result analysis where character is not identified.

Table 3. Confusion Matrix for Gujarati Characters

	ણ	શ	ગ	લ	જ	NR
ણ	147	0	0	3	0	0
શ	4	121	0	0	0	25
ગ	8	0	142	0	0	0
લ	16	0	0	131	0	3
જ	0	10	0	0	125	15

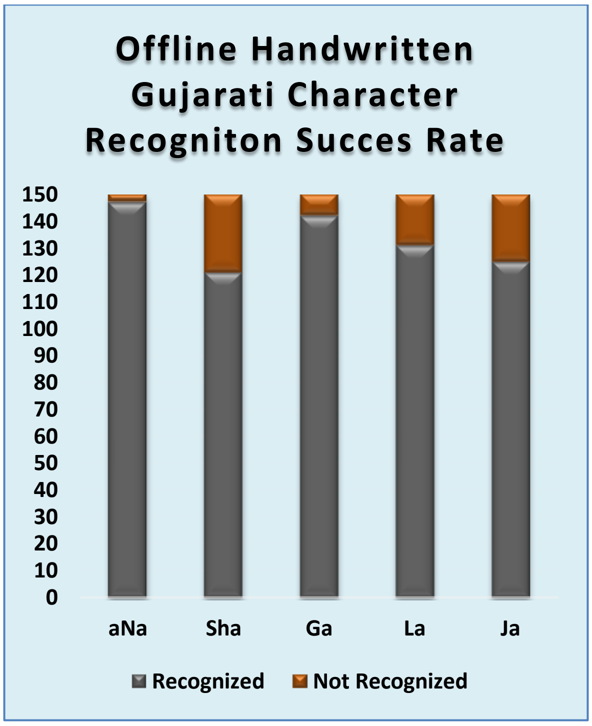


Fig 11 : Success Ratio of Gujarati character recognition

Fig (11) represents graphical representation of success ratio achieved to identify offline handwritten Gujarati Characters i.e. ‘aNa’, ‘Ga’, ‘Sha’, ‘La’ and ‘Ja’.

9. SAMPLE OUTPUT

Fig (12) represents sample output of proposed experimental work for sample handwritten character ‘Ga’ and ‘Sha’.

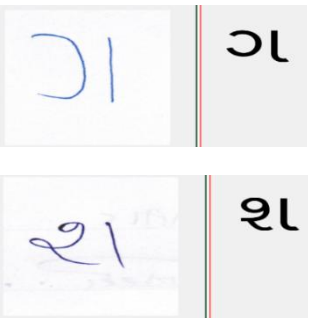


Fig 12 : Sample output

10. CONCLUSION

Structural feature are useful for offline handwritten character recognition as features extraction process doesn’t rely on size of a character. This paper proposed recognition of five Gujarati characters based on their structural features and decision tree classifier and authors are able to achieve 88.78% of average recognition rate. This paper is a first step toward recognizing offline isolated handwritten Gujarati character. In future this work can be extended further to classify other characters of Gujarati script. Proposed approach is also useful for recognizing machine printed Gujarati characters.

11. ACKNOWLEDGEMENT

Authors are thankful to all the writer who have contributed by providing handwritten input for proposed experimental work.

12. REFERENCES

[1] A. D. Chhaya Patel, "Gujarati Handwritten Character Recognition Using Hybrid Method Based On Binary Tree-Classifier And K-Nearest Neighbour," *International Journal of Engineering Research & Technology*, vol. 2, no. 6, 2013.

[2] D. U. R. S. P. Jayashree R. Prasad, "Offline Handwritten Character Recognition of Gujrati Script using Pattern Matching".

[3] R. P. S. P. J. M. Lipi Shah, "Handwritten Character Recognition using Radial Histogram," *International Journal of Research in Advent Technology*, vol. 2, no. 4, 2014.

[4] M. f. Kamal moro, "Gujarati Handwritten Numeral Optical Character through neural network and skeletonization," *jurnal of sistem komputer*, vol. 3, no. 1, pp. 40-43, 2013.

[5] J. Prasad, U. Kulkarni and R. Prasad, "Template Matching Algorithm for Gujarati Character," in *In Proc. Of 2nd International Conference on Emerging Trends in Engineering and Technology (ICETET)*, 2009.

[6] S. S. Rejean Plamondon, "Online and off-line Handwriting Recognition : A comperehensive survey," *IEEE transaction on pattern analysis and machine intelligence*, vol. 22, no. 1, 2000.

[7] B. C. U.Pal, "Indina Script recogniton: a survey," *Pattern Recognition*, vol. 37, pp. 1887-1899, 2004.

[8] "Babu Suthar - Gujarati English learner's dictionary".

[9] A. A. Desai, "Gujarati handwritten numeral optical character reorganization therough neural network," *Pattern Recognition*, vol. 43, 2010.

[10] S. R. V. D. G. K. Avani R. Vasant, "Gujarati Character Recognition : The state of the art comprehensive survey".

[11] D. C. Hetal R. Thaker, "Preprocessing and Segregating Offline Gujarati Handwritten Datasheet for Character Recognition," *International Journal of Computer Applications*, vol. 97, no. 18, pp. 43-47, 2014.