Volume 3, Issue 1, January 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Feature Extraction of Gurmukhi Script and Numerals: A Review of Offline Techniques

Gurpreet Singh *CSE,NIT JALANDHAR India*
Chandan Jyoti Kumar *CSE,NIT JALANDHAR India*
Rajneesh Rani *CSE,NIT JALANDHAR India*
Dr. Renu Dhir *CSE,NIT JALANDHAR India*

ABSTRACT: ***Offline Isolated handwritten Gurmukhi character recognition has been a very intensive area of research during last decades due to it is wide range of solution to real world problems. A lot of work has been done in languages like Chinese, Arabic, Devnagari, Urdu and English [1-3]. Research on the different stages of OCR of Gurmukhi script is being carried out by the authors and their M.Tech students at Punjabi University, Patiala. A preliminary work was done by Sanjeev Kumar [4] and Khushwant Kaur [5] under the guidance of one of the authors, Lehal, developing a feature based Gurmukhi recognition script system. The count and location of local features such as endpoints, T-points, cross points and loops were used to identify isolated Gurmukhi characters A neural networks based Gurmukhi recognition system has been developed by Goyal et.al. [6]. Range free skew detection algorithms for de-skewing Gurumukhi machine printed text skewed at any angle, have been developed by Lehal and Madan [7] and Lehal and Dhir [8]. If different classifiers cooperate with each other, group decisions may reduce errors drastically and achieve a higher performance. In this survey, focuses on the various techniques used for recognition of isolated offline handwritten characters in Gurmukhi script. The whole process consists of two stages. The first, feature extraction stage analyzes the set of isolated characters and selects a set of features that can be used to uniquely identify characters. The performance depends heavily on what features are being used.***

Keywords: OCR, Handwritten Gurmukhi Script, projection histogram feature extraction techniques, survey, isolated handwritten character recognition.

1. Introduction to Gurmukhi script

Gurmukhi script is used primarily for Punjabi language, which is the world's 14th most widely spoken language. The Character set of Gurmukhi script is as in Fig3.

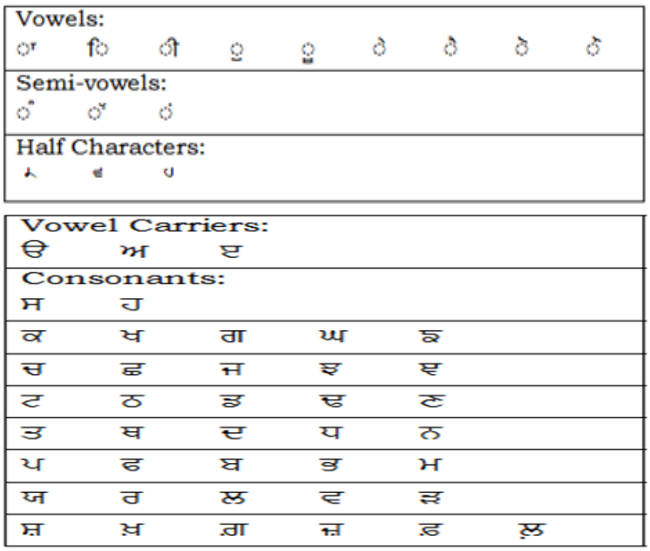


Fig 3: Character set of Gurmukhi Script

Some of the properties of Gurmukhi script are: Gurmukhi script is cursive and the character set consist of 41 consonants 9 vowels, 3 sound modifiers(semi-vowels) and 3 half characters, lie at the feet of consonants. Most of the Gurmukhi characters have a horizontal line at the upper part. The characters of words are connected mostly by this line called head line and so there is no vertical inter-character gap in the letters of a word.

For example:

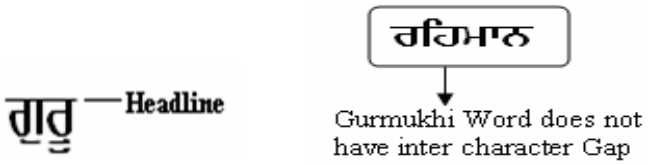


Fig 4: Headline and missing inter-character gap

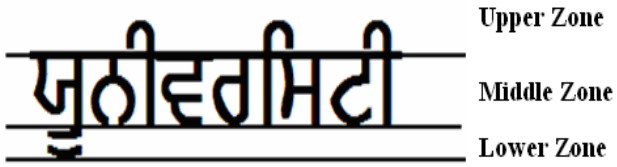


Fig 5: Horizontal Zones

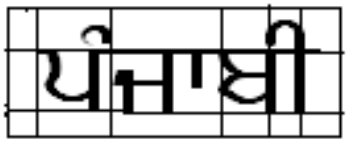
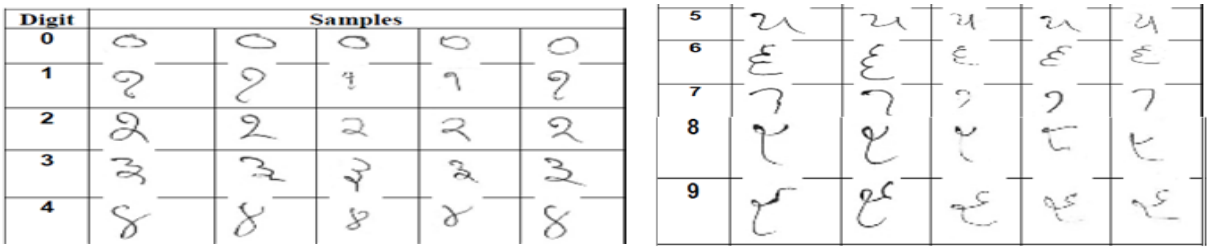


Fig 6: Intersecting/overlapping characters

A word in Gurmukhi script can be partitioned into three horizontal zones, as shown in Figure5. The upper zone denotes the region above the head line, where vowels reside, while the middle zone represents the area below the head line where the consonants and some sub-parts of vowels are present. The middle zone is the busiest zone. The lower zone represents the area below middle zone where some vowels and certain half characters lie in the foot of consonants. But there is no concept of upper and lower zones in Gurmukhi digits.

The bounding boxes of 2 or more characters in a word may intersect or overlap vertically. For example bounding boxes of ਪ and ੰ intersect. There are lots of topologically similar character pairs in Gurmukhi script. Some similar pairs are

ਖ and ਖ਼, ਵ and ੜ, ਖ and ਖ਼, ਤandੜ etc



Handwritten samples of gurmukhi numerals

Many techniques like median filtration, dilation, isolated pixels removal and many other morphological operations to bridge unconnected pixels and to remove spur pixels etc. Before extracting the features normalized the pre-processed numeral images.

2. Problem Description

Recognition of isolated handwritten characters is the process of identifying individual characters. It is useful in wide range of real world problems like documentation analysis, mailing address interpretation, bank check processing, signature verification, documentation verification and many others. The problem is to recognize the isolated handwritten characters in Gurmukhi script.

The major difficulties are:

1. The variability of writing styles, both between different writers and between separate examples from the same writer overtime. For example:

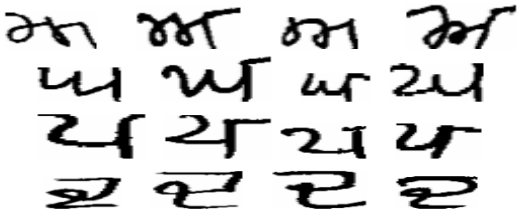


Fig 7: Varying writing styles

2. The similarity of some characters. Table1 shows examples of similar characters

ਖ and ਖ਼	ਵ and ੜ	ਜ andਜ਼
ਗ and ਘ	ਤ and ੜ	ਬ and ਥ
ਨ andਠ	ਤandਡ	ਥ and ਖ
ਫ and ਫ਼	ਟand ਦ	ਪ and ਧ
ਅ andਘ	ਹ and ਰ	ਪ and ਖ
ਸ andਮ	ਏ and ਟ	ਗ andਗ਼
ਸ and ਸ਼	ਵ andੜ	ਚ andਦ
ਲ and ਲ਼		

Table .1

3. The possible low quality of the text image as in fig. 8

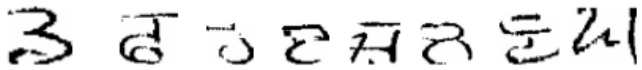


Figure 8: Low quality of text images

4. The unavoidable presence of background noise and various kinds of distortions (such as poorly written, degraded, or overlapping characters) can make the recognition process even more difficult. For example:

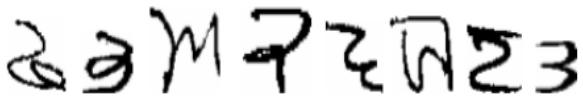


Figure 9: Poorly written characters

In the problem of recognition of isolated handwritten characters the input is isolated characters. Word segmentation provides isolated characters. Characters can be in upper zone, middle zone or lower zone.

Some of the examples are shown in figures below:

Upper zone:

Fig 10: Upper zone characters

But in case of vowels ੀ and ਿ, ੁ lie in middle zone and ੂ lie in upper zone.

Middle zone:

ਓ	ਅ	ਏ	ਸ	ਹ	ਕ	ਖ	ਗ	ਘ	ਙ	
ਚ	ਛ	ਜ	ਝ	ਞ	ਟ	ਠ	ਡ	ਢ	ਣ	
ਤ	ਥ	ਦ	ਧ	ਨ	ਪ	ਫ	ਬ	ਭ	ਮ	
ਯ	ਰ	ਲ	ਵ	ੜ	ਸ਼	ਜ਼	ਖ਼	ਫ਼	ਗ਼	ਲ਼

Fig 11: Middle zone characters

But in case of character ਓ, ◠ lies in upper zone and ਓ lies in middle zone. Vowel lies in middle zone.

Lower zone:

Fig 12: Lower zone characters

3. FEATURE EXTRACTION

We have used following three sets of features extracted to recognize gurmukhi numerals. These approaches are adopted from our earlier practice [11] used to recognize isolated Gurmukhi handwritten characters.

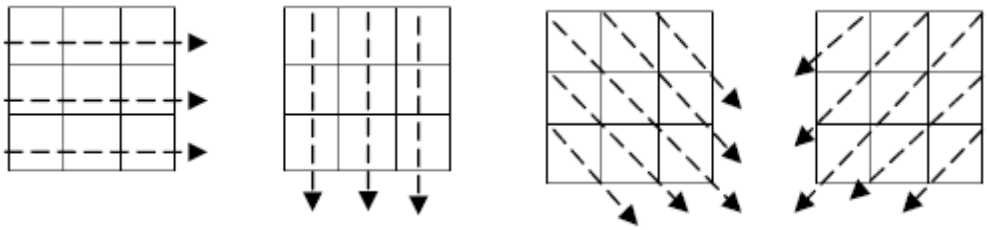
1. Distance Profile Features
2. Projection Histogram Features
3. Zoning density
4. Zernike moments feature extraction
5. Hybrid feature

3.1 Distance Profile Features

In our approach we have used distance profiles using distance computation from bounding box to outer edges of character from four sides- two in horizontal direction from left and right sides and other two in vertical direction from top and bottom side. Left and right profiles are traced by horizontal traversing of distance from left bounding box in forward direction and from right bounding box in backward direction respectively to outer edges of character. Similarly, top and bottom profiles are traced by vertical traversing of distance from top bounding box in downward direction and from bottom bounding box in upward direction respectively to outer edges of character. The size of each profile in our approach is 32 similar to number of pixels in each row or column forming total 128 features by all four types of profiles.

Projection Histogram Features

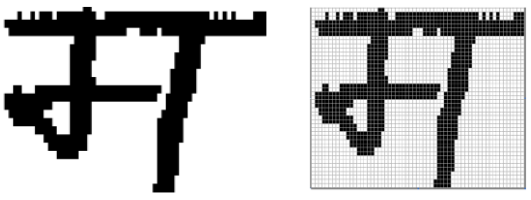
Projection histograms count the number of foreground pixels in specified direction. In our approach we have used four directions of horizontal, vertical and both diagonal traversing.



(a) Horizontal Histogram **(b) Vertical Histogram** **(c) Diagonal-1 Histogram** **(d) Diagonal-2 Histogram**

3.2 Zoning density (ZD)

The frame containing the character is divided into several overlapping or non-overlapping zones and the densities of object pixels in each zone are calculated. Density is calculated by finding the number of object pixels in each zone and dividing it by total number of pixels. Densities are used to form a representation. For binary images, value of each pixel is either 1 or 0. We have considered pixels having value BLACK (0) as object pixels. This feature is extracted from the scaled (normalized) character matrix of the character. The original character image (matrix) is first scaled to Normalized window of size 48*48. The Zoning feature set consists of 64 values. The values in feature vector are normalized in the range 0 to 1.Normalization is done by dividing all the values by the largest value in the feature set.



Original Image of sassa Scaled Image of sassa (ਸ)

3.3 Zernike moments feature extraction

Zernike [12] introduced a set of complex polynomials $\{V_{nm}(x, y)\}$ which forms a complete orthogonal set over a unit disk $x^2 + y^2 \leq 1$. The form of the polynomial is:

$$V_{nm}(x,y) = V_{nm}(p,\theta) = R_{nm}(p).\exp(jm\theta)$$

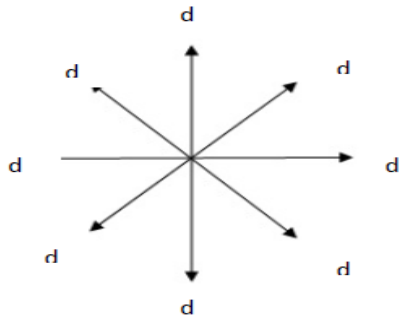
$$\text{where, } p = \sqrt{x^2+y^2} \text{ and } \theta = \tan^{-1}\frac{y}{x} \text{ and } j=\sqrt{-1}$$

here, n is positive integer or zero, m is an integer subject to constraints n-$|m|$ is even, and $|m| \leq n$, p is the length of the vector from the origin to the pixel (x,y), θ is the angle between the vector p and x -axis in counter clockwise direction. $R_{nm}(p)$is radial polynomial defined as:

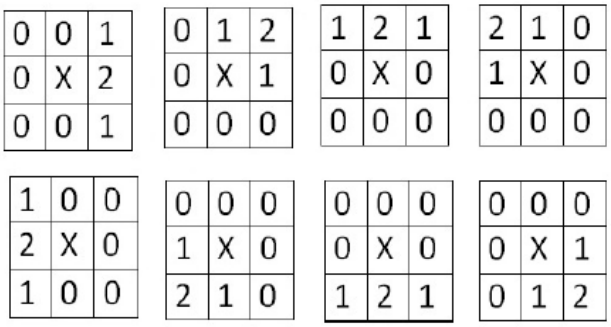
$$R_{nm}(p) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n+|m|}{2}-s\right)!} p^{n-2s}$$

3.4 Hybrid features [ZD and Background Directional Distribution (BDD)]

In zoning, the character image is divided into N×M zones. From each zone features are extracted to form the feature vector. The goal of zoning is to obtain the local characteristics instead of global characteristics. By dividing the number of foreground pixels in each zone by total number of pixels in each zone obtained the density of each zone. Thus we obtained different zoning density features. We have considered the directional distribution of neighboring background pixels to foreground pixels. To calculate directional distribution values of background pixels for each foreground pixel, we have used the masks for each direction shown in figure 2(b). The pixel at center 'X' is foreground pixel under consideration to calculate directional distribution values of background. The weight for each direction is computed by using specific mask in particular direction depicting cumulative fractions of background pixels in particular direction.



(a) 8 directions



(b) Masks used to compute different directional distributions

4. Advancement in feature extraction offline handwritten Gurmukhi Scripts and numerals

4.1 Kartar Singh Siddharth et.al. [13] have used three different feature sets. First feature set is comprised of distance profiles, second feature set is comprised of different types of projection histograms and third feature set is comprised of zonal density and Background Directional Distribution (BDD). The SVM classifier with RBF (Radial Basis Function) kernel is used for classification. We have obtained the 5-fold cross validation accuracy as 99.2 %.

4.2 Anuj Sharma et al. [18], [19] have presented the implementation of three approaches: elastic matching technique, small line segments and HMM based technique, to recognize online handwritten Gurmukhi characters and reported 90.08%, 94.59% and 91.95% recognition accuracies respectively.

4.3 Dharam Veer Sharma et al. [20] first extracted Gurmukhi digits from printed documents and then recognized. They have used many structural features like loops, entry points, curve, line, aspect ratio, and statistical features like zoning, directional distance distribution for recognition and observed 92.6% recognition rate for Gurmukhi digits. For offline handwritten Gurmukhi character recognition two approaches are reported. First one is proposed by Puneet Jhajj et al. [14] and second one by Ubika Jain et al. [15]. A little more detailed survey on Gurmukhi recognition is presented in [6] and [19].

4.4 Puneet Jhajj et.al. [14]. used a 48*48 pixels normalized image and created 64 (8*8) zones and used zoning densities of these zones as features. They used SVM and K-NN classifiers and compared the results and observed 72.83% highest accuracy with SVM kernel with RBF kernel.

4.5 Ubeeka Jain [15] et al. created horizontal and vertical profiles, stored height and width of each character and used neocognitron artificial neural network for feature extraction and classification. They obtained accuracy of 92.78% at average.

4.6 Dharamveer Sharma et.al. [14] use zoning feature extraction method with different types of classifiers KNN ,SVM (linear,poly kernel,RBF kernal) and results was not too satisfactory.Ubeeka Jain et.al. [15] has recognition system of isolated handwritten characters by using Neocognitron overall recognition accuracy for both learned and unlearned Gurmukhi characters are 92.78 %.

4.7 Munish Kumar [16] presents an efficient offline handwritten Gurmukhi character recognition system based on diagonal features and transitions features using *k*-NN classifier.Diagonal and transitions features of a character have been computed based on distribution of points on the bitmap image of character. In *k*-NN method, the Euclidean distance between testing point and reference points is calculated in order to find the *k*-nearest neighbors. It achieves a maximum recognition accuracy of 94.12% using diagonal features and *k*-NN classifier.

4.7 G. G. Rajput et.al. propose a novel method towards multi-script identification at block level. The recognition is based upon features extracted using Discrete Cosine Transform (DCT) and Wavelets of Daubechies family. The proposed method is experimented on handwritten documents of eight Indian scripts that include English script and yielded encouraging results.

Tri-script Group	Tr-scripts	Recognition %
1	Kannada, English and Hindi	98%
2	Malayalam, English and Hindi	99.2%
3	Punjabi, English and Hindi	93%
4	Tamil, English and Hindi	99.2%
5	Gujarati, English and Hindi	90%
6	Telagu, English and Hindi	99%

4.8 Naveen Garg et al. [8] have recognized offline handwritten Gurmukhi characters using neural network and obtained 83.32% average recognition accuracy. Puneet Jhajj et al. [9]used a 48×48 pixels normalized image and created 64 (8×8) zones and used zoning densities of these zones as features.

4.9 Ashutosh aggrawal has proposed recognition of Gurmukhi handwritten characters using gradient feature with svm classifier with RBF kernel shows 97.38% accuracy. Anita rani has used BDD and zonal density based features for Gurmukhi numerical on SVM with RBF kernel shown accuracy of 99.4%.

5. Conclusion

Methods of extracting different features for character recognition have developed remarkably in the last decade. In this paper we have proposed an organization of these methods under four basic strategies, with hybrid approaches also identified. It is hoped that this comprehensive discussion will provide insight into the concepts involved, and perhaps provoke further advances in the area.

The paper has concentrated on an appreciation of principles and methods. We have not attempted to compare the effectiveness of algorithms, or to discuss the crucial topic of evaluation. It would be very difficult to assess techniques separate from the systems for which they were developed. We believe that wise use of context and classifier confidence has led to improved accuracies, but there is little experimental data to permit an estimation of the amount of improvement to be ascribed to advanced techniques. Perhaps with the wider availability of standard databases, experimentation will be carried out to shed light on this issue. We have included a list of references sufficient to provide more detailed understanding of the approaches described. We apologize to researchers whose important contributions may have been overlooked.

References

[1] J. Mantas, "An overview of character recognition methodologies", *Pattern Recognition*, Vol. 19, pp 425-430 (1986).

[2] V. K. Govindan and A. P. Shivaprasad, "Character recognition – A survey ", *Pattern Recognition,* Vol. 23, pp 671-683(1990).
[3] B. Al-Badr and S.A. Mahmoud, "Survey and bibliography of Arabic optical text recognition", *Signal Processing,* Vol.41, pp. 49-77(1995).
[4] S. Kumar, "A technique for recognition of printed text in Gurmukhi script", M.Tech. thesis, Punjabi University, (1997).
[5] K. Kaur, "An approach towards the recognition of machine printed Gurmukhi script", M.Tech. thesis, Punjabi University, (1999).
[6] A K Goyal, G S Lehal and J Behal, "Machine Printed Gurmukhi Script Character Recognition Using Neural Networks", Accepted for publication in Proceedings 5th International Conference on Cognitive Systems, Delhi,India,(1999)
[7] G S Lehal and S. Madan, "A New Approach to Skew detection and Correction of Machine Printed Gurmukhi Script", Proceedings 2nd International Conference on Knowledge Based Computer Systems, Mumbai, India, 215-224 (1998)
[8] G S Lehal and R. Dhir, "A Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents" , In Proceedings 5th International Conference of Document Analysis and Recognition, IEEE Computer Society Press, California, pp. 147-152, (1999)
[9] G S Lehal and P. Singh, "A Technique for Segmentation of Machine Printed Gurmukhi Script", Proceedings 4th International Conference on Cognitive Systems, Delhi, India, 283-287 (1998)
[10] A. K. Goyal, G S Lehal and S S Deol, "Segmentation of Machine Printed Gurmukhi Script" , Proceedings 9th International Graphonomics Society Conference, Singapore, pp. 293-297 (1999)
[11]Kartar Singh Siddharth, Mahesh Jangid, Renu Dhir, Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", *International Journal of Computer Science and Engineering (IJCSE),* Vol. 3, No. 6, June 2011.
[12]J. Tripathy, "Reconstruction of Oriya alphabets using Zernike Moments"*IJCA,* Vol. 8 (8), pp. 26-32, 2010.
[13]Kartar Singh Siddharth, Renu Dhir Rajneesh Rani " Handwritten Gurmukhi Numeral Recognition using Different Feature Sets ", published International Journal of Computer Applications (0975 – 8887)Volume 28– No.2, August 2011
[14]Dharamveer Sharma, Puneet Jhajj "Recognition of Isolated Handwritten Characters in Gurmukhi Script**",** International Journal of Computer Applications (0975 – 8887) Volume 4– No.8, August 2010
[15]Ubeeka Jain, D. Sharma, "Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron", International Journal of Computer Applications (IJCA),Vol. 4, No. 8, 2010.
[16]Munish Kumar, M. K. Jindal, R. K. Sharma, "Classification of Characters and Grading Writers in Offline Handwritten Gurmukhi Script ", accepted for publication in 2011 International Conference on Image Information Processing (ICIIP 2011).
[17] Pritpal Singh, Sumit Budhiraja / International Journal of Engineering Research and Applications (IJERA) Vol. 1, Issue 4, pp. 1736-1739 1736
[18]Anuj Sharma, Rajesh Kumar, R. K. Sharma, "Online Handwritten Gurmukhi Character Recognition Using Elastic Matching," *Image and Signal Processing, 2008. CISP '08. Congress on,* vol.2, no., pp.391-396, 27-30 May 2008
[19]Anuj Sharma, R.K. Sharma, Rajesh Kumar, "Online Handwritten Gurmukhi Character Recognition", Ph.D. Thesis, Thapar University, 2009[Online].
[20]D. Sharma, G. S. Lehal, Preety Kathuria, "Digit Extraction and Recognition from Machine Printed Gurmukhi Documents", MORC Spain, 2009

About the authors

Gurpreet Singh received his B.Tech degree in Computer Science and Engineering from Punjabi University, in 2011. He is persuing his M.Tech degree from NIT Jalandhar. His areas of current research interest are OCR of handwritten text, Punjabi script, pattern recognition and image processing.

Chandan jyoti Kumar received his B.Tech degree in Computer Science and Engineering from NIT Silchar, in 2011. He is persuing his M.Tech degree from NIT Jalandhar. His areas of current research interest are OCR of handwritten text, Bengali script, pattern recognition and image processing.

Rajneesh rani is working as assistant professor in Computer Science and Engineeringin NITJ. She received her M.Tech degree from Punjabi university. Her areas of current research interest are OCR of handwritten text, Punjabi script, pattern recognition and image processing.

Dr. Renu Dhir is working associate professor in Computer Science and Engineering in NITJ. Her areas of current research interest are OCR of handwritten text, Punjabi script, pattern recognition and image processing

