International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-3, February 2013

# A Survey of Feature Extraction and Classification Techniques Used In Character Recognition for Indian Scripts

**Aditya Raj, Ranjeet Srivastava, Tushar Patnaik, Bhupendra Kumar**

***Abstract*— The Constitution of India has recognized 22 languages as official languages of India. Among these most of the recognition research work has been done for Devanagari, Gurumukhi, Telugu, and Bangla scripts etc. OCR system development for Indian script has many application areas like preserving manuscripts and ancient literatures written in different Indian scripts and making digital libraries for the documents. Feature extraction and classification are essential steps of character recognition process affecting the overall accuracy of the recognition system. This paper gives a detailed overview of different feature extraction and classification techniques for recognition process of different Indian scripts by the researchers over the past few decades***

***Index Terms*— Optical Character Recognition (OCR), Feature Extraction, Classification.***

## I. INTRODUCTION

The history of OCR can actually found back in 1923 Tausheck [22] and 1933 Handel [23] gave the first idea of the concept of the OCR. Optical Character Recognition deals with the problem of recognizing optically processed characters. Optical recognition is an offline process i.e. the recognition starts after writing or printing has been completed. Hand printed and machine printed characters both can be recognized, but the performance is directly dependent upon the input parameters like quality of Input image. Over the last few decades character recognition research has gained a considerable attention, because preserving the handwritten/machine printed text in to digitized format has become prevalent.

Out of around 33 different languages and 2000 dialects that have been identified in India, 22 are officially recognized languages. The use of multi-lingual documents has increased which necessitates the intelligent extraction of features and use of classification techniques for achieving maximum accuracy and performance and minimum error rate in recognition process. Optical Character recognition system comprises of following 5 steps.

### *1.1 Image acquisition: -*

Digital imaging or digital image acquisition is the creation of digital images. With the help of scanning process digital image of the document is captured.

**Manuscript published on 28 February 2013.**
* Correspondence Author (s)
**Aditya Raj,** M.Tech(CSE), Department of CSE, C-DAC, Noida, India.
**Ranjeet Srivastva**, M.Tech(IT), Department of IT, C-DAC, Noida, India.
**Tushar Patnaik**, Sr. Lecturer/ Sr. Project Engineer, Department of CSE, C-DAC, Noida, India.
**Bhupendra Kumar**, Sr. Technical Officer, Department of CSE, C-DAC, Noida, India.



### 1.2 Pre-processing:-

Digital image obtained from scanning may contain some amount of noise depending upon the quality of scanner. The lines might be skewed or characters may be smeared or broken. Thus, pre-processing is required which involves elimination of noise, Binarization of the image and segmentation (line, word and character level).

### 1.3 Feature Extraction:-

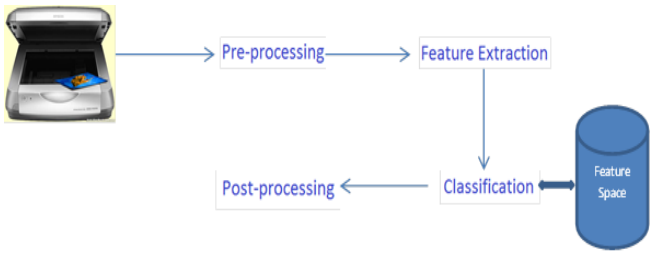
The objective of feature extraction is to capture the essential characteristics of the symbols. Being the most important and crucial step of the recognition process, selection of the feature extra00ction technique becomes important factor in achieving the high recognition performance. It can be said to be one of the most difficult problems of pattern recognition. Some feature extraction methods are Template matching, Deformable Templates, Zoning, Projection Histogram, Contour Profile, Moments calculation (Ex- Geometrical, Hu-moments, Zernike).

### 1.4 Classification:-

Classification is the process of assigning the sensed data to their corresponding class with respect to groups with homogeneous characteristics, with the aim of discriminating multiple objects from each other within the image. Classification is carried out on the basis of stored features in the feature space, such as structural features, global features etc. It can be said that classification divides the feature space into several classes based on the decision rule. Some classification techniques used in previously developed Optical character recognition systems are Neural Network, Support Vector Machine, K-Nearest Neighbors, Bayesian Classification, and Decision Tree Classification.

### 1.5Post-processing:-

Post-processing step involves grouping of symbols. The process of performing the association of symbols into strings is referred to as grouping.



## II. FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUES USED IN THE OCR SYSTEMS FOR INDIAN SCRIPTS

Veena Bansal and R.M.K Sinha [1] presented a complete OCR for printed Hindi text written in Devanagari script. The system used following features: Coverage of the region of the core strip, Vertical bar feature, Horizontal zero crossings, Number of positions of the vertex points, Moments, Structural descriptors of the characters for classification, Tree classifiers are used. Overall accuracy obtained at the character level is 93%. Sinha and Mahabala [2] designed a syntactic pattern analysis system for Devanagari script recognition. The system stores structural descriptors for each symbol of the script. They achieved 90% accuracy. Reena, Lipika and Chaudhury [14] have tried to exploit information about stylistic variations, similarity between numerals and style invariant features. They presented a approach for recognition of handwritten Devnagari numerals using multiple neural classifiers. Sandhya Arora [15] have used Intersection features with Neural Network for Devanagari script and achieved 89.12% accuracy

Singh and Budhiraja [3] presented an OCR system for handwritten isolated Gurumukhi script using Zoning, Projection histogram, Distance profile features, and Background directional features and used Support Vector Machines (SVM) for classification and thus obtained 95.04% of overall accuracy. Further Geeta and Rani [4] represented an OCR system for Gurumukhi numerals using Zone Distance features and SVM classifier and achieved 99.73% accuracy. G. S. Lehal and Chandan Singh [16] directed their efforts towards development of OCR system for Gurumukhi. They used Local features (concave/convex parts, number of endpoints, branches, joints) and Global features (connectivity, projection profiles, number of holes etc.). For classification hybrid classification technique, binary decision tree and nearest neighbour was used. They achieved a recognition rate of 91.6%. Dharamveer Sharma and Puneet Jhajj [17] used zoning feature with hybrid classification technique using KNN and SVM classifier and achieved 72.7% accuracy.

A very influential attempt made by the Jalal, Feroz and Choudhuri [5] for Bangla script. They represent neural network classifier by using Bounded rectangle calculation, Chain code generation, Slope distribution generation features. They achieved 96% system accuracy. Chaudhuri and Paul [6] represent an OCR system to recognize Bangla and Devanagari using stroke and shaded portion feature with tree classifier. U. Bhattacharya, M. Shridhar, and S.K. Paruil [18] implemented Neural network classifier for isolated Bangla characters with chain code features and achieved 92.14% accuracy on testing sets and 94.65% on training sets.

Negi and Chakravarthy [7] represent an OCR system with 92% performance using template matching, fringe distance for Telegu script. Another attempt was made by Lakshmi and Patvardhan [8] for Telegu script. They used neural classifier by using directional features and they achieved 92% accuracy. Arun K Pujari, and C Dhanunjaya Naidu [19] implemented an adaptive character recognizer for Telugu scripts using Multi resolution Analysis. They represented DNN (Dynamic Neural Network) using Wavelet analysis and achieved 93.46 % success rate.

In south India, Telegu and Kannada have similar scripts. R Sanjeev and R D Sudhakar[9] represent an OCR system for printed Kannada Script using two stage

Multi-Network(Neural Network) classification technique employing wavelet feature and achieved 91% accuracy at character level. M Sagar, Shobha and Ramakanth [10] designed a syntactical analysis system using Ternary Tree based classification for isolated Kannada characters. They have given more emphasis on Post-processing step, using dictionary based approach to increase the OCR accuracy. T V Ashwin and P S Sastry [20] represents a font and size-independent OCR system for printed Kannada documents using support vector machines (SVM).

B Chaudhuri U Pal and Mitra [11] gave a prototype OCR system for Oriya script. They use Directional features and Global Features and classified them using Decision tree classifier and achieved 96.03% accuracy at character level.

Junaid, Umar, and Muhammad Umair [12] attempted to make an OCR system for isolated Urdu characters using NN classifier using structural features like width, height and checksum of the character. Their prototype gained the accuracy of 97.43%. Another good attempt was made by Jhuwair and Abdul [13] for Urdu script. They achieved the 97.12% recognition rate using Sliding window and Hu-moment feature using KNN classifier.

## III. CONCLUSION WITH COMPARISON TABLE

Survey represents a study of feature extraction methods with different classifiers implemented in OCR systems for different Indian scripts .Variance between the features should be clearly discriminative and specific so

that system can classify the characters with maximum efficiency and minimum error rate. This survey paper helps researchers and developers to understand history of the OCR research work for Indian scripts. OCR for Indian scripts that works under all possible conditions and gives highly accurate results still remains a highly challenging task to implement.

| Sno | Language | Feature extraction Methods | Classification Methods | Recognition Rate(in %) |
|---|---|---|---|---|
| 1 | Devanagari | Vertical Feature Bar ,Horizontal Zero ,crossing , Number of positions on the vertex point, Moments Structural descriptors[1] | Tree Classifier | 93 |
| 2 | Devanagari | Syntactical Analysis[2] | Tree Classifier | 90 |
| 3 | Devanagari | Intersection Features[15] | Neural Network | 89.12 |
| 4 | Gurumukhi | Local features(concave/convex parts, number of endpoints, branches, joints, Global features(connectivity, projection profiles, number of holes[16] | binary decision tree and nearest neighbour | 91.6 |
| 5 | Gurumukhi | zoning feature[17] | KNN and SVM classifier | 72.7 |
| 6 | Gurumukhi | Zoning, Projection Histogram, Distance Profile Features, Background Directional Features[3] | Support Vector Machines(SVM) | 95.05 |
| 7 | Gurumukhi | Zone Distance Features[4] | SVM | 99.73 |
| 8 | Bangla | Bounded Rectangle Calculation, Chain code generation, Slope Distribution generation[5] | Neural Network | 96 |
| 9 | Bangla | Directional Features[6] | Decision Tree | --- |
| 10 | Bangla | Chain code[18] | Neural network | 92.14 |
| 11 | Telugu | Template Matching using Fringe Distance[7] | Neural Network | 92 |
| 12 | Telugu | Directional Features[8] | Neural Network | 92 |
| 13 | Telugu | Wavelet Analysis[19] | DNN | 93.46 |
| 14 | Kannada | Wavelet Features, Syntactical Analysis[9] | Neural Network, Ternary Tree | 91 |
| 15 | Oriya | Directional Features Global Features[11] | Decision Tree | 96.73 |
| 16 | Urdu | Structural Features(width, height, and checksum)[12 | Neural Network | 97.43 |
| 17 | Urdu | Sliding Window, Hu-moment [13] | KNN | 97.12 |


Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

## IV. ACKNOWLEDGEMENT

Our thanks to Madhuri Yadav(MTECH) and Kirti Singh (MCA) for many suggetions and contributions.

## REFERENCES

[1] Veena Bansla and R M K Sinha, "A Complete OCR for printed Hindi Text in Devanagari Script", IEEE 800 - 804 2001 .
[2] Sinha. M. K., Mahabala., "Machine Recognition of Devnagari Script", IEEE T. SYST. MAN Cyb., vol.. 9,pp.435-449,1979.
[3] Pritpal Singh and Sumit Budhiraja, "Feature Extraction and Classification Techniques in O.C.R. Systems for Handwritten Gurmukhi Script". International Journal of Engineering Research and Applications (IJERA), Vol.1, ISSUE 4,pp.1736-1739.
[4] Gita Sinha Rajneesh Rani Renu Dhir , "Handwritten Gurmukhi Numeral Recognition using Zone-based Hybrid Feature Extraction Techniques", International Journal of Computer Applications(0975-8887), Volume 47- No. 21 June 2012.
[5] Jalal Uddin Mahtnud, Mohammed Feroz Raihan and Chowdhury Mofizur Rahman, "A Complete OCR System for Continuous Bengali Characters",IEEE 1372 - 1376 Vol. Oct. 2003 .
[6] B. B. Chaudhuri and U. Pal, "An OCR System to Read Two Indian Language Scripts: Bangla and Devnagari (Hindi)", IEEE 1011 - 1015 vol.2, Aug 1997.
[7] Atul Negi, Chakravarthy Bhagvati, B. Krishna, "An OCR system for Telugu", IEEE 1110 – 1114 -2001.
[8] Vasantha Lakshmi, C. Patvardhan, C. , "A high accuracy OCR system for printed Telugu text", IEEE 725 - 729 Vol.2 , Oct-2003.
[9] R Sanjeev Kunte, R D Sudhaker Samuel, "An OCR System for Printed Kannada Text Using Two - Stage Multi-network Classification Approach Employing Wavelet Features", IEEE 349 – 353,Dec-2007.
[10] Sagar, B.M. , Shobha, G. , Kumar, P.R. , "Complete Kannada Optical Character Recognition with syntactical analysis of the script", IEEE 1 – 4 , Dec. 2008.
[11] Chaudhuri, B.B, Pal, U. , Mitra, M. , "Automatic recognition of printed Oriya script", IEEE 795 – 799, 2001.
[12] Tariq, J, Nauman, U, Naru, M.U , "Softconverter: A novel approach to construct OCR for printed Urdu isolated characters", IEEE V3-495 - V3-498 ,April 2010.
[13] Sardar, S, Wahab, A, "Optical character recognition system for Urdu",IEEE 1 - 5 , June 2010.
[14] Reena Bajaj, Lipika Dey and Santanu Chaudhury, "Devnagari numeral recognition by combining decision of multiple connectionist classifiers", Vol. 27, Part 1, February 2002, pp. 59–72.
[15] Sandhya Arora , "A Two Stage Classification Approach for Handwritten Devanagari Characters", IEEE 399 - 403 vol 2.
[16] G. S. Lehal and Chandan Singh, "Feature Extraction and Classification for OCR of Gurmukhi Script".
[17] Dharamveer Sharma, Puneet Jhajj, "Recognition of Isolated Handwritten Charactersin Gurmukhi S Volume 4– No.8, August 2010cript", International Journal of Computer Applications (0975 – 8887),
[18] U. Bhattacharya1, M. Shridhar, and S.K. Parui1, "On Recognition of Handwritten Bangla Characters".
[19] Arun K Pujari, C Dhanunjay Naidu, "An Adaptive Character Recognizer for Telugu Scripts using Multiresolution Analysis and Associative Memory",
[20] T V Ashwin and P S Sastry, "A font and size-independent OCR system for printed Kannada documents using support vector machines", Vol. 27, Part 1, February 2002, pp. 35–58.
[22] G. Tauschek, "Reading machine," U.S. Patent 2 026 329, Dec. 1935.
[23] P. W. Handel, "Statistical machine," US. Patent 1915 993, June 1933

**Mr. Aditya Raj** received his B.Tech in IT from U.P.T.U Lucknow, Uttar Pradesh, India in 2010.Currently, he is doing M.Tech in CSE from C-DAC Noida(Affiliated to G.G.S.I.P.U New Delhi), India. He is working on the project "OCR of machine printed Oriya script". His interest areas are Digital Image Processing, Theory of Computation, Data Structure, Operating Systems.

**Mr. Ranjeet Srivastva** received his B.Tech in IT from U.P.T.U Lucknow, Uttar Pradesh, India in 2008.Currently, he is doing M.Tech in IT from C-DAC Noida(Affiliated to G.G.S.I.P.U New Delhi), India. He is working on the project "Separation of Machine Printed and handwritten text in Hindi". His interest areas are Digital Image Processing, Theory of Computation, Data Structure, Computer Networks, Network Security and Operating Systems.

**Mr. Tushar Patnaik** (Sr. Lecturer/Sr. Project Engineer) joined C-DAC in 1998. He has thirteen years of teaching experience. His interest areas are Computer Graphics, Multimedia and Database Management System and Pattern Recognition. At present he is leading the consortium based project "Development of Robust Document Image Understanding System for Documents in Indian Scripts".

**Mr. Bhupendra Kumar** (Senior Technical Officer) joined C-DAC in 2005, he received his M.Tech.degree from IIIT Allahabad with the specialization in wireless communication and computing. His interest areas are Advanced Image processing, pattern recognition, computer network, wireless network, MANETs. Currently he is involved in project "Development of Robust Document Image Understanding System for Documents in Indian Scripts".


*Published By:*
*Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)*