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, Gurmukhi, 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. 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.

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 extraction 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.

1.5 Post-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 Buddhijra [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.


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-Netwrok (Neural Network) classification technique employing wavelet feature and achieved 91% accuracy at character level. M. Sagar, Shobha and Ramakanth [10] designed a syntactic 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).


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.
IV. ACKNOWLEDGEMENT

Our thanks to Madhury Yadav(MTECH) and Kirti Singh (MCA) for many suggestions and contributions.

REFERENCES


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 CS 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. 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 CS 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 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 IIT 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”.