Abstract

Aiming at the continuation, the development and innovation in different fields, usually the knowledge is preserved in handwritten or printed form, so that it may be circulated from generation to generation. Surprisingly, due to the development in the printing technology, the quantity of printed material has been increasing with every passing day. Searching and retrieving information from the printed and handwritten documents is time consuming and tedious task. With the help of optical character recognition technology, now it has become possible to scan document as an image and to make it editable and searchable for further information processing. The fundamental objective of character recognition system is converting non electronic documents into electronic documents which can be manipulated by word processing software. This technology enables a machine to recognize characters automatically with the help of characters’ features. It makes machine capable to mimic the human function of reading.

But still for many languages, robust character recognition system are not available and there is no efficient way to search through printed or handwritten scanned material quickly and efficiently. Urdu is one the languages which needs robust Character Recognition System to convert huge handwritten as well as printed data into editable form. Development of robust Urdu OCR system is lagging behind due to word and character segmentation dilemma. In this research work attempts have been made to study this problem domain and achieve few of the essential outcomes.

The availability of standard dataset is a primary requirement for developing any character recognition system. To the best of our knowledge and reviewed literature, no standard datasets of handwritten Urdu character is publicly available. Thus, a dataset of 107325 handwritten Urdu characters and 15900 handwritten Urdu
digits have been developed with the help of 159 writers. The Urdu characters’ shapes vary according to their position in the word i.e. initial, medial and ender. That’s why the designed dataset consists of the isolated Urdu characters, along with all possible shapes of each character.

The Skew Correction, Normalization and Noise Reduction techniques have been applied on scanned Handwritten Document Image. Document image farthest points, Centroid and moments are used for Skew Angle Detection, To remove the noise of document image, both spatial and frequency domain filters are applied. Max and Median filters, Histogram Equalization and frequency domain Gaussian Low-pass filter are used in combinations.

In Urdu Script there is hardly a difference between inter-word and intra-word vertical gap. Consequently the Urdu word segmentation is considered as a non-trivial task by most of the researcher. It is very difficult to judge if the two adjacent ligatures belonging to same or different words. For segmentation purpose, connected component analysis and projection profile methods are used together and it is observed that in case of Urdu script, most of the segmentation techniques work inadequately. It is because of horizontal overlapping, touching characters, slants, complex secondary strokes and cursive script.

Usually the Urdu Characters are written using two types of strokes, primary strokes and secondary strokes. The primary strokes are the main body of the characters and secondary strokes further describe the characters. The Urdu characters’ shapes change as per their position in word. Thus, the total shapes to be recognized are up to 135. With a wide range of character, the recognition becomes more challenging and needs extra computational requirement and much execution time. That’s why in this research work, all the characters have been grouped into two categories – 46 Single Stroke Characters and 88 Multi-Stroke Characters. They are again classified as isolated, initial, middle and ender.

In case of Multi Stroke character, primary and secondary component are separated thorough Connected Component Analysis. The primary stroke is said as a Ghost like Character, and it is common mostly in all the characters. The Primary component has normalized into 60 X 60 and further divided into 3 horizontal zones. Similarly secondary component is normalized into 22 X 22 and divided into 2 horizontal zones. Seven Moment Invariant features were calculated for each zone of
primary as well as secondary components. In this way 28 Moment Invariants features are extracted from primary stroke and 21 features from secondary stroke. The position of secondary stroke component (Above, Below or middle) is also used to recognize the character.

In case of single stroke character, an image is normalized into 60 X 60 and divided into three horizontal zones for features extraction. From each zone 7 Moment Invariant features and from whole image 7 Moment Invariant features were computed, in this way 28 Moment Invariant features were extracted from single image. Once MI features are extracted, the Support Vector Machines is used for classification and character is put into appropriate class. The Ghost Character based method has been found much adequate, due to which there was no need to compare each character with all 134 Urdu characters. Partitioning of character image into zones also leads toward better recognition rate.

In case of Offline Handwritten Urdu Isolated Single Stroke Characters 92.20% recognition rate has been achieved. It is observed that for non-isolated characters, the recognition accuracy decreases. For initial, middle and ender single stroke characters, the obtained recognition accuracy rates are 79.74%, 76.21%, 79.31% respectively. In case of Multi Stroke Isolated characters, 84.95% recognition accuracy is found. On the basis of the position of the Multi Stroke character, obtained recognition accuracy is 68.12%, 75.13% and 85.66% of initial, middle and ender respectively. For handwritten Urdu digits, the obtained recognition accuracy is 93.01%.