# Abstract

Pattern recognition is one of the most popular and useful field of research in computer science. Optical character recognition (OCR), especially hand-written character recognition, is one very interesting branch of pattern recognition. In hand-written character recognition, many inherent problems are encountered which are due to the differences in writings by different people resulting in similarity between characters belonging to different classes, variations among characters belonging to the same class and the complexities in the character patterns.

Script identification is a special type of character recognition problem. In order to choose the appropriate OCR algorithm, it is first required to determine the script used in the document. So, it plays a key role in automatic processing of document images in an international environment. The problem becomes difficult when more than one class of script are similar to each other. Further difficulty in script identification arises when the characters are hand-written. In this research work, we attempt to develop a sophisticated method to identify the script from only a string of few hand-written characters and so, making it applicable in a task-specific reader. The proposed scheme is based on

- shape description of character patterns for feature extraction using morphological bi-variate pattern spectrum moments.
- combined possibilistic (soft) and crisp approach for clustering, followed by detecting appropriate number of clusters for each character through examination of the proposed cluster validity functional and finally, fine tuning of cluster parameters.
- assigning possibilistic class label for each character in the input string and integration of class labels over all characters for final decision.

A large database which incorporates every possible variations, is required for training and testing of any hand-written character or script recognition system. In our work, we have described an analytic approach for character generation from given generative models. A character is split into several lines and arcs. A set of equations and parameters describing these lines and arcs, completely defines the character. The parameter values are randomly modulated to create variations in the structure of the character. The pattern is then subjected to affine transform and dilation by different transform matrices and structuring elements, respectively to bring variations in size, aspect ratio, orientation, position and thickness.

The efficiency of any pattern recognition system largely depends on the method used for extraction and selection of pattern features. In this thesis, we propose to use the concept of bi-variate pattern spectrum for feature extraction in a hand-written character/script recognition system. The bi-variate pattern spectrum, so obtained, is translation invariant while moment calculation from these pattern spectrum will yield size invariant features.

In our research, we have proposed a new algorithm for clustering that may be applied to a set of same characters to generate clusters for all the different structures present in that set. The proposed integration of the possibilistic Kohonen self-organizing feature map with the CM algorithm (PKSFM-CM) provides a scheme to overcome the underutilization problem in hard clustering and on the other hand, unlike soft clustering algorithms, is capable of forming crisp disjoint clusters. The proposed algorithm also has the provision to escape any local minimum, that arises when more than one cluster centers move towards a common point, through proper partitioning of the input data set. Thus, the algorithm may lead to the global minimum. We have also developed some fuzzy rules that control the rate of convergence in the soft mode of clustering. This results in an optimum number of iterations in the soft mode as necessary. Further computational savings can be achieved by using our proposed fast nearest neighbor search algorithm in the crisp phase of clustering. The optimum number of clusters that may be formed from the set is decided by our proposed cluster validity functional that is based on the total volume of the hyperellipsoids that enclose each of the clusters, average intra-cluster distances and the minimum of all the inter-cluster separation.

In this research work, we have proposed to use possibilistic classifier for our purpose. Given an input, the aim is to determine the possibility of it to belong to a particular class on the basis of the membership grade of the input in each cluster. However, an input may be rejected by the classifier if there is any ambiguity or the pattern does not belong to any class at all. Therefore, it is desired that a pattérn inside a cluster will have high membership in that cluster while that outside the cluster have a low membership. This can be achieved by making suitable choice of the parameters that control the slope of the membership function in a cluster. A tuning algorithm, that maximizes the recognition rate for the given set of training patterns, is developed for the the same.

Once the set of training patterns are available at hand, the recognition system may be trained to develop some classification rules. In our script recognition problem, the training patterns are available with known class labels (both script and character classes). As mentioned earlier, patterns belonging to the same character class may be different from each other and may exhibit different structures. The task is to find the different structures present in the set of characters belonging to the same class and may be accomplished using our proposed clustering technique. Thus, we may obtain one or more clusters from a class of character patterns. The union of all the clusters, corresponding to each of the characters of a given script, forms the complete set of clusters for that script. The centroids of all these clusters are representatives of that script class. In the classification phase, an input is fed to the classifier and its distances from all the prototypes are measured. The degree of belongingness of the input to each of the script class is the maximum of the membership grades of the input in all its representative clusters. The pattern is classified as belonging to that script class for which it has the maximum membership grade. However, sometimes it is not possible for the classifier to assign a class to an input document just by looking at a single character. This is due to the similarities in some characters belonging to different script or may be due to distortion in the character pattern. In such a case, we look for successive characters in the given character string and the average membership grades to different classes are calculated until the classifier is able to make a clear decision. The classifier may reject the input without classification if even with all the characters in the string the average membership grade is low or reveals ambiguity. Experiments on different sets of characters with different string length was carried out to demonstrate the performance of our scheme. Simulation results show that we can identify the script from just a small set of characters, although better classification rate with lower rejection rate is achieved when the number of characters in the string is increased. So, our proposed scheme proves to be very useful for handwritten script identification in task-specific readers.