A REVIEW PAPER ON CLASSIFICATION TECHNIQUES OF PATTERN RECOGNITION

Himani Gupta[1], Asia Mashkoor[2]

[1]*School of Computing Science and Engineering, Galgotias University,Uttar Pradesh, India*

[2]*School of Computing Science and Engineering, Galgotias University, Uttar Pradesh, India*

ABSTRACT- **Optical character recognition (OCR) is the conversion of handwritten, typed or impressed word images into a form that the computer can manipulate. There are various steps in OCR. One of them is classification. To do the Classification, we must have to compare a training data with many feature vectors. A classifier is needed to compare the feature vector of input and the feature vector of training data. This paper discusses about some classification techniques, which are use to recognize character or word, and about some related work which has been done.**

Keywords: *Optical Character Recognition, Segmentation, Classification*.

I. INTRODUCTION

An optical character recognition (OCR) is a system that involves reading text from scanned document images and translating the scanned document images into text. OCR is the conversion of handwritten, typed or impressed text images into a arrangement that the computer can manipulate. Optical Character Recognition, or OCR, is a procedure use to translate distinct types of records, such as images captured by a digital camera, PDF files or scanned paper documents into editable and searchable data. It takes the text character-by-character for photo scanning, and then it analyse scanned image, and then it translate the character image into character codes. OCR follows a process that includes six steps:

1) Image acquisition, 2) Pre-processing,
3) Segmentation, 4) Feature Extraction,
5) Classification and 6) Post-processing.

The aim of Image acquisition in image processing is to get an image from some source, usually a hardware-based source. The Pre-processing step is use to provide an improvement of the scan image data which contain unwanted deformation or improves some image attributes which are essential for next step. In Segmentation step, an image of a succession of characters is decomposing into sub images of single characters. These sub images is taken as input for next step of OCR, ie feature extraction. feature extraction step extracts the attributes of symbols. Feature extraction stage gives us the attributes, called feature vector, that is used for classification. Classification step takes the output of previous step as input which are feature vectors. To do the Classification we must have a training data to compare with many feature vectors. A classifier is needed to compare the feature vector of input and the feature vector of training data. The Post-processor step takes output of the classifier as input to decide on the recommended action. In Optical Character Recognition, there are various problems occurs. The issue in character identification can be categorized based on some criteria. One is based on the type of the text which is printed or hand written. The other is based on the acquisition process which can be on-line or off-line [2].

II. CLASSIFICATION TECHNIQUES

OCR systems follow procedures for assigning an unspecified symbol to a predefined class. There are various techniques to recognize a symbol. Some of them are as follows:

A. Template Matching.
B. Statistical Techniques.
C. Neural Networks.
D. Support Vector Machine(SVM)
E. Combination classifier.

Sometimes these approaches works as separately and sometimes works together. Sometimes combination of classifiers gives better result.

A. *Template Matching*

Template Matching is the easiest method of character recognition. It performs matching between the predefine symbols and the character or word to be recognized. It compares the feature vector of predefined dataset with the feature vector of input symbol. And this matching procedure discovers the level of resemblance between these two vectors (structure, set of pixels, curvature etc.) for applying this technique, firstly we need two components first is binary input character and second is predefined dataset. This technique is flexible and straight forward. Similarities and dissimilarities between a gray-level or binary input symbol and a standard set of predefined prototypes are estimated by this approach. This estimation is done by matching operation. corresponding to a similarity measure and match strength, a template matcher assign the input symbol to a class. The calculated recognition rate of this approach is very sensitive to distortion and image restoration. Some modification was performed in template matching for getting improvement in the result of classification. After modification, some other approaches were found, called Deformable Templates and Elastic Matching.

B. *Statistical Techniques*

Statistical decision approach is related to statistical decision consequences and a group of optimality measures, which increases the occurrence of the perceived shape given the structure of a certain category. These procedures are depend on some hypothesis. These are:

- Distribution of feature group is uniform in the trounce instance, otherwise Gaussian,
- There are adequate statistics accessible for every category,
- Given set of images is allowed to take out a group of attributes which shows every different category of symbols.

A symbol contains various features. And feature extraction step of OCR provides feature vector for input image. And in classification step each feature of input image is compare with the feature vector of training dataset. This measurements from n-features of input image can be use to show n-dimensional feature vector volume and a vector whose correlation represents original character or word unit. There are some significant statistical methods, which are beneficial for getting good result of OCR. They are Likelihood or Bayes classifier, Hidden Markov Modelling (HMM), Quadratic classifier, Nearest Neighbour (NN), Clustering Analysis, Fuzzy Set Reasoning.

C. *Neural Networks*

As human read the paragraph on his computer area, his eyes and brain bring OCR without his even attention. His eyes are identifing the shape of light and dark, those build the symbols (numbers, characters and word) printed on the area and his brain is using those to understand what I want to indicate (sometimes via taking single symbols or characters yet mainly via figure out whole letters and complete sets of letters suddenly).

Problem with Character identification is belonging to heuristic argumentation as person can identify symbol, characters, and records by their learning, knowledge and practice. Therefore neural networks with more or less heuristic in essence are exceedingly acceptable for this type of issue. There are various categories of neural networks that are utilized for OCR recognition.

A neural network is a framework that contains of analogous interconnection of adjustable 'neural' processors. Neural network can execute calculations at a admirable rate juxtapose to the classical techniques due to its parallel nature. Neural network can accommodate for modification in the data and assimilate the attributes of input signal because of its adaptive nature. One node takes the output of previous one as input in the network and the final resolution based on the compound interaction of all vertices.

Artificial Neural Network (ANN) models involves of the following some components:

i) Topology – the process in which an ANN model is assembled into layers and arranges in a way in which these layers are interlinked;

ii) Learning – the procedure by which detail is accumulated in the grid; and

iii) Recall – how the accumulated detail is recovered from the grid.

The basic construction of an ANN model involve artificial neurons. The neurons are also called as nodes, processing elements (PEs), neurodes, units, etc. And neurons are similar to biotic neurons in the person brain, which are assembled into layers (also called slabs). The simple ANN construction involves one or more than one hidden layers, an input layer, and an output layer.

D. *Support Vector Machine Classifier*

The SVM technique was first arrived at erstwhile AT&T Bell Laboratories by Vapnik and his group. Support vector machine is also called two-class classifier. It works on distance between the classes. This distance is called width of the margin. This distance is the separation to the closest training symbols. This is the free space throughout the decision boundary. Classification function is define by these training patterns. Theses training patterns called support vectors. Their number is decreased by increasing the perimeter. The support vectors exchange the originals with the main dissimilarity between support vector machine and an existing template matching procedures. they assign the categories by a decision boundary. This classifies use a linear differentiating hyperplane. This hyperplane increases the seperation between two categories for creating a classifier. This classifier deals with linear and non-linear separation as well. The working of SVM is similar with the working of a statistical learning converter that merges points of distinct classes from n-dimensional area into a higher dimensional area where the two classes are more distinct. SVM works for finding a best hyperplane in high dimensional area so that it provides best separation for the two classes of points. The hyperplane is find by the points that are located nearest to the hyperplane, called support vectors. It is not necessary to have only one support vector, it may have more than one support vector on every side of the plane.

The limitation of SVM classifier is that it classifies only two classes at a time. But practically, there are multiple classes to classify, so to classify multiple classes we require multiclass SVM. Multiclass SVM includes the construction of binary SVM classifiers for each pair of classes.

E. *Combination Classifier*

A classification method has its own advantages and disadvantages. A classifier does not give best result for all classification problem. So sometimes, for solving classification problem, combination of classifiers required. different classifiers follows different procedure so they are differ in their local and global performances. Each classifier, has in the feature space has its own area where it gives best result. Some classifiers like neural networks gives distinct outcomes for distinct initializations because of undirected inherent in the training process. Combination of various networks is use rather than selecting best network. So the advantages of all the attempts are taken out for solving classification problem. A set of classifiers contain different feature sets, different classification methods and different training sets. The output of all classifier is combined to improve overall classification accuracy. But it is not necessary that we will get improved result. There are various

combination schemes are use. A difficult combination procedure involves a group of seperate classifiers and a merger which combines the outcomes of the seperate classifiers to give the final conclusion.
Sometimes only two classifiers can solve classification problem, no need of third classifies. Sometimes three classifiers are required, sometimes more than three. So there are various methods to combine multiple classifiers. These methods are categories according to their structure. They are:

a) parallel, b) hierarchical (tree-like) and c) cascading (or serial combination)

i) Selection and Training of single Classifiers:

If the separate classifiers are largely autonomous then a classifier combination is mostly helpful. To get improved the classification rate, various techniques like bootstrapping and rotation can be use to create difference of training sets.

ii) Combiner

Combiner is a module which is use to combine classifiers when individual classifiers have been selected. Each combiner has some attributes, like adaptively, trainability. So each combiner can be different from another. Combiner also depends on the outcome of single classifiers. Some mergers are trainable and others are not. Combiners which does not require training are static, like averaging (or sum), voting, and borda count. The trainable combiners give better development than static mergers. The price of supplementary training and also the demand of supplementary training data is less in trainable combiners. Some amalgamation strategies are adjustive in the nature that the merger assesses (or weighs) the conclusions of separate classifiers based on the input sample. In contrast, no adjustable combiners behave same with all the input samples.
There are various composite classifiers. Some of them are used in Indian scripts. These are HMM and ANN, SVM and K-means, SVM and MLP, NN, fuzzy neural network, fuzzy logic and genetic algorithm, MLP and minimum edit.

III. LITERATURE SURVEY

There are various research papers, we have studied, on character recognition. Many classification techniques are used in those papers. An OCR system is developed to convert the scanned documents into text form. Various methods are use in each step of OCR. In many papers, various techniques for Feature Extraction and for character classification are explained. Our motive is to study and select the best classification technique approach. But not only one technique is best, because one technique can give the best result for one language but not for others. So we Observed results and other relevant issues for any latest research task on character identification. After our accurate analysis we were able to identify the theme of our suggested task incorporating the techniques which we have incorporated in different phases, to judge our outcome and conclusions.

Two pre-processing steps were discussed for an online autography identification structure, which are taking the size to its normal stage and maintaining the difference between two lines (slope correction) from their point of view while taking the size to its normal stage, the base line and the middle-line are required to be determined. The region encircled by the bottom-line and the mid-line is the only portion of any word which is every time non-empty. Formally exact determination of the base-line and the mid-line are acquired, a magnification element can be figure out from the ratio of the normal mid-portion size and that of input. The complete entered data may than be scaled using the acquired magnification element. Then the slope correction is done which depends on the valuation of the mean velocities x and y aiming. The angle between these velocity vectors is used to calculate the angle from which the words should be resolved. For specific cases such as in words with all cubic capital letters, specific provisions was picked out to keep out of improper calculation of the revolving angle.[1]

A technique is proposed for pre-processing that involves interpolation, restoration, smoothing and normalization or modification of strokes. For normalizing or modify a stroke authors first discover the bounding box of the whole stroke. After this, they cleave measurements of the stroke by the „height‟ of the horizontal block. This conserve the relative dimension of strokes in a horizontal block. The strokes are then metamorphoses onto curve length base and then smoothed seperately along t-axis with a Gaussian filter.[2]

A combination of various feature extraction procedures was used for handwriting Devnagari character recognition. Histograms of Chain code and unswerving line fitting attributes and weighted majority voting procedures for combining the recognition conclusion obtained from non-identical Multilayer Perception (MLP) establish classifier as well as they also utilized chain code histograms and moment dependant attributes to classify handwritten Devnagari characters. The procedure of chain code was discovered by identifing the orientation of the succeeding pixel in that line in the scaled shape image. Moment attributes were retrieved from scaled and less densed character image.[4]

A method was proposed for character recognition with 98.56% accuracy by using a simple neural network. In this method, data are initially rotated and normalized in the pre-processing step. Then, the image is divided into two parts in horizontal direction and the features are extracted by whisking data in the direction of row and column in each part of image. SOM neural network is used for recognizing and classifying characters in this paper [5].

A character recognition method on ID card using Template Matching Approach is also proposed in which text is detected on Indonesia ID card consist of some phases, which are pre-processing, extract the text regions, segmentation step and recognize the segmented area. In the segmentation stage, approximately 93% of character can be cut off correctly [6].

A multiple classifier system is presented for the recognition of offline Malayalam characters. Gradient features and density features were extracted from pre-processed character images to form feature vectors which were fed as input to two feed forward neural networks. The final results were obtained by combining these two neural networks using 4 combination strategies: Max rule, Borda count method, Sum rule and Product rule. The proposed system achieves a recognition accuracy of 81.82% using the Product rule combination scheme[7].

A Rectangle Histogram Oriented Gradient representation is use as the basis for withdrawal of attributes. These algorithms need a few easy mathematical actions per image pixel which formulates rules which are appropriate for real-time applications. In this paper, dataset contains of 8000 representatives each of 40 basic handwritten Marathi characters. All sample images of handwritten Marathi characters are modified to 20 × 20 pixel size. The explanation of the step by step rules and investigation with some data set is suggested in this paper. Trial outputs using Support Vector Machines (SVM) and feed-forward Artificial Neural Network (FFANN) classification techniques are presented. Outcomes demonstrate high interpretation of these attribute when classified using feed-forward Artificial Neural network, classification[8].

An efficient approach was proposed for document layout analysis for Indian newspapers. In this bottom up approach and top-down approach were used for segmentation. The words bounding box classification system for Hindi newspaper Pre-processing techniques utilized in Hindi chronicle images as an primary step in words classification systems were dispensed[14].

IV. CONCLUSION

There are various classification techniques. It is not necessary that every technique is best for every language. So these techniques give different results for different-different languages. The words bounding box classification system for Hindi newspaper Pre-processing techniques utilized in Hindi document images as a primary step in words classification systems were dispensed. The feature extraction pace of optical character recognition is very indispensable. It can be utilized with existing OCR procedures, mainly for English word. The words of Hindi newspaper can be classify by using inverse support vector machine before it need to apply canny edge detection to detect the edge of character (often involved in a digital image) more similar to alphanumeric or other symbols. The process of OWR(optical words recognition) contains various steps containing segmentation, feature extraction, and classification. This process use Image Processing Toolbox to get all interrelated techniques which are stated in citations are examined and the limitations are being decreased and therefore obtaining an upgraded model of the earlier task.

REFERENCES

[1] Homayoon, S.M., Beigi, K. N., Gregory J. Clary, and Subrahmonia. J., 1994. "*Sizenormalization in on-line unconstrained handwriting*", IEEE Journal 0-8186-6950-
0194.

[2] Aparna, K.H., Subramanian, V., Kasirajan, M., Prakash, G. V., Chakravarthy, V.S., and Madhvanath, S., 2004. "*Online handwriting recognition for Tamil. Ninth International Workshop on Frontiers in Handwriting Recognition*", pp. 438-443.

[2] Araki, N., Okuzaki M., Ishigaki, K. H., 2008. "*A Statistical Approach for Handwritten Character Recognition Using Bayesian Filter*", 3rd International Conference on Innovative Computing Information and Control (ICICIC), pp. 194-198.

[3] Hosny, I., Abdou, S. and Fahmy, A., 2011. "*Using Advanced Hidden Markov Modelsfor Online Arabic Handwriting Recognition"*, First Asian Conference on Pattern Recognition, pp.565-569.

[4] Arora, S., Bhattacharjee, D., Nasipuri, M., Basu, D. K. and Kundu, M., 2008. "*Combinig Multiple Feature Extraction Techniques for Handwriting Devnagari Character Recognition*", Industrial and Information Systems, IEEE Region 10 Colloqium and the Third ICIIS, pp. 1-6.

[5] Najmeh Samadiani and Hamid Hassanpour *"A neural network based approach for recognizing multifont printed English characters",* Journal of Electrical Systems and Information Technology (JESIT),2015.
[6] Michael Ryan, Novita Hanafiah *"An Examination of Character Recognition on ID card using Template Matching Approach",* International Conference on Computer Science and Computational Intelligence (ICCSCI 2015).
[7] Anitha Mary M.O. Chacko, Dhanya P.M. *"Multiple Classifier System for Offline Malayalam Character Recognition",* International Conference on Information and Communication Technologies (ICICT 2014).
[8] Parshuram M. Kamble, Ravinda S. Hegadi *"Handwritten Marathi character recognition using R-HOG Feature",* International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

[9] Dungre, V. J. *et al.*, 2010. "A *review of Research on Devnagari Character recognition*", International Journal of Computer Applications, vol. 12, No, 2, pp. 8- 15.

[10] Guerfali, W. and Plamondon, R., 1993. "*Normalizing and restoring online handwriting*", Pattern Recognition, Vol. 26, No. 3, pp. 419, 1993.

[11] Kumar, A. and Bhattacharya, S., 2010. "*Online Devnagari Isolated Character Recognition for the iPhone using Hidden Markov Model*". Proceedings of the IEEE Students‟ Technology Symposium, pp. 300-304.

[12] Lehal, G. S., Singh, C., 2000. "*A Gurmukhi Script Recognition System*", Proceeding of International Conference on Pattern Recognition, Vol. 2, pp. 557-560.

[13] Lehal, G. S., Singh, C., 2002. "*A post Processor for Gurmukhi OCR*", *Sadhana,* Vol. 27, Part 1, pp. 99-111.

[14] Vijay singh and Bhupendra kumar "*document layout analysis for Indian newspapers using contour based symbiotic approach",* 2014 International Conference on Computer Communication and Informatics (*ICCCI* - 2014),IEEE.