JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-2, Issue-7 E-ISSN: 2347-2693

Recognition of Degraded Printed Gurmukhi Numerals- A Review

Nishu Goyal[1*] and Seema Garg[2]

[1,2]*Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India*
nishugoyal89@gmail.com; garg_seema238@yahoo.co.in
www.ijcaonline.org

Received: 26 /06/ 2014 Revised: 8 /07/ 2014 Accepted: 15 /07/ 2014 Published: 30 /07/ 2014

Abstract- OCR is optical character recognition. It is the prominent area of research in the world. It is translation of scanned images of handwritten, typewritten or printed document into machine encoded form. This machine encoded form is editable text and compact in size. OCR is a common method of digitizing printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as machine, text-to-speech and text mining. Many OCR's have been designed which correctly identify fine printed documents both in Indian and foreign scripts. But little reported work has been found on the recognition of degraded Gurmukhi script. The performance of standard machine printed OCR system working for fine printed documents decreases, if it is tested on degraded documents [8]. The degradation in any document can be of many types. A major issue that leads in degraded printed numerals is heavily printed character, broken character, and background noise problem and shape variance character [10]. Although humans can read these documents easily, it is far complicated for computers to recognize them. So, our main focus will be to make the system recognize degraded printed Gurmukhi numerals.

Keywords- Optical character recognition, Degraded Gurumukhi Numerals, Printed Documents.

I. INTRODUCTION

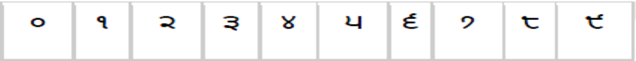
1.1 Introduction to OCR

Optical character recognition is the important area of research in the world. OCR is generally an "offline" process, which analyzes a static document. It is the conversion of scanned images OCR is optical character recognition. It is the prominent area of research in the world. It is translation of scanned images of handwritten, typewritten or printed document into machine encoded form. This machine encoded form is editable text and compact in size. OCR is a common method of digitizing printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as machine, text-to-speech and text mining. Many OCR's have been designed which correctly identify fine printed documents both in Indian and foreign scripts. But little reported work has been found on the recognition of degraded Gurmukhi script. The performance of standard machine printed OCR system working for fine printed documents decreases, if it is tested on degraded documents [8]. The degradation in any document can be of many types. A major issue that leads in degraded printed numerals is heavily printed character, broken character, and background noise problem and shape variance character [10]. Although humans can read these documents easily, it is far complicated for computers to recognize them. So, our main focus will be to make the system recognize degraded printed Gurmukhi numerals. of handwritten, printed document into machine encoded form. Numeral recognition can be applied on printed, type-written or handwritten text. Recognition for degraded numeral is more complex due to various noises, background problem, touching, broken, etc. Most commercial Optical character recognition systems are designed for well-formed business documents [6]. The basic mechanism of character recognition consists of following phases: Image Pre-processing, Feature Extraction, Classification and Post Processing [5].

1.2 Introduction to Gurmukhi Script [15]

Gurmukhi Script is used primarily for Punjabi language, which is the world's 14th most widely spoken language. Following are the properties of Gurmukhi Script[4] are:

i. Writing style is from left to right.
ii. No concept of upper and lower case characters.
iii. Gurmukhi script is cursive.



੦	੧	੨	੩	੪	੫	੬	੭	੮	੯

Figure 1: Gurmukhi numerals [14]

Gurmukhi Script has following challenges [16]:
i. Variability of writing style, both between different writers and between separate examples from the same writer overtime.
ii. Similarity of some characters.
iii. Low quality of text images
iv. Unavoidable presence of background noise and various kinds of distortions.

1.3 Degraded Printed Numerals

On analysis of the degraded printed numerals we have observed following problems [10]:

1. Broken Numeral Problem.
2. Heavy Printing Problem.
3. Shape Variance.
4. Background Noise Problem.

1.4 Feature Extraction [5]

The OCR engine consists of two stages, feature extraction and classification. Transforming the input data into the set of features is called *feature extraction*. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.The feature extraction stage analyses a text segment and selects a set of features that can be used to uniquely identify the text segment. The selection of a stable and representative set of features is the heart of pattern recognition system design.

1.5 Classification

The second step of OCR engine, is classification stage, in which objects are recognized, differentiated, and understood. It is the main decision making stage of an OCR system and uses the features extracted in the previous stage to identify the text segment according to preset rules. Classification is concerned with making decisions concerning the class membership of a pattern in question. The task in any given situation is to design a decision rule that is easy to compute and will minimize the probability of misclassification. If we assume that d features are observed on a pattern or object, then we can represent the pattern by a d-dimensional vector $X = (x1, x2... xd)$ and usually refer to X as a *feature vector* and the space in which X lies as the *feature space*. Patterns are thus transformed by feature extraction process into points in d-dimensional feature space.

II. LITERATURE SURVEY

Singh and Budhiraja (2012) presented an OCR (optical character recognition) system for the handwritten Gurmukhi numerals where recognition system the feature vector has lesser elements as compared to other OCR systems developed so far. The result obtained is comparable with similar works reported earlier. In this recognition system an average recognition rate of 88.8% has been obtained. It has been found that db1 and coif1 wavelets have given the highest recognition accuracy.[14]

Siddharth et al. (2011) proposed handwritten Gurmukhi character recognition for isolated characters. They have used some statistical features like zonal density, projection histograms (horizontal, vertical and both diagonal), distance profiles (from left, right, top and bottom sides). In addition, they have used background directional distribution (BDD) features. Their database consists of 200 samples of each of basic 35 characters of Gurmukhi script collected from different writers.[7]

Singh et al. (2011) created database by implementing pre-processing on the set of training data. Then by the use of Principal Component Analysis they extracted the features of each image, some researchers have also used density feature extraction. Since different people have different writing style, so here they are trying to form a system where recognition of numeral becomes easy.[13]

Rajashekararadhyag et al. (2009) proposed Zone and projection distance metric based algorithm on feature extraction system. The character /image (50x50) are further divided in to 25 equal zones (10x10 each). For each zone column average pixel distance is computed in Vertical Downward Direction (VDD) (one feature). This procedure is sequentially repeated for entire zone/grid/box columns present in the zone (ten features). [16]

Jindal et al. (2008) developed an OCR system for recognizing high quality machine-printed text can recognize words at a high level of accuracy. However, given a degraded text page, performance usually drops significantly. In this paper, author has discussed efficient structural features selected for recognizing degraded printed Gurmukhi script characters containing touching characters and heavy printed characters. These features are very much tolerant to noise. Author has identified some projection and profile features like directional distance distribution and transition features which handle noisy characters. For classification purpose K-NN and SVM classifiers are used. It is observed that maximum accuracy of 91.54% using SVM Classifier has achieved. [9]

Garg et al. (2007) provided a new feature set for handwritten digit recognition, which has structural features different from the features taken by most of the researchers like number of junctions, number of loops and number of endpoints etc. Firstly, they explained by experiments that slant invariant and size invariant features help in developing general software, which is free from some of the pre-processing steps. Secondly, they confirm that pixel counting technique is very useful for deformed images than contour following technique. SVM and Tree classifier are used for classification. Overall 90.3% handwritten digit recognition rate is achieved. [11]

III. STEPS FOLLOWED FOR OCR SYSTEM

Step 1 Pre-Processing: [10]

Pre-processing is representing the scanned image in binary format for processing the image for recognition. In this step we actually perform Binarization i.e. extracting foreground from original image. It converts the image in binary form i.e. 0 or 1. 0 means white and 1 means black pixel. Pre-processing aims to produce data that are easy for the OCR systems to operate accurately. It reduces noise and distortion, removes Background problem to much extent and also enhance the image and performs skeletonising of the image thereby simplifying the processing the rest of the stages.

Step 2 Feature Extraction Techniques [9],[12]

Feature extraction and selection can be defined as extracting the most representative information from the raw data, which minimizes the within class pattern variability while enhancing the between class pattern variability. For this purpose, a set of features are extracted for each class that helps distinguish it from other classes, while remaining invariant to characteristic differences within the class. A good survey on feature extraction methods for character recognition can be found in [14]. Various feature extraction methods are classified in two major groups:

1. Statistical Features [6][7]

2. Geometrical and Topological Features

Step 3 Classifiers [5]

In an OCR process classification stage assigns labels to character images on the base of features extracted and the relationships among them. In simple terms, this part of OCR recognizes individual characters and returns the output in character processing form.

The two basic phases of any classification problem are training and testing. In training phase, the classifier learns the relationship between samples and their labels from samples that been labelled whereas, in testing phase analyzing of errors in the samples is performed in order to evaluate classifier's performance. For better performance it is desirable to have a classifier with minimal test error. An introduction to these classifiers has been given in this section.

3.1 Support Vector Machines (SVM) [2]

SVM are based on statistical learning theory that uses supervised learning [3]. In supervised learning, a machine is trained instead of programmed, to perform a given task on a number of input-output pairs. According to this concept, training means choosing a function which best describes the relation between the inputs and the outputs. The standard SVM classifier takes the set of input data and predicts to classify them in one of the only two distinct classes. SVM classifier is trained by a given set of training data and a model is prepared to classify test data based upon this model. The central problem in statistical learning theory is how well it estimates the output for previously unseen inputs. In general, any learning problem in statistical learning theory will lead to a solution of the type

$$f(x)=\sum_{i=1}^{m} c_i \quad K(x,x_i)$$ [2]

Where, x_i, i = 1, …, m are the input examples, K a certain symmetric positive definite function named kernel, and c_i a set of parameters to be determined from the examples. A practical guide for SVM and its implementation is available at [3, 4]. Commonly used kernels are: *Linear kernel*, *Polynomial kernel*, Gaussian *Radial Basis Function* (RBF) and *Sigmoid* (*hyperbolic tangent*). The effectiveness of SVM depends on kernel used *K*, kernel parameters c_i. We have considered 2 kernel parameters that are linear and polynomial kernel. [9]

Working of SVM is mapping points of different categories from *n*-dimensional space into a higher dimensional space by using discriminant functions so that the two categories are further separable [9]. A discriminant function represents a surface that separates the patterns so that the two class's patterns lie on the opposite sides of the surface (*i.e.,* To find an optimal hyperplane in that high dimensional space that best separates the two categories of points.) as in Figure we consider square and dots are two classes and line separating the two classes with a support vector shown in circle.

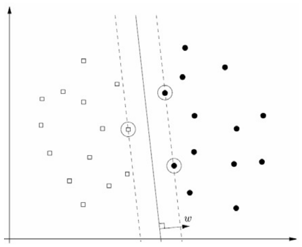


Figure2:Separating hyperplane with support vector [9]

3.2 K- Nearest Neighbour (K-NN) classifier [1],[8]

K-NN classifier uses the instance based learning by relating unknown pattern to the known according to some distance or some other similarity function. It classifies the object by majority vote of its neighbour. Because it considers only neighbour object to a particular level, it uses local approximation of distance function. 'K' specifies the number of nearest neighbours to be considered and the class of majority of these neighbours is determined as the class of unknown pattern and by default, it is '1'. [1]

IV. CONCLUSION

In this paper we conclude that recognition can be done by first representing image in binary form i.e. o or 1.Then features are extracted by statistical methods (like zoning, crossings and distances etc) and geometrical and topological features include (extracting and counting topological structures, coding, graphs and trees etc.)

REFERENCES

[1] A. Antonacopoulos and C. Casado Castilla "Flexible Text Recovery from Degraded Typewritten Historical Documents", *Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong*, pp. 1062-1065, 2006.

[2] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition", *Knowledge Discovery and Data Mining*, Vol. 2 No. 2, pp. 121-167, 1998.

[3] C.W. Hsu, C.C. Chang, and C.J. Lin, "A Practical Guide to Support Vector Classification", [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[4] D.Sharma and U.Jain "Recognition of Isolated Handwritten Characters of Gurumukhi Script using Neocognitron*", International Journal of Computer Applications,* Vol. 10 No.8, pp. 10-16, 2010.

[5] G. S.Lehal, and C.Singh, "Feature Extraction and Classification for OCR of Gurmukhi Script", *Vivek*, Vol. 12, No.2, pp. 2-12, 1999.

[6] K.S. Siddharth, R.Dhir and R.Rani, "Handwritten Gurumukhi Character Recognition Using Zoning Density and Background Directional Distribution Features", *International Journal of Computer Science and Information Technologies,* Vol. 2, pp. 1036-1041, 2011.

[7] K. S. Siddharth, M. Jangid, R. Dhir and R. Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features", *International Journal on Computer Science and Engineering (IJCSE)*, Vol. 3 No. 6, pp. 2332-2345, 2011.

[8] M. K. Jindal, R. K. Sharma and G. S. Lehal, "A Study of Different Kinds of Degradation in Printed Gurmukhi Script", *Proceedings of the IEEE International Conference on Computing: Theory and Applications (ICCTA'07), IEEE Computer Society USA*, pp. 538-544, 2007.

[9] M.K. Jindal, R.K. Sharma., G.S. Lehal ., "Structural Features for Recognizing Degraded Printed Gurmukhi Script", *International conference on Information Technology: New Generation, IEEE Computer Society*, pp. 668-673, 2008.

[10] M. Kumar, "Degraded Text Recognition of Gurmukhi Script", *Ph. D. Thesis, Thapar University Patiala*, 2008.

[11] N. K. Garg, and S.Jindal, "An Efficient Feature Set For Handwritten Digit Recognition", *15th International Conference on Advanced Computing and Communications, IEEE computer Society*, pp. 540-544, 2007.

[12] O.D.Trier, A.K. Jain and T.Taxt," Feature Extraction methods for character recognition – A survey", *appeared in Pattern Recognition*, Vol. 29, No. 4, pp. 641-662, 1996.

[13] P.Singh and N.Tyagi, "Radial basis function for handwritten devanagari numeral recognition", *International journal of advanced computer science and applications,* Vol. 2 No. 5, pp. 126-129, 2011.

[14] P.Singh and S.Budhiraja ., "Offline handwritten gurmukhi numeral recognition using wavelet transforms", *International journal modern education and computer science*, Vol. 8, pp. 34-39, 2012.

[15] P.Jhajj and D. Sharma, "Recognition of Isolated Handwritten Characters in Gurmukhi Script", *International Journal of Computer Applications*, Vol. 4 No.8, pp. 9-17, 2010.

[16] S. V. Rajashekararadhya, and P. V. Ranjan, "Zone based feature extraction algorithm for handwritten numeral recognition of kannada script", *Proceedings of International advance computing conference*, pp. 525-528, 2009.