



Improve The Character Detection System Based On Feature Extraction Algorithm

Jaswinder Kaur, Mrs. Rupinder Kaur

CSE & PTU, India

jass1153@gmail.com, rupy.wahla@gmail.com

Abstract— The character recognition is the major important part in the area of document analysis. Character Recognition could be evaluated on printed text and handwritten text. Printed texture could be from a good quality image. In this research work, we implemented in the OCR approach to improve the recognition of character with Classification approach. We work on filtration techniques to improve the pixel quality of the punjabi character images. In bilateral filter works binary pixels could be close to single another, i.e occupies nearby locations they could be same to another that is having close values possibly in a perceptually meaningful manner. An inverse filter is used to recompense for the effect of unwanted structure filtering of signals. The quality of the character is improving and increase the accuracy rate due to presence of characters in the image. To enhance the accuracy with the help of a classification approach i.e FFNN. In this approach work in two phases i.e Training State and Testing State. In training State to evaluate the performance based on trained features in the punjabi character image. We decided the epoch value is 100 means, train the punjabi characters in this given reputations. After that train the features, then simulation model for analysis the features in testing stage. From the existing research, it is clear that detect the punjabi characters are accuracy better than text detection systems in terms of security, accuracy, performance and image efficiency and quality. Existing research proposed optical character recognition systems that are detecting the punjabi character, but doesn't improve the accuracy parameter. Tool used for simulation is MATLAB.

Keywords— Optical Character Recognition (OCR), Classification, Digital Image process, Feed Forward Neural Network(FFNN),Support Vector Machine(SVM).

I. INTRODUCTION

Image Processing is a technique for changing an image into a digital form and then performs some operations, so as to urge some important useful data from it. Nowadays, image processing is one among the fastest growing technologies. It is becoming the area of research in computer science and engineering disciplines [5].

Following steps are the basic and important steps in image processing:

- Firstly Import an image with the help of some image acquisition tool;
- Analysing and manipulation of an image;
- Output can be modified image or in the form of report based on analyse of image.

It has various uses in different fields which are as:

1. Medical Imaging
2. Textiles

3. Military
4. Film Industry
5. Document processing
6. Printing Industry etc.

Character Recognition

Abbreviated to optical character recognition & shortened to OCR, is a mechanical/electronic transformation of hand-written, printed text which is usually captured with scanner into a machine editable text. For document input tasks, character recognition is becoming more efficient and speedy. Different examples of OCR applications are here. OCR is the people’s first wish for converting a text document to some digital representation.

1. People desire for scanning documents and making that text available as a word document.
2. In Recognition of license plate number [6].
3. In Post Offices for recognition of zip-codes.

Procedure:

- Learning from corresponding text files and an image files
- Extracting and isolating individual character in an image
- Determine the attributes of characters extracted.
- Comparison between the properties of learned character and extracted character
- If no good match found then some additional operations on extracted characters are done.

A. Optical Character Recognition

Optical Character Recognition deals with the matter of recognizing optically processed characters. Optical recognition is performed off-line once the writing or printing has been completed, as critical on-line recognition wherever the pc acknowledges the characters as they're drawn. Each hand written and written characters could also be recognized, however the performance is directly dependent upon the standard of the input documents.

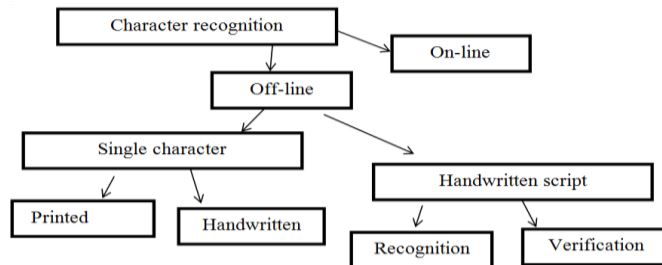


Figure no. 1 Different areas of character

A lot of affected the input is, the higher can the performance of the OCR system be. However, once it involves whole at liberty handwriting, OCR machines are still an extended approach from reading still as humans. However, the pc reads quick and technical advances are regularly transferral the technology nearer to its ideal [12].

Although, OCR machines became commercially offered already within the 1950’s, solely a couple of thousand systems had been oversubscribed worldwide up to 1986. The most reason for this was the value of the systems. However, as hardware was obtaining cheaper, and OCR systems began to become offered as package packages, the sale multiplied significantly. These days variety of some thousand is that the number of systems oversubscribed hebdomadally, and also the price of AN Omni font OCR has born with an element of 10 each alternative year for the last half dozen years

B. FEED FORWARD NEURAL NETWORK

The feed forward neural network was the primary and humblest kind of artificial neural network devised. During this system, the information moves in barely a way, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There aren't any cycles or loops within the network [11].

C. Feature Extraction using Principle Component Analysis

It is how of classifying patterns in data, and articulates the knowledge in such how on underline their resemblances and changes. 1st of all, we tend to had to create the knowledge set. The be after is to make a

decision a glorious amount of images and a decent purpose of those so as to possess the foremost glorious appreciation with the fewest record. Then, consecutive step is to deduct the mean from every of the information dimensions. The denote take from is largely the conventional across each dimension. The step 3 is to calculate the variance matrix of the record. we tend to couldn't confirm the variance matrix of the primary medium, since it had been too vast. thus we tend to had to seek out a technique to get the foremost eigenvectors barren of artful the massive variance matrix. The technique consists in selecting a unique variance matrix. Our variance middle on behalf of A was referred to as D and D is outlined by[10]:

$$C = D^* D'$$

The Manfred Eigenvectors & the Eigen standards of Care the foremost mechanisms of our information set.

PCA Algorithm: Principal part associate analysis (PCA for small) involves a pure mathematics method that transforms variety of presumably correlate variables into a smaller figure of unrelated variables referred to as main mechanism. The primary principal part, accounts for the maximum amount of the inconsistency within the data as probable, & each following part accounts for the maximum amount of the remaining variability as attainable. For an information matrix, by zero mean, wherever each row represents a distinct repetition of the experiment, & each support offers the marks from associate exacting investigate, the PCA transformation.

D. SVM CLASSIFIER

Support Vector Machine (SVM) also called Support Vector Networks are overseen learning models that examine data and recognize patterns. SVM models represent examples as point in space mapped in method that separate groupings examples are divided by a gap thereby performing linear organization. Apart from this SVMs can also achieve nonlinear classification using Kernel trick. The main idea of SVM is that; it discovers the optimal unraveling hyper plane such that error for unseen patterns is minimized. Consider the problematic of separating the set of training vectors going to two separate classes [39].

The SVM embodies many important principles. It solves the problem of classification directly without trying to solve the much harder problem of estimating the distribution of data samples. It provides efficient means of trading the training error for generalization error. Furthermore, even in the nonlinear case, the very central minimization task is stated as a convex optimization problem for which efficient numerical methods of finding the globally optimum solution exist. The SVM uses two main ideas. First, kernel functions are used to transform the problem from the original input space into a highly dimensional one, called the feature space, where linear separation of training samples belonging to different classes is possible. Second, to find the best separating hyperplane, the concept of maximum margin is introduced. Finally, the optimization problem which defines the SVM is convex and quadratic, and therefore it can be solved efficiently.

II. METHODOLOGY EXPLANATION

The major steps concerned in recognition of characters are:-

Step 1: First, we search the dataset form the internet. We select the punjabi characters database from the UCI machine Learning Repository.

Step 2: Upload the Image which we are using text base image.

Step 3: In gray scale (or gray level) image is simply one in which the only colors are shades of gray. The reason for differentiating such images from any other sort of color image is that less information needs to be provided for each pixel. In fact a 'gray' color is one in which the red, green and blue components all have equal intensity in RGB space, and so it is only necessary to specify a single intensity value for each pixel, as opposed to the three intensities needed to specify each pixel in a full color image.

Step 4: After gray scale conversion, we applied the canny property used for edge detection. The edge detection operator that uses a multi-stage algorithm to detect a wide range of edges in images.

Step 5: Remove noise with the help of inverse filter. In inverse filter can create a good model of the blurring function that corrupted an image, the quickest and easiest way to restore that is by inverse filtering. Unfortunately, since the inverse filter is a form of high pass filer, inverse filtering responds very badly to any noise that is present in the image because noise tends to be high frequency.

Step 5: Apply Optical character recognition scheme. OCR is the recognition of printed or written text characters by a computer. This involves photo-scanning of the text character-by-character, analysis of the scanned-in image, and then translation of the character image into character codes, such as ASCII, commonly used in data processing.

Step 6: Feature Extraction: we implement the component- based approach used for feature extraction means find unique properties (Eigen values and Vectors) in the image.

Step 6 : We implement the feed forward neural network used for classification approach and detect the Gurumukhi characters from the image.

In FFNN generates the two phases:

- Training Phase
- Testing Phase

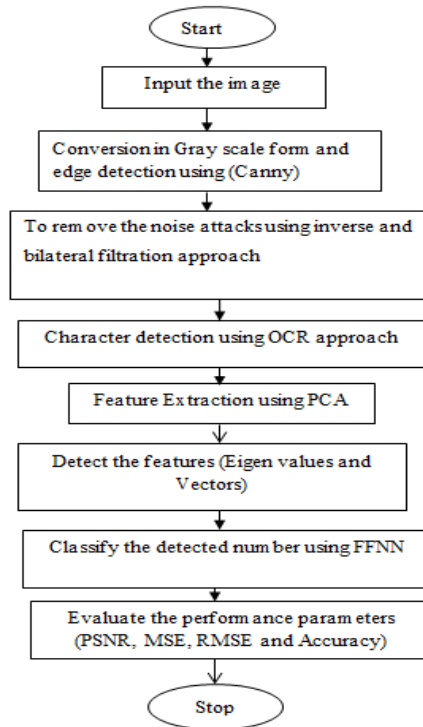


Figure no. 2 Flowchart of proposed work

III.RESULT

These parameters False rejection rate, Accuracy, false acceptance rate, Mean Square error Rate and Peak Signal to Noise Ratio are used as below:-

- A. *False rejection rate* is one of the result parameter which used to measure the performance of any algorithm. The high rate of false rejection rate degrades the performance of an algorithm. Here in this it reduced with the use of hybrid approach to enhance the accuracy rate and other performance parameters.

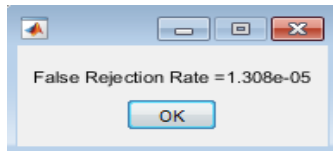


Figure no:3 False rejection rate

- B. *Accuracy* parameter used to measure the working capability of system in terms of positive classification. Here in this figure accuracy level is 99% which shows the enhancement from the other existing approaches.

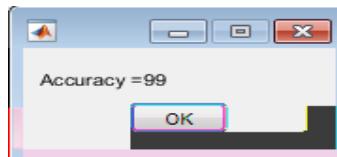


Figure no:4 Accuracy

- C. *The false acceptance rate* is used to define the working accuracy of the proposed approach. Here in this figure the performance of proposed approach is higher than other algorithms. The less amount of FAR provide high accuracy rate.

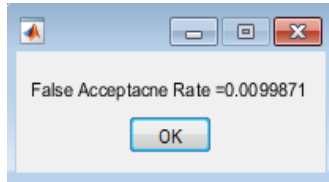


Figure no:5 False acceptance rate

D. *Mean Square error Rate* :The figure defined that the performance of the Mean Square error Rate in OCR-FFNN.

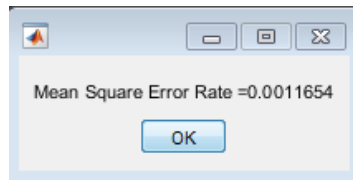


Figure no. 6 Means Square Error Rate in OCR-FFNN

E. *Peak Signal to Noise Ratio* :The figure defined that the quality of the image of the Peak Signal to Noise Ratio in OCR-FFNN.

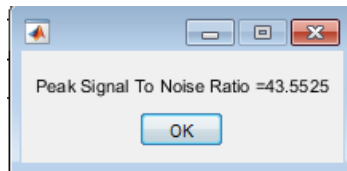


Figure no. 7 Peak Signal to Noise Ratio in OCR-FFNN

The table shows that the comparison of the proposed and exiting work . In proposed work we implement the feed forward neural network and existing we used to support vector machine types (RBF, LINEAR and POLYNOMIAL). We improve the accuracy rate with Feed Forward Neural network,

TABLE 1 COMPARISON OF ACCURACY

Number of Iterations	SVM-RBF	SVM-Linear	SVM – Polynormial	FFNN (Proposed)
10	7.8	8.7	8.9	9
20	18	17	17.5	18
30	29	28	28	29
40	37	37	36	37
50	48	47	48	49
60	58	57	57	59
70	67	66	69	66
80	78	77	78	78
90	87	86	89	90
100	97	95	96	99

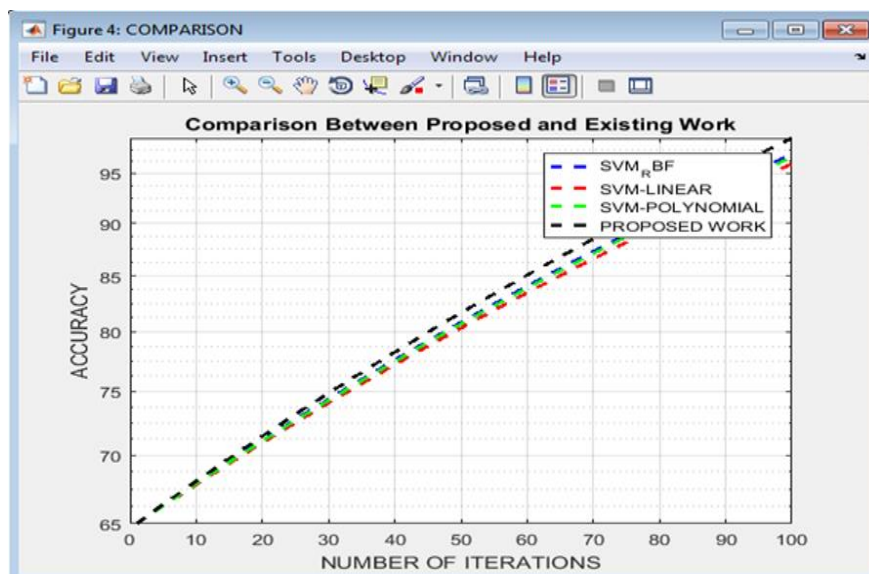


FIGURE NO 8: COMPARISON BETWEEN PROPOSED AND EXISTING WORK

IV. CONCLUSION

The complete solution for recognizing character for the degraded printed punjabi Character has been proposed in this thesis work. Various typed of deprivations are found in punjabi characters, normally in not clear character. Degradation might be due to defects in thesis work, issues introduced during handwritten during copying. Dissimilar types of degradations available in punjabi Character have been discussed by us in this research work. Sourced of each type of degraded character in punjabi script with the corresponding degraded character in Roman Script.

Normally, three phases, adding filtration, edge detection, Feature Extraction and Classification have been analyzed in searched in detail for recognition degraded character of punjabi in this work. To check the noise level in the punjabi characters. Applied the filtration process based on two processes i.e Bilateral and Inverse filter. Then, after that we implemented feature extraction approach to identify the unique features i.e eigenvector and values. PCA algorithm to use extract the features in matrix forms.

The Feed Forward Neural Networks, have been used for classification purposes. After that we evaluate the performance i.e FAR, FRR, MSE and Accuracy.

REFERENCES

- [1] "A Review on the Various Techniques used for Optical Character Recognition", Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi/International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 2, Issue 1, Jan-Feb 2012.
- [2] "Pre-processing Techniques in Character Recognition", Yaseer Alginahi, 17, ISBN 978-953-307-105-3, August, 2010.
- [3] Singh, Guriqbal, and Vikas Mongia. "Recognition Of Punjabi Script Character And Number For Multiple Fonts."
- [4] "Document Analysis and Recognition", Alon, Jonathan, 2005. Eighth International Conference on 29 Aug.-1 Sept. 2005.
- [5] "Personal Authentication Using Signature Recognition", Diana Kalenova, Department of Information Technology, Laboratory of Information Processing, Lappeenranta University of Technology.
- [6] "An introduction to biometric recognition", A.K. Jain, A. Ross, and S. Prabhakar, IEEE Trans. on Circuits and Systems for Video Technology, vol. 14, no. 1, pp. 4–20, January 2004.
- [7] "Signature Verification: Increasing Performance by a Multi-Stage System", Sansone and Vento, Pattern Analysis & Applications, vol. 3, pp. 169–181, 2000.
- [8] Rhead, Mke, "Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems." IEEE International Carnahan Conference on Security Technology (ICCST), 2012.
- [9] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." IEEE International Conference on Document Analysis and Recognition, 2012.

- [10] Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network", IEEE, 2012.
- [11] Ntirogiannis, Konstantinos, Basilis Gatos, and Ioannis Pratikakis. "A Performance Evaluation Methodology for Historical Document Image Binarization." IEEE International Conference on Document Analysis and Recognition, 2013.
- [12] Malakar, Samir, et al. "Text line extraction from handwritten document pages using spiral run length smearing algorithm." IEEE International Conference on Communications, Devices and Intelligent Systems (CODIS), 2012
- [13] Lim, Jae S., "Two-Dimensional Signal and Image Processing", Englewood Cliffs, NJ, Prentice Hall, 1990, pp. 469-476.
- [14] P. K. Sahoo, S. Soltani, A.K.C Wong and Y C Chen, "A survey of Thresholding Techniques", Computer Vision, Graphics and Image processing, vol 41, pp 233-260, 1988.
- [15] Otsu.N, "A threshold selection method from gray level histograms", IEEE Trans. Systems, Man and Cybernetics, vol.9, pp.62-66, 1979
- [16] L. Lam, S.W. Lee and C.Y.Suen, "Thinning Methodologies: A Comprehensive Survey", IEEE Trans. Pattern Analysis and Machine Intelligence, vol.14.pp 869-885,1992
- [17] Richard G. Casey And Eric Lecolinet " A Survey Of Methods And Strategies In Character Segmentation", IEEE Trans. On Pattern Analysis And Machine Intelligence, Vol 18, Pp 690-706,1996
- [18] Trier.O.D, Jain.A.K and Taxt.J, "Feature extraction methods for character recognition - A survey", Pattern Recognition, vol.29, no.4, pp.641-662, 1996.