ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Character Recognition: A Signature Approach

Arusa Firdous
M. Tech Student
Department of Computer Sciences,
Swami Devi Dyal Inst. of Engg. & Technology
Kurukshetra University, Kurukshetra, India

Neha Pawar
Assistant Professor
Department of Computer Sciences,
Swami Devi Dyal Inst. of Engg. & Technology
Kurukshetra University, Kurukshetra, India

Muheet Ahmed Butt
Scientist, PG Department of Computer Sciences, University of Kashmir, Srinagar, India

Majid Zaman
Scientist, Directorate of IT&SS. University of Kashmir, Srinagar, India

Abstract: Optical Character Recognition includes interpretation of meaningful information pertaining to a character from a digitized image in which scanned images of handwritten, typewritten text are converted into relevant machine text. The Character Recognition of both computer typed and handwritten characters has still a long way to go in terms of research. Although significant success has been achieved in type written characters but in handwritten it is still to touch an appreciable level. Most of the methods that have been proposed in this regard have huge computational complexity. The proposed research introduces anapproach of character recognition which besides producing better results based on signatures of histograms for each character to be recognized has very less computational complexity as comparted to the other methods. The proposed research provides segmentation, classification and recognition of characters which are independent in size and texture and the method proposes methods that are able to accommodate character styles which have slight variations and also does not require thinning and other preprocessing measures as is required in other approaches.

Keywords : Character set, projections, recognition, features, document processing, Histogram, Digitization.

INTRODUCTION

The character recognition has been a very important area of research of Image Processing field over the past three decades. The prominence of the character recognition has presumeda very high significance ever since the office automation research projects have been taken up. Presently the character recognition forms one of the most important activities in document processing in any organization. Considering the fact that different languages have different character sets therefore intense research has been going on for automating their recognition during document processing [1][2] using various standard techniques.

The conventional methods used for the recognition of the characters mostly use a matrix based approach [3] where each character is divided into a predefined number of rows and columns. Then depending upon the character under process a particular set of cells in horizontal and vertical direction is selected. Similarly another approach called connected component [4] traces the character under process from one end to another to find its different parts. Approaches of this nature involve excessive computations and are mostly time consuming. In these conventional approaches some pre-processing like thinning is required before actually taking up the actual character for recognition [5].

This paper presents an efficient procedure for character recognition, besides being fast is simple to implement using reasonable computational facility. The method used, transforms the character set into a new one by way of projections and the technique is called as Signature Analysis which a unique signature is identified for each character. These projections are based on pixel density and every character has a different pixel density value. The character recognition decision is made on the basis of the signatures obtained from the images based on pixel density. Similarly the objective of drawing the projections should be to get as small set of distinct features as possible. A knowledge base is created which stores the features of different characters and is always referred while making the character recognition decision.

THE CLASSIFICATION PROCESS

Classification in general for any type of classifier which is used in training a testing of extracted features of characters. These main steps of Training and Testing can be further broken down into sub-steps.

1. Training involves
   - i Pre-processing – In prepressing data image is kept ready for the OCR process eliminating noise if any in the image.
   - ii Feature extraction – Involves extraction of various dominant features which in turn reduce the amount of data extracting relevant information in the form of vector and scaler values. Further normalization can also be performed on features for distance measurements.
   - iii Model Estimation – from the finite set of feature vectors, need to estimate a model (usually statistical) for each class of the training data.
2. Testing
   - i. Pre-processing – In prepressing data image is kept ready for the testing process by performing certain basic operations.
   - ii. Feature extraction - Involves extraction of various dominant features that are involved in testing purpose.

iii. Classification – Compare feature vectors to the various models and find the closest match by using distance measures.

Figure 1: The pattern classification process

OCR – PRE-PROCESSING

These are the pre-processing steps often performed in OCR

i Binarization – Usually presented with a grayscale image, binarization is then simply a matter of choosing a threshold value.

ii Morphological Operators – Remove isolated specks and holes in characters, can use the majority operator.

iii Segmentation – Check connectivity of shapes, label, and isolate. We have used usedMatlabbwlabel and regionprops functions.

iv Segmentation is by far the most important aspect of the pre-processing stage. It allows the recognizer to extract features from each individual character.

LITERATURE SURVEY

Claudiu et al. [6] has investigated using simple training data pre-processing gave us experts with errors less correlated than those of different nets trained on the same or bootstrapped data. Hence committees that simply average the expert outputs considerably improve recognition rates.

Georgios et al. [7] has presented a methodology for off-line handwritten character recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the character image so that the resulting sub-images at each iteration have balanced (approximately equal) numbers of foreground pixels, as far as this is possible. Feature extraction is followed by a two-stage classification scheme based on the level of granularity of the feature extraction method.

Devlin et al. [8] has discussed that when performing handwriting recognition on natural language text, the use of a word-level language model (LM) is known to significantly improve recognition accuracy. The most common type of language model, the n-gram model, decomposes sentences into short, overlapping chunks.

Al-Khaffaf et al. [9] has presented the current status of Decapod's English font reconstruction. The Potrace Algorithm with its parameters affect glyph shape is examined. The visual fidelity of Decapod's font reconstruction is shown and compared to Adobe ClearScan.

Rhead et al. [10] has considered real world UK number plates and relates these to ANPR. It considers aspects of the relevant legislation and standards when applying them to real world number plates. The varied manufacturing techniques and varied specifications of component parts are also noted.

Majida Ali Abed Hamid Ali Abed Alasadi [11] considers a new approach to Simplifying Handwritten Characters Recognition based on simulation of the behaviour of schools of fish and flocks of birds, called the Particle Swarm Optimization Approach (PSOA). The input samples and database samples which improves the final recognition rate. Experimental results show that the PSOA is convergent and more accurate in solutions that minimize the error recognition rate.

Mohammed Z. Khedher, Gheith A. Abandah, and Ahmed M. Al-Khawaldeh [12] describe that Recognition of characters greatly depends upon the features extracted and the generation of the knowledge base. The research proposes an off-line recognition system based on the selected features was built. The proposed system was trained and tested with realistic samples of handwritten Arabic characters. Evaluation of the importance and accuracy of the selected features is made.

Amir BahadorBayat [13] Automatic recognition of handwritten characters has long been a goal of many research efforts in the pattern recognition field. This paper investigates the design of a high efficient system for recognition of handwritten digits. First it proposes an efficient system that includes two main modules: the feature extraction module and the classifier module.

PROPOSED OCR PROCESS

The characters that were to be to be recognized are digitized by using any commonly used scanner using an image processing software. Initially a knowledge base is created which contains the features of each character of the language that needs to be recognized. The selection of the features set of every character under observation forms one of the most important tasks in the modern character recognition system. The proposed research tries of optimize the features selection so that the recognition process does not become complex which leads to slow recognition. The selection of large number of features also results in un-necessarily results in a complicated recognition procedure whereas an incomplete set of features may result in incorrect recognition system.

Therefore selection of a set of features from histogram signature projections obtained after binarization of the segmented character is an important factor in the proposed research for developing an accurate character recognition system. The aim in the selection of the features should be to develop an efficient recognition system involving minimum computations and comparisons. The different potential attributes in the proposed research for selection of feature set is can be considered are as under:

a) Various histogram projections for various characters
b) Feature extraction set for Various Sizes of characters
c) Uniqueness in feature set

On the basis of the above attributes the different characters are classified into different feature groups. The classification of characters is based on generation of unique feature attributes of the resultant character images obtained afterbinarization and histogram projects. The figure 2 represents the histogram projections of the all the English characters. The projection parameter groups so formed are

on the basis of intensity of zero and non zero attributes in the character images which have been unique for each character.

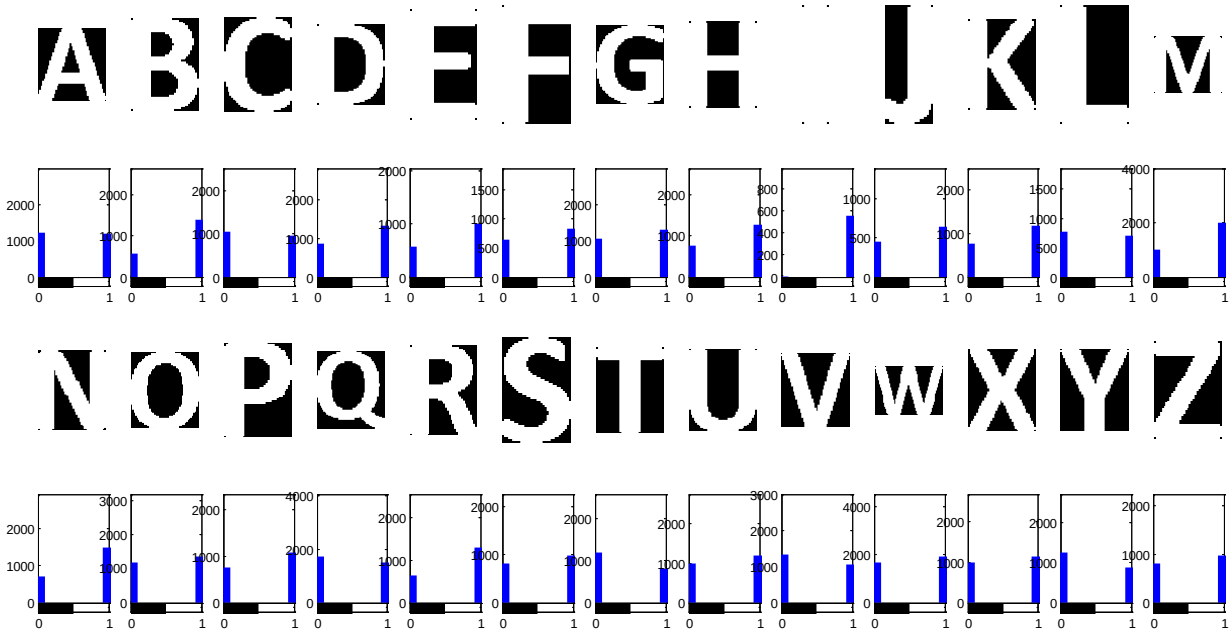

Figure 2: Histogram Representation of the Binary Character Image

FORMATION OF THE KNOWLEDGE BASE AND RECOGNITION

This is the most essential step in developing anefficient character recognition system. The knowledge base or training set is developed to store the features of characters or various sizes and styles and is always referred during the proposed recognition process. Therefore the development of a wider inclusive knowledge base is required for achieving accurate results during recognition of characters. In order to develop the required knowledge base, the basic character set of the language is digitized and the features are stored in a knowledge base by creating an efficient training set. A proper mapping of training values is also performed in the created training set which will ensure the proper outputs of the recognition process. A two dimensional array using Matlab is used to store the basic and mapped feature values of each character. The values so obtained are stored in an array of structures where each record is used to contain the information of one character. This ensures that the size of the character set and array are thus optimizing the features set and lowering the complexity in computation and comparison.. The following algorithm is used for the creation of the knowledge base:

1. Start of Algorithm
2. Load the image in a Array Data Structure
3. Binarizationthe complete character set image
4. Perform Image Segmentation so that each character is isolated in a different data structure.
5. Count the Number of 0's and 1's for every segmented Image
6. Use Histogram projections to show projections for every character
7. Generate a Training Set for Each Size and Texture of Characters using 0's and 1's pixel density in the image.
8. Now store the actual characters to be recognized in the separate image file and load the image in separate array.
9. Binarization of the input character image for recognition and perform the segmentation process for the image so the each character is isolated from the main image.
10. Retrieve the feature set as done in Step 7.
11. Compare the feature set with the knowledge Base/ Training Set.
12. Store the output Test in the String where the characters are mapped with the training dataset.
13. Display the String values.

The process of creation of the knowledge base is shown as under.

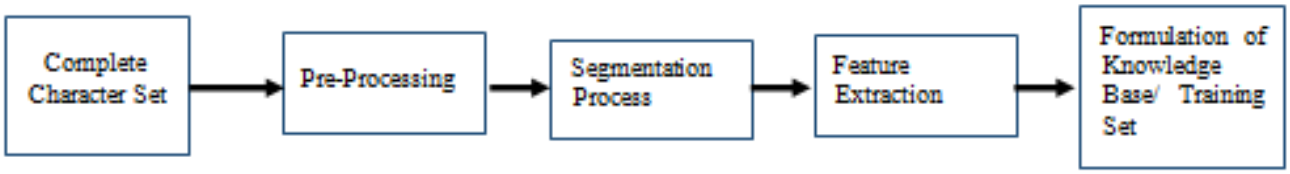

Figure 3: Creation of a Knowledge Base/ Training Set

In the system proposed the characters to be recognized are assumed to be are typedalphabets and the process can be extended to written or handwritten alphabets also. The handwritten characters have to be assumed to be written in

isolation in which the characters are not connected. The proposed system does not require any preprocessing in respect of thinning [5] and hence in the first instance the characters are digitized. Figure 3 show various steps of creation of Knowledge Base/ Training Set for the proposed character recognition research. Every character consists of a set of lines, curves, loops or a combination of these primitives which is not the matter of concern in the proposed research. The projections of the characters result in an image with quite different attributes as compared to the actual character under observations. The histogram projection contains the spatial histogram of the character when scanned including the number of black and white pixels. These spatial histograms signature analysis are mostly stored in a two dimensional arrays with columns of one row containing the pixel position and the corresponding columns of the other row containing the number of pixels. The number of columns of the array is directly proportional on the width of the character. The array so formed is processed to extract the features required for the recognition. The following equation illustrates the process generation of Training Set.

$$Training\ Set(i) = \sum_{k=1}^{n} \left(\frac{P(nz)}{Pn}\right)$$

Where P(nz) → Total Nonzero Pixels in the Segmented Image.

Pn→ Total Number of Pixels in the Segmented Image

The features so obtained are compared with the entries forming the knowledge base and the one matching returns the character containing these features.

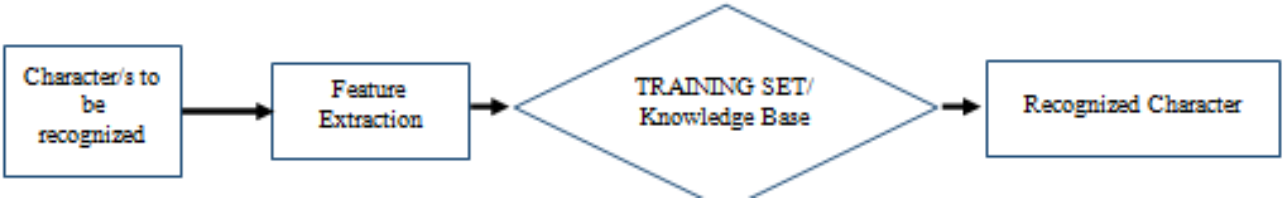

Figure 4: Recognition Process

EXPERIMENTAL RESULTS

The proposed OCR development and experimentation has been carried out on an Intel I7 processor and proposed procedure has been applied on documents containing English characters of various sizes and styles. It is mentioned here that the recognition process demands creation of a new training set/ knowledge base for every size and texture of the characters to be recognized. The Figure 4 shows the comprehensive recognition process of the proposed research where extracted features are compared with the training set so that actual character/s is/are recognized. The characters were separated by using segmentation process and for that cell structure of Matlab [14] is used where each character is stored in separate arrays within a same cell. The features that are extracted from the cell locations are separated in a separate Nx2 array where each feature extracted from the segmented image is mapped with the asci value of the character which is unique for every texture and size of the character. The results were highly satisfactory with a success rate of about 99.6% for Typed and 82.3% for handwritten characters. The success rate is dependent on the segmentation as well as on quality of the typed and handwritten characters. More comprehensive data set is required for handwritten characters as each handwritten character can have various writing styles of each size.

CONCLUSION

The character recognition using signature analysis is an active area of research and Automatic Recognition of characters of different languages has been going on since decades and is still a very important research area. Initially the proposed research is mostly concentrated on the recognition of typed English characters. However the proposed research can be extended for handwritten characters where training set has to be more elaborative without changing the algorithm. The proposed research focuses in the formation of a knowledge base/Training Set where initially the features of the character set of a language are stored and later during the recognition process the features of the characters to be recognized are extracted and compared with those of the knowledge base. A unique match found in the knowledge base recognizes the character/s recognition is said to be successful. The proposed research could be extended efficient recognition of for handwritten English characters by modifying the knowledge base.

REFERENCES

[1] Jie Zhou, QiiangGan and Ching Y Suen "A High Performance Hand–printed Numeral Recognition System with Verification Module", ICDAR pp 293, 1997.

[2] Park and Lee "Off line recognition of large-set handwritten characters with multiple hidden Markov models", Pattern recognition, vol 20, no 2, 1996.

[3] II-Seok Oh and Ching Y Suen "A Feature for Character Recognition Based on Directional Distance Distribution", ICDAR pp 288, 1997.

[4] Kim and Park Off line recognition of handwritten Korean and alphanumeric characters using hidden Markov models", Pattern recognition, vol 29, no 5, 1996.

[5] Raymond Yee –MianTeo&RajjanShinghal "A Hybrid Classifier for Recognizing Handwritten Numerals", ICDAR pp 283, 1997.

[6] Dan ClaudiuCires¸an and Ueli Meier and Luca Maria Gambardella and JurgenSchmidhuber, "Convolutional Neural Network Committees for Handwritten Character Classification", 2011 International Conference on Document Analysis and Recognition, IEEE, 2011.

[7] GeorgiosVamvakas, Basilis Gatos, Stavros J. Perantonis, “Handwritten character recognition through two-stage foreground sub-sampling” ,Pattern Recognition, Volume 43, Issue 8, August 2010.
[8] Devlin, Jacob, "Statistical Machine Translation as a Language Model for Handwriting Recognition." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.
[9] Al-Khaffaf, Hasan SM, et al. "On the performance of Decapod's digital font reconstruction." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.
[10] Rhead, Mke, "Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems." Security Technology (ICCST), 2012 IEEE International Carnahan Conference on. IEEE, 2012.
[11] MAJIDA ALI ABED HAMID ALI ABED ALASADI Simplifying Handwritten Characters Recognition Using a Particle Swarm Optimization Approach EUROPEAN ACADEMIC RESEARCH, VOL. I, ISSUE 5/ AUGUST 2013 ISSN 2286-4822
[12] Mohammed Z. Khedher, Gheith A. Abandah, and Ahmed M. Al-Khawaldeh Optimizing Feature Selection for Recognizing Handwritten Arabic Characters PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY VOLUME 4 FEBRUARY 2005 ISSN 1307-6884
[13] Amir BahadorBayat Recognition of Handwritten Digits Using Optimized Adaptive Neuro-Fuzzy Inference Systems and Effective Features Journal of Pattern Recognition and Intelligent Systems Aug. 2013, Vol. 1
[14] https://in.mathworks.com/

