Allograph discovery on Machine Printed Text

Palvi Gupta
Computer Science and Engineering Department
IET, Alwar
Rajasthan, India
erpalvigupta@gmail.com

Dr Basant Verma
Computer Science and Engineering Department
IET, Alwar
Rajasthan, India
bitspilani.basant@gmail.com

Anamika Mitra
Computer Science and Engineering Department
Sharda University
Greater Noida, India
anamika.mitra@sharda.ac.in

Md.Ansari
Computer Science and Engineering Department
Sharda University
Greater Noida,India
Mohiuddinansari@sharda.ac.in

***Abstract*— This paper presents a novel definition for allograph models and recognizing them in machine printed text. The aim of allograph extraction method is to find and collect the allograph from the set of characters. The objective of this paper is to give feature definition of such character primitives which can be used to distinguish the shape characteristic. This paper also proposes a new design for extracting the structural properties of the pattern in machine printed text (only English alphabets from A-Z). The recognition method is tested using multi character random letter sequence**

***Keywords*—Optical Character Recognition, allograph, feature extraction.**

I. INTRODUCTION

Character recognition is an old yet active topic of research where handwritten, typewritten or printed text images are converted into machine encoded code or text [3].Character recognition is the a field of research in pattern recognition, artificial intelligence and also it is widely used in computer vision [2].Our approach does not depend on a priori intuitive shape definition. On the basis of the scanning method, it gives the inherent shapes which constitute the fundamental shape. To represent the text, it uses 5 state method and pixel ratio.

The input image can be read/scanned in two ways: Horizontal (raster) or Sequential. Here in this paper the body region is being scanned from top to bottom (Sequential). In sequential scan, the body region can be in 5 different states i.e. start, split, merge, continue and termination state [1].With these 5 states, allograph of primitive characters can be well defined. The objective of retrieving such states and finding the characteristics of the patterns on state basis can be used to distinguish one character from another. Therefore, importance of these states can be used to recognize the components and help in solving many practical pattern recognition problems.

In pattern recognition, feature extraction of character plays a crucial role. Each character has many such particular features. Feature extraction stage describes the relevant shape information contained in a pattern is extracted so that the task of classifying the pattern is made easy by a formal procedure. Every character has different shape or model. These shapes are known as Allograph. (Parizeau & Plamondon, 1995) (Duneau & Dorizzi, 1994)[5].Allograph models or simply allograph in machine printed text are similar for different styles, for example, upper/lower case letters and for numerical numbers also. The approach relies on the concept that machine printed characters are spatially distributed and can be recognized using states with reference to these models.

In this paper, we apply this allograph model to machine printed text type identification. A careful and extensive survey of the literature shows that this is a newer approach to apply the 5-state allograph extraction model.

The remainder of this paper is subdivided as follows. In Section II and III, goal of OCR and definition of Artificial Intelligence is given.In section IV preliminary definitions of allograph is presented , followed up by literature review in section V.And in section VI a methodology overview of pattern recognition are provided. Section VI also gives details of feature definition of the characters and explicates the feature extraction phase.Experimental results are presented and discussed in Section VII, and concluding remarks are made in Section VIII.

II GOAL OF OCR

The goal of Optical Character Recognition (OCR) is the conversion of scanned images of handwritten, typewritten or printed text into encoded text. It is widely used as a form of data entry from some sort of original paper as one data source.OCR is one of the oldest ideas in the history of pattern

recognition using computers. Many researchers have been done on character recognition in last 56 years. Many surveys [11, 1] have been published on the character recognition.

III Artificial Neural Network

An Artificial Neural Network (ANN) is a computational model that is inspired by biological nervous systems [10].It is much similar to the way human brain process information. The important element of this model is the original structure of the information processing system

IV. What is an allograph?

The study of shapes has always been an active topic of research in pattern recognition. Many typewritten/handwriting models have been proposed for analyzing or generating pieces of handwritten/typewritten. Most researchers in the field of character recognition follow the decision theoretic approach, which is generally only for classification of patterns and ignores structural information. Most of the previous works stated different definition for allograph for the same letter.

IV .Literature review

The study of shapes has always been an active topic of research in pattern recognition. Many typewritten/handwriting models have been proposed for analyzing or generating pieces of handwritten/typewritten. Most researchers in the field of character recognition follow the decision theoretic approach, which is generally only for classification of patterns and ignores structural information. Most of the previous works stated different allograph for the same letter. According to the researchers, the extraction of allograph may involve as an interest area as it may help the recognition of characters easily and establishing the relationship between character instances and hence such an allograph can help to advance the recognition capability. The desirable objective is in the creation of small size allograph dictionary.
The work presented by Marc Parizeau and Ritjean Plamondon et al. [5] describes a new model based on Attributed Handwriting Primitive (AHP) used in a fuzzy syntactic allograph modeling approach for cursive script recognition [6].This model is used as a tool for the extraction of attributed primitives, themselves used in shape grammars. In this paper Marc Parizeau and Ritjean Plamondon et al. defined an operational handwriting model for on-line syntactic recognition of cursive script.Handwriting could be represented and defined as a 'characteristic points' which can be linked together by segments of uniform curvature.

The major contribution in recognizing allographs were made by Marc Parizeau and Ritjean Plamondon et al [7]. In their work they presented an new and advanced method for creating allograph models. And subsequently that can be recognized within cursive handwriting. The proposed method concentrates on the morphological aspect of cursive script recognition. In their work they have used fuzzy-shape grammars in order to define the morphological characteristics of allograph. And using this, grammer can be viewed as basic knowledge for developing a writer independent recognition system.

Segmentation and recognition work [8] based on, intrinsic models of cursive letters (allographs) was proposed by Parizeau,Plamond and Lorette[8].According to the ref [8],have proposed the definition of
allograph models using stratified context-free shape grammars that permit the definition of both syntactic and semantic attributes. These attributes synthesize pertinent morphological characteristics of allographs that are then used for recognition. Their main work concerns with the parsing process developed for allograph segmentation, which uses fuzzy-logic to evaluate the likelihood of segmentation hypotheses. This process was marked as the first step in their recognition method and thus leaded to the construction of a graph.In the graph, each nodes represented segmented allographs and arcs linked to the adjacent nodes. And finally, it was suggested that this analysis of segmentation graph can be successfully carried out for submitting possible letter sequences to higher linguistic evaluation modules. On experimentation an average recognition rate of 91.7% was obtained for a test database containing cursive samples of 10 different writers, Recognition is non personalized, that is, cursive samples of all writers are treated with the same algorithm parameters.

Miguel L. BOTE-LORENZO, Yannis A. DIMITRIADIS and Eduardo GÓMEZ-SÁNCHEZ [9] presented a new allograph extraction method. They proposed a technique which is based on two clustering phases. In the first phase, a rough clustering of handwritten data is made taking into account both characters' global and local information. And in the second phase, the clustering is refined in order to obtain clusters of characters belonging to the same allograph that is finally computed.

VI. METHODOLOGY OVERVIEW

The steps involved in this model are as follows/methodology used are:

A.Preprocessing is an initial operation that is carried out on scanned input image. The reliability of the whole recognition system depends on the quality of preprocessed image. In this stage, the input pattern is segmented from the background. Techniques applied are noise removal, filtering, skeletonization, so that character image is easy to extract relevant features. This step also includes:

(a) Data Acquisition: The input images are acquired from documents containing English text by using scanner as an input device. Scanned images are then stored in some picture file such as BMP, JPG etc.

(b) RGB to gray conversion: In the pre-processing 1st stage is to convert the input RGB image into gray scale image.

B.Binarization: is the process of converting a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by selecting a threshold value in between 0 to 255(here threshold value is 102).In this thesis the character skeleton is made up of ones and background pixels of zeroes after binarization.

C.Segmentation: is a process to simplify or change the representation for easier detection and labeling of the states. While performing the sequential scan, 5 state parameters are detected and correspondingly labels are assigned. Segmentation is the process that is performed in three different steps

a. Identifying start-of-character feature
b. Identifying the end-of character feature, and
c. Identifying the split and merge of the edges in between (a and b).

D.Object labeling: The object labeling method proposed here is similar to the method discussed in[1]. Starting from the bottom left corner, every image is scanned column by column from bottom to top[1]. Information about the starting and ending coordinates of every component and object label encountered at the ith scan is recorded in a tuple (Tik (j),j=1,2,3,4) where k denotes the component number and the last field is used to store the objects labels.

E.Shape Recognition: Shape recognition is the process to analyze the shape i.e allograph of a character. In this section we present the and discuss the Five state shape recognition method for offline classification of characters. In the survey on character recognition application, Suen et al.[1980] recommended the syntactic approach as a possible solution to achieve the optimal recognition result.[1]

Feature Extraction

In this phase, the features of the characters that are crucial for classifying them at recognition stage are extracted. This is an important stage as its effective functioning improves the recognition rate and reduces the misclassification. The segmented English characters are converted into feature vector that is used to find the characteristic of the pattern. Each character has particular characteristic; one of them is its shapes features. Feature extraction describes the relevant shape information contained in a pattern so that the task of classifying the pattern is made easy by a formal procedure. Feature extraction stage in OCR system analyses these character segment and selects a set of features that can be used to uniquely identify that character segment.[1] Mainly, this stage is heart of OCR system because output depends on these features.

Algorithm1: Five state

Step1: The machine printed data is converted into digital form by scanning.

Step2: The image of that data is fed to the pre-processing phase.

Step3: The digital image is converted into the gray scale image.

Step4: Convert a gray scale image (0 to 255 pixel values) into binary image (0 and 1 pixel values) by selecting a threshold value in between 0 to 255(here threshold value is 102).

Step5: Each character image is then filtered using 5 state based.

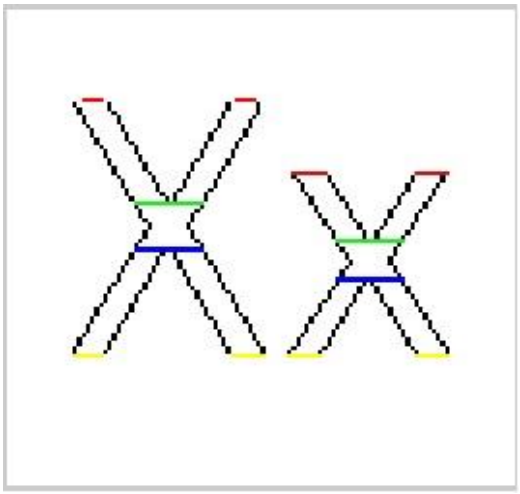

Figure1: Representation of Five States

Here, in the above figure, four states are clearly shown. Each state is distinguished by a color where red represent the start state, green represents merge state, blue is for split state condition and yellow color is representing the stop state.

Algorithm 2:

The extraction of the image “E” into a corresponding matrix is shown above. Here 0= Numerical value corresponding to white and 1= Numerical value corresponding to black. From this matrix we find the element which occurs the most and therefore this will

correspond to the most outstanding pixels, and hence the background. This proves to be helpful in distinguishing the character from its background of the paper on which it has been written or printed.

Therefore the number of pixels constituting the background will be less than the number of pixels that make up the character. Henceforth, the assumption is that the numerical values other than the background constitute the character.Next step is to store the pixels belonging to the character in a table as shown on the next page. This table is called as the initial X-Y table of the unknown character. The number of rows in the table is equal to the number of pixels of the image belonging to the character. This table will differ even for two exactly identical characters at two different positions in the image.

VI. Experimental Results

Data Collection Allograph models for all 26 letters has been collected from different source and optimized using the algorithm In this research machine printed English Character sets are taken. These steps are followed to obtain best accuracy of input printed text English character image. First of all, training of system is done by using different data set or sample. And then system is tested for few of the given sample and hence accuracy is measured. The data set was partitioned into two parts. The first piece is used for training the system and the second was for testing purpose. For each character, the proposed feature were computed and stored for training the network to classify. The results obtained from the proposed algorithms are displayed in the below table.

Letters	Training samples	Testing samples	Correctly Recognized	% of Recognition accuracy
A	15	5	5	100
B	15	5	2	40
C	15	5	2	40
D	15	5	5	100
E	15	5	4	80
F	15	5	4	80
G	15	5	2	40
H	15	5	4	80
I	15	5	5	100
J	15	5	5	100
K	15	5	4	80
L	15	5	3	60
M	15	5	5	100
N	15	5	4	80
O	15	5	4	80
P	15	5	3	60
Q	15	5	3	60
R	15	5	4	80
S	15	5	4	80
T	15	5	3	60
U	15	5	3	60
V	15	5	3	60
W	15	5	5	100
X	15	5	3	60
Y	15	5	2	40
Z	15	5	5	100

Table1: Result Obtained

VII. Recognition problem

Shape discrimination of similarly shaped characters is difficult for machine recognition of typewritten characters. Some characters have similar shapes, such as U and V, C and L, V and Y. There may also exist certain difficulties for shape distinctions between certain characters and numbers. These misclassifications errors were confined to one or more similar characters and they usually occurred within the same subgroup. For that, some more allographs can be discovered for clearly distinguishing their characteristic features.

VIII. Conclusion

Machine printed character recognition based on character feature is discussed here. State based feature algorithm and pixel ratio are implemented .There are lots of ways to implement this project but recognition is done based on character feature. Recognition approaches heavily depend on the nature of the data to be recognized. The system was tested on upper case characters of two different typewritten fonts. The result indicates with approximately 80.23 % accuracy for Typewritten English Character recognition. The confusion among characters are arisen that can be solved by including more allographic feature. Introduction to more allographic feature can advance the system and can encounter wherever such ambiguities will occur.

IX.Scope for further improvement:

We can include the power of this recognition method for lower-case alphabets, numerals recognition also. This approach can be extended for the recognition of handwritten documents. The discovered allograph definition can be also

used in graphology system. Neural network can also be implemented in the near future to make the system learn and adapt; this would eliminate the need of database. The database structure shall be improved by associating with every characteristic vector, the corresponding alphabet and font, this would enhance the understandability. Better user interface can be developed. This will improve the accuracy of the system considerably as in that case semantic checking can also be incorporated.

References

[1].P.Ahmed, C Y Suen,"Computer Recognition Of Totally Unconstrained handwritten Zip Codes", September 1987,International Journal of pattern Recognition and Artificial Intelligence Vol. I No.1(1987) 1-15

[2].Priya Sharma and Randhir Singh "Performance of English Character Recognition with and without Noise" International Journal of Computer Trends and Technology- volume4Issue3-2013

[3].Schantz, Herbert F. (1982). The history of OCR, optical character recognition. [Manchester Center, Vt.]: Recognition Technologies users Association. ISBN 9780943072012

[4].M. Hanmandlu, K. R. M. Mohan and H. Kumar, "Neural-based Handwritten character recognition", in
Proceedings of Fifth IEEE International Conference on Document Analysis and Recognition, ICDAR'99,
Bangalore, India, (1999), pp. 241-244.

[5].Marc Parizeau and Ritjean Plamondon, A Handwriting Model for Syntactic Recognition of Cursive Script

[6].Parizeau M., "Systsme de reconnaissance d'6criture cursive et bloc-notes electronique" , Ph.D. Thesis, Ecole Polytechnique de Montreal, to be published summer 92.

[7]Marc Parizeau and Rejean Plamondon,"A fuzzy-syntactic Approach Modelling for Cursive Script Recognition, IEEE Transactions On Pattern Analysis And Machine Intelligence, VOL. 17. NO. 7, July 1995

[8].Marc Parizeau, Rejean Plamondon Guy Lorette,"Fuzzy shape grammers for cursive script Recognition ", Advances in Structure & Synt. Pattern Recogn.,H. Bunke(ed), World Scientific,P. 320-332,1993

[9].Miguel L. BOTE-LORENZO, Yannis A. DIMITRIADIS and Eduardo GÓMEZ-SÁNCHEZ," Allograph extraction of isolated handwritten characters"

[10].S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR research and development." Proceedings of the IEEE, Vol. 80(7), pp. 1029-1058, 1992.

[11].U. Pal and B. B. Chaudhuri, "Indian script character recognition", Pattern Recognition, Vol.37(9), pp. 1887-899, 2004.