iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VII **Month of publication:** July 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A survey on feature extraction based on allograph technique

Anamika Mitra[1], Palvi Gupta[2], Nishu Singh[3]
Assistant Prof, Sharda University, Greater Noida

Abstract—This survey describes the state of art of feature extraction based on allograph in the field of character recognition.It is based on extensive review of the literature on various definitions given by researchers on shape recognition algorithms particularly based on allograph.
Keywords—Optical Character Recognition, allograph, feature extraction.

I. INTRODUCTION

This literature survey is based on the previous work and related approaches about character recognition. The most advanced and efficient OCR(Optical Character Recognition) systems are designed for English, Chinese and Japanese like scripts and languages. In our literature survey we have studied many research approaches that are practiced using allograph on many languages and scripts .Therefore, our emphasis is to study and analyse the feature extraction approaches mainly based on allograph, observed or reported results and many other relevant issues considerable to a new research work on OCR. After our detailed literature survey we were able to discover the theme of our proposed work including the modified definition for the feature extraction stage based on new allograph definition and also the techniques we have incorporated in various phases, to implement it and to evaluate our results and conclusions.
The history of OCR research, like that of speech recognition, is comparatively old in the field of pattern recognition, according to the survey done [1]. During survey, it was found that pattern recognition had attracted attention of many researchers mainly in the OCR domain. Reason behind this, was that the characters were found very handy to deal with and were regarded as a problem which could be solved easily. However, against what was the expectation of many researchers, after some initial easy progress, great difficulty in solving this problem surfaced. On the other hand, it was found that dedicated and successful research cannot exist without its applications in engineering. Fortunately, market demand for OCR has grown tremendously and had became very strong even though word processors are prevalent[1].
English language scripts are fundamentally non-cursive in nature, where the characters are written independently, separated by space or pen-lifts. The intensive research effort in the field of CR was not only because of its challenge on simulation of human reading but also because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines, and web interface to paper documents.

II. ONLINE VS OFFLINE CHARACTER RECOGNITION

Character recognition is a process, which relates a symbolic meaning with object (letters, symbols and numbers) drawn on an image. It may be classified as offline and online. Online character recognition deals in real time recognition of characters. Whereas the case of Offline character recognition, the typewritten/handwritten character is typically scanned in form of a paper document and made available in the form of a binary or gray scale image to the recognition algorithm. The major difference between Online and Offline Character Recognition is that Online Character Recognition has real time contextual information but offline data does not.

III. GOAL OF OCR

The goal of Optical Character Recognition (OCR) is the mechanical or electronic conversion of scanned images of handwritten, typewritten or printed text into much in encoded text. It is widely used as a form of data entry from some sort of original paper data source, whether documents, sales receipts, mail, or any number of printed records.OCR is one of the oldest ideas in the history of pattern recognition using computers. In character recognition, the process starts with reading of a scanned image of a series of characters, determines their meaning, and finally translates the image to a computer written text document. Many researchers have been done on character recognition in last 56 years. Some books [1-3] and many surveys [11, 1] have been published on the character recognition. Most of the work on character recognition has been done on Japanese, Latin, Chinese characters in the middle of 1960s.

IV. ARTIFICIAL NEURAL NETWORK

An Artificial Neural Network (ANN) is a computational model that is inspired by the way biological nervous systems [1].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. The system is composed of a large number of highly interconnected processing elements called as neurons working in unison to solve specific problems.

An ANN is configured for a particular application, such as pattern recognition or data classification through a learning process and have been used to solve those tasks that are comparatively hard to solve using ordinary rule based programming. The way learning in biological systems involves

adjustments to the synaptic connections that exist between the neurons. The same way ANNs works as well. An Artificial Neural Network is a network of many very simple processors units(neurons), each unit have a small amount of local memory. Subsequently, these units are connected by unidirectional communication channels connections, which carry numeric as opposed to symbolic data. These units therefore, operate only on the local data and on the inputs they receive.

V. 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.

According to the researchers, the extraction of allographs 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. [2] describes a new model based on Attributed Handwriting Primitive (AHP) used in a fuzzy syntactic allograph modeling approach for cursive script recognition [3].This model acts as a guide 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. To represent handwriting, they defined a 'characteristic points' that gets linked together by segments of uniform curvature. The handwriting model is evaluated by human readers in a comparative analysis of original cursive letter sequences versus their reconstructed traces. The performance was measured in terms of mean reconstruction error and data compression rate [2].

The major contribution in recognizing allographs were made by Marc Parizeau and Ritjean Plamondon et al [12]. In their work they presented an original method for creating allograph models and subsequently recognizing them within cursive handwriting. Their method concentrates on the morphological aspect of cursive script recognition. In their work they have used fuzzy-shape grammars to define the morphological characteristics of conventional allograph. And using this, grammar can be viewed as basic knowledge for developing a writer independent recognition system.

As accordance with [11], their recognition method is further tested using multi-writer cursive random letter sequences. For testing, a dataset containing a handwritten cursive text 600 characters in length written by ten different writers, average character recognition rates of 84.4% to 91.6% were obtained, These results were achieved without any writer-dependent tuning. The same dataset was used to evaluate the performance of human readers. An average recognition rate of 96.0% was reached, using ten different readers, presented with randomized samples of each writer. The worst reader-writer performance was 78.3%. Moreover, results show that system performances are highly correlated with human performances.

Segmentation and recognition work [4] based on, intrinsic models of cursive letters (allographs) was proposed by Parizeau, Plamond and Lorette[4].According to the ref [4],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 [8] 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. The result of the experiment on allograph extraction made by using characters collected from UNIPEN international database yielded an allographs per character rate up to 7,2 in upper-case characters. Also the quality of allographs is tested using them to initialize a handwriting recognizer achieving an average recognition rate of 90,15%.

Kazutaka Yamasaki has proposed an automatic method for categorizing writing styles of characters (allographs) in his work [6].In his work, he allowed the construction of a recognition dictionary that includes various writing styles. Allograph categorization, first step, handwritten strokes contained in training data is categorized to obtain prototype strokes. These strokes are used to categorize handwritten characters and thus obtain their respective allographs. In this

approach, allographs share common prototype strokes. This makes it possible to reduce the dictionary size and computation time needed for recognition. A method of automatic allograph categorization has been proposed. Allographs are defined by using prototype strokes, which are obtained by a proposed method of stroke clustering. Comparing these allograph allows us to find stroke connections and stroke order permutations.

The problem of allograph categorization for on-line English words has been discussed in the context of style recognition [7], systematic naming schemes for allographs [8], segmentation of cursive script into letters [9]. our approach to allograph categorization is based on a stroke categorization that allows us to reduce the dictionary size and computation time needed for recognition.

The work of Dinesh Dileep[16] describes a geometry based technique for feature extraction applicable to segmentation-based word recognition systems. The work [16] is concerned with feature extraction based on local and global geometric features of the character skeleton .The proposed system extracts the geometrical features of the character contour. This feature is based on the basic line types that form the character skeleton. And as a result their system shows a feature vector as its output. The feature vectors so generated from a training set, were then fed to train a pattern recognition engine based on Neural Networks so that the system can be benchmarked. The method proposed was tested after training a Neural Network with a database of 650 images. On experimentation with a set of 130 sample images collected , 6 of them were detected erroneously.

Preprocessing stage includes contour smoothing, binarization, skew detection and noise reduction of a digital image. Subsequently, this algorithms leads to final classification can be made simple and more accurate. Reference [6] presents a good work on preprocessing and segmentation .In this paper, preprocessing stage consists of 4 steps - compression, skew correction, binarization, noise removal. Here, the nearest neighbor interpolation method is used for scaling down the original image for fast processing and accurate results.

The work of Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre[17] focused on recognition of English alphabet in a given scanned text document with the help of Neural Networks. Using Matlab Neural Network toolbox, tried to recognize and written characters by projecting them on different sized grids. In the first step, image acquisition is done, which acquires the scanned image followed by noise filtering, smoothing and then normalization of scanned image.And also,rendering image suitable for segmentation where image is decomposed into sub images.Thus,feature extraction improves recognition rate and misclassification. In their work character extraction and edge detection algorithm for training the neural network is used to classify and recognize the handwritten characters[3].

The work of Lament Duneau and Bernadette Dorizzi[6] presents a system able to recognize English words dynamically written on a digitizing tablet[10]. Lament DUNEAU and Bernadette Dorizzi described the different modules of a prototype of a system that would be able to recognize on-line handwritten words, with a limited number of scriptors and a large vocabulary. Their work was based on a statistical approach, in which letters are first identified with certain likelihood. A score marked then given to each word in the dictionary for the final recognition. Subsequently, to optimize the allographs' identification phase, a hierarchical classification algorithm was applied to group the allographs corresponding to a given letter and a given size. Their experimental result was very good as the effort was made in the use of appropriate normalization coefficients in the classification algorithm.

Reference [4] presents a system using hidden Markov models (HMMs) for identification and recognition of handwritten and typewritten text from document images. The text type identification uses OCR decoding to generate word boundaries followed by word-level handwritten/typewritten identification using HMMs. Their work showed that the contextual constraints from the HMM significantly improves the identification performance over the conventional Gaussian mixture model (GMM)-based method. Type identification is then used to estimate the frame sample rates and frame width of feature sequences for HMM OCR system for each

type independently -dependent approach to computing the frame sample rate and frame width shows significant improvement in OCR accuracy over independent approaches. The use of contextual constraints in the HHM was proved to be highly effective for reducing type identification errors. We also showed that the word error rate on document images of mixed types of text can be reduced significantly using our type identification method.

Another segmentation [10] work based on the decision to segment a character on a combination of feature extraction and character-width measurements. In their extensive investigation of some methods for isolating or segmenting characters during the reading of machine printed text by optical character recognition systems. Two new segmentation algorithms using feature extraction techniques were presented; both are used in the recognition of machine-printed lines of l0-, 11- and 12-pitch serif-type multifont characters. One of the methods, known as quasi-topological segmentation, bases the decision to "section" a character on a combination of feature extraction and character-width measurements. And second method, topological segmentation, involves feature extraction alone. The algorithms have been tested with an evaluation method that is independent of any particular recognition system. Test results are based on application of the algorithm to upper-case alphanumeric characters gathered from print sources that represent the existing world of machine printing. The topological approach demonstrated better performance on the test data than did the quasi-topological approach.

In the work of Louis Vuurpijl and Lambert Schomaker[13], have introduced a variant of agglomerative hierarchical clustering techniques. According to their proposed work, their new technique is used for categorizing character shapes (allograph) into a large set of data of handwriting into a hierarchical structure. Hence, this technique was used as a basis for the systematic naming scheme of character shapes. Their work reflects much improvement in the existing system. After application of the method to a very large set of characters, separately for all the letters of the alphabet, relevant clusters are identified and given a unique name. Each cluster represents an allograph prototype.

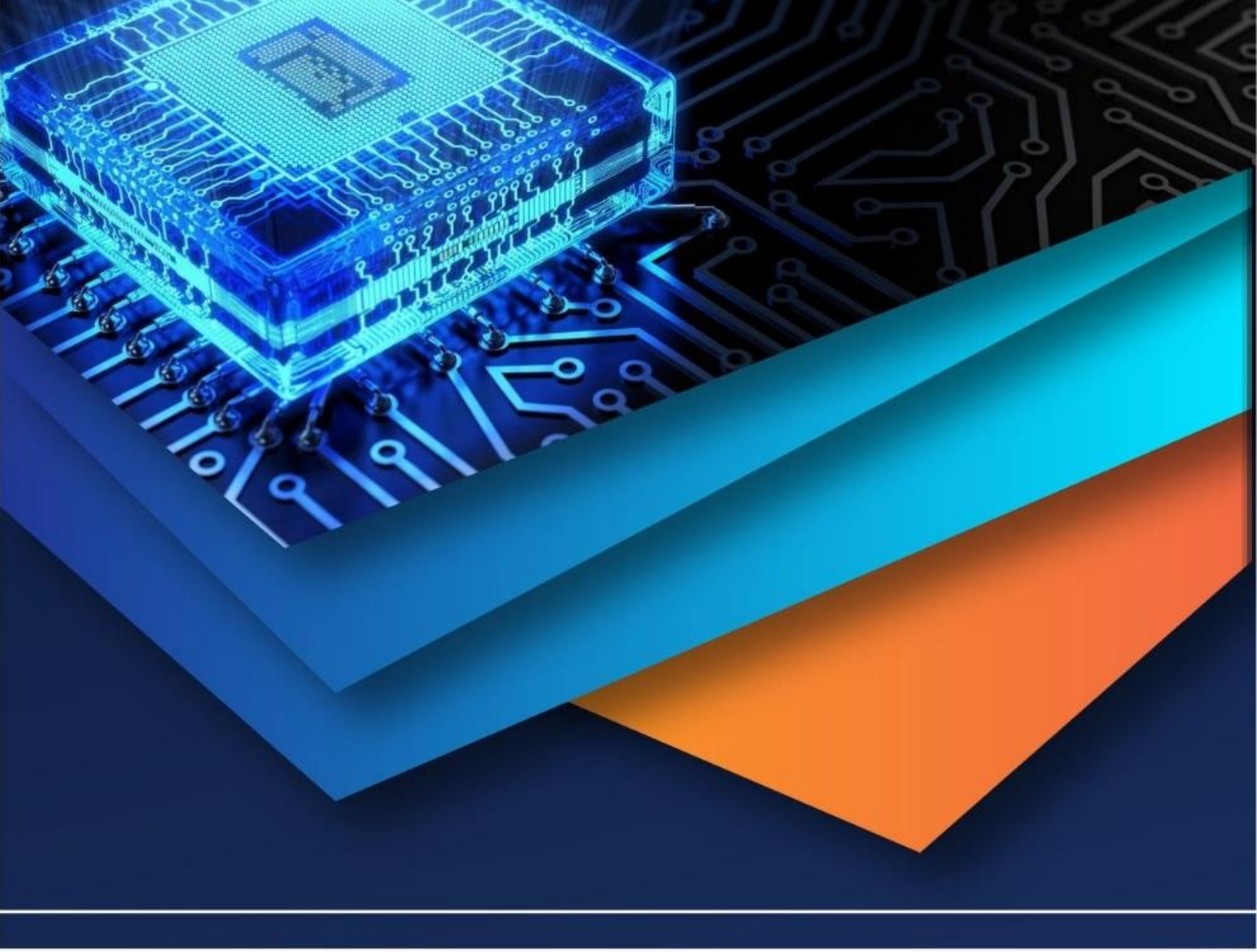
The initial results of the proposed modified agglomerative hierarchical clustering methods are very promising. In their experiment they have shown that the split-half consistency of the procedure proves to be quite good. The benefit of their method lies highest in the lower branches of the cluster dendrogram, where a large number of members may belong to single nodes.

A segmentation-free approach to OCR is presented as part of a knowledge-based word interpretation model. This new method is based on the recognition of sub graphs homeomorphic to previously defined prototypes of characters [14]. Gaps were identified as potential parts of characters by implementing a variant of the notion of relative neighborhood used in computational perception [14]. In this system, each subgraph of strokes that matches to previously defined character prototype is recognized anywhere in the word even if it corresponds to a broken character or to a character touching another one. The characters are detected in the order defined by the matching quality. Each sub graph that is recognized is introduced as a node in a directed net that compiles different alternatives of interpretation of the features in the feature graph. A path in the net represents a consistent succession of characters in the word. The method allows the recognition of characters that overlap or those are underlined. A final search for the optimal path under certain criteria gives the best interpretation of the word features. The character recognizer uses a flexible matching between the features and a flexible grouping of the individual features to be matched. Broken characters are recognized by looking for gaps between features that may be interpreted as part of a character. Touching characters are recognized because the matching allows non-matched adjacent strokes. The recognition results of this system for over 24, OOO printed numeral characters belonging to a USPS database and on some hand-printed words confirmed the method's high robustness level.

REFERENCES

[1] 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.
[2] Marc Parizeau and Ritjean Plamondon, "A Handwriting Model for Syntactic Recognition of Cursive Script",
[3] Parizeau M., "Systsme de reconnaissance d'6criture cursive et bloc-notes electronique" , Ph.D. Thesis, Ecole Polytechnique de Montreal, to be published summer 92.
[4] Marc Parizeau, Rejean Plamondon Guy Lorette,"Fuzzy shape grammers for cursive script Recognition ",1997
[5] Miguel L. Bote-lorenzo, Yannis A. Dimitriadis and Eduardo GÓMEZ-SÁNCHEZ," Allograph extraction of isolated handwritten characters"
[6] K.H.Aparna and V.S. Chavravarthy,"A complete OCR system development of Tamil Magazin documents" Tamil Internet 2003, Chennai, August, 22-24, 2003.
[7] J. Crettez. A set of handwritten families: Style recognition. In Proc. of the Third ICDAR, pages 489–494, Montreal, Canada, August 1995.
[8] L. Vuurpijl and L. Schomaker. Finding structure in diversity: A hierarchical clustering method for the categorization of allographs in handwriting. In Proc. of the Fourth ICDAR,pages 387–393, Ulm, Germany, August 1997.
[9] H. Teulings and L. Schomaker. Unsupervised learning of prototype allographs in cursive script recognition. In S. Impedovo and J. Simon, editors, From Pixels to Features III: Frontiers in Handwritten Recognition, pages 61–73. North-Holland, 1992.
[10] R.L Hoffman and J.W.McCullough," Segmentation Methods for Recognition of Machine-printed Characters,
[11] U. Pal and B. B. Chaudhuri, "Indian script character recognition", Pattern Recognition, Vol.37(9), pp. 1887-899, 2004.
[12] Marc Parizeau,and Rejean Plarnondon,," A Fuzzy-Syntactic Approach to Allograph Modeling for Cursive Script Recognition" IEEE transactions on pattern

analysis and machine intelligence, vol. 17. no. 7, july 1995
[13] Louis Vuurpijl and Lambert Schomaker,"Finding Structure in diversity:A Heirarchicalclustering method for the categorization of allographs in handwriting
[14] J. Rocha and T. Pavlidis, "A shape analysis model with applications to a character recognition system, IEEE Trans. Patrent Analysis and Machine Intelligence, vol. 16, no. 4, pp. 393-404, Apr. 1994.
[15] Jerome R Bellegarda, David Nahamoo, Krishna S. Nathan, and Eveline I. Bellegarda," supervised hidden markov modeling for on-line handwriting recognition "
[16] Dinesh Dileep," A Feature Extraction Technique based on character geometry for character recognition
[17] Kauleshwar Prasad, Devvrat C. Nigam, Ashmika Lakhotiya and Dheeren Umre," Character Recognition Using Matlab's Neural Network Toolbox, International Journal of u- and e- Service, Science and Technology Vol. 6, No. 1,February, 2013



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY