International Journal of Scientific Research and Engineering Development-– Volume 2 Issue 4, July – Aug 2019

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Comparative Analysis of Statistical-GA Writer Identification

Bhumika Baboria*, Prabjot Kaur**, Bhanu Gupta***

*(Department of ECE, IKGPTU/SSCET, Badhani)

** (Department of ECE, IKGPTU/SSCET, Badhani)

*** (Department of R&D, FCSC, Jammu Tawi)

--************************----------------------------------

Abstract:

The paper presents a new method of handwriting recognition using a unique and robust combination of Hybrid statistical edge-hinge features and genetic algorithms. On programming and testing the modules a high efficiency has been noted. The handwriting recognition is indeed a tough task which can be easily done with the help of the technique described here. A high efficiency reflects the accuracy of segmentation as well as the recognition using the statistical edge hinge features that has been optimized by genetic algorithms. The concept here has abridged handwriting recognition partially with artificial intelligence after the application of both statistical and genetic algorithms. For all correct matches the Euclidean Distance which was used as the matching comparator was calculated to be 0. The % accuracy and % error values for IAM Writer (Words) Dataset was evaluated to be 92 % and 8% where as for IAM Writer (Lines) Dataset it was evaluated to be 97.5% and 2.5 % respectively which is considerably higher in accuracy and lower in error than the latter one. The overall combined results for the both databases comprise of 220 images out of which 209 were accurate matches and 11 were inaccurate matches thus leading to overall %accuracy of 88.5% for the application and %error of 11.5%.

Keywords **— Edge-Hinge features, Genetic Algorithms, Writer, MATLAB, , IAM Writer DB**

--************************----------------------------------

I. INTRODUCTION

The problem of writer recognition is related to that of handwriting recognition. Handwriting recognition aims at eliminating the writer-dependent variations between writings and thus identifying the individual characters and words. Writer recognition on the other hand relies on these writer-specific variations between character shapes (see figure below) which allow characterizing its writer's hand.

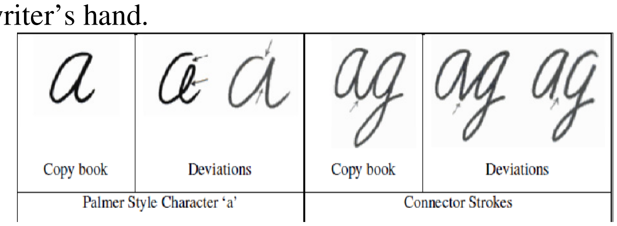

Fig.1 Copy book norms and individual [1]

Despite this contradiction between the two approaches, writer recognition can be handy in handwriting recognition, exploiting the principle of adaptation of the system to the type of writer.

There are four factors or identifiers responsible for variations in handwriting (see figure below).

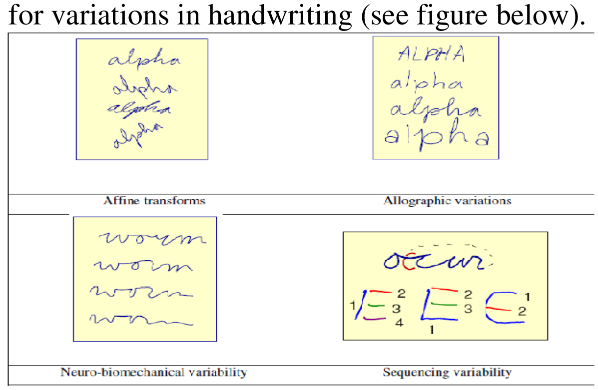

Fig.2 Factors causing handwriting variability [1].

These are affine transforms (rotation, translation, scaling etc.), allographic variations (character shapes employed by a writer), neuro-biomechanical variability and sequencing variability (variable order of stroke production). Among these factors, the allographic variations provide the most useful information for automatic writer recognition. These variations result in the first place from the copy book style taught and then from the writer-specific preferences in drawing these shapes, developed over time [1].

Writer recognition is generally divided into writer identification and verification. Writer identification involves a one-to-many search where given a handwritten sample s of unknown authorship and a database with samples of N known authors, the objective is to find the writer (or a likely list of writers) of s in the database.

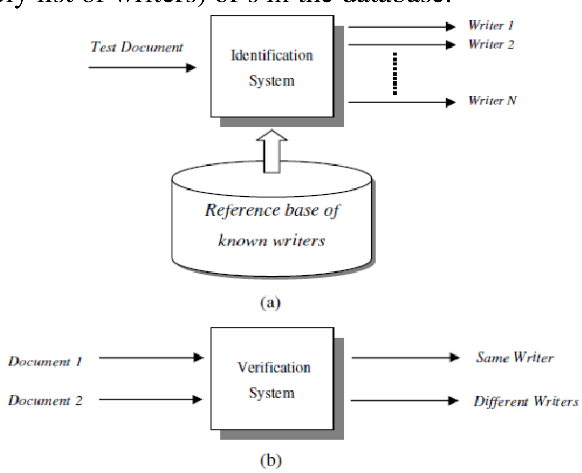

Fig. 3 Writer Identification and Verification models [2]

Writer identification is generally carried out by calculating a similarity index between the questioned writing and all the writings of known writers and sorting the retrieved results in a hit list with an increasing distance from the query. Choosing appropriate acceptance thresholds on these similarity indices can then be used to perform writer verification. The techniques for writer identification and handwriting classification are traditionally categorized into two broad classes: text-dependent and text-independent methods. In text dependent methods the writing samples to be compared require to contain the same fixed text. The text independent methods on the other hand identify the writer of a document independent of its semantic content. These methods use features extracted from the entire image of a text or from a region of interest [2].

Handwriting recognition is classified into offline handwriting recognition and online handwriting recognition. If handwriting is scanned and then understood by the computer, it is called offline handwriting recognition. In case, the handwriting is recognized while writing through touch pad using stylus pen, it is called online handwriting recognition. Each word is represented as a set of global features, e.g. ascender, loops, cusp, etc. Whereas segmentation based approach; each word/ligature is segmented into subunits either uniform or non-uniform and subunits are considered independently. Normally handwritten recognition is divided into six phases which are image acquisition, pre-processing, segmentation, feature extraction, classification and post processing. The block diagram of the basic recognition system is shown in figure below [3].

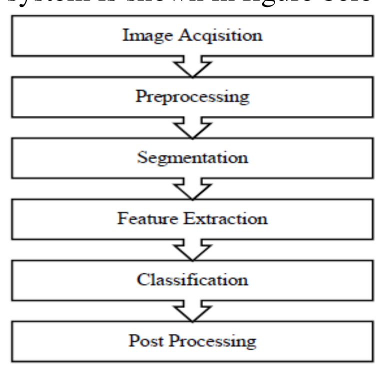

Fig.4 Basic Steps in Writer Recognition System [3]

The proposed system evaluates the performance of edge-based directional probability distributions as features in writer identification in comparison to a number of non-angular features. It is noted that the joint probability distribution of the angle combination of two “hinged” edge fragments outperforms all other individual features. Combining features may improve the performance. Limitations of the method pertain to the amount of handwritten material needed in order to obtain reliable distribution estimates.

II. LITERATURE REVIEW

Chandranath Adak et al., publish that the handwriting of an individual may vary substantially with factors such as mood, time, space, writing speed, writing medium and tool, writing topic, etc. It becomes challenging to perform automated writer verification/identification on a particular set of handwritten patterns (e.g., speedy handwriting) of a person, especially when the system is trained using a different set of writing patterns (e.g., normal speed) of that same person. Here, they work on writer identification/verification from offline Bengali hand- writing of high intra-variability. To this end, they use various models mainly based on handcrafted features with SVM (Support Vector Machine) and features auto-derived by the convolutional network [4].

E. N. Zois et al., present, a feature extraction method for offline signature verification that harnesses the power of sparse representation in order to deliver state-of-the-art verification performance in several signature datasets. The obtained state-of-the-art results on the most challenging signature datasets provide a strong indication towards the benefits of learned features, even in writer dependent (WD) scenarios with a unique model for each writer and only a few available reference samples of him/her [5].

Fredrik Wahlberg proposes a statistical machine learning approach for constructing a metric separating unseen writer hands. An unsupervised feature learning approach, based on dense contour descriptor sampling, was combined with a novel way of learning a general space for clustering writer hands, in a forensic setting. The metric learning inference was based on multiclass Gaussian process classification. This paper builds on earlier work from our group on building a system for estimating the production dates of medieval manuscripts, and act as a foundation for future use of writer identification techniques on our historical data [6].

Vivek Venugopal et al., propose a system to identify the authorship of online handwritten documents. They represent the trace of handwriting of a writer with descriptors that are derived from a set of dictionary atoms obtained in a sparse coding framework. The descriptors for each dictionary atom encode the error while using it alone for reconstruction. The use of sparse representation offers flexibility in describing each of the segmented handwritten sub-strokes of a writer as a combination of more than one atom or prototype. The descriptor is constructed by considering the attributes obtained from sets of histograms extracted at a sub-stroke level. In addition, an entropy based analysis for the bin size to be used for obtaining the feature sets is proposed [7].

Victor L. F. Souza et al., publish that the use of features extracted using a deep convolutional neural network (CNN) combined with a writer-dependent (WD) SVM classifier resulted in significant improvement in performance of handwritten signature verification (HSV) when compared to the previous state-of-the-art methods. In this work it is investigated whether the use of these CNN features provide good results in a writer-independent (WI) HSV context, based on the dichotomy transformation combined with the use of an SVM writer-independent classifier. The experiments performed are in the Brazilian & GPDS datasets [8].

Mohammad Abuzar Shaikh et al., propose an effective Hybrid Deep Learning (HDL) architecture for the task of determining the probability that a questioned handwritten word has been written by a known writer. HDL is an amalgamation of Auto-Learned Features (ALF) and Human-Engineered Features (HEF). To extract auto-learned features we use two methods: First, Two Channel Convolutional Neural Network (TC-CNN); Second, Two Channel Auto-encoder (TC-AE). Furthermore, human-engineered features are extracted by using two methods: First, Gradient Structural Concavity (GSC); Second, Scale Invariant Feature Transform (SIFT) [9].

Sourajit Saha et al., publish that there are several tasks that human excel at and computers do not and vice-versa. However, in the past 6 years with the boon in artificial neural network, labeled data and computation power; machines have started becoming smart at tasks like recognizing images, detecting different objects in images, captioning images, understanding and summarizing videos, detecting semantic actions in videos and so on. Deep learning researchers and practitioners have started demonstrating notable performance of AI (Artificial Intelligence) on many different tasks that pushes the boundaries and as a continuation of that process, we took one specific problem to solve using deep learning that humans can't solve [10].

III. METHODOLOGY

The proposed framework is illustrated in this section. Hand writing Recognition enables a person to scribble something on a piece of paper and then convert it into text. If taken the practical reality there are enumerable styles in which a character may be written. These styles can be self combined to generate more styles. Even if a small child knows the basic styles a character can be written, he would be able to recognize characters written in styles intermediate between them or formed by their mixture. This motivates the use of proposed modification of Genetic Algorithms for the problem. In order to prove this, it is required to make a pool of images of characters and convert them to graphs. The graph of every character was intermixed to generate styles intermediate between the styles of parent character. Writer recognition involved the matching of the graph generated. It is aimed to test this code for accuracy and reliability on available writer database for example IAM Handwriting Database.

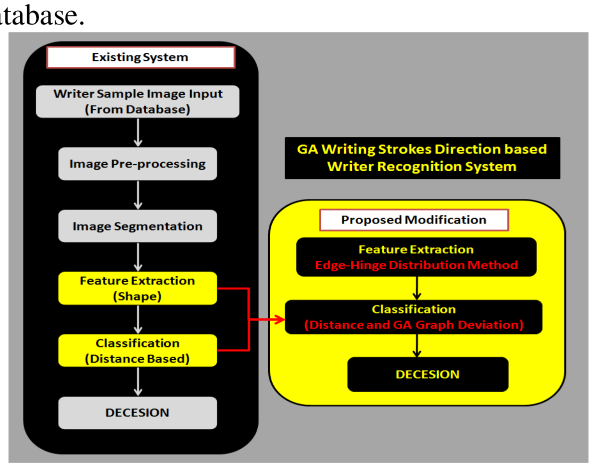

Fig. 5 Edge Hinge and GA Graph Deviation Writer Recognition System

The **edge hinge distribution** is a statistical feature, which outperforms all the other statistical approaches. First, by using the same generic edge detection approach as before, the edge pixels are extracted and considered. Next, a sliding window technique is applied, but in this method the entire window is quantized in directions. The main difference in the edge hinge distribution is to consider, not one, but two edge fragments in the neighborhood, emerging from the central pixel, and subsequently compute the joint probability distribution of the orientations of the two fragments. This feature concerns the direction changes of a writing stroke in handwritten text. The edge-hinge distribution is extracted by the use of a window that scans an edge-detected binary handwriting image.

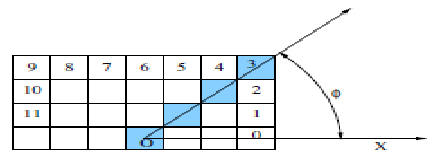

Fig. 6 Extraction of Edge-Direction Distribution

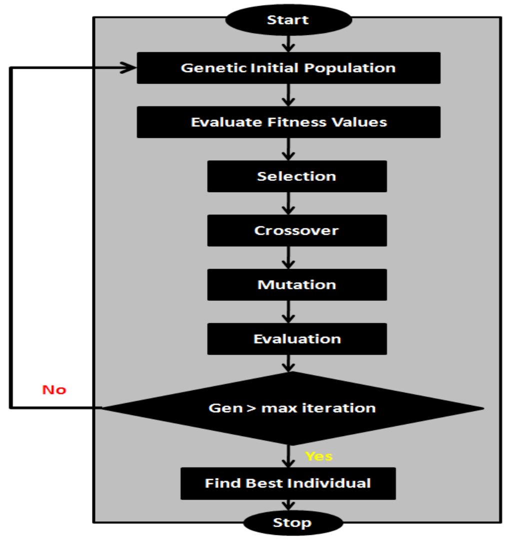

Fig. 7: Flowchart Depicting Simple Genetic Algorithm

Genetic Algorithms are very good means of optimization to recognize offline characters. They optimize the desired property by generating hybrid solutions. These hybrid solutions are added to the solution pool which is used to generate more hybrids. These solutions generally give better results than the already generated solutions. This is all done by the genetic operators. The system will use set of graphs generated from training data for any character. Genetic Algorithms are used to mix two such graphs and generate new graphs. This newly generated graph may happen to match the character better than the existing graph. So we can say that Genetic Algorithms are good means of optimization. To find deviation of a GA based graph easily, it is first required to know the deviation of an edge with another edge of a graph which is based on edge hinge distribution based feature. By pairing up of edges and iterating through all the edges, we can found this deviation. The distance between two graphs is also measured for evaluation of optimized writer features. As a preparation to start the optimization process, the following method is applied for handwriting recognition:

1. Provide initial inputs of sample handwritten letters to train the four artificial neural networks with different parameters.
2. Start the process of training the networks with different sets of letters.
3. Store the weight matrices and bias values obtained after training as files.
4. Read the file containing the input matrix.
5. Feed this as the input to all the four neural networks.
6. Send the outputs of the neural networks to the Genetic Algorithm.

The following method is applied for applying genetic algorithm to the outputs of the statistical equations:

Initialization: Select the output of the statistical equation with the indexes comprising of “1’s”. This corresponds to the initial population for the Genetic Algorithm.

Selection: Select the indexes from the neural network that has minimum number of “1’s”.

Fitness function: Compute the correlation coefficients of the selected indexes.

Mutation & Crossover: For the correlation coefficients less than the threshold value 0.50 repeat the step of the fitness function for a different training set. Discard the indexes that have coefficient values less than 0.3.

Evaluation: Select the index which has the maximum correlation coefficient with the input matrix.

Result: Output the selected character.

IV. RESULTS AND DISCUSSIONS

The results are obtained for Hybrid Edge-Hinge (Statistical) - Genetic Algorithm (GA) of a robust user-interface in MATLAB using GUIDE Tool. The application GUI’s for hybrid Statistical-GA Writer Recognition System using a combination of Edge-Hinge Features and Genetic Algorithm is developed and tested using database entry samples from two datasets from IAM Writer Image Databases. The two databases are namely IAM Writer Dataset (Words) and IAM Writer Dataset (Lines)/(Sentences). The application tested on a total of 220 images combined from the two datasets and results are recorded and compared for performance analysis of hybrid Statistical-GA technique for one-to-many (1: N) matching criteria after GA Optimization and application testing.

TABLE I
WRITER IMAGE DATASETS

Dataset	File Format	No. of Classes	Images per Class	Total Images per Dataset	Total Ear Images
Dataset 1 IAM Writer (Words)	.png	20	05	100	220
Dataset 2 IAM Writer (Lines)	.png	20	06	120	

The Statistical-GA Writer Recognition Process starts and the progress is shown in MATLAB Command Window. The MATLAB Command window showing writer recognition results with Euclidean Distance and Nearest Class.

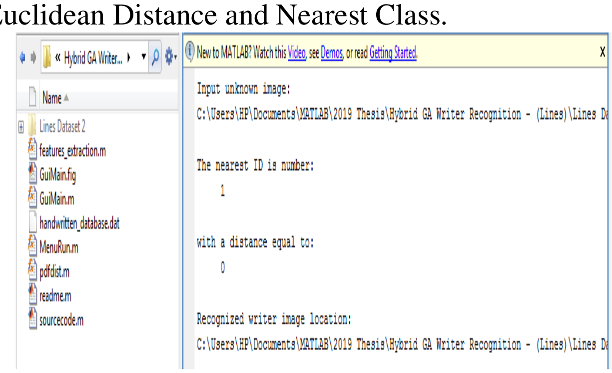

Fig. 8 MATLAB Command Window: Writer Recognition Result in MATLAB Command Window

On Completion of Task the Recognized ID is also displayed in the MATLAB command window.

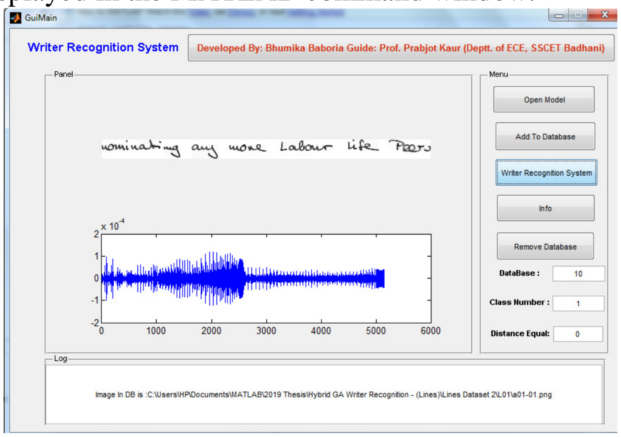

Fig. 9 Statistical-GA Writer Recognition Result in MATLAB GUIDE GUI

The Performance Analysis is done by statistical and graphical comparison and evaluation of parameters like Euclidean Distance, Nearest Class Number Recognition, % Accuracy, % Error, Accurate and In-accurate matches. These parameters belong to the Edge-hinge Statistical technique in combination with Genetic Algorithm optimization for one-to-many (1: N) matching criteria for a total of 220 input writing images including, 100 images from IAM Writer (Words) Dataset1 and 120 images from IAM Writer (Lines) Dataset2.

Analysis Graph 1: Comparison of Writer Image Datasets: IAM Writer Image Database is used to create two different Datasets. IAM Writer Image (Words) is used to create the Dataset1 comprising of .png images and IAM Writer Image (Lines) is used to create Dataset2 also comprising of .png images.

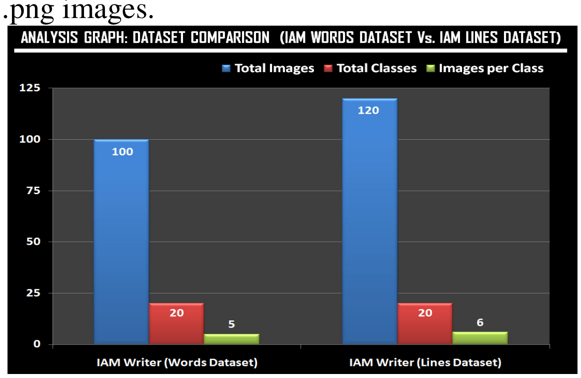

Fig. 10 Hybrid Statistical-GA Writer Database Comparison and Evaluation

Analysis Graph 2: Comparison of Euclidean Distance: The **Euclidean distance** or **Euclidean metric** is the "ordinary" straight line distance between two points in Euclidean space. The Euclidean Distance corresponds to the nearness of the input writer image to be matched to the writer image stored in the database.

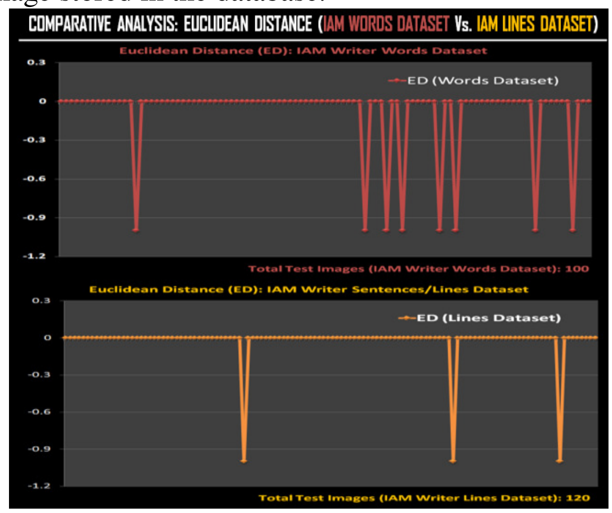

Fig. 11 Hybrid Statistical-GA Writer Database Comparison and Evaluation (Euclidean Distance)

Analysis Graph 3: Application Performance Evaluation Accurate Matches Vs. Inaccurate Matches: The IAM Writer (Words) Dataset1 has 92 accurate image class matches out of 100 and 08 inaccurate matches. The IAM Writer Image (Lines) Dataset2 has 117 accurate image class matches out of 120 and 03 inaccurate matches.

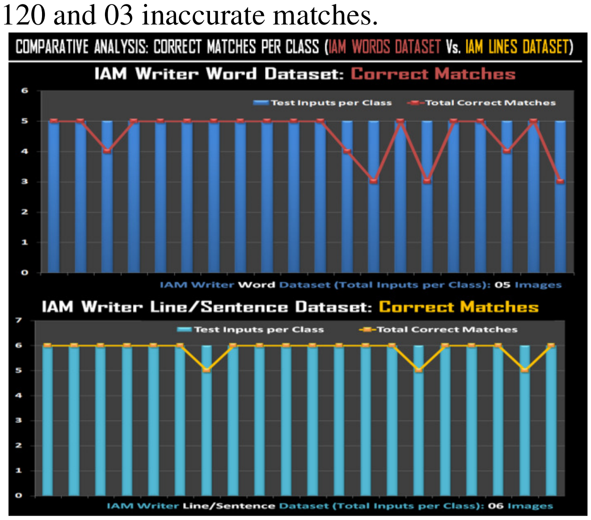

Fig. 12 Performance Comparison and Evaluation (Accurate Matches per Class)

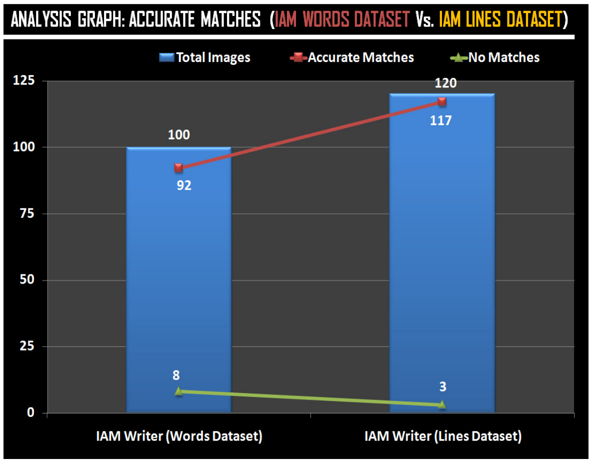

Fig. 13 Performance Comparison and Evaluation (Accurate Matches Vs. No Matches)

Analysis Graph 4: Application Performance Evaluation %Accuracy per Class and % Error per Class: The % accuracy per class and %error per class on the basis of number of images matched is calculated and show in the graph.

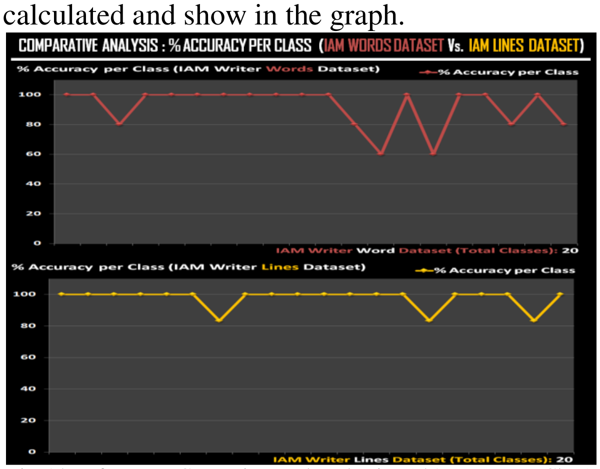

Fig. 14 Performance Comparison and Evaluation (%Accuracy per Class):

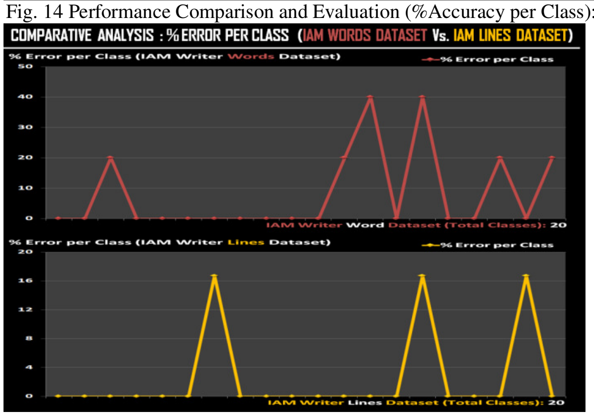

Fig. 15 Performance Comparison and Evaluation (%Error per Class)

Analysis Graph 5: Application Performance Evaluation %Accuracy Vs. % Error: The IAM Writer (Words) Dataset1 gives a considerate % accuracy performance of 92% and %error 8%. The IAM Writer (Lines) Dataset2 leads to a considerably higher % accuracy performance of 97.5% and a lower %error of 2.5%.

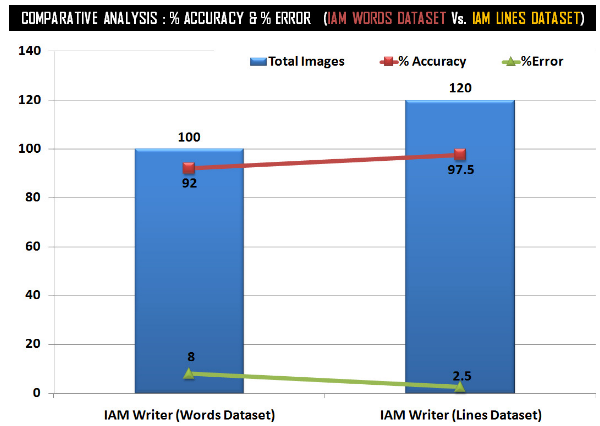

Fig. 16 Performance Comparison and Evaluation (%Accuracy and %Error)

Analysis Graph 7: Application Performance Evaluation Overall Analysis: The overall % accuracy is calculated to be 95% and the overall % error is 5%.

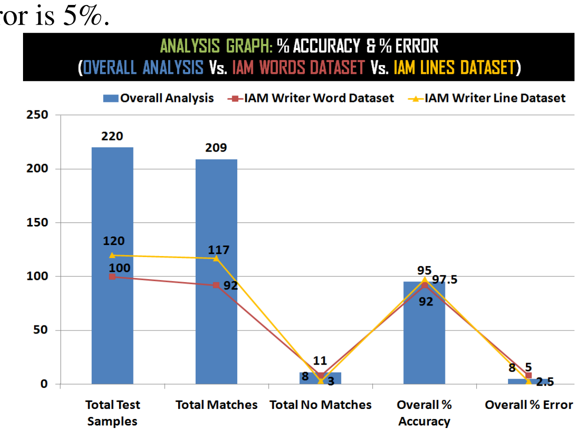

Fig. 17 Performance Comparison and Evaluation (Overall Analysis)

V. CONCLUSION

The objective of this research was to address the problem of automatic writer identification and verification from offline images of handwriting, a problem that enjoys a renewed interest of the community due to its applications in forensic document analysis, indexing and retrieval of document bases and recognition of handwriting.

This objective was met by developing an original method that exploits two different facets of handwriting: the existence of certain redundant patterns in writing and the visual attributes of orientation and curvature characterizing the writer of a handwritten text. The Application GUI for the Statistical-GA Writer Recognition System has been developed along with initial stages of Writing Image Acquisition, Writing Image Pre-processing and Writing Feature Extraction and Writer Recognition. The system has presented a new method of handwriting recognition using a unique and robust combination of Hybrid statistical edge-hinge features and genetic algorithms. On programming and testing the modules a high efficiency has been noted.

REFERENCES

[1] Haruna Chiroma, Ahmad Shukri Mohd Noor, Sameem Abdulkareem, Adamu I. Abubakar, Arief Hermawan, Hongwu Qin, Mukhtar Fatihu Hamza, Tutut Herawan, "Neural Networks Optimization through Genetic Algorithm Searches: A Review", Appl. Math. Inf. Sci. 11, No. 6, 1543-1564 (2017), pp. 1543-1563.

[2] Rehman, A., Naz, S. & Razzak, M.I. Multimed Tools Appl (2018). https://doi.org/10.1007/s11042-018-6577-1

[3] Sheng He and Lambert Schomaker. 2017. Writer identification using curvature-free features. Pattern Recogn. 63, C (March 2017), pp. 451-464. DOI: https://doi.org/10.1016/j.patcog.2016.09.044

[4] Chandranath Adak, Bidyut B. Chaudhuri, Michael Blumenstein, "An Empirical Study on Writer Identification and Verification from Intra-Variable Individual Handwriting", IEEE Access, Volume 7, 2019, pp. 24738-24758.

[5] E. N. Zois, D. Tsourounis, I. Theodorakopoulos, A. L. Kesidis and G. Economou, "A Comprehensive Study of Sparse Representation Techniques for Offline Signature Verification," in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 1, no. 1, pp. 68-81, Jan. 2019. doi: 10.1109/TBIOM.2019.2897802

[6] F. Wahlberg, "Gaussian Process Classification as Metric Learning for Forensic Writer Identification," 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, 2018, pp. 175-180. doi: 10.1109/DAS.2018.76

[7] V. Venugopal and S. Sundaram, "Online Writer Identification With Sparse Coding-Based Descriptors," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 10, pp. 2538-2552, Oct. 2018. doi: 10.1109/TIFS.2018.2823276

[8] V. L. F. Souza, A. L. I. Oliveira and R. Sabourin, "A Writer-Independent Approach for Offline Signature Verification using Deep Convolutional Neural Networks Features," 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, 2018, pp. 212-217. doi: 10.1109/BRACIS.2018.00044

[9] M. A. Shaikh, M. Chauhan, J. Chu and S. Srihari, "Hybrid Feature Learning for Handwriting Verification," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, 2018, pp. 187-192. doi: 10.1109/ICFHR-2018.2018.00041

[10] Saha, Sourajit & Khaled, Md. Asif & Saiful Islam, Md., Saha Puja, Nisha & Hasan, Mahady. (2018). Detecting Sex from Handwritten Examples. 1-7. 10.1109/ICSCAN.2018.8541214.

