Author Identification of Handwritten Text: A Review

Mr. Nikhil R Shrivastva
JSPM's RSCOE,
S.P. Pune University,
Pune, India.
nikhil43874@hotmail.com

Dr. Seema. V. Kedar
JSPM's RSCOE,
S.P. Pune University,
Pune, India.
seemaahkeddar@gmail.com

ABSTRACT

Author identification is the process of identifying the author of the document based on their handwritten. Recent advances in artificial intelligence (AI), image processing, data mining, pattern recognition and machine learning have shown that it is possible to automate author identification. This research paper gives a review and summary of various papers for author identification based on handwritten text. Features are extracted by using scanned images of handwritten words and trained using various classification algorithm.

Keywords—Feature Extraction, Support Vector Machine, Writer Identification, Author Identification

I. INTRODUCTION

Identification of an author is highly essential in areas like forensic expert decision-making systems, network security, digital rights administration, biometric authentication in information and, document analysis systems. In forensic science author identification is used to authenticate documents such as records, diaries, wills, signatures and also in criminal justice.

The digital rights administration system is employed to protect the copyrights of electronic media. Two broad classes of biometric modalities are: physiological biometrics that perform person identification based on measuring a physical property of the human body and behavioral biometrics that use individual characteristics of a person's behavior to identify. Author identification is the category of behavioral biometrics. Handwritten document analysis is used either textually or graphically in the field of information retrieval. Author identification mode can be generally classified into two types as online and offline. In online, the author's behavior is directly captured from the author and converted to a sequence of signals using a transducer device, but the handwritten text is used in offline to identify scanned images

The identification of the offline author is considered to be more challenging than online because it can have more information about a person's writing style, such as pressure, speed and angle, which is not available in the offline. The approaches to authors' identification can be divided into two types: text and text methods. An author must write the same text in text methods to perform identification, but in text-independent methods, any text can be used to determine the author's identity. This system shows the use of computer intelligence in the development of a discriminatory model. The scanned images are divided into words which process and extract tasks. Features such as edge-based features, word measurements, moment invariants used in the current research work are considered. Edge-based functions are calculated using the detected edge image. The directional distribution based on the edges and the distribution of the edges is two edge-based functions. Features such as length of the word, height of the word, height from baseline to upper edge, height from baseline to lower edge, ascender and descended are measurement characteristics of the word.

Invariant moment calculates a set of seven moments for a given image. It uses additional features not taken into consideration. They are character-level characteristics such as aspect ratio, loops, connections and end points. The ratio of aspect is given by the ratio between width and height. Loops identify the length of the loop, area etc. Junctions are defined as a point at which two strokes meet or intersect. The end points are those that only contain one pixel in their neighborhood of 8 pixels. These are calculated by passing through the thin image. By applying these character level functions to the identification of text-dependent authors, better results are obtained.

In this review paper, we have tried to identify the different means for assessment of handwriting and also the amount of work that is done on author identification till date.

II. MOTIVATION

The handwriting of an individual may vary considerably with factors such as mood, time, space, speed of writing, writing medium and tool, writing theme, etc. Automated author verification / identification on a certain set of hands-written patterns are difficult for a person, especially when the system is trained using a different set of the same person's writing patterns. It would be interesting, however.

III. VARIOUS OTHER APPROACHES FOR AUTHOR IDENTIFICATION

In this section, we reviewed the previous work on the identification of the author, which focuses on the various functions used to represent the text.

A. GENERALIZED APPROACH IN AUTHOR IDENTIFICATION

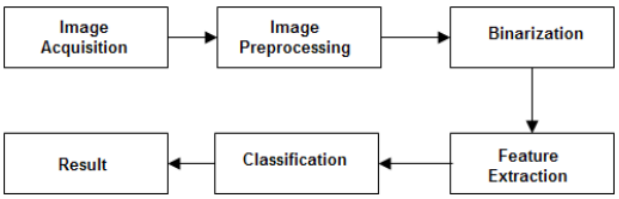

Fig 1. Steps in Author Identification

The author authentication is performed in six steps as follows:

1. **Image Acquisition:**

The image acquisition phase generally involves pre-processing, such as scale.

2. **Image preprocessing:**

The purpose of image preprocessing is to remove the redundancy in captured images without affecting the details that play a key role in the overall process. This process is called as sharpening the image.

3. **Binarization:**

A binarization method is used to binarize an image by extracting lightness (shining, density) from the image as a feature.

4. **Feature Extraction:**

Feature extraction is a low-level image processing operation which is usually performed as the first operation on an image. A feature can be defined as the important part of an image. The desirable property for a feature detector is repeatability; i.e. whether or not the same feature will be detected in different images of the same scene. Step edges, width, height, slant lines and junctions usually convey the most relevant information of an image; hence it is important to detect them in a reliable way.

5. **Classification:**

Support vector machine is supervised learning algorithm which can be used for classification.

6. **Result**:

Based upon the classification, final result are evaluated.

B. LITERATURE REVIEW

In paper [1] presents a author identification system. Twelve characteristics are extracted from handwritten English text lines, which are used to recognize people based on their handwriting. The extracted characteristics mainly correspond to the visible characteristics of the writing. Three main focused writing areas with width,

slant and height. Functions based on the writing's fractal behavior, also two classifiers were used with these features: K-nearest neighbor of the classifier and feeds the neural network forward: 100 pages of English text written by 20 different experiments Authors are used. The average recognition accuracy is 87.8 % by classifying text lines.

Authors focusd on pre-processing and extraction of features in paper [2] and presented standardization and normalized techniques for subsequent extraction of features. Normalization of shear angles, rotation, stroke width and size are considered for feature extraction

A robust system for off-line signature verification is presented in paper [3]. They have used simple features, different cell resolutions and multiple codebooks in an HMM framework. The simple and random error rates have been demonstrated to be low and close. Thus, demonstrated the system’s potential in real terms.

In paper [4], author proposed a Fuzzy modelling system. The sparkling modelling uses the Takagi–Sugeno (TS) model which divides the image of the input signature into eight sections and resizes each section accordingly based upon the signature size, then the angle and distance features are extracted in each section. These extracted features are provided to the forgery detection model of the TS model. 200 samples were tested and experimented which resulted in 77.5% accuracy.

In paper [5] author proposed a grid-based and centroid-based system. In a grid-based approach, the pre-processed image is divided into 240 cells and 100 pixels each cell. Then signature information is extracted and stored in an array and signature verification using column matching score is done. The pre-processed signature image is divided into three sections and the center of each section is calculated in the case of a centroid approach. The author proposed use of Bayesian classifier over 50 samples that resulted in maximum accuracy of 94%.

In paper [6] author describes the review of various methods for the analysis of authorship and the identification of a set of texts provided. Research in the analysis and identification of authors will certainly continue and continue to increase over decades. They present a vision of future authorship analysis and identification with high performance and solution for the extraction of behavioural features from text documents and used SVM for classification.

The author presents a new method for text classification in paper [7] that achieves significantly faster results than most existing classifiers. They extract computer-efficient and more costly features used by many other text classifiers, notably n grams (contiguous sequences of n words). They have analyse the feature vectors using a hybrid SVM technique, which classifies them sequentially with a one-to-all classifier and a binary classifier sequence. Overall method finishes execution on large novels within seconds, with accuracy comparable to that of standard classifiers but a much shorter runtime.

In paper [8], classification and verification of signatures based on descriptors derived from the theory of causal information. The proposal uses the entropy of Shannon, the statistical complexity and the fishing information evaluated through the symbolisation of the horizontal and vertical signature coordinates of Bandt and Pompe. These features are easy and quick to calculate and are given as input to a support vector support classifier. The results are better than online technologies using higher-dimensional functional spaces that often require specialized software and hardware. They evaluate the consistency of proposed work in the size of the training sample and use it to classify the signatures into meaningful groups.

Various texts from various authors are selected in paper [9], then these texts are tokenized and stemmed. The frequency of each word is determined in the stemmed text and the top k element is selected from the dataset available. Other features are also extracted, such as the number of characters, words, phrases and their ratios. The number of different punctuation types and symbols is specified. These features were later analyzed to conclude that each author has certain features that are unique to him. These functions were then used to train the fluorescent and SVM classifier and the conducted experiments have shown that SVM is more precise than the Fuzzy classifier. The combined classifier was later found to be more accurate than the two other classifiers.

In paper [10] the author applies an approach based on the grammar of dependency for the identification of Chinese authors. Their approach consists of four steps: data collection, extraction of features, optimization of features and identification. For the feature set, they proposed dependency as a new syntactic level feature combined with three additional features: empty word, voice part and punctuation to form the whole set. Principle component analyses (PCA) are used to reduce features and SVM classification.

In paper [11] author presents a fully automated approach to Turkish text authors identification by adapting a set of style markers to text analysis. 35 style markers are defined for a set of 20 different authors to identify an author. They tested and compared multiple classification machine learning algorithms. Maximum success rate obtained with Naïve Bayes Multinomial is 80%.

In paper [12] author examines the use of SVM as a text classifier and identifies it. The word stems are used to represent text characteristics where words are considered characteristics if they are not stop words and if they occur at least 3 times in training data. Information gain is used for the selection of features. The experiments compared SVM with 4 other methods for learning.

In paper [13] author introduces an Arabic text classification automatic approach based on two classification algorithms: SVM and C5.0. To cover different subject domains, they use 7 different data sets. Lexical features (single word) are extracted and square statistics are used to calculate the dependence of the term and class for the selection of features. The experiments are carried out using RAPIDMINER to implement the SVM algorithm and Clementine for the algorithm of the decision tree C5.0. The results show that C5.0 exceeded the SVM by about 10 percent. The authors analyzed the relationship between personality and various types of Twitter users in their paper. They collected data from 335 Twitter users and predicted the relationship between the Big Five and five Twitter users.

In paper [14] author studied the feature extraction and recognition operations on Arabic text. Implementation of the system on the basis of a dataset containing 32,000 Arabic text images in 16 different words, repeated 20 times each and written by 100 people using the same pen. In training 75 percent of the words were used for K-Nearest Neighbor Training, while the other 25 percent were used for testing. The performance measurements used were the Top 10 rates.

In paper [15] author proposed the use of genetic algorithms and graph theories to solve the problem of offline handwriting recognition. Input is given in the form of images. The algorithm has been trained on the training data in the database. The training data consisted of at least two sets of training data per language character, the graph theory and geometry of the coordinates were used to convert the images to graphs. They saw that these conversions changed the whole handwriting problem to the graph matching problem. When a pure graph match was made, the results were sufficiently fine. In all, they have an efficiency of 98.44 percent, which proves that the algorithm works correctly in most cases and correctly matched the unknown character input**.**

In paper [16], the author proposes Gaussian Mixture Models (GMMs) to deal with the task of automatic offline text identification of text lines. The resulting system is compared to a system using an approach based on the Hidden Markov model (HMM). The GMM-based approach is conceptually much simpler and faster to train that HMM-based system, on a data set of 4,103 text lines from 100 authors it achieves a significantly higher author identification accuracy of 98.46 percent**.**

In paper [17] author proposes a method based on author identification texture features in a handwritten document image. The co-occurrence histogram-based texture features are extracted using the correlation between sub bands with the same resolution of the decomposed image of the wavelet, indicating that the information is important in characterizing the authors based on the manual document image. The experimental results demonstrate the effectiveness of the proposed method and the potential of such a global approach for the identification of authors in the analysis of the document image, which is important in biometrics and forensics.

In paper [18] several functions based on statistics and model are presented. In particular, an improvement in the statistical feature, the distribution of the edge hinge, is intended. In addition, the features is combined with a model-based function based on a graphical codebook. The Fire maker DB, made up of 250 authors, including 4 pages per author, was used for the evaluation. The best result for the statistically proposed approach, the distribution of the skeleton hinge, achieved 90.8 percent accuracy, while the combination of the method with the graphical codebook reached 96 percent.

In paper [19] author presented a new approach to generating codebooks based on the segmentation of skeletons. Graphs are categorized according to their grid. This approach achieved a 90% identification accuracy for the data set for the ICDAR 2011 Writer Identification Contest.

In paper [20] author has proposed a author identification system based on a recovery mechanism that reduces the identifying process search space. The probability distributions of run-length and edge-inge features were used to characterize handwritten documents. Two databases containing Arabic, German, English, French and Greek samples are used to assess the effectiveness of the proposed approach and the experimental results show that a recovery mechanism is useful prior to identification.

Table 1. Comparison of various approaches used for author identification

Paper. No.	Approach used for author identification	Dataset	Accuracy Achieved
1.	Mathematical features using KNN and NN	100 pages of 20 different authors	87.8%
3.	HMM using simple static and pseudodinamic features	2400+1200 words	94.15%
4.	Takagi- Sugeno (TS) model	200 signatures	77.5
5.	Bayesian Classifier	50 signatures	94%
7.	Hybrid SVM	38 training examples and 7 test examples from 3 separate authors.	Approximately 70% true positive classification
8.	One-Class Support Vector Machine classifier	MCYT Dataset	82%
9.	SVM and Fuzzy Logic algorithm	20 different texts of each 10 different authors.	76%
10.	Principle component analysis (PCA) is used for feature reduction and SVM for classification.	The Online Books Page at the University of Pennsylvania edited by John Mark Ockerbloom	80%
11.	Naïve Bayes	35 style markers are determined for a set of 20 different authors.	80%.
14.	KNN	32000 words	93.8%
15.	Genetic Algorithm	69 characters from 385 input text	98.44%
16	GMM and HMM	4,103 text lines from 100 authors	98.46%
18	Skeleton hinge and Graphical Codebook	250 authors *4 pages = 1000	96%
19.	Codebook Generation	208 texts	90%

IV. FUTURE RESEARCH DIRECTIONS

The directions for further research related to author identification are given below:

i. Designing a fully automated language independent, and best accurate system for author identification.
ii. Developing a system to analyze author based on individual's their handwriting samples
iii. We also plan to study different method for identification accuracy.

V. CONCLUSION

In this paper, we discuss different approaches to the identification of authors, types of features extraction approaches, the classifiers and the different data sets. The literature was grouped by research papers based on similarities in the characteristics and classifiers used. This shows the researchers improvements. The classifiers, the databases used, the best identification accuracy for each publication, the number of authors and the year of publication are tabulated in order to compare the different research features. The table has been included for the databases used, the number of authors, samples etc. This specifies the large number of publications and the growing number of researchers in this field. We also reviewed that Skeleton hinge and Graphical Codebook classifier gives maximum accuracy of 96% in author identification.

REFERENCES

[1] U.V. Marti, R. Messerli, H. Bunke, "Writer Identification Using Text Line Based Features", Proc. ICDAR'01, Seattle (USA), pp 101-105, 2001

[2] T. Caesar, J.M. Gloger, E. Mandler "Preprocessing and Feature Extraction for a Handwriting Recognition System" 0-8186-4960-7/93 $3.00 Q 1993 IEEE.

[3] E. Justino, E. Bortolozzi, R Saburin "Off-line Signature Verification Using HMM for Random, Simple and Skilled Forgeries" International Journal of Computer Applications (0975 8887) Volume 10 No.2, November 2010.

[4] Madasu Hanmandlua, Mohd. Hafizuddin Mohd. Yusofb, Vamsi Krishna Madasuc "Off-line signature verification and forgery detection using fuzzy modelling" M. Hanmandlu et al. / Pattern Recognition 38 (2005) 341 356.

[5] Imran Ahmed, "Writer Identification in Handwritten Documents " ICDAR, Brasil, 2007.

[6] Mubin Shaukat Tambol "Authorship Analysis and Identification Techniques: A Review" International Journal of Computer Applications (0975 8887) Volume 77 No.16, September 2013.

[7] Sean Stanko, Devin Lu, Irving Hsu "Using Machine Learning to Identify the Author of Unknown Texts" Computer Science Department Stanford University, Stanford, CA 94305, 2013

[8] Osvaldo A. Rosso, Raydonal Ospina , Alejandro C. Frery "Classification and Verification of Handwritten Signatures with Time Causal Information Theory Quantifiers" Article in PLoS One, January 2016

[9] M. Sudheep Elayidom, Chinchu Jose, Anitta Puthussery , Neenu K Sasi, "Text Classification For Authorship Attribution Analysis" Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September 2013.

[10] Jing, W. Y. "Authorship Identification for Chinese Texts Based on Dependency Grammar." Journal of Convergence Information Technology 6(6):317-328 June 2011

[11] Tufan Tas, Abdul Kadir "Author Identification for Turkish Texts" Journal of Arts and Sciences Say: 7, Mays 2007.

[12] Joachims, T "Text categorization with support vector machines: learning with many relevant features." springer 2005.

[13] Al-Harbi, A Almuraheb, Abdulmohsen Al-thubaity, Mohammad Khorseed, A Al-Rajeh . "Automatic Arabic Text Classification" JADT 2008: 9es Journes internationales dAnalyse statistique des Donnes Textuelles. 2014

[14] Al-Ma'adeed S, Mohammed E, AlKassis D, Al-Muslih F, "Writer identification using edge-based directional probability distribution features for Arabic words," IEEE/ACS International Conference on Computer Systems and Applications, pp.582-590, 2008.

[15] Rahul Kala, Harsh Vazirani , Anupam Shukla and Ritu Tiwari "Offline Handwriting Recognition using Genetic Algorithm" IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 1, March 2010.

[16] Andreas Schlapbach and Horst Bunke "Off-line Writer Identification Using Gaussian Mixture Models" The 18th International Conference on Pattern Recognition (ICPR'06) 0-7695-2521-0/06 $20.00 © 2006.

[17] Hiremath, P. S, Shivashankar, S, Pujari, J. D., & Kartik, R. K. "Writer identification in a handwritten document image using texture features." International Conference on Signal and Image Processing. 978-1-4244-8594-9/10/$26.00 c 2010 IEEE.

[18] Paraskevas, D., Stefanos, G., & Ergina, K. "Writer Identification Using a Statistical and Model Based Approach." 14th International Conference on Frontiers in Handwriting Recognition, 2167-6445/14 $31.00 © 2014 IEEE

[19] Al-Maadeed, S., Hassaine, A., & Bouridan, A. (2014). "Using codebooks generated from text skeletonization for forensic writer identification". 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA),978-1-4799-7100-8/14/$31.00 ©2014 IEEE.

[20] Djeddi, C, Siddiqi, I, Souici-Meslati, L, & Ennaji, A. (2012). "Multi-script Writer Identification Optimized with Retrieval Mechanism". 2012 International Conference on Frontiers in Handwriting Recognition." 978-0-7695-4774-9/12 $26.00 © 2012 IEEE.