Naivigyan Journal of Multidisciplinary Research (NJMR), 2025; 1(1): 15-22
Original Article | eISSN (O):
Published January 31, 2025

Different Data Mining Techniques to Recognize Handwritten Characters and Gender Identity

Bibitha Baby*, Anusha Sivanandhan, Neenu Thomas

Assistant Professor, Naipunnya Institute of Management and Information Technology,
Pongam, Thrissur- 680308, Kerala, India

*Corresponding Author's Email: bibitha@naipunnya.ac.in

Abstract

Data mining is the process of discovering patterns, correlations, trends, and useful information from large datasets using various techniques from statistics, machine learning, and database systems. The goal of data mining is to extract meaningful insights from data that can be used for decision-making, forecasting, and improving business operations. It also helps to solve business problems through data analysis. Research in other fields can be accelerated by using handwriting to determine gender. Furthermore, the study can be applied to any field that requires gender detection. This study fulfils two objectives Finding out if a writer can recognize their handwriting is the first step. The second goal is to use computer sciences and graphology to determine the gender of a text's author. In which each sample has been defined by a set of features, composed of 67 geometrical, statistical, and temporal features. The study's impact is demonstrated by the fact that its conclusions can be applied in domains where gender detection is required and that it is carried out using the help of expert and intelligent systems. The goal was to determine the individual's gender by a character analysis of the handwriting using data mining techniques for decision tree creation.

Keywords: Pattern recognition, Handwritten recognition, Character identification, Gender identification, Offline handwritten recognition, Text recognition.

Introduction

There are different methods used to identify different persons: Character identification as well as gender identification. This behavioural analysis has gained popularity in recent years due to widespread applications across diverse fields, such as psychology, education, medicine, criminal detection, marriage guidance, commerce recruitment etc. These identified handwritings reveal the inner feelings of persons though such characteristics are invisible from a person's behaviors. Therefore, traditional methods that use visible facial/biometric features or human actions to identify personal behaviours may not be effective. This analysis is used as an objective tool for studying a person's behaviours without depending on appearance-based features of persons to make a system independent of fields, data, gender, age of a person, applications, etc. Furthermore, characteristics will be sensitive to individual behaviours because graphology concentrates on individual letters, strokes, and portions of characters rather than the entire character, word, or document. Help in predicting a person's behaviours as well as gender. Several methods have been proposed for predicting personal behaviours using graphology-based handwriting in the literature.

Methods and Materials

Pattern recognition

Pattern recognition is a data analysis method that uses machine learning algorithms to identify patterns in the input data. There are different types of pattern recognition: (1) statistical pattern recognition (2) neural pattern recognition (3) template matching (4) syntactic pattern recognition. The following are just a few of the many uses for pattern recognition: (1) Image processing: Image processing uses pattern recognition and frequently a particular classification scheme to learn how to recognize patterns in images; (2) Video processing: Pattern recognition helps analyses videos to identify people, detect objects, and enable autonomous driving; (3) Speech/audio recognition: Text-to-speech converters and digital assistants like Apple’s Siri use pattern recognition to analyses voice cues and understand what different words and phrases express; (4) Natural language processing: Pattern recognition can be used to teach a computer how to speak and comprehend human language;(5) data mining: Pattern recognition is essential for extracting useful information and patterns from large quantities of data.

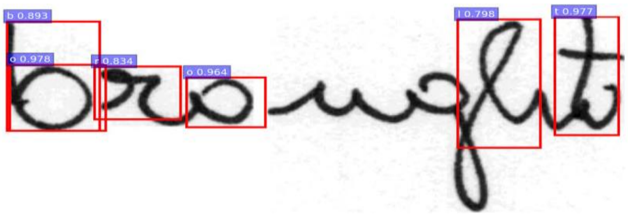
Data mining

Data mining is the process of removing significant information from enormous amounts of data. While many consider data mining to be the same as the widely used phrase knowledge discovery from data, or KDD, others view it as a crucial advancement in the interaction of information disclosure. Seven steps are included in the knowledge _finding process from data in data mining:

1. Data cleaning is the first stage in removing unnecessary and noisy data from the raw data that has been collected.
2. Data integration: Various data sources are combined into significant and valuable data at this stage.
3. Data Selection: Information needed for the study is gathered from several sources in this section.
4. Data transformation: Using various techniques, such as smoothing, normalization, or aggregation, data is transformed or integrated into the necessary forms for mining in this stage.
5. Data Mining: Various cunning methods and instruments are combined at this stage to extract data patterns or principles.
6. Pattern evaluation: At this stage, distinguishable, visually appealing patterns that convey knowledge are made based on predetermined metrics.
7. Knowledge representation: Perception and knowledge representation techniques are applied in this final step to help people comprehend and interpret the knowledge or result of data mining.

Handwritten recognition

A great attempt of research workers in machine learning and data mining has been contrived to achieve efficient approaches for approximation of recognition from data. The variety and distortion of the handwritten character set is one of the biggest obstacles to fully recognizing handwritten characters. This is because distinct communities may use a diverse style of handwriting and control to draw similar patterns of the characters of their recognized script. In Reference (S M Shamim et al.,2018), Identification of digits from where the best discriminating features can be extracted is one of the major tasks in the area of digit recognition systems. The primary objective of feature extraction in digit recognition is to eliminate redundant information from the data and obtain a more efficient representation of the word image using a set of numerical attributes. Additionally, unlike the printed characters, the curves are not always smooth. Additionally, the dataset of characters can be displayed in a variety of sizes and orientations, although they should always be written in an upright or downward position according to guidelines. Therefore, by taking these constraints into account, an effective handwritten recognition system can be created. It is quiet exhausting sometimes to identify handwritten characters as it can be seen that most of human beings cannot even recognize their own written scripts As a result, there are restrictions on what a writer can write that seem to be for handwritten document recognition. This image shows that recognizing of handwritten characters;



Character Identification Using Data Mining

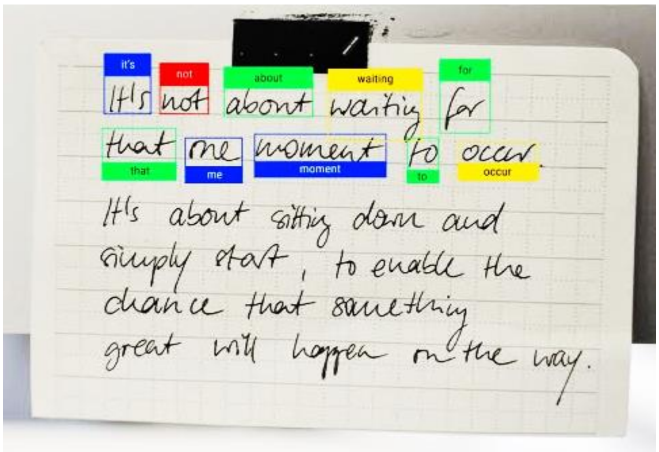
In order to create a system that is independent of fields and data, graphology-based handwriting analysis is utilized as an objective method for examining human behaviour without relying on aspects based on appearance, gender, age of a person, applications, etc.(Subhankar Ghosh et al.,2020), Additionally, because graphology concentrates on individual letters, strokes, and portions of characters rather than the entire character, phrase, or document, features will be sensitive to individual behaviours, which help in predicting person behaviours, (Robert P. Tett , Cynthia A. Palmer ,1997). In Reference (Nesrine Bouadjenek, Hassiba Nemmour, Youcef Chibani, 2017), the authors propose a system that uses the same features like topological pixel distribution and the gradient feature gradient local binary patterns. As test records, IAM, KHATT, and IAM+ KHATT these three database were used. This combined system gives 4% results in comparison with individual methods.

Gender identification

In reference (Ashish Mishra; Neelu Khare ,2015), Using Fingerprints to identify gender is one of the important techniques, in the recognition for gender identification methods done through various data mining techniques that include support vector machines, neural networks, and fuzzy-c means. Fingerprint data is indubitably the most dependable and acceptable proof till date in the court of law. Due to the enormous potential of fingerprints as an effective method of identification. Association rule mining and classification methods for gender identification and found some encouraging result. There is a need of a well-organized method for fingerprint recognition systems which will reduce computational time and increase efficiency. Gender identification using handwriting is a technique that is used to analyse handwriting samples like images to determine if the writer is male or female. Handwriting characteristics that can be analyzed include letter spacing, pen pressure, line quality, and slant. In reference (Najla AL- Qawasmesh,Muna Khayyat ,Ching Y.Suen,2023), some automatic handwritten analysis systems have been developed to detect the gender of the writer. Gender-related features have been extracted using machine learning techniques.

Text Recognition

In reference (U.Karthikeyan ,Dr. M Vanitha, 2019), text recognition is a technique that recognizes text from the paper document. If it is a writing name, signature or something else written on the paper for identifying the gender and character. The text recognition process involves several steps, that include (1) pre-processing of initial data (2) segmentation, in this step segment the image given in online and segment each character of the segmentation line (3) feature extraction, in this step convert the content of a paper document into a machine readable format (4) classification of current data (5) and finally post- processing .the final stapes was post-processing stage where an image is to convert a grayscale image. In the feature extraction stage, the paper analysis and compare the technical challenges, methods and it perform the text detection and recognition studies in the images.



It's not about waiting for
that me moment to occur.
It's about sitting down and
simply start, to enable the
chance that something
great will happen on the way.

Offline Handwritten Recognition (OHR)

Offline handwritten recognition is the process for converting handwriting image into a form that computer can use. In this process, an optical scanner converts the handwritten text into image. Then, the image is processed by a machine. The machine converts the image into characters that the computer can recognize.

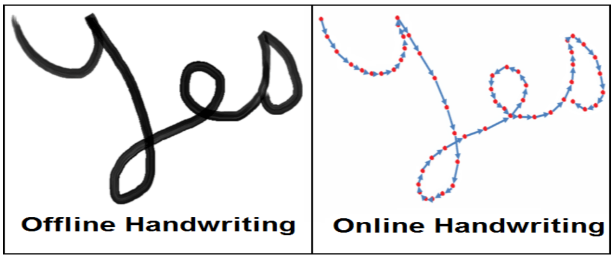


Table 1. Data mining steps

Sl.no	Stage/phases	Definition	Explanation
1	Data cleaning	It helps remove noisy or incomplete data from the data collection.	Data cleaning carries out in two major steps ; 1. Filling the missing data 2. Remove the noisy data
2	Data Integration	When multiple data source are combined for analysis, such as database, data cubes, or files.	This step enhances the accuracy and speed of the mining process. Data integration is performed using migration tool such as Oracle Data service integration and Microsoft SQL.
3	Data Reduction	It helps obtain only the relevant data for analysis from data collection.	Data reduction is performed using Naïve Bays, Decision trees, Neural network etc.
4	Data Transformation	Transforming the data into a form suitable for the mining process.	Data transformation involves mapping of the data and a code generation process.
5	Data Mining	Is the process of identifying the patterns and extracting knowledge from an extensive database.	The data is represented in patterns, models that are structured by classification and clustering.
6	Pattern Evaluation	Is the process that involves identifying interesting patterns representing the knowledge based on some measures.	Data summarization and visualization methods make the data understandable to the user.
7	Knowledge Representation	Is the process of organizing and presenting data that can be used by a system or understood by humans.	It involves creating a frame work to transform large amounts of data into a form that can be used to make decision.

Table 2. Different steps involved in text recognition

Sl No.	Stages	Definitions	Methods

1	Image acquisition	Capture the image	Resizing , Binarization, Digitalization, compression
2	Pre-processing	Enhanced the quality	Noise removal , filtering ,skew, edge detection and correction
3	Segmentation	Splitting image into characters or words	Character based , word- based, sequence based
4	Feature extraction	Extracting characteristics of an image	Statistical and geometrical features
5	Classification	Extracting characters are in a category	Decision tree, SVM, nearest neighbor , distance –based methods
6	Post processing	Increased the performance accuracy of text prediction	Confusion matrix, contextual approaches , dictionary based approaches

Table3. Handwriting recognition methods in reference (Salma Shofia Rosydaand Tito Waluyo Purboyo, 2018).

Sl. No.	Stages	Definition
1	Convolutional neural network	That uses deep learning to identify patterns in images, audio, and other data
2	Semi-incremental segmentation	For reducing waiting time and improve recognition accuracy
3	Incremental	Any new character class can be instantly learned by the system
4	Lines and words	The word segmentation into letters is a usable approach. One line segmentation is detected by scanning the written image that has been inputted horizontally
5	Parts	It use multiple key points to represent a single image
6	Slope and correction slant	Is used to reduce the style variation in writing
7	Ensemble	Is to generate multiple classifier form one base class base automatically.

Results and Discussions

The hybrid method of the handwritten character recognition system results in a rich mosaic of discoveries and discussions, demonstrating the accomplishment of significant character recognition benchmarks. Increased recognition precision is achieved by this hybrid technology, which combines recurrent neural networks and convolutional neural networks to capture the fine-tuning between sequential and spatial information. The output from the HCR system shows how well it can read a variety of handwritten characters. This hybrid design is excellent at representing the temporal connections observed in cursive writing, while also meeting the needs of various handwriting styles. Already we have an automatic handwritten analysis technique developed to detect the gender of the writer. Machine learning algorithms have been applied to extract the set of gender-related attributes. To test the gender detection methods, a sizable dataset was generated. A graphologist and a psychologist were consulted in order to select a novel set of attributes. The suggested detection mechanism was compared to the work of another researcher using benchmark data.

Conclusion

The paper aims to facilitate for identification of gender and characters using handwriting using standard classification techniques. Hybrid techniques using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have improved character recognition accuracy in Handwritten Character Identification Systems. By combining spatial and sequential data, these systems adapt better to various handwriting styles. They demonstrate robustness through careful pre- and post-processing and effective training on datasets. In gender identification from text, researchers have found that Support Vector Machines (SVM) are more effective than Bayesian logistic regression in determining an author's gender. Gender differences are most notable in personal writing but can also be seen in news articles, despite the common use of neutral language.

References

Giones, F. and A. Brem "Digital technology entrepreneurship: a definition and research agenda", Technology Innovation Management Review, 2017, Vol. 7, No. 5, pp. 44–51.

Gruppa Vsemirnogo Banka. Tsifrovaya povestka Evraziyskogo ekonomicheskogo soyuza do 2025 goda: perspektivy i rekomendatsii. Obzor. URL: http://mosopen

Hansen, B. "The digital revolution – digital entrepreneurship and transformation in Beijing", Small Enterprise Research, 2019, Vol. 26, No. 1, pp. 36–54.

Martinez Dy, L. Martin, and S. Marlow, "Emancipation through digital entrepreneurship? A critical realist analysis", Organization, 2018, Vol. 25, No. 5, pp. 585–608.

Montiel-Campos, H. and Y.M. Palma-Chorres, "Technological entrepreneurship: A multilevel study", Journal of Technology Management & Innovation, 2016, Vol. 11, No. 3, pp. 77–83.

Nambisan, S. "Digital entrepreneurship: toward a digital technology perspective of entrepreneurship", Entrepreneurship Theory and Practice, 2017, Vol. 41, No. 6, pp. 1029–1055.

Nazarov, N. Butryumova and D. Sidorov, "Development of technology entrepreneurship in a transition economy: an example of the Russian region with high scientific potential", DIEM: Dubrovnik International Economic Meeting, 2017, Vol. 3, No. 1, pp. 89–104.

Pergelova, T. Manolova, R. Simeonova-Ganeva and D. Yordanova, "Democratizing entrepreneurship? Digital technologies and the internationalization of female-led SMEs". Journal of Small Business Management, 2019, Vol. 57, No. 1, pp. 14-39.

Rippa, P. and G. Secundo, “Digital academic entrepreneurship: The potential of digital technologies on academic entrepreneurship”, Technological Forecasting and Social Change, 2019, Vol. 146, No. 9, pp. 900–911.

Song, A.K. “The Digital Entrepreneurial Ecosystem – a critique and reconfiguration”, Small Business Economics, 2019, Vol. 53, No. 3, pp. 569–590.

Venkatesh, K.A. and N. Pushkala, “Digital entrepreneurship: the technology deployment in internationalization speed in the digital entrepreneurship era and Opportunities-Tirumala Tirupati Devasathanam (TTD)”, International Journal on Recent Trends in Business and Tourism, 2018, Vol. 2, No. 4, pp. 39–42