***Naivigyan Journal of Multidisciplinary Research* (NJMR)**
Volume 1, Issue 1, April 2025, Pages: 9-16
Original Article | ISSN (O): 3108-1487

Different Data Mining Techniques to Recognise 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

Finding patterns, correlations, trends, and important information in big datasets through a variety of statistical, machine learning, and database system techniques is known as data mining. Finding valuable insights in data that may be applied to forecasting, decision-making, and enhancing corporate operations is the aim of data mining. Data analysis is also useful for resolving corporate issues. Identifying gender from handwriting can speed up research in other areas. Additionally, the study can be used in any industry where gender detection is necessary. This research accomplishes two goals. The first stage is to determine whether a writer can identify their handwriting. The second objective is to identify the gender of a text's author using graphology and computer science. wherein a set of features, consisting of 67 geometrical, statistical, and temporal features, has been used to define each sample. The impact of the study is evidenced by the fact that its findings are applicable in fields where gender detection is necessary and that it is conducted with the assistance of knowledgeable and sophisticated technologies. Through character analysis of the handwriting and the use of data mining techniques for decision tree development, the objective was to ascertain the person's gender.

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

Introduction

Character and gender identification are two of the various techniques used to identify individuals. Due to its extensive applicability in various sectors, including psychology, education, medicine, criminal detection, marriage counselling, and recruitment in the commercial sector, among others, this behavioural analysis has gained popularity in recent years. Even though these traits are not obvious in a person's actions, these recognised handwritings provide insight into their innermost thoughts and emotions. As a result, conventional techniques that identify individual behaviours using observable facial or biometric traits or human actions might not work. To create a system that is independent of fields, data, gender, age, applications, and other factors, this analysis is utilised as an objective instrument for examining people's behaviours without relying on their

outward appearance. Furthermore, because graphology focuses on individual letters, strokes, and character segments rather than the complete character, phrase, or document, features will be sensitive to individual behaviours. aid in forecasting a person's gender and behaviour. In the literature, a number of techniques have been put to use graphology-based handwriting to predict individual behaviours.

Methods and Materials

Pattern recognition

Pattern recognition is a data analysis method used to identify patterns in the data received through the use of machine learning algorithms. 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

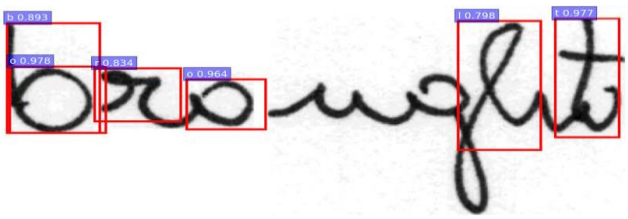
Data mining is the process of removing significant information from enormous amounts of data. While many consider data mining to be the same as the widely used phrase knowledge discovery from data, or KDD, others view it as a crucial advancement in the interaction of information disclosure. Seven steps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

Handwritten recognition

The data mining and machine learning researchers have strived to come up with practical approaches to approximating data recognition. The variation and distortion of the handwritten character set are one of the major issues that make it difficult to recognise handwritten characters completely. This can be explained by the fact that various communities can use various styles of handwriting and have control over creating similar character shapes in their familiar script. One of the primary matters of a digit recognition system is finding the digits based on the best distinguishing features, says the Reference (S M Shamim et al., 2018). In the digit recognition, the primary objective of the feature extraction is to eliminate the unneeded data in the input and apply a set of number qualities in forming a more proficient representation of the word picture. Moreover, the curves are not necessarily as clean as the characters in print. An array of characters can also be presented at various sizes and orientations, although it is usually advisable to write the characters in an upright or down-facing position. It is thus possible to come up with an effective handwritten recognition system by considering these constraints. It is a bit burdensome at times to trace handwritten characters because most people are not in a position to even recognise their own handwriting. Therefore, the restrictions on what a writer can write exist and seem to be aimed at the acknowledgement of handwritten documents. The following image shows the recognition of handwritten characters:



Character Identification Using Data Mining

The application of graphology-based handwriting analysis is an objective method of analysing human behaviour without relying on aspects such as appearance, gender, age, or other factors, to create a system that is not dependent on specific fields or data (Subhankar Ghosh et al., 2020). Moreover, features will be responsive to personal habits, as the study of graphology is based on personal letters, lines, and parts of the character, rather than the entire character, phrase, or document. This helps in the forecasting of individual behaviour (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 as topological pixel distribution and the gradient feature gradient local binary patterns. As test records, IAM, KHATT, and IAM+ KHATT, these three databases were used. This combined system gives 4% results in comparison with individual methods.

Gender identification

According to Ashish Mishra and Neelu Khare (2015), one of the significant methods of recognition involved in gender identification is using fingerprints to identify gender, which is extracted through several data mining methods, such as support vector machines, neural networks, and fuzzy-c means. It is undoubtedly true that the most valid and recognised evidence before the court of law as of now is the fingerprint data, since fingerprints can be a very potent means of identification. In identifying gender, the association rule mining and classification techniques were employed, and some excellent outcomes were promising. In the case of fingerprint identification systems, it needs a systematic approach is needed to reduce the time of calculations and increase the effectiveness. Gender identification using handwriting is the process of examining handwriting samples, including photographs, to determine their writer as either a male or a female. Slant, line quality, pen pressure, and letter spacing are some of the aspects that can be analysed in handwriting. There are also certain automatic handwritten recognition tools made to determine the gender of the writer (Najla AL-Qawasmesh, Muna Khayyat, Ching Y. Suen, 2023). Gender-related features have been extracted with the use of machine learning technology.

Text Recognition

U. Kartikan and Dr M. Vanitha (2019) state that one of the ways through which a paper document may be identified is text recognition. It may be a name, signature, or some other writing on the page that signifies the gender of the character. The steps involved in the text recognition process are pre-processing of the original data, segmentation (division of the online image and each character on the segmentation line), feature extraction (conversion of the material of a piece of paper into a machine-readable one), classification of the existing data, and post-processing. The final step was the post-processing step, whereby an image is changed into grayscale. In the feature extraction stage, the paper analyses and compares the technical challenges, methods, and it performs the text detection and recognition studies in the images.

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.

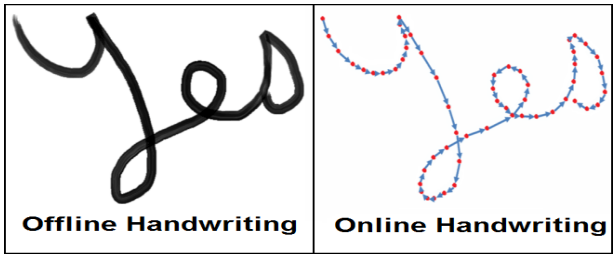


Table 1. Data mining steps

Sl.no	Stage/phases	Definition	Explanation
1	Data cleaning	It helps remove noisy or incomplete data from the data collection.	Data cleaning aids in eliminating noise or missing data in the data collection. There are two significant steps of data cleaning: 1. Filling the missing data 2. Remove the noisy data
2	Data Integration	When integrating data sources to analyse them, multiple data sources may be involved: databases, data cubes or files.	This step increases the speed and accuracy of the mining process. Data is integrated using migration tools such as Microsoft SQL and Oracle Data Service Integration.
3	Data Reduction	It helps in reaping out of data, collecting only the relevant data to be analysed.	Naive Bayes, decision trees, neural networks and others are employed in reducing the data.
4	Data Transformation	Data Mining is the process of identifying patterns and making conclusions from a large database.	Data mapping and the development of code are part of data transformation.
5	Data Mining	It is the procedure for finding trends and drawing conclusions from a large database.	The data are represented in patterns and models arranged in classification and clustering.
6	Pattern Evaluation	It is the process that involves the identification of interesting patterns that demonstrate the information based on some metrics.	Data visualisation and data summarising techniques assist the user in comprehending the data.
7	Knowledge Representation	The process of organising and presenting data in a way that it may be comprehended by a system or an individual.	It involves coming up with a system for transforming large amounts of data into a decision-making format.

Table 2. Different steps involved in text recognition

Sl No.	Stages	Definitions	Methods
1	Image acquisition	Capture the image	Resizing, Binarisation, Digitalisation, compression
2	Pre-processing	Enhanced the quality	Noise removal, filtering, skew, edge detection and correction
3	Segmentation	Splitting an 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 is 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 Rosyda and 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 improving 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 input horizontally
5	Parts	It uses multiple key points to represent a single image
6	Slope and correction slant	It is used to reduce the style variation in writing
7	Ensemble	Is to generate multiple classifiers form one base class automatically.

Results and Discussions

The hybrid method of the handwritten character recognition system offers a plethora of results and discussions and demonstrates that key character recognition standards were reached. The hybrid approach is a fine-tuning of the sequential and spatial information between recurrent neural

networks and the convolutional neural networks, which leads to a higher level of recognition. The output of the HCR system can reveal the capabilities of the system to analyse a great variety of handwritten characters. This hybrid design has been found to do even better at preserving the temporal associations of what is observed in cursive handwriting, besides supporting the needs of various handwriting styles. We already have a way to automatically analyse the handwriting and determine the gender of the writer. Machine learning algorithms have been applied in order to extract the set of gender-related features. A huge sample was developed to test the gender detection methods. A psychologist and a graphologist were consulted to select a new set of characteristics. The proposed system of detection was compared with the work of another researcher using benchmark data.

Conclusion

The idea behind the study is that common classification algorithms can be used to ease the process of locating characters and gender using handwriting. The use of hybrid methods based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) has allowed Character Identification Systems to identify characters with greater accuracy in Handwritten Character Identification Systems. It is these systems that would be more adaptive to the variations in handwriting by combining sequential and spatial information. They are robust through careful pre- and post-processing and effective data training. In the context of textual gender identification, researchers have found that Support Vector Machines (SVM) are more efficient when compared to Bayesian logistic regression in the recognition of the gender of an author. Gender differences are evident most in personal writing, although most of the time neutral language is used in news articles.

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