An Automated Optical Character Recognition of Handwritten English Letters using Decision Trees & Random Forests.

1) Introduction and Overview

- Project idea in detail

Optical character recognition or optical character reader (OCR) is the automated conversion of images of typed, handwritten, or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (e.g., the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (e.g., from a television broadcast). the goal of a character recognition system is to transform handwritten text document on paper into a digital format that can be manipulated by word processor software. The system is required to identify a given input character form by mapping it to a single character in a given character set.

- Similar applications in the market

1- Evernote

This app can quickly capture all your handwritten ideas and notes with Evernote's built-in camera on Android and Apple devices. It was already possible to digitize texts in Evernote with the help of the Penultimate app , but an update also allows it to be done from your own application thanks to the Evernote Scannable function. Among its various options, it allows you to digitize the text you find, expanding the possibilities and being a perfect ally for, for example, passing notes between colleagues.

2- Pen to print

In a variation on the handwriting recognition concept, Pen to Print reads scanned handwritten documents and converts them into editable, searchable digital text that can be stored on your device or within a cloud service. Pen & Print allows you to turn handwriting to text on the iPad or iPhone without an Apple pencil. The app's handwriting OCR (optical character recognition) engine extracts text from paper documents, like letters, school notes, meeting notes, and grocery lists, allowing those who prefer to write in longhand the freedom to continue.

3- Goodnotes 5

If you seek a powerful app with handwriting recognition, check out the updated GoodNotes 5 for searchable notebook and document creation. Available for iOS devices, it is

suitable for drawing and illustrating. In addition, the written text becomes vector and has the option of being digitized. It also incorporates a search engine to locate the desired information among the notes, both those that are at hand and those that have already adopted a digital format.

4- Notability

A perfect iPhone and iPad app for converting handwriting to text. And not just only that, it is a full package to go completely paperless. Apart from basic handwriting features, users can create a full document by including images, text, GIFs, web pages, and whatnot. The app also provides an option called multi-note in which the users can work on more than one note side by side. And all of your documents will be automatically saved to iCloud. When counting all the features, this is surely the best handwriting app for iPad and iPhone users.

5- Write for iPad

If you prefer to write longhand but need to see your text in digital format, consider Write Pad for iPad. You can configure a host of options to recognize input forms and predefined commands, or you can input lettering with your finger or a stylus.

- A Literature Review of Academic publications(papers/books/articles) relevant to the problem

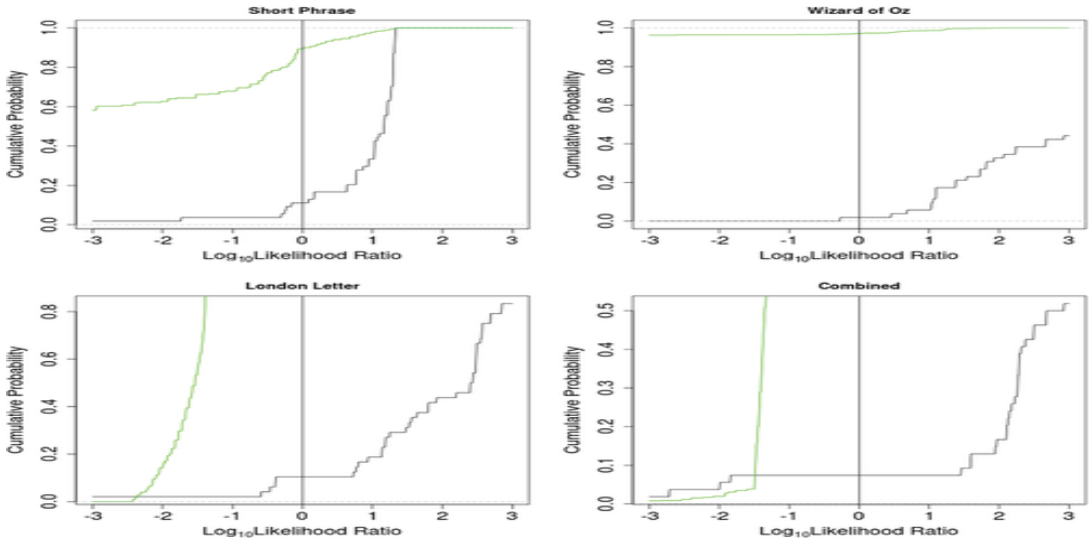
Forensic handwriting analysis has traditionally been conducted by trained forensic examiners who rely on visual inspection to compare writing samples. The assumption that underlies forensic handwriting examination is that each person has a unique writing style that is developed over years and that may depend on cultural, demographic, and physiological factors [1]. After reviewing the evidence, forensic document examiners summarize their findings using categorical conclusions such as “identification,” “strong probability,” “probable,” “indicate,” and “indeterminable,” among others [2]. The 2009 National Research Council's report “Strengthening Forensic Science in the United States” includes an outline of current practices and concludes that the scientific basis of handwriting analysis must be strengthened [3]. In recent years, methods to quantify the similarity between two handwritten samples have been proposed, and software to implement those methods is now available. FLASH ID is a software tool developed by Sciometrics that uses the topology and “geometric features” in handwriting samples from a closed set of writers. FLASH ID then provides a ranked list of the writers in that set who are most likely to have written the questioned document [4]. Another approach proposed by Hepler et al. numerically evaluates the similarity in handwriting with score-based likelihood ratios (SLRs) [5].

Earlier work performed by researchers in the Center for Statistics and Applications in Forensic Evidence (CSAFE) focused on estimating the probability that two documents have the same source [6]. This is known as forensic identification of a common source. In this context, the goal is to determine whether the two documents were written by the same person, even if the identity of that person is unknown [7]. Furthermore, Crawford and collaborators considered only the closed set scenario, which requires all of the potential sources of the handwriting to be known [8].

We extend the scope of the research to the open set case, where the writer of the documents can be anyone in a defined population subgroup. More specifically, the objective is to estimate the probability that two samples were authored by the same or by a different writer without the requirement of knowing all of the potential sources. The SLR approach provides an open-set solution to the question of whether the similarity observed between a questioned and a known document supports the proposition that the documents were written by the same person. To compute the SLR, several decisions regarding the distance/scoring function and the method for estimating score densities, for example, need to be made. We explore both common source and specific source SLR approaches using simple distance measures between two documents as inputs for a random forest which outputs a score between 0 and 1 which can be loosely interpreted as the empirical probability of same source. We then use a kernel density framework to estimate the densities of the similarity scores among pairs of documents known to have been written by the same or by different persons. These approaches are applied to a handwriting data set collected by CSAFE.

Handwriting samples were collected in a study conducted by CSAFE at Iowa State University [9]. Participants were asked to copy three prompts in their natural handwriting with a ballpoint pen on unlined paper supplied by CSAFE. The prompts include “The London Letter,” an excerpt from the book *The Wizard of Oz*, and the short phrase: “The early bird may get the worm, but the second mouse gets the cheese.” The longest of the prompts is “The London Letter” which has been used in other studies because it contains the numbers 0 through 9 and all of the letters in the alphabet in both uppercase and lowercase. Here, we use the samples obtained from 90 participants, most of who provided three replicates of each prompt.

The results of applying the trained random forest to the CSAFE data showed promising discrimination for the common source question using the similarity scores as the length of the prompt increased. As the length of the writing sample increases, the same-writer and different-writers KDEs become more separated. However, the KDEs for the shortest prompt are not well separated. The degree of separation between the two KDEs is directly related to the performance of an SLR. When the comparison is known to come from different writers, a well-behaved SLR system would result in a small log-SLR value (less than 0), whereas a comparison known to come from the same writer should result in a large log-SLR value (greater than 0). For a closer inspection of the



performance of the common source SLR system to the CSAFE data, Tippett plots were utilized (Figure **6**). Tippett plots assess the performance of the SLR system as a whole by splitting the data into known same-writer and known different-writer comparisons. Then, the empirical cumulative distribution function (ECDF) within each group of comparisons is plotted (black for the known same-writer and green for the known different-writers comparisons). Ideally, the ECDF for the known different-writers (green) would be to the left of the vertical line at zero and the ECDF for the known same-writer (black) would be to the right of the vertical line at zero. This would indicate that the SLR system made no errors on the CSAFE data set. We can see that the green line is to the left of zero only for the London Letter and combined prompts, whereas none of the black lines are completely to the right of zero. This indicates that it is more difficult for the SLR system to properly support the prosecution proposition when it is true than it is to properly support the defense proposition when it is true. The area below the black line and to the left of the vertical line at zero indicates the rate of SLR values that support the defense proposition when the prosecution proposition is actually true (Type 1 error). The results in Figure **6** show the largest Type 1 error rate for the combined prompt. Similarly, the area above the green curve and to the right of the vertical line at zero indicates the rate of SLR values that support the prosecution proposition when the defense proposition is actually true (Type 2 error). The results in Figure **6** show the largest Type 2 error rate for the short phrase prompt.

Resources

- 1R. Huber and A. Headrick, *Handwriting identification: Facts and fundamentals*, CRC Press, Boca Raton, FL, 1999.

  Crossref Google Scholar

- 2 Expert Working Group for Human Factors in Handwriting Examination, *Forensic handwriting examination and human factors: Improving the practice through a systems approach*. U.S. Department of Commerce, National Institute of Standards and Technology. NISTIR 8282, 2020.

  Google Scholar

- 3 National Research Council Committee on Identifying the Needs of the Forensic Sciences Community, *Strengthening forensic science in the United States: A path forward*. The National Academies Press, Washington, DC. https://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward, 2009.

  Google Scholar

- 4J. J. Miller, R. B. Patterson, D. T. Gantz, C. P. Saunders, M. A. Walch, and J. Buscaglia, A set of handwriting features for use in automated writer identification, *J. Forensic Sci.* **62** (2017), no. 3, 722– 734. https://doi.org/10.1111/1556-4029.13345

Wiley Online LibraryPubMedWeb of Science®Google Scholar

- 5A. B. Hepler, C. P. Saunders, L. J. Davis, and J. Buscaglia, Score-based likelihood ratios for handwriting evidence, *Forensic Sci. Int.* **219** (2012), no. 1, 129– 140. https://doi.org/10.1016/j.forsciint.2011.12.009

  Crossref PubMedWeb of Science® Google Scholar

Using Random Forests for Handwritten Digit Recognition

https://www.researchgate.net/publication/4288221_Using_Random_Forests_for_Handwritten_Digit_Recognition

Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)

https://www.academia.edu/61998976/Handwritten_Optical_Character_Recognition_OCR_A_Comprehensive_Systematic_Literature_Review_SLR_

Off-line Arabic Handwriting Recognition Using Dynamic Random Forests

https://www.academia.edu/83138119/Off_line_Arabic_Handwriting_Recognition_Using_Dynamic_Random_Forests

Isolated Printed Arabic Character Recognition Using KNN and Random Forest Tree Classifiers

https://link.springer.com/chapter/10.1007/978-3-319-13461-1_2

Handwritten Urdu character recognition via images using different machine learning and deep learning techniques

https://pdfs.semanticscholar.org/d529/56f2dfc1f503cc15108a603e1f6b3825ea29.pdf

---

1) Proposed solution & dataset

- Main Functionalities

1- Draw character

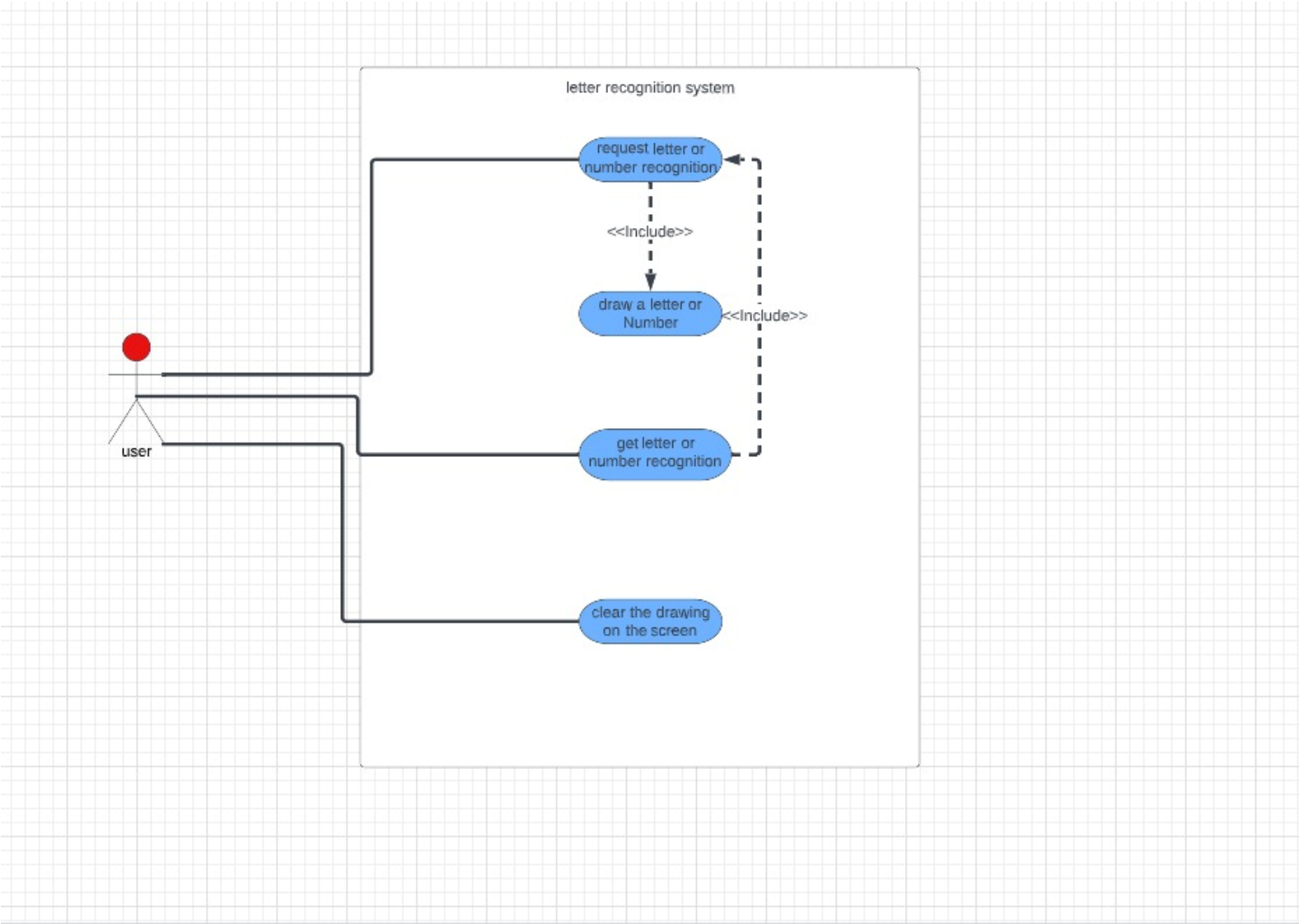
2- Save image of the character

3- Upload Image

4- Convert Image to pixels

5- Predict the pixels to generate the letter

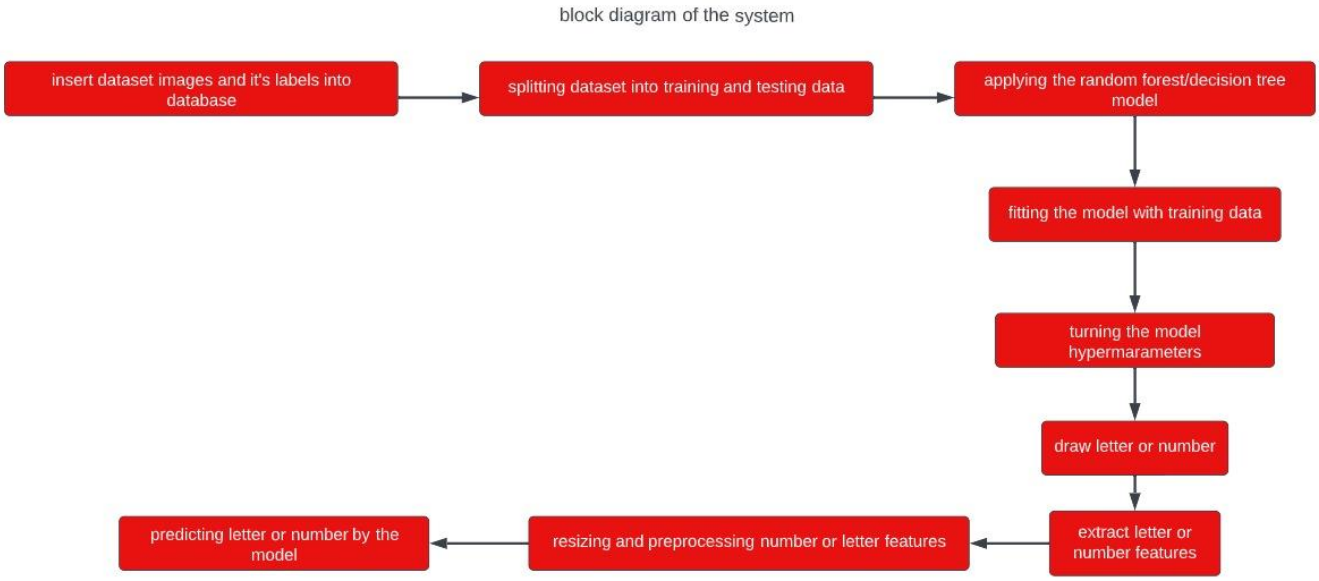
6- Clear drawing.



- Dataset

Already in GitHub: https://rb.gy/0dbf7k

3- Applied Algorithms



4- Experiments & Results

Experiments:

We tried the decision tress algorithm and random forests algorithm. The decision tress algorithm was 59% accuracy in the model, and the random forests algorithm was 79% accuracy.

Results:

	0	1	2	3	4	5	6	7	8	9	10
0	239	0	0	0	0	0	1	0	1	0	0
1	0	250	0	0	0	0	0	0	0	0	0
2	0	0	257	1	0	0	0	4	2	0	0
3	0	0	4	312	0	2	0	2	3	0	0
4	0	0	0	0	297	0	0	0	0	2	10
5	0	2	0	5	0	267	1	0	2	2	0
6	0	0	0	0	0	0	312	0	0	0	0
7	0	1	0	0	1	0	0	334	0	5	0
8	1	2	0	4	0	0	1	0	269	1	0
9	0	0	0	1	3	1	0	8	0	327	2
10	0	0	2	0	1	0	0	0	0	1	297

5- Analysis, Discussion and Future work

- By analyzing the results of both algorithms, we got insights about the efficiency of both algorithms then we found that performance of results of Random Forest algorithm is better than Decision tree.

- The advantage of random forest algorithm in this solution is the testing accuracy percentage which is very good while the disadvantage is the training accuracy percentage which may be overfitting and may not predicting accurately the results of the data which was not trained.

- The future modification for solving the problem can be by applying another machine learning algorithm which can has better efficiency and performance for the classification of character recognition.

6- Development Platform

Tools: Jupyter Notebook.

Programming Languages: Python.

Python Libraries:

- Model:
  - pandas
  - numpy
  - matplotlib.pyplot
  - warnings
  - sklearn.model_selection
  - sklearn.ensemble
  - sklearn.tree
  - sklearn.metrics
  - cv2
  - joblib
- GUI:
  - Os
  - PIL
  - cv2
  - numpy
  - tkinter
  - Joblib
  - pandas