FIT: Report 1

Handwriting detection and Note Conversion using OCR

Pietka Ł, Barot H, Krzyszosiak M, Markowicz J, Milewski M, Sobczak M

March 12 2025

Contents

1 Re-evaluating Report 1

In report 1, the aim of the project was stated to be creation of a program that utilizes OCR technology, which allows for recognition and digitization of handwritten notes into easily readable and storable text documents on a computer. However, the project was not clearly defined and the problem statement was not well articulated. The target population was not specified, and the aim of the project was not clearly defined. The project lacked a clear focus and direction, which made it difficult to understand the purpose and goals of the project.

Due to the target population being vague, multiple areas of this project were hard to define, and ambiguous as to what direction the project is headin in. This was a major roadblock because when we discussed questions like "Do the samples include math equations?" or "Do the samples include letters?" "Are the letters allowed to be geometrically malformed or the the samples have to be clean and perfectly written?". Our group memmbers had different approaches to these questions, and this led to a lot of confusion and miscommunication.

Hence, this report would like to revisit and address the flawed initial approach. Define the who the project is for, which will make it simpler to find source of raw data, and implement a model that will best suit the use case defined.

1.1 Problem Statement

Certain two members of our group particularly enjoyed writting english literature. Wether it was analitical essays, or fantasy stories. They both agreed that writting on a physical paper felt the most comfortable and natural. However, they couldn't organize their raw notes and digitizing them was really time consuming. Hence, we decided to create a program that would help them with this problem. The specific aim of this project is to create a program for the the two members of our group, that could assist in digitization of their english literature notes.

2 Data Collection Procedure:

In the previous report, we decided to use the GNHK dataset [1], as it best suited the freely available OCR model that we chose. However, this was a large dataset, that contained a variety of different samples. The sample types included math equations, plain english text, distorted handwritting and much more. Additionally these datasets conainted a the different types of english (british, american, asian) used all over the world.

This essentially ment, that this dataset ranged over a wider domain than what was needed for our project. This is displayed using a venn diagram in figure 1. Set A (red circle) is the collectioon of what GNHK has to offer, meanwhile Set B (blue circle) is what is required for our project. The intersection of the two sets is what we need.

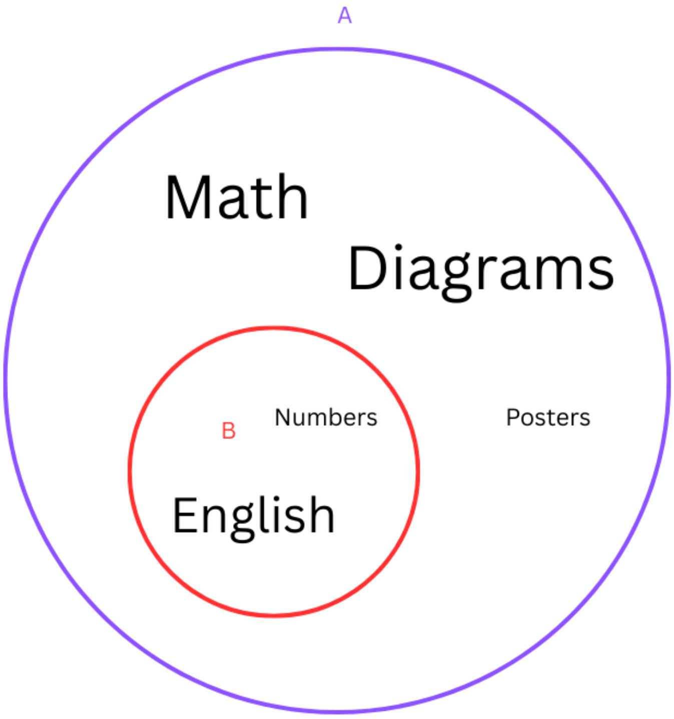

Figure 1: Venn Diagram of Dataset Characteristics vs. Project Needs

3 Raw Data (samples and syntatic characteristics):

4 Normalization of Raw Data:

5 Extraction of words/letters:

6 Closing Thoughs:

References

[1] Alex W. C. Lee, Jonathan Chung, and Marco Lee. *GNHK: A Dataset for English Handwriting in the Wild*. 2021.