Project Topic

Your Name

October 23, 2025

1 Introduction

Currently many different computer vision systems are used to interpret handwriting as digital text, but these systems are constrained usually to one style or lack the complexity to comprehend more challenging handwriting. My goal is to create a series of computer vision systems that maintain accuracy while circumventing some of these constraints by having an AI redirect the given image of text to the appropriate computer vision system to accurately interpret it. By providing a single interface able to interpret and return text to users from diverse styles, I have created a functional multipurpose AI system implementing computer vision systems that can be implemented for many different purposes from a single source. This approach allows for greater adaptability and flexibility, as it enables the AI to recognize a wide variety of handwriting styles without being limited to a single computer vision system. The system is designed to evolve over time, allowing for the continual input of new computer vision systems to expand the AI models capabilities in reading and transcribing different kinds of text. In addition, by integrating multiple specialized vision modules, the AI can effectively balance precision while also ensuring that even uncommon or low quality handwriting can be interpreted with a reasonable degree of accuracy. For the implementation of this project, I created a simple application allowing image and optional text input. The image input will take the image of the text while the optional text box would allow for a description of the writing style, skipping the optional analysis from a large AI model to detect a specific writing style. From here, the file will be run through the required computer vision system or systems to convert it to plain text by an MCP Client sending the image to the MCP server for the determined writing style. This will be returned to the user in the form of a text file that they can download and if the size permits, it will also be displayed onto the page itself. The images chosen are diverse Latin alphabet handwriting styles including the generic modern handwriting found in the EMNIST data sets. I convert the images of this text using the PIL library to convert it to a similar size to that of the individual image files in the EMNIST database. From here, if they are a part of a larger or more popular writing style, the images will be fed into a convolutional neural network module which will perform the training and measure accuracy as it goes. If the accuracy is satisfactory, the resulting computer vision system will be made accessible to the AI implemented by MCP. Small fluctuations in accuracy do occur in each run of the program. If the text is from a smaller dataset then I perform the necessary labelling myself and attempt to process text with the unmonitored clustering method. This provides far less accuracy,

but the system is still provides an output. This involves the same image processing steps as convolutional neural networks and also may be used by the AI. The text is output on the simple user interface in the same manner regardless of whichever computer vision system the AI chose to use. By connecting these systems through the Model Context Protocol, this project demonstrates how new AI architectures can expand beyond the limited capabilities of individual datasets and models to form expandable systems for complex computer vision tasks. Future developments may include the implementation of more complex output with potential areas of inaccuracy, such as typos analyzed and identified by a program linked to a dictionary dataset, and the continuous integration of more and more computer vision systems.

2 Related Works

Computer Vision Methods

Within the AI field of computer vision, one popular goal is the transcription of handwritten text. By different organizations, this is called handwriting recognition technology or Handwritten Text Recognition (HTR). Researchers implemented a convolutional neural network system was utilized to build a computer vision system able to recognize digits from the MNIST and achieve accuracy of up to 100% on the number 1 in the test data[5]. Essentially a convolutional neural network uses a combination of a single input layer then convolution layers that create feature maps, pooling layers that decrease data, and fully connected layers to create and process feature maps from images and fully connected layers that link the neurons to each of the potential final output results[5]. The created convolutional neural network is able to receive further input images and can assign a label to them based on what it interprets the image to be. These are very effective systems when a large, thoroughly labeled, dataset is available, but according to the research focused on obscure texts, specifically a medieval book with a Runic script known as the *Codex Runicus*, there are issues with methods like convolutional neural networks, and recurrent neural networks because they require deep learning that involves these vast labeled datasets inaccessible to some less common scripts and styles[4]. Alternative options that did not involve deep learning were studied and implemented to accurately transcribe segments of the *Codex Runicus*. While doing this, they used learning free methods such as unsupervised clustering which creates clusters which are subdivided, then the labels of each cluster are propagated through the other symbols, and the few-shot classification method which seeks to represent each individual symbol as a node in a graph and compare similarity between each pair of symbols[4]. With modern methods, it is very possible to perform highly accurate machine learning for popular fonts and styles but the pursuit of handwritten text recognition for any obscure text may result in lower accuracy systems that avoid machine learning methods. These methods may all be used to accomplish similar tasks but at different scales, when attempting to transcribe a frequently used writing style with a large, labeled database available. a convolutional neural network should always be used.

Datasets

There are several different frequently implemented datasets for handwritten text recognition computer vision. As was seen in the implementation of a high accuracy convolutional neural network, the MNIST is a dataset used in computer vision systems[5]. This is a popular data set used for testing computer vision systems provided by the National Institute of Standards and Technology consisting of 70,000 grayscale images of 28 by 28 pixel digits[5]. Because the text is labelled accurately and organized, it allows for easy implementation of computer vision systems, but to add further data to a dataset like this or use the output computer vision system to analyze another image, it would have to be processed similarly, a task that is much more complicated. On the NIST website for another dataset, the EMNIST, documentation clarifies that this new data set includes images of characters from a-z in the same grayscale 28 by 28 pixel format used by the MNIST dataset[2]. Each character in the dataset is labelled with its ASCII value. This makes many methods described for the other computer vision system very easily applicable to this extended version of the dataset they had used. More broadly, datasets are accessible for many written styles and languages, and complications only seem to arise when pursuing data sets for languages and styles that are rarely used. The preparation of images for computer vision is a complex process involving simplifying images and often storing them on a gray scale. As was seen with the EMNIST images, they were represented in gray scale with gray scale pixels. Additionally, a gaussian blur smoothed the image, ROI Extraction Centered Frame, and resizing and resampling were applied to ensure that the images were the correct consistent size and had the proper shading[2]. The final result was images with white text with pixels blurred into a black background along the edges of the letters. In implementations of interactive image measurement software using Tkinter and PIL libraries from Python, researchers showed how tools may be used to perform the individual tasks implemented for the EMNIST database[3]. They explain that the Python PIL libraries may be used to perform image processing such as grayscale conversion, Gaussian blur, image sharpening, binarization and more[3]. The PIL package in Python could effectively be used for the preparation of images for EMNIST computer vision systems. The same methods could also be easily implemented to feed images into the output computer vision systems during their implementation.

Model Context Protocol

To connect AI models with the tools that they require to function optimally and perform all tasks that may be required of them, there are several different popular methods. One recently released was the Model Context Protocol (MCP) and according to its own website, this protocol is a standardized way to connect AI applications to other systems[1]. This is primarily described in the context of agentic AI applications. The Model Context Protocol simplifies the integration of algorithm models, platform tools, and data sources to an AI agents[6]. Using the model context protocol, AI systems would be able to access and use the computer vision systems that are created using different datasets. It would become the AI's task to select the appropriate computer vision system to analyze an image that is fed to it.

3 Methodology

While seeking to implement the necessary computer vision systems to complete this project, I first explored the computer vision systems frequently used by other researchers in the field. I discovered that among them one of the most popular was the MNIST[5]. From there, further research soon revealed the existence of the EMNIST which included the full set of characters that I required for my project. Thus, I used the dataset's parameters to create a convolutional neural network (CNN), a multi-layer architecture designed for image classification tasks. To perform the machine learning tasks involved, I utilized the Python PyTorch library. The model begins with an input of a single-image, 28 by 28 pixels. From here, kerneling with 10 filters with a 5 by 5 size and stride of 1 are applied in the first convolutional layer. This operation extracts 10 feature maps with emphasized details of the image. From here, a pooling layer halves the dimensions of each feature map reducing them to 12 by 12 squares. Next, another convolutional layer expands the feature map depth to 20 using the same kernel size and stride. From here, a second pooling layer simplifies the feature map to dimensions of 4 by 4 for each map. Finally, three linear fully connected layers map the 320 inputs to 104 neurons, then reduce that to 52, and finally bring that down to 26, fully aligned with the number of potential characters in the dataset.

4 Results

5 Conclusion

References

[1] What is the model context protocol (mcp)?, 2025.

[2] Cohen, G., Afshar, S., Tapson, J., and van Schaik, A. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)* (2017), pp. 2921–2926.

[3] He, X., Wang, Z., and Zhan, S. *Design and implementation of interactive image measurement software based on Tkinter and PIL library.* Association for Computing Machinery, New York, NY, USA, 2025, p. 564–568.

[4] Souibgui, M. A., Bensalah, A., Chen, J., Fornés, A., and Waldispühl, M. A user perspective on htr methods for the automatic transcription of rare scripts: The case of codex runicus. *J. Comput. Cult. Herit. 15*, 4 (Mar. 2023).

[5] Yu, Y., and Tian, Y. Research application of computer vision-based convolutional neural network in handwriting recognition technology. In *Proceedings of the 4th International Conference on Computer, Artificial Intelligence and Control Engineering* (New York, NY, USA, 2025), CAICE '25, Association for Computing Machinery, p. 177–181.

[6] Zhang, X., Dong, X., Wang, Y., Zhang, D., and Cao, F. A survey of multi-ai agent collaboration: Theories, technologies and applications. In *Proceedings of the 2nd*

Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence (New York, NY, USA, 2025), DEAI '25, Association for Computing Machinery, p. 1875–1881.