Data set choice

This project intends to create a program that can predict the style of a Chinese character in image form after training itself on PNG files of traditional Chinese calligraphy characters written with a brush labeled with their styles. This program is interesting as it can analyze Chinese characters written by anyone and try to deduce the style it is most similar to, even if the writer did not intend to write in any style.

Methodology

1. **Data preprocessing**
    a. The input data(the PNG files) will first need some noise polishing. The image should be filtered such that only the portions that contained texts are used for training. To extract more information from the image, I may have to create a 2D array of RGB values for the pixels of the image. Cropping and rescaling may be necessarily to make the dataset more feasible to train on. The usage of CNN as an image recognition technique may be something that I will need to research more about.
2. **Machine learning model**
    a. This machine learning model will be a classifier that may assign one of many calligraphy styles to an input image. Perhaps I may emulate the *TypeFont model* by Vasile Peste which achieved the identification of the font of English characters presented in the form of an image.
3. **Presentation**
    a. I would like to use this opportunity to design a web/local application where users may submit a photo of their written Chinese character (preferably with a Chinese brush) and the program utilizes this trained model to predict the style this character resembles the most of.

References

1. **https://blog.usejournal.com/making-of-a-chinese-characters-dataset-92d4065cc7cc**
2. https://github.com/Vasile-Peste/Typefont.git