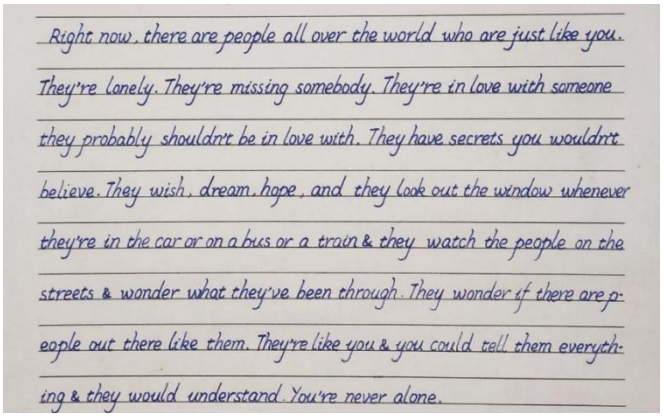
Handwritten Text Recognition

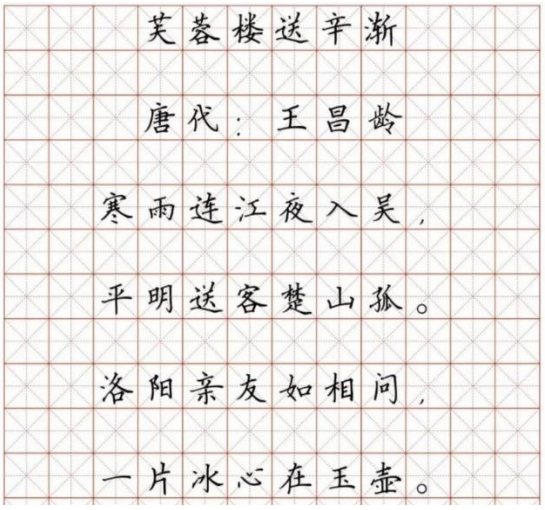
Motivation: Text recognition, also known as Optical Character Recognition (OCR), is the process of converting printed or handwritten text into machine-readable format. The significance of text recognition lies in its ability to extract meaningful information from physical documents, enabling a wide range of applications such as document digitization, text data analysis, and improved accessibility. In this project, you will implement text recognition on handwritten text.

a) Handwritten English composition (handwritten_English.jpg). 424 characters. (50 pts)



Expected result: "Right now there are people……
They re lonely They……"
(No requirement for punctuations (like ",", "&", "-").)

b) Handwritten Chinese poetry (poetry_Chinese.jpg). 39 characters. (50 pts)



Expected result:

"芙蓉楼送辛渐
唐代 王昌龄
寒雨连江……"

Hint for this project: Please try to avoid calling built-in functions of your core algorithms, otherwise your rating will be negatively affected. Deep learning algorithms are prohibited. For most cases, you can simply determine whether a built-in function can be called based on the implementation difficulty and the correlation of the algorithm.

Checkpoints:

1) How do you eliminate the influence of unrelated factors? (Such as lines, noise, etc.)
2) How do you preprocess the text?
3) How do you segment the characters?
4) How do you perform text recognition?
5) How is the final effect? (Show your recognition accuracy.)

$$recogniton\ accuracy = \frac{Number\ of\ correctly\ recognized\ characters}{Number\ of\ total\ characters}$$