Data preprocess

What we tried?

1. Return raw data back into images.

Use raw data released form website, transfer them into jpg pictures.

Why we did that?

1. It’s easy to find something and check the result from image process with our eyes.
2. It’s faster to do some image process than raw text data.

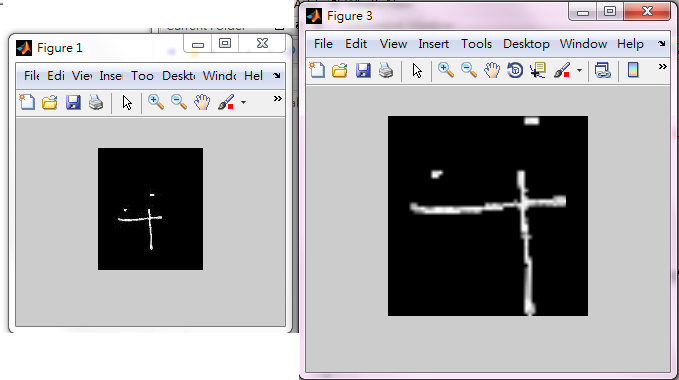
2. Simple Filter

We make an absolute value filter. We cleaned all pixels which are smaller than 100/255, because we thought them as noises.

3. Word normalization

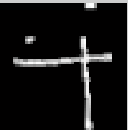
Raw data are very messy. For example, some words with are on the right-up corner of whole picture, with most of picture is black. Some words are on the other corner, some are at middle. Besides, some words are big, and some are small.

Therefore, we want to make all data become “normalized”. We make all words are in the same size, and all at middle. We also make picture form 105\*122 to 200\*200 for further image process.



4. Image dilation and Gaussian mask

We want to make words more blur, so we use image dilation. We tried different dilation process, like Gaussian mask, disk, and ball.



5. Shrink image

In Word normalization part, we make picture form 105\*122 to 200\*200. However, 200\*200 is too large for most classifiers (with 40000 pixels, and it’s too slow to train), so after image processing, we shrink all picture to 100\*100, 10000 pixels.

6. Data output

We output data in two ways. First, we use 0/1 output. We make a threshold, if each pixel is larger than the threshold, output value is 1, otherwise is 0. Second, we also make a threshold, but output value is original double value.