A Multimodal Network on Handwritten Chinese Character Error Correction: Supplementary Material

Haizhao Sun*
sunhaizhao@bupt.edu.cn
Yu Ning*
ningyuv@bupt.edu.cn
Xu Ji
jixv@bupt.edu.cn
Chuang Zhang
zhangchuang@bupt.edu.cn
Ming Wu
wuming@bupt.edu.cn

Beijing University of Posts and Telecommunications
China

1 More Experiment Details of Adopted Datasets

In this article, we use HWDB1.2[1] as the error dataset and ICDAR2013[2] as the correct dataset. Figure 1 shows some examples from these datasets.

The left side of Figure 1 shows some results of testing using two methods in the HWDB1.2 dataset. It can be seen from this that when the CCR-CLIP[3] method encounters unseen characters, it will look for similar characters in the learned character set based on the recognized features. Our method generates the corresponding IDS strictly according to the recognized character structure and radical information.

The right side of Figure 1 shows some results of the two methods tested on the ICDAR2013 dataset. We have selected the Chinese characters corresponding to the prediction results of the CCR-CLIP method in the picture on the left for display. It can be seen that in the 3755 categories of first-level Chinese characters, both methods show excellent performance.

2 Visualization of Attention

We selected three character images with different glyph structures in the typo dataset and performed a visual analysis of their attention maps shown in Figure 2. For these images, the decoder is able to generate the next part by focusing on different areas of the image based on the current predicted sequence.

*These authors contributed equally to this work.

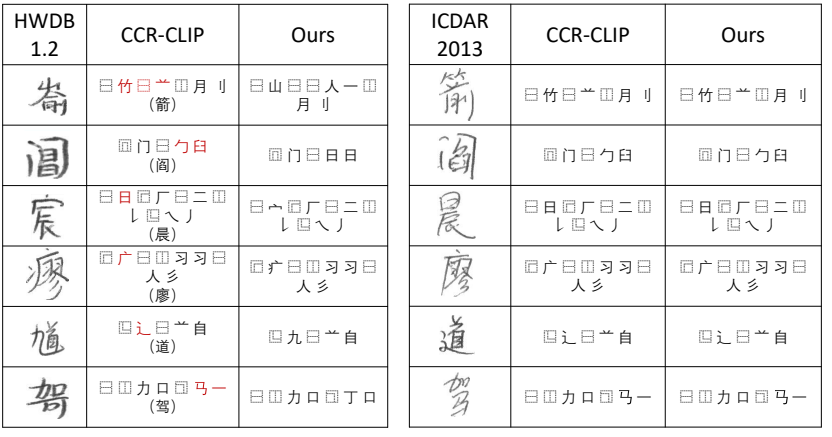

Figure 1: Comparison of some prediction results between the CCR-CLIP method and Ours on HWDB1.2 (error set) and ICDAR 2013 (correct set).

Figure 2: Some attention visualizations from the IDS decoder.

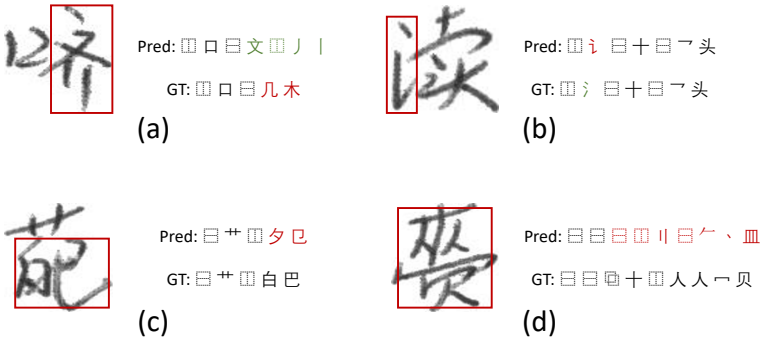

Figure 3: Results of four main types failure cases.

3 Failure Cases

Figure 3 shows the four main types of recognition failures. The parts enclosed by red boxes are where our prediction results are inconsistent with the ground truth (GT). The right side of each image shows the IDS of our predicted IDS and GT. The IDS elements marked in red are the IDS elements that are inconsistent with the characters. The green elements in Figure 3 (a) are the parts of our prediction results that match the character images but do not match the GT. Next, we analyze the reasons for these four main failure types.

As can be seen from Figure 3 (a), GT is different from the character image, but our prediction results are consistent with the characters on the image. This shows that since the HWDB1.2 dataset, we used as an error dataset contains some rare characters, when an incorrect Chinese character image appears, HWDB1.2 chooses to label it as a similar rare Chinese character. The result is that there are some incorrect labels in the dataset, causing us to predict the characters in the image correctly, but it is judged as a recognition error.

As can be seen from Figure 3 (b), our method incorrectly predicts " ⺗ " as " 讠 ". Since the shapes of " ⺗ " and " 讠 " in handwritten Chinese characters are very similar, it is difficult for even native Chinese speakers to tell whether the part in the red box of the character image in Figure 3 (b) is " ⺗ " or still " 讠 " without context. This error is caused by the prediction radical being too similar to the GT radical.

As can be seen from Figure 3 (c), our method predicts "葩" as "苑". "苑" is a character category that appears in the training set. It can be observed that the radicals in the right half of the red box are very similar to the IDS elements at the corresponding positions in GT and the IDS elements at the corresponding positions in the prediction. This shows that although we have used various methods to reduce the hallucinations of the model, the model still cannot eliminate the hallucinations when the target character image has too many similarities with the character categories that have appeared in the training set.

As can be seen from Figure 3 (d), after our method predicted the first two character structures correctly, because the structure of the target character was too confusing, our model also fell into chaos, forgetting the previously predicted parts, and began to re-according to A

series of predictions are made for the outline of the character image. So it is obvious that our predicted IDS cannot even successfully form a character.

References

[1] Cheng-Lin Liu, Fei Yin, Qiu-Feng Wang, and Da-Han Wang. Icdar 2011 chinese handwriting recognition competition. In *2011 International Conference on Document Analysis and Recognition*, pages 1464–1469, 2011. doi: 10.1109/ICDAR.2011.291.

[2] Fei Yin, Qiu-Feng Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Icdar 2013 chinese handwriting recognition competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1464–1470, 2013. doi: 10.1109/ICDAR.2013.218.

[3] Haiyang Yu, Xiaocong Wang, Bin Li, and Xiangyang Xue. Chinese text recognition with a pre-trained clip-like model through image-ids aligning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11943–11952, October 2023.