Reading Chinese in Natural Scenes with a Bag-of-Radicals Prior: Supplementary Materials

Yongbin Liu[1, 2]
liuyongbin@buaa.edu.cn
Qingjie Liu [*1, 2]
qingjie.liu@buaa.edu.cn
Jiaxin Chen[1]
jiaxinchen@buaa.edu.cn
Yunhong Wang[1, 2]
yhwang@buaa.edu.cn

[1] State Key Laboratory of Virtual Reality Technology and System
Beihang University
Beijing, China

[2] Hangzhou Innovation Institute of Beihang University
Hangzhou, China

1 Non-Latin Transcript Recognition Performance

In the main paper, we give the recognition performance on all transcripts. In this part, we separately give the prediction accuracy on non-Latin transcripts in Tab. 1.

Table 1: The table lists the performance of 10 selected models evaluated by word accuracy and 1-normalized edit distance on six datasets (only the non-Latin samples are taken into computation).

Model	Word accuracy on non-Latin transcripts							1-Normalized edit distance on non-Latin transcripts						
	ArT	CASIA	ctw	LSVT	RCTW	ReCTS	All	ArT	CASIA	ctw	LSVT	RCTW	ReCTS	All
GRCNN	42.9	42.9	31.1	37.7	44.6	51.9	41.8	72.2	70.0	61.3	64.2	68.9	70.5	67.2
R2AM	63.9	51.0	53.2	57.9	55.7	66.2	57.4	78.8	72.3	69.8	74.3	72.9	78.9	74.2
CRNN	52.4	46.7	38.5	46.8	47.9	60.1	48.7	78.9	72.2	65.3	70.2	70.8	76.9	71.7
Rosetta	65.1	60.6	54.8	62.6	63.5	74.5	63.7	86.0	81.8	76.8	81.3	81.6	86.5	82.0
RARE	74.7	61.9	62.8	67.2	66.5	75.1	67.2	87.5	80.9	78.2	82.0	81.4	85.4	82.0
STAR-NET	51.5	53.1	49.5	53.7	57.4	69.4	56.6	75.8	75.1	69.0	72.8	75.6	82.8	75.3
TRBA	77.7	62.6	63.8	68.5	67.2	77.1	68.5	89.2	81.0	78.7	82.5	81.9	86.7	82.7
RobScan	**85.2**	67.0	66.8	71.8	70.4	79.7	72.0	**93.6**	84.6	82.4	85.7	84.9	88.6	85.8
SAR	74.2	66.7	65.2	70.9	69.7	79.8	71.3	87.5	84.2	81.5	84.9	84.2	88.7	85.3
GCAN	74.7	68.1	65.2	71.6	70.7	80.5	71.8	88.1	85.0	81.5	85.3	84.7	89.1	85.5
Ours	79.2	**69.3**	**67.5**	**73.6**	**72.5**	**81.5**	**73.5**	89.9	**85.3**	**82.5**	**86.1**	**85.4**	**89.5**	**86.1**

2 Case Study

In order to intuitively analyze the improvement of our CVFM module and BCE loss, we give some successful cases predicted by our model and some failure cases to analyze the defect

* Corresponding author.

of our model. As shown in Fig. 1a, each line demonstrates a sample, followed by the predictions of GCAN (our baseline model), predictions of GCAN + CVFM module (G+CVFM), and predictions of GCAN + CVFM module + BCE loss (G+CVFM+BCE). The predictions and edit distances between predictions and groud truths are also given. It is quite clear that the CVFM module and BCE loss can improve the structural awareness of the characters. For example, 桥 and 栋 are similar characters, and the model with CVFM + BCE successfully predict it, but GCAN does not (this result is located in line #3).

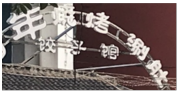
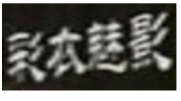
Besides successful cases, we also dive into the failure cases and give some reasons that the model fails to recognize. The first line shows that the model fails to recognize a Chinese character of an ancient shape, since the shape of the first character 偕 is in its ancient style, which is different from the modern usage. The third character in the 2nd text line seems very blur, which is the reason causing the failure. The 3rd and 4th line are two examples with low resolution. These samples are difficult to recognize even by human. The 5th example is a decorated Chinese character, and there are some ornamental symbols around them, making difficult to recognize. The last failure case is due to the low contrast ratio, since the foreground color and background color are all red.

3 Performance Analysis on the Latin Subset

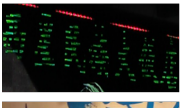
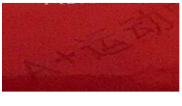
We now give the model performance on Latin datasets in this section.

The model performance on Latin transcripts achieves 75%+, and it seems that the results do not outperform the counterpart on non-Latin transcripts (71%+) by a large margin. We argue it is because the train datasets are not language-wise balanced (only 17% samples are Latin, as previously described). To validate this, we train GCAN and our model using two experimental settings. The first one is to train models from scratch (GCAN (Latin) and Ours (Latin)) using the pure Latin part of our datasets, and the second is to finetune models given in Tab. 2 (in the main paper) on Latin transcripts (GCAN (FT) and Ours (FT)). The experimental results are listed in Tab. 2. The *ctw* column is removed because there are only two Latin samples in the test dataset. Compared with the non-Latin results, the performance on Latin transcripts are much higher. Take GCAN (FT) as an example, the performance on CASIA-Latin is 76.4, 8.3 higher than CASIA non-Latin; and on LSVT-Latin is 78.4, 6.8 higher than LSVT non-Latin. It is worth noting that in these two experiments, the radical embedding is set to zero and no BCE loss is used because there are no radicals for Latin characters.

Another interesting observation is that the performance on Latin transcripts also rises after we add CVFM module. As previously described, the radical embedding of Latin characters is set to zero vector, so only the original learnable embedding is used during training. We argue that the improvement mainly credits to the scaler module operated on the extracted learnable embedding. According to the results in 2, adding CVFM brings an performance improvement of approximately 1%, (GCAN (Latin) *vs* ours (Latin), and GCAN (FT) *vs* ours (FT)). The improvement on the previous Latin benchmark (refer to Tab. 3) also echoes the argument.

Image Samples	Ground Truth	GCAN Prediction	G+CVFM Prediction	G+CVFM+BCE Prediction
	年烧烤涮肚	年级烤料盘 (Edit Dist=3)	年烧烤锅鱼 (Edit Dist=2)	年烧烤涮肚 (Edit Dist=0)
	广州普兰内特概念酒店	广州普兰肉精机念酒店 (Edit Dist=3)	广州兰兰内特概念酒店 (Edit Dist=1)	广州普兰内特概念酒店 (Edit Dist=0)
	长江委二桥小区	长江委工桥中区 (Edit Dist=2)	长江委二栋小区 (Edit Dist=1)	长江委二桥小区 (Edit Dist=0)
	祛黑点	桂寨点 (Edit Dist=2)	祛黑盒 (Edit Dist=1)	祛黑点 (Edit Dist=0)
	师同学们	你同学行 (Edit Dist=2)	师同学门 (Edit Dist=1)	师同学们 (Edit Dist=0)
	彩衣魅影	彩衣肤歇 (Edit Dist=2)	我衣魅影 (Edit Dist=1)	彩衣魅影 (Edit Dist=0)

(a) Samples predicted successfully predicted by our model.

Image Samples	Ground Truth	GCAN Prediction	G+CVFM Prediction	G+CVFM+BCE Prediction
	偕只有冷流片破	缙只有给流代玻 (Edit Dist=4)	缙只有给流片被 (Edit Dist=3)	缙只有冷流片破 (Edit Dist=1)
	中国厨电	加多 (Edit Dist=4)	加盟电 (Edit Dist=3)	中国网 (Edit Dist=2)
	原价:40元	京价TAON (Edit Dist=5)	京价100元 (Edit Dist=3)	京价140元 (Edit Dist=2)
	西星民族大饭店	昌星民美大厦 (Edit Dist=4)	嘉星民族大促店 (Edit Dist=2)	西星民泰大饭店 (Edit Dist=1)
	中国智造	焖风福造 (Edit Dist=3)	中国驾吉 (Edit Dist=2)	中国福造 (Edit Dist=1)
	A+ 运动	中国福奇 (Edit Dist=5)	全场运动 (Edit Dist=3)	A+运运动 (Edit Dist=1)

(b) Failure cases. As shown in the figure, the fail reasons include low resolutions, blur, different type and decorated characters.

Figure 1: We give successful cases and failure cases in Fig.1a and Fig.1b, respectively. The wrongly predicted characters are marked as red color.

Table 2: Recognition accuracy on pure Latin transcripts.

Model	Word Accuracy on Latin Transcripts					
	ArT	CASIA	LSVT	RCTW	ReCTS	All
GRCNN	41.8	34.3	36.5	30.9	60.8	41.7
R2AM	58.5	60.0	62.9	54.8	72.6	62.1
CRNN	41.1	40.5	40.2	38.5	63.8	45.0
Rosetta	64.4	54.9	60.1	53.7	75.2	62.2
RARE	71.0	71.4	72.2	67.2	83.3	73.4
STAR-NET	62.7	52.2	59.3	51.8	76.4	61.0
TRBA	74.8	74.6	73.8	71.1	84.3	76.1
RobScan	75.3	73.3	74.0	69.8	83.8	75.7
SAR	75.0	74.5	73.5	70.7	84.2	76.1
GCAN	75.6	74.8	74.4	71.0	84.4	76.5
Ours	77.8	76.4	75.6	72.6	85.1	78.1
GCAN (Latin)	76.2	74.0	75.4	70.9	85.1	76.8
Ours (Latin)	77.8	74.9	75.8	71.3	84.9	77.5
GCAN (FT)	78.8	76.4	78.4	74.1	86.5	79.2
Ours (FT)	80.0	78.0	78.5	74.9	86.6	80.1

Table 3: Recognition accuracy on the previous Latin benchmark. The performance of GCAN is reported in [4].

	IIIT5K	SVT	IC13	IC15	SVTP	CT80	All
GCAN	94.4	90.1	93.3	77.1	81.2	85.6	87.8
Ours	95.3	90.1	93.0	79.1	82.6	89.6	88.9

4 Compatibility with Text Detectors

The scene text spotting is to localize and recognize text boxes from panoramic images, and the recognizer is a downstream processor of the detector. The upstream detector may generate bounding boxes of all sizes, and the border of these boxes may not cover ground truth boxes exactly. In order to evaluate our model in the realistic setting, we employ the box discretization network [2, 3] as our detector and use our recognition model as the downstream recognizer. We give an example of the text spotting results on ReCTS test datasets in Fig. 2, and the end-to-end text spotting performance is shown in Tab. 4. The performance of the text detector is evaluated by *recall, precision* and *hmean*, and **1-NED** is to evaluate the performance of the recognition performance. It is worth noting that we do not utilize the ensemble tricks during the detection stage, so the detection performance is slightly below the results given by [2, 3]. The results shown in Tab. 4 validate the compatibility of our model with the text detector, and demonstrate superiority to the baseline (GCAN).

Table 4: End-to-end performance on the ReCTS-19 test dataset.

Models	Recall	Precision	Hmean	1-NED
GCAN	87.9	95.7	91.6	78.6
Ours				79.0

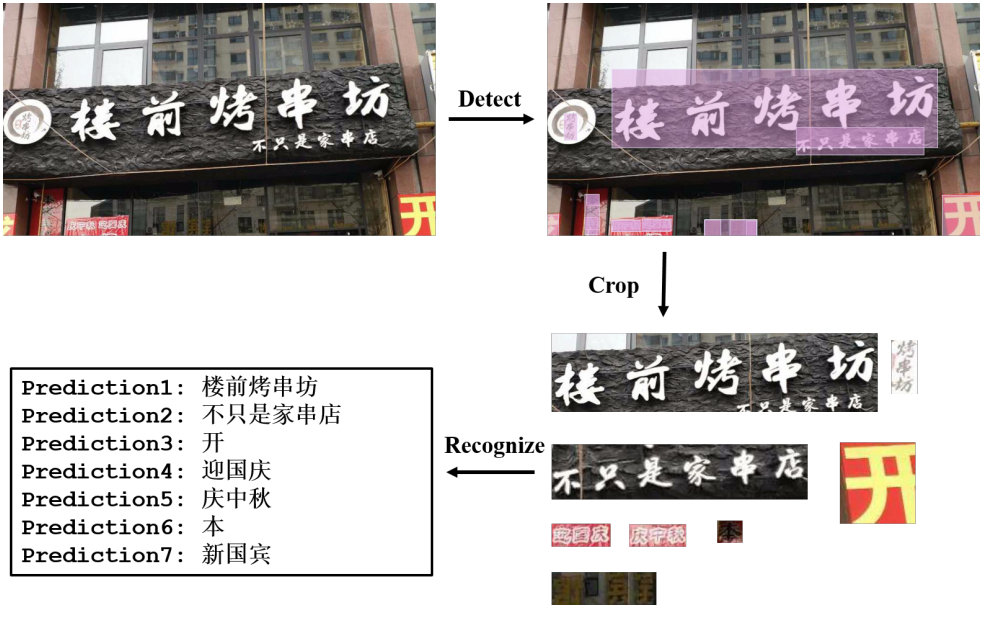


Figure 2: Our text spotting system.

5 A Demonstration of BCE Loss Masks

Since there are no radicals in Latin characters, we use a mask to label zero for the them and one for Chinese characters, and the mask is multiplied by the loss matrix, as shown in Fig.3.

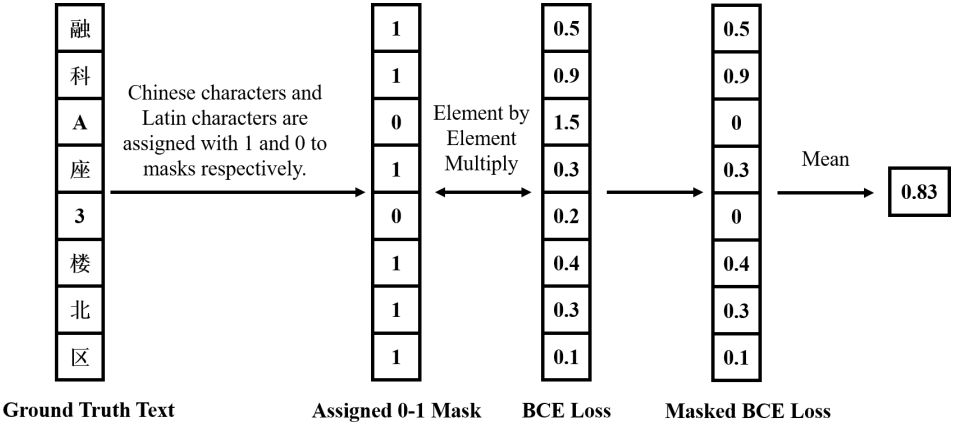


Figure 3: The figure shows how we assign 0-1 mask for the BCE loss of each word. The Latin characters are not taken into the computation of BCE loss.

6 Discussions about Related Chinese OCR Work

In this section, we give a discussion about the differences between our use of Chinese radicals and previous papers.

In most of the related work [1, 5, 6, 7, 8, 9], Chinese radicals are taken as a kind of intermediate code (IC), while we use it as a kind of embedding to fuse with visual features. Hence there are fundamental differences between them. We use Fig. 4 to compare these two different methodology.

For single-character recognition, using radicals as IC is feasible. However, for a sequence of characters, the code would be very long, making it impractical for recognition. As shown in Fig. 5, a transcript usually generates 4x $\sim$ 5x the length of radical-based IC. In STR datasets, transcripts with 10 $\sim$ 20 characters are common, which means the length of IC can reach 50 $\sim$ 100.

Hence, we conclude that long radical-based or stroke-based intermediate code may (1) make sequential modeling more difficult; (2) increase the time complexity. In contrast to radical-based IC, our method is not influenced by the length of transcripts since they are translated into radical-embedding with the same dimension and fused with visual features. As a result, our model benefits from low time complexity and shorter sequence to prediction.

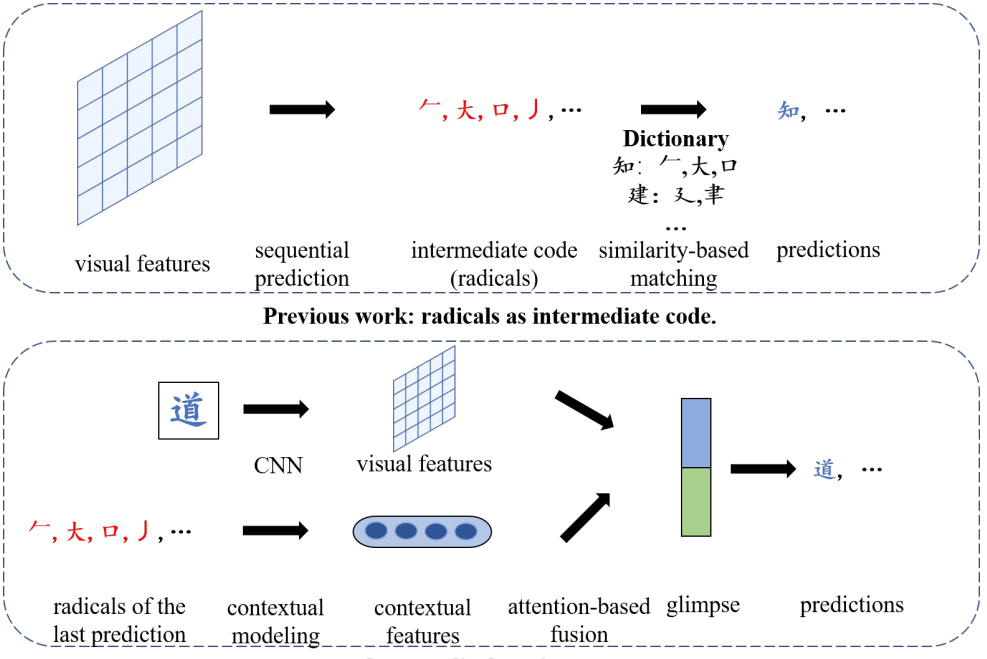


Figure 4: The top figure demonstrates the previous work that takes radicals as a kind of intermediate code for character matching, and the bottom shows how we use radicals as a kind of features for classification.

Transcript	Radical-based IC	IC/TSC
西星民族大饭店	一, 丩, 儿, 口, EOC, 民, EOC, 方, ⺈, 矢, 大, 一, 人, EOC, 一, 人, EOC, 饣, 反, 厂, 又, EOC, 广, 占, 口,EOC	3.71
融科资讯中心	鬲, 虫, 一, 口, EOC, 禾, 斗, EOC, 次, 贝, 冫, 欠, EOC, 讠, 孔, 十, EOC, 口, 丨, 心, EOC	3.50
物美超市	牛, 勿, 勹, 丿, EOC, 羊, 大, 土, EOC, 走, 召, 土, 辶, 刀,口, 冂, 丿, EOC, 亠, 巾, 丶, 一, EOC	5.75

Figure 5: Some examples of radical-based intermediate code (IC). We calculate the ratio of the length of IC to the length of transcripts (IC/TSC), and find it is approximately $4 \sim 5$. EOC means *end of a character*.

References

[1] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[2] Yuliang Liu, Tong He, Hao Chen, Xinyu Wang, Canjie Luo, Shuaitao Zhang, Chunhua Shen, and Lianwen Jin. Exploring the capacity of sequential-free box discretization network for omnidirectional scene text detection. *arXiv preprint arXiv:1912.09629*, 2019.

[3] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. *IJCAI*, 2019.

[4] Zhi Qiao, Xugong Qin, et al. Gaussian constrained attention network for scene text recognition. In *ICPR*, pages 3328–3335. IEEE, 2021.

[5] Tianwei Wang, Zecheng Xie, Zhe Li, Lianwen Jin, and Xiangle Chen. Radical aggregation network for few-shot offline handwritten chinese character recognition. *Pattern Recognition Letters*, 125:821–827, 2019.

[6] Tie-Qiang Wang, Fei Yin, and Cheng-Lin Liu. Radical-based chinese character recognition via multi-labeled learning of deep residual networks. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 579–584. IEEE, 2017.

[7] Wenchao Wang, Jianshu Zhang, Jun Du, Zi-Rui Wang, and Yixing Zhu. Denseran for offline handwritten chinese character recognition. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 104–109. IEEE, 2018.

[8] Changjie Wu, Zi-Rui Wang, Jun Du, Jianshu Zhang, and Jiaming Wang. Joint spatial and radical analysis network for distorted chinese character recognition. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 5, pages 122–127. IEEE, 2019.

[9] Jianshu Zhang, Yixing Zhu, et al. Radical analysis network for zero-shot learning in printed chinese character recognition. In *ICME*, pages 1–6. IEEE, 2018.