# Joint Spatial and Radical Analysis Network For Distorted Chinese Character Recognition

Changjie Wu, Zi-Rui Wang, Jun Du, Jianshu Zhang, Jiaming Wang
*National Engineering Laboratory for Speech and Language Information Processing*
*University of Science and Technology of China*
Hefei, Anhui, P. R. China
wucj@mail.ustc.edu.cn, cs211@mail.ustc.edu.cn, jundu@ustc.edu.cn, xyysszjs@mail.ustc.edu.cn, jmwang66@mail.ustc.edu.cn

***Abstract*—Recently, a novel radical analysis network (RAN) has been proposed for Chinese characters recognition (CCR). The key idea is treating a Chinese character as a composition of radicals rather than a single character class. Compared with traditional learning ways, two serious issues in CCR, i.e., enormous categories and limited training data, can be effectively alleviated. In this paper, we further excavate the potential capability of RAN. First, we validate RAN can reduce the equivariant requirement of regular convolutional neural network (CNN) owing to finer modeling and a local-to-global recognition process, especially considering the rotation transformation. This modeling approach of RAN can be regarded as one instance of compositional models. Second, we propose a joint spatial and radical analysis network (JSRAN) to handle more general situation in which the test data includes kinds of affine transformations. No matter for rotated printed Chinese characters or natural scene, JSRAN can outperform RAN and traditional CNN-based classifier. Finally, according to visualization analysis, we empirically explain why JSRAN can yield a remarkable improvement.**

***Index Terms*—Chinese characters, Radical analysis, Spatial transformer, Attention**

## I. Introduction

Chinese characters recognition (CCR) has a widespread application and has been studied for decades. All research efforts can be divided into two categories, namely character-based CCR (CCCR) [1], [2] and radical-based CCR (RCCR) [3]–[5].

Due to the large number of characters in Chinese dictionary, the character-based approaches usually focus on 3755 common Chinese characters [2] and can achieve excellent recognition accuracy with deep learning [6]. However, this kind of approach has obvious shortcomings: 1) Training data need exponential growth with the increasing of classes; 2) No capability to handle unseen situations. The shortcomings can be described as combinatorial explosion [7], which leads to a more strict requirement for our models, e.g., the equivariance property [8]. For regular convolutional neural network (CNN) [9], it is easy to observe that the convolutions are equivariant to translation, i.e., a translation followed by a convolution is the same as a convolution followed by a translation [8], which means CNN can still capture the corresponding features to correctly recognize the input image if the input image is changed by a translation operator. But it is a pity that CNN is not equivariant to most common operations, even for simple rotation operator, which makes CNN based CCCR fail.

As a more reasonable approach, the RCCR regards a Chinese character as a composition of radicals. In [10], the authors over-segmented characters into candidate radicals and searched an optimal path for recognition based on these candidate radicals. Recently, the authors [3] used a deep residual network with multi-labeled learning to detect position-dependent radicals. More recently, in [4], [5], the authors proposed radical analysis network (RAN) to identify radicals and analyze their corresponding two-dimensional spatial structures simultaneously, which is a successful application under the encode-decode framework [11].

The proposed RAN and other RCCR approaches are realizations of the compositional model [7], in which the algorithm focuses on the basic components and their structural relationships. Similar to human-learning, the potential idea of the compositional model is the infinite changes of samples can be addressed if the model can learn the intrinsic representation. From another view, a local mistake has larger probability to be corrected in the compositional model. For example, although a rotated character is never seen in the training data, it still can be recognized by RAN due to the possibility that the rotated radicals have been learned, which makes the model relax the equivariant requirement.

To further improve the generalization capability of RAN, a spatial transformer mechanism (ST) [12] is equipped in RAN, yielding the proposed joint spatial and radical analysis network (JSRAN). ST is an effective solution for the affine-transformed image, which is also helpful to reduce the equivariant requirement for models. Intuitively, JSRAN can generate double effects. We will observe this phenomenon in the experimental section and analyze it. It should be noted that the rotation is a special transformation in affine transformation and the proposed JSRAN is an end-to-end system that is free to any extra assumptions [12].

The structure of JSRAN is illustrated in Fig. 1. The spatial transformer mechanism and DenseNet [13] are used as the encoder to extract high-dimensional features. The extracted high-dimensional features are sent to the decoder. The decoder consists of an attention block and GRUs. The attention block performs weight redistribution of the high-dimensional feature vectors to generate a new representation for current prediction. The new generated feature vector is sent to the GRUs [14]. Finally, the decoder predicts the current result.

978-1-7281-5054-3/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDARW.2019.40092

To detect the radicals and internal two-dimensional structure simultaneously, a coverage based spatial attention model is built in the decoder [15], [16]. The main contributions of this study are summarized as follows:

- We further explore the potential capability of RAN from both new opinions and massive experiments.
- The proposed JSRAN can generate double effects for relaxing the equivariant requirement of CNN. Compared with CNN and RAN, JSRAN can yield a remarkable improvement.
- Visualization analysis is conducted to experimentally explain recognition results.
- The source codes of our own JSRAN are publicly available[1].

The rest of this paper is organized as follows. In Section II, we illustrate the structure of JSRAN in detail. In Section III, we present and analyze the experimental results. In Section IV, we conclude our study.

## II. NETWORK ARCHITECTURE OF JSRAN

The architecture of our proposed JSRAN includes two parts: CNN encoder with spatial transformer (Section II-A) and RNN decoder with spatial attention (Section II-B). The encoder takes the image as input and performs two operations on it. First, the image is rectified by the spatial transformation mechanism, and then the FCN is used to extract the high-dimensional features of the rectified image to generate a fixed-length context vector. Then the decoder takes the context vector as input to generate a sequence of spatial structures and radicals. The details are described in the following sections.

### A. Encoder with spatial transformer

*1) Spatial transformer:* As shown in Fig. 1, the spatial transformer mechanism consists of 3 parts: localisation network, grid generator and . A localisation network, which is a simple CNN, takes the Chinese character image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ and outputs a 6-dimensional vector $\boldsymbol{\theta}$. Then, the grid generator uses $\boldsymbol{\theta}$ to generate a sampling grid $\mathbf{G} = \{G_i\}$, where each element $G_i$ is the target coordinate $(x_i^t, y_i^t)$ of the regular grid in the output image. We can get the set of sampling points $A_{\boldsymbol{\theta}}(\mathbf{G})$, where $A_{\boldsymbol{\theta}}$ is a $2D$ affine transformation. Finally, the sampler rectifies the original Chinese character image:

$$\mathbf{O} = S\left(\mathbf{I}, A_{\boldsymbol{\theta}}\left(\mathbf{G}\right)\right) \tag{1}$$

where $\mathbf{O} \in \mathbb{R}^{H' \times W' \times C}$ and $S$ means a bilinear interpolation function for the input image $\mathbf{I}$ by using the sampling grid $A_{\boldsymbol{\theta}}(\mathbf{G})$. This sampling function is differentiable, which implies that ST can be trained with gradient descent methods. More details can be found in [12].

*2) FCN:* The convolutional neural network is capable of extracting high dimensional features from an image. Since a spatial attention mechanism is equipped in the decoder, we do not add any fully connected layers following the last convolution layer and we call this kind of CNN structure fully convolutional network (FCN). Through the feature extraction of FCN, we can obtain a three-dimensional array of size $H'' \times W'' \times D$. And then the array is reshaped into a two-dimensional array of size $L \times D$, where $L = H'' \times W''$. The row vectors in matrix can be denoted as:

$$\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}\ ,\ \mathbf{a}_i \in \mathbb{R}^D \tag{2}$$

where $\mathbf{a}_i$ is the $i^{\text{th}}$ row of the matrix. Each element of $\mathbf{A}$ corresponds to a local area of the input image.

### B. Decoder with spatial attention

The decoder generates a sequence of spatial structures and radicals of input Chinese character. The output sequence can be represented by $\mathbf{Y}$ as a sequence of one-hot vectors.

$$\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}\ ,\ \mathbf{y}_i \in \mathbb{R}^K \tag{3}$$

where $K$ is the number of total spatial structures and basic radicals. $C$ is the length of the output sequence.

It is should be noted that $C$ is with a variable length for different characters while $L$ is constant in the training and decoding stages. In the decoder, the attention mechanism has two important functions: First, it can solve the problem of converting a sequence from a fixed length to a non-fixed length sequence of vector $\mathbf{c}_t$. Second, the attention mechanism is able to focus on the most relevant information instead of all features at each step of prediction by increasing the weight of useful information and reducing the weight of useless information. Given the vector $\mathbf{c}_t$, the decoder adopts two unidirectional GRU layers to calculate the hidden state $\mathbf{h}_t$:

$$\hat{\mathbf{h}}_t = \mathrm{GRU}_1\left(\mathbf{y}_{t-1}, \mathbf{h}_{t-1}\right) \tag{4}$$

$$\mathbf{c}_t = f_{\text{att}}\left(\hat{\mathbf{h}}_t, \mathbf{A}\right) \tag{5}$$

$$\mathbf{h}_t = \mathrm{GRU}_2\left(\mathbf{c}_t, \hat{\mathbf{h}}_t\right) \tag{6}$$

where $\mathbf{h}_{t-1}$ denotes the previous $\mathrm{GRU}_2$ hidden state, $\hat{\mathbf{h}}_t$ denotes the current $\mathrm{GRU}_1$ hidden state, and $f_{\text{att}}$ denotes the attention mechanism.

Considering the spatial relationship between radicals, we use a spatial attention mechanism. At every step $t$ of prediction, the vector $\mathbf{c}_t$ is calculated by the following equations:

$$\mathbf{c}_t = \sum\nolimits_{i=1}^{L} \alpha_{ti} \mathbf{a}_i \tag{7}$$

where $\alpha_{ti}$ denotes the spatial attention coefficient of $\mathbf{a}_i$ at time step $t$. And $\alpha_{ti}$ can be calculated:

$$\mathbf{F} = \mathbf{Q} * \sum\nolimits_{l=1}^{t-1} \alpha_l \tag{8}$$

$$e_{ti} = \boldsymbol{\nu}_{\text{att}}^{\mathrm{T}} \tanh(\mathbf{W}_{\text{att}} \hat{\mathbf{h}}_t + \mathbf{U}_{\text{att}} \mathbf{a}_i + \mathbf{U}_f \mathbf{f}_i) \tag{9}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{10}$$

where $\mathbf{F}$ is a coverage vector based on the summation of past attention probabilities $\alpha_l$. Let $n'$ denote the dimension of $\boldsymbol{\nu}_{\text{att}}$ and $m'$ denote the number of feature maps of filter $\mathbf{Q}$, then $\mathbf{W}_{\text{att}} \in \mathbb{R}^{n' \times n}$, $\mathbf{U}_{\text{att}} \in \mathbb{R}^{n' \times D}$ and $\mathbf{U}_f \in \mathbb{R}^{n' \times m'}$.

[1] https://github.com/ChangjieWu/JSRAN

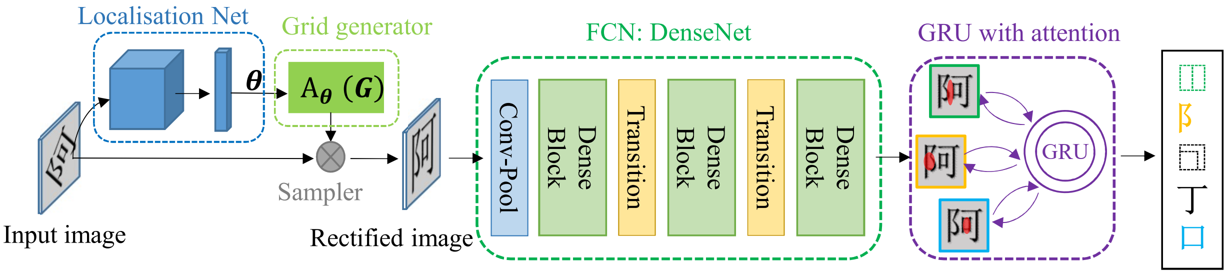


Fig. 1. Architecture of JSRAN.

Finally, the probability of target word $\mathbf{y}_t$ is computed by the following equation:

$$\mathbf{P}(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{I}) = f_s\left(\mathbf{W}_o f_{\max}\left(\mathbf{W}_c\mathbf{c}_t + \mathbf{W}_s\mathbf{h}_t + \mathbf{E}\mathbf{y}_{t-1}\right)\right) \tag{11}$$

where $f_s$ denotes a softmax activation function, $f_{\max}$ denotes a maxout activation function, $\mathbf{W}_o \in \mathbb{R}^{K\times\frac{m}{2}}$, $\mathbf{W}_c \in \mathbb{R}^{m\times D}$, $\mathbf{W}_s \in \mathbb{R}^{m\times n}$, and $\mathbf{E}$ denotes the embedding matrix, $m$ and $n$ are the dimensions of embedding and GRU decoder, respectively.

## III. Experiments

In order to more clearly explore the principles, advantages and disadvantages of JSRAN, we first designed several sets of experiments on rotated Chinese characters. Then we designed a set of experiments on *Chinese Text in the Wild* (CTW) [17] to prove the practicality of RAN. All experiments are presented by answering the following questions:

- What are the advantages of RAN in recognizing rotated Chinese characters (Section. III-B)?
- Is JSRAN effective in recognizing rotated Chinese characters when the number of rotated samples in training set is small (Section. III-C2)?
- What is the rotation range when JSRAN is still able to effectively recognize rotated Chinese characters (Section. III-C3)?
- Why are spatial transformer in the encoder and radical analysis in the decoder strongly complementary in recognizing rotated Chinese characters (Section. III-D)?
- Is JSRAN effective in recognizing Chinese characters in the wild (Section. III-E)?

### A. Training and testing details

In the experiment, the localisation network consists of three convolutional layers and two fully connected layers. The first convolutional layer has 8 convolution kernels of size $9\times9$. The second has 16 convolution kernels of size $5\times5$ and the third has 32 convolution kernels of size $3\times3$. There is a $2\times2$ max pooling layer between the first and the second convolutional layers. The number of units in the two fully connected layers are 64 and 6 respectively. The full convolutional network employs DenseNet [13] mainly composed of DenseBlocks and Transitions. We set the depth of each DenseBlock to 22, which is connected directly to all subsequent layers. In DenseBlock, the number of channels for each layer is 24. A transition layer is inserted between every two DenseBlocks. Each transition layer is a 1x1 convolution layer with halved channels to reduce the computational cost and storage overhead. The first layer of DenseNet is a convolutional layer with 48 convolution kernels of size $7\times7$. We employ batch normalization [18] after each convolution layer and the activation function is ReLU [19].

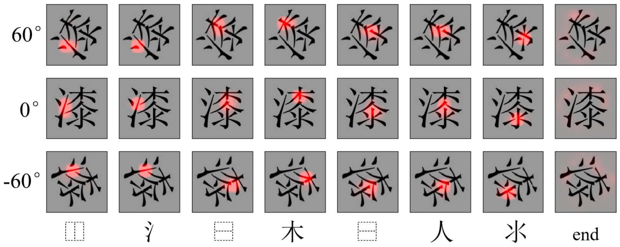


Fig. 2. Attention visualization of recognizing one Chinese character with different rotation angles.

The decoder employs two unidirectional GRU units. The dimension of GRU units $n$ and the dimension of embedding $m$ are set to 256. The convolution kernel of Q is set to $3\times3$ and the number of feature maps $M$ is set to 512. The loss function is cross entropy loss and the optimization algorithm is adadelta [20] with gradient clipping.

### B. Visualization of recognizing rotated Chinese characters

In order to better analyze the advantages of RAN in recognizing rotated Chinese characters, we do not add the ST mechanism in this experiment. The dataset is generated by 3755 Chinese character categories in Song font style. Each Chinese character category has three angles: 0°, 60° and $-60°$. Among them, 3755 samples of 0° Chinese characters, the first 3000 samples of 60° Chinese characters and the first 3000 samples of $-60°$ Chinese characters are used as the training set. The remaining 755 samples of 60° Chinese characters and 755 samples of $-60°$ Chinese characters are adopted as the test set. In Fig. 2, we show the recognition results and process of recognizing one character with the pronunciation "qi" by visualization of attention. As shown in Fig. 2, in each step of predicting the spatial structures or radicals, the attention

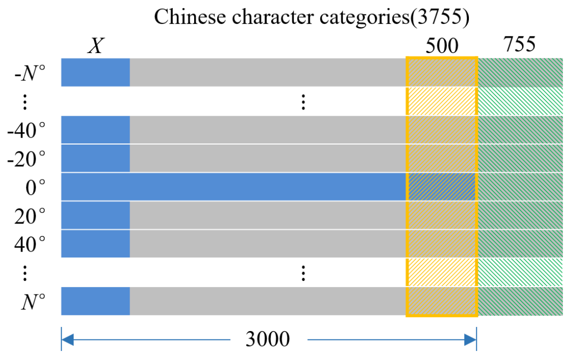


Fig. 3. Description of dividing training set and testing set.

mechanism focuses on the local regions with strong correlation in the feature map. The red color in the attention maps represents the spatial attention probabilities, where the darker color describes the higher attention probabilities. Interestingly, we observe that when Chinese characters are at different angles, the attention area will also change accordingly. Although a rotated sample of character has not been learned by RAN in the training stage, the rotated radicals of this sample can be learned from other Chinese characters, which indicates that the equivariance property of the radicals can still be satisfied and the requirement of equivariance of the whole Chinese character is relaxed. Therefore, the unseen characters with 60° and $-60°$ rotations in Fig. 2 can be correctly recognized by RAN.

### C. Experiments on rotated Chinese characters

To prove that JSRAN is more effective in recognizing rotated Chinese characters, we design other three networks for comparison, namely RAN without ST, whole character classifier based on DenseNet and STN-DenseNet (DenseNet equipped with ST). To perform a fair comparison, the same DenseNet structure and configuration is used in the four networks.

*1) Datasets:* We select 20 fonts from [1] to generate our datasets. As shown in Fig. 3, $N$ represents the maximum rotation angle. The blue area represents the training set, in which, the number of 0° Chinese character category is 3000 and the number of rotated Chinese character category with different angles is $X$. We have two test sets marked with yellow and green colors. The yellow shaded area represents the testing set that contains 500 seen Chinese character categories. The green area section contains 755 unseen Chinese character categories used to validate the ability to recognize unseen Chinese character categories. These 3,755 Chinese characters were split into 406 radicals and 10 spatial structures. During the training process, all data with different angles are randomly rotated between $-10°$ and $+10°$.

*2) Experiments on limited rotated Chinese characters:* As shown in TABLE I, when the number of rotated Chinese character categories in the training set is 200, the recognition accuracies on the test set with yellow color in Fig. 3, are 55.52%, 80.31%, 59.12%, and 91.56% respectively. The experimental results of STN-DenseNet and JSRAN are significantly better than those of DenseNet and RAN, implying that the ST mechanism can yield significant improvements for a small amount of rotation training set. When the number of rotated Chinese characters increases from 200 to 500, the recognition accuracies are increased to 65.03%, 89.39%, 88.19% and 97.27% respectively. We can observe that the performance of RAN has been greatly improved, which means that RAN needs about 500 Chinese character categories to cover most of the radicals. When the number increases to 1000, the recognition accuracies are 84.49%, 95.52%, 96.50% and 99.01% respectively. For all datasets, JSRAN can consistently outperform other networks, which indicates that JSRAN is able to combine the advantages of ST and RAN. Moreover, RAN and JSRAN can recognize unseen Chinese character categories by recognizing the radicals and spatial relationships. For example, the accuracies of unseen characters with green color in Fig. 3 are 0%, 0%, 47.4% and 63.45% for DenseNet, STN-DenseNet, RAN, and JSRAN, respectively. We can increase the accuracy of unseen Chinese characters by increasing the number of training categories [4].

TABLE I
THE RECOGNITION ACCURACY (IN %) OF DIFFERENT APPROACHES ON LIMITED ROTATED CHINESE CHARACTERS.

| $N$ | 60 | 60 | 60 |
|---|---|---|---|
| $X$ | 200 | 500 | 1000 |
| DenseNet [13] | 55.52 | 65.03 | 84.49 |
| STN-DenseNet | 80.31 | 89.39 | 95.52 |
| RAN [4] | 59.12 | 88.19 | 96.50 |
| JSRAN | 91.56 | 97.27 | 99.01 |

TABLE II
THE RECOGNITION ACCURACY (IN %) OF DIFFERENT APPROACHES IN DIFFERENT ROTATION RANGES.

| $N$ | 20 | 40 | 60 | 80 | 120 |
|---|---|---|---|---|---|
| $X$ | 500 | 500 | 500 | 500 | 500 |
| DenseNet [13] | 99.80 | 93.52 | 65.03 | 38.11 | 23.60 |
| STN-DenseNet | 99.85 | 97.61 | 89.39 | 65.32 | 29.78 |
| RAN [4] | 99.95 | 95.50 | 88.19 | 73.67 | 49.44 |
| JSRAN | 99.98 | 98.83 | 97.27 | 93.13 | 60.56 |

*3) Experiments in different rotation ranges:* The rotation range is determined by $N$ as shown in Fig. 3. In TABLE II, when $N = 20$, the four networks can achieve high accuracies. Obviously, the recognition task becomes more challenging with the increasing of $N$, yielding the decline of the recognition accuracies for the four networks. When $N = 80$, except for JSRAN, which still maintains a high recognition accuracy rate of 93.13%, the recognition accuracies of the other three networks have dropped significantly. When $N = 120$, the performance of JSRAN also becomes poor. In this case, it is difficult for ST to learn the correct parameter $\boldsymbol{\theta}$ used to rectify the rotated image to be frontal due to the large rotation range. These experiments prove that JSRAN is robust and effective in recognizing Chinese characters within a reasonable rotation range.

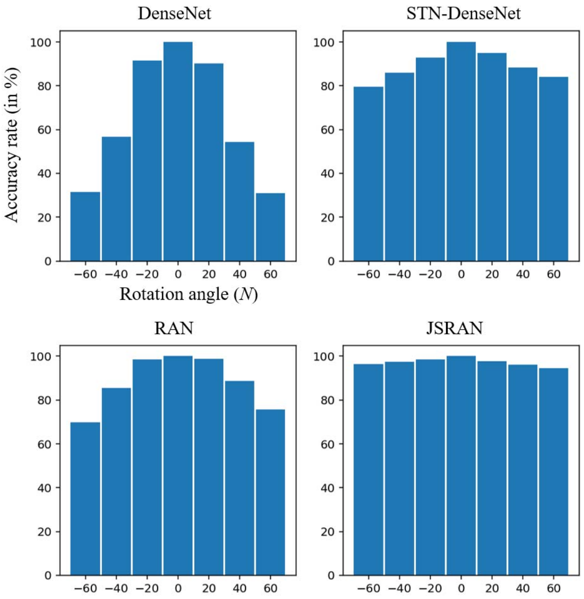


Fig. 4. Comparisons on different angle intervals.

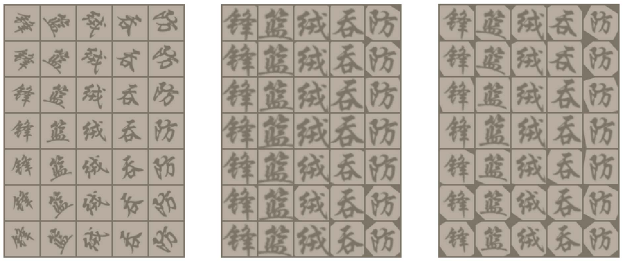
### D. Complementarity between ST and RAN

We use the partial results in Section III-C2 as an example to further analyze the complementarity between ST and RAN.

*1) RAN is benefit from ST:* Fig. 4 shows the recognition accuracy of Chinese characters in different rotation intervals. The recognition accuracy of DenseNet severely degrades when the rotation angle increases, which proves that the conventional CNN-based classifier needs a higher requirement for equivariant and can not learn the rotated information effectively. The trends of sharp decrease of accuracies with large rotations in STN-DenseNet and RAN have been alleviated, but their performances are still worse than that of JSRAN. JSRAN maintains high performance when the rotation angle increases. ST can rectify the rotated Chinese characters to a smaller rotation range. This demonstrates that the combination of ST and RAN can improve the performance of recognizing rotated Chinese characters.

*2) ST is benefit from RAN:* The main role of ST is to perform affine transformation for input images. The affine transformation includes rotation, scaling, and translation. These transformations help extract useful information from input images and reduce the interference caused by edges and positions. Fig. 5 shows images rectified by ST in STN-DenseNet and JSRAN. We can find that Chinese characters rectified by JSRAN is bigger and more frontal than those by STN-DenseNet, which indicates that the rectified images in JSRAN retain more accurate and useful information. Experiments show that the combination of RAN and ST can improve the rectification effect for rotated Chinese characters.

### E. Experiments on CTW

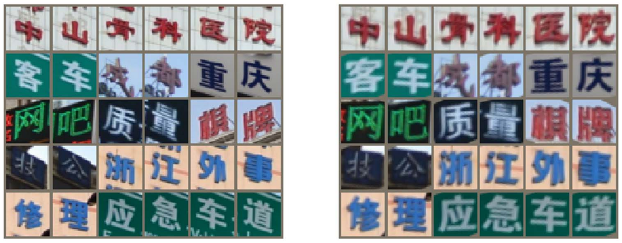


Rotated images

Rectified images (JSRAN)

Rectified images (STN-DenseNet)

Fig. 5. Rectification of rotated Chinese characters by ST.



Original images

Rectified images

Fig. 6. Distorted images in CTW and the rectified images by ST in JSRAN.

TABLE III
RESULTS ON CTW

| Network | Accuracy on distort | Accuracy on all |
|---|---|---|
| AlexNet [9] | 71.64% | 76.43% |
| OverFeat [21] | 72.11% | 76.84% |
| ResNet50 [23] | 76.05% | 79.46% |
| ResNet152 [17] | 77.52% | 80.94% |
| Inception-v4 [22] | 79.03% | 82.28% |
| DenseNet [13] | 76.79% | 79.88% |
| STN-DenseNet | 77.83% | 80.81% |
| RAN [4] | 81.56% | 85.56% |
| **JSRAN** | **83.97%** | **87.57%** |

In order to prove the superiority and practicability of JSRAN, we choose a more challenging dataset named as *Chinese Text in the Wild* (CTW) [17]. CTW is a large dataset of Chinese text in natural scene containing planar text, raised text, text in cities, text in rural areas, text under poor illumination, distant text, partially occluded text, etc. In CTW, the proportion of distorted text is about 26%. As shown in Fig. 6, we selected some representative images to show the correction effect of ST in JSRAN.

In addition to DenseNet, STN-DenseNet, RAN and JSRAN, we train several other typical convolutional neural networks on CTW, including: AlexNet [9], OverFeat [21], Inception-v4 [22], ResNet50 [23] and ResNet152. Same as [17], we only consider recognition of the top 1000 frequent Chinese character categories. TABLE III presents detailed accuracies of these networks. Furthermore, we also compute the recognition accuracy of distorted Chinese characters in the test set. We can observe that the recognition accuracy of JSRAN is much better than other networks for both distorted Chinese characters and all Chinese characters.

## IV. CONCLUSION AND FUTURE WORK

In this paper, we propose a joint spatial and radical analysis network to recognize distorted Chinese characters. Through massive experiments on rotated Chinese characters, we prove that ST and RAN have complementarity and the coupling of them can greatly improve the recognition accuracy. Furthermore, we demonstrate JSRAN outperforms RAN and traditional CNNs on *Chinese Text in the Wild*, which shows its great practicability. In the future, we will continue to explore how to solve more complex distorted Chinese characters, such as excessive distortion.

## ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Huawei Noah's Ark Lab.

## REFERENCES

[1] Z. Zhong, L. Jin, and Z. Feng, "Multi-font printed chinese character recognition using multi-pooling convolutional neural network," in *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE, 2015, pp. 96–100.

[2] X.-Y. Zhang, Y. Bengio, and C.-L. Liu, "Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark," *Pattern Recognition*, vol. 61, pp. 348–360, 2017.

[3] L. CL, C.-L. Liu *et al.*, "Radical-based chinese character recognition via multi-labeled learning of deep residual networks," 2017.

[4] J. Zhang, Y. Zhu, J. Du, and L. Dai, "Radical analysis network for zero-shot learning in printed chinese character recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[5] W. Wang, J. Zhang, J. Du, Z.-R. Wang, and Y. Zhu, "Denseran for offline handwritten chinese character recognition," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 104–109.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[7] A. L. Yuille and C. Liu, "Deep nets: What have they ever done for vision?" *arXiv preprint arXiv:1805.04025*, 2018.

[8] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*, 2016, pp. 2990–2999.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] L.-L. Ma and C.-L. Liu, "A new radical-based approach to online handwritten chinese character recognition," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[15] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," *Pattern Recognition*, vol. 71, pp. 196–206, 2017.

[16] J. Zhang, J. Du, and L. Dai, "Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 221–233, 2019.

[17] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, and S.-M. Hu, "Chinese text in the wild," *arXiv preprint arXiv:1803.00085*, 2018.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[20] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.