CNN-Based Text Image Super-Resolution Tailored for OCR

Haochen Zhang, Dong Liu, Zhiwei Xiong
CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China
zhc12345@mail.ustc.edu.cn, {dongeliu,zwxiong}@ustc.edu.cn

***Abstract*—Since low-resolution images may hamper the performance of optical character recognition (OCR), text image super-resolution (SR) has become an increasingly important problem in computer vision. Convolutional neural network (CNN) has been proposed for generic image SR as well as text image SR, but the previous works concern more on the objective quality (e.g. PSNR) rather than the OCR performance. In this paper, we propose a new loss function when training CNN for text image SR to facilitate OCR, and conduct model combination to further improve the performance. Also, we propose a simple yet effective image padding method to refine the image boundaries during SR. Experimental results show that we achieve an OCR accuracy of 78.10% on the ICDAR 2015 TextSR dataset, which is comparable with that of using the original high-resolution images (78.80%), and also exceeds the state-of-the-arts.**

***Index Terms*—Convolutional neural network, Loss function, Model combination, Optical character recognition, Super-resolution.**

I. Introduction

Optical character recognition (OCR) refers to the process by which an electronic device checks the characters printed on the paper, determines its shape by detecting the color patterns, and then translates the shapes into machine-encoded text. The ultimate goal is to convert the text image into text. However, the diversity of the input content makes accurate identification of characters a challenge for the OCR system. In particular, the systems being tuned for high quality high-resolution (HR) text images may produce many errors when recognizing low-resolution (LR) text images. It is because the LR text image lacks high-frequency image details, making it difficult for the OCR system to retrieve text information correctly. Thus, performing a super-resolution (SR) preprocessing on input text images is a direct way to achieve higher OCR accuracy, which has been verified in the previous researches [1].

Image SR, or termed resolution enhancement, is a classical problem in computer vision. In recent years, convolutional neural network (CNN) has been proposed for image SR and is known to be the best performer in the task of single natural image SR. Also, CNN is proposed for text image SR and achieves remarkable performance [1]. However, it is worth noting that previous works concern more on the objective quality (e.g. PSNR) of the super-resolved images, but higher objective quality does not necessarily lead to better OCR performance. For example, Table I shows the results of different methods on the ICDAR 2015 TextSR dataset [2], where the evaluation metrics include root-mean-squared-error (RMSE), PSNR, mean structural similarity (MSSIM) and OCR accuracy. It can be observed from the table that the method achieving the highest PSNR does not achieve the highest OCR accuracy. Therefore, how to optimize text image SR method to pursue better OCR performance instead of better objective quality, is an important issue to investigate.

TABLE I
Results of different methods on the ICDAR 2015 TextSR dataset

Method	RMSE	PSNR	MSSIM	OCR (%)
Bicubic	19.04	23.50	0.879	60.64
Lanczos3	16.97	24.65	0.902	64.36
Orange Labs [3]	11.27	28.25	0.953	74.12
Zeyde *et al.* [4]	13.05	27.21	0.941	69.72
A+ [5]	10.03	29.50	0.966	73.10
Synchromedia Lab [6]	62.67	12.66	0.623	65.93
ASRS [7]	12.86	26.98	0.950	71.25
SRCNN-1 [8]	7.52	31.75	0.980	**77.19**
SRCNN-2 [8]	**7.24**	**33.19**	**0.981**	76.10

This paper studies CNN-based text image SR methods tailored for OCR, and our contributions can be summarized as follows. First, we propose a new loss function when training CNN for text image SR. The new loss function is inspired by the intuition that high-frequency image details play a more important role in OCR, and thus puts more weights on edge regions. Second, we propose a simple yet effective image padding method to refine the image boundaries during CNN-based SR. Third, we conduct model combination to further improve the performance.

Experimental results show that our proposed method achieves an OCR accuracy of as high as 78.10% on the ICDAR 2015 TextSR dataset, which exceeds the state-of-the-art results as given in Table I. Our achieved accuracy is quite close to that of using the original HR images to perform OCR (78.80%), which demonstrates the effectiveness of our method.

II. Related Work

Image super-resolution. Image super-resolution, which

This work was supported in part by the Natural Science Foundation of China under Grant 61390512 and Grant 61331017, and in part by the Fundamental Research Funds for the Central Universities under Grant WK3490000001. *(Corresponding author: Dong Liu.)*

978-1-5386-0462-5/17/$31.00 ©2017 IEEE.

uses a single or a group of LR images to produce HR image, is a classical problem in computer vision. According to the utilized image priors, single image SR algorithms can be categorized into four types: prediction models, edge-based methods, image statistical methods, and example-based methods. These methods have been thoroughly investigated and evaluated in [9]. Recently, CNN has been proposed for SR and achieves significant improvement [8]. Following the CNN approach, Kim *et al.* proposed the very deep super-resolution (VDSR) network [10], which adopts residue learning and adaptive gradient clipping to train a very deep CNN (20 layers), and reported the state-of-the-art performance for single natural image SR. In [1], Dong *et al.* adopted CNN to perform text image SR and also reported better performance.

Loss function. CNN is trained to minimize a given loss function, which determines the optimization objective of CNN training. In other words, the loss function determines how the CNN is supposed to learn from training data. Previous works of CNN-based image SR often concern the objective quality of super-resolved images, and then adopt mean-squared-error (MSE) as the loss function. Recently, it is proposed to change the loss function into a generative adversarial loss, and thus the resulting images have better visual quality, but worse PSNR [11]. As for text image SR, we believe the OCR performance is the ultimate goal and then we propose a new loss function in this paper.

Model combination. Model combination has been widely used in machine learning problems. Dong *et al.* have investigated CNN-based model combination for text image SR in [1], which shows the PSNR can be improved from 31.99 dB (single best CNN) to 32.80 dB (combination of 14 CNNs). In that paper the CNNs for combination differ in their network structures, but in this paper we investigate a simple manner where the models for combination are identical in network structure but are trained by different random initializations.

Image padding. Convolution is the basic processing in CNN, which is sensitive to the boundary conditions. Padding is a common practice in CNN, for example padding with a predefined value (e.g. zero), periodical padding (assuming the signal is periodical) and duplication padding (duplicating the boundary values). Advanced padding methods have also been studied before [12]. In the existing CNN-based SR methods, zero padding is widely used [10]. This paper presents a simple yet effective image padding method to refine image boundaries.

III. The Proposed Method

In this paper, we adopt the VDSR network structure as the basic network to perform CNN-based image SR. The VDSR network has 20 layers and adopt residue learning and adaptive gradient clipping during training, which are all inherited in our experiments. For more details of VDSR please refer to [10].

A. New Loss Function

Most of the previous works adopted MSE as the loss function to train CNN. MSE minimization is equivalent to PSNR optimization, but as mentioned above, PSNR is not a good indicator of the OCR accuracy. In this paper, we propose a new loss function to replace MSE. Our key idea is that the OCR accuracy depends highly on the high-frequency image details in the text regions of text images, text regions usually feature high contrast edges, and edges also represent high-frequency component of an image. Thus, we need to concern more on the edge regions of an input text image. Using MSE is to assume each pixel is equally important, but we can use a weighted MSE (WMSE) to emphasize some pixels more than others. In short, our designed loss function is a WMSE,

$$\mathrm{WMSE} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|I(i,j)-\hat{I}(i,j)\right\|^2 \times f[\mathrm{grad}(i,j)]}{mn} \tag{1}$$

where I and $\hat{I}$ are the original and super-resolved images, respectively, grad is the gradient magnitude map of the original image, which is obtained by using Sobel operator. m, n are the height and width of images and i, j are pixel indexes. $f[\cdot]$ is a certain function to convert gradient magnitude into weight. Intuitively f should be a monotonously increasing function. Our experimental results also verify that using monotonously increasing functions is better than other functions. In this paper we report results of using three forms of weighting function, i.e. $f[x] = x^2$, $f[x] = x$, and $f[x] = \sqrt{x}$. A special note is, in order to avoid zero weight (i.e. one pixel is totally ignored in the loss function), we change the zero gradient magnitude into a small value ϵ before applying the weighting functions.

During implementation, we set $g[x] = \sqrt{f[x]}$, and then

$$\mathrm{WMSE} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|I(i,j)\times g[\mathrm{grad}(i,j)]-\hat{I}(i,j)\times g[\mathrm{grad}(i,j)]\right\|^2}{mn} \tag{2}$$

which means, both the original and the super-resolved images are multiplied by a weight map g element-wise, and then the normal MSE is calculated between them.

B. Model Combination

In [1], combination of different trained networks to pursue better performance is proposed. However, the combination objective is to optimize the resulting PSNR rather than OCR accuracy. In this paper, we propose to combine different trained networks to achieve higher OCR accuracy. For example, if we have multiple trained networks $N_i, i = 1, \ldots, M$, we can train a set of combination weights $\{w_i\}$ to fuse the outputs of these networks as the final super-resolved image, i.e.

$$\hat{I} = \sum_{i=1}^{M} w_i \times N_i(I_{LR}) \tag{3}$$

where I_{LR} is the LR input image. We want the OCR accuracy of $\hat{I}$ to be as high as possible. However, the combination weights $\{w_i\}$ cannot be calculated analytically as the optimization target is OCR accuracy. We perform exhaustive search from a finite set of weights to identify the optimum.

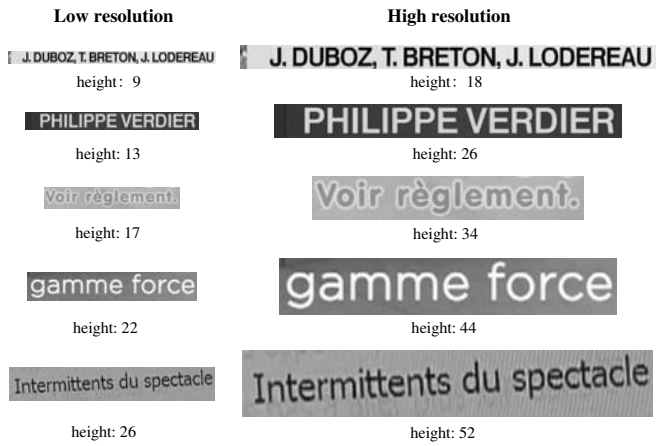


Fig. 1. Examples of LR-HR image pairs in the training data of the ICDAR 2015 TextSR dataset.

C. Image Padding

Padding is a common practice in CNN. For example, the VDSR network [10] adopts zero padding before each convolutional operation to ensure the output size is the same as the input size. Although this manner is claimed to be effective in [10], we observed the results of using VDSR on the text images are not satisfactory, especially at image boundaries. The reasons are twofold. On the one hand, text images in the TextSR dataset are much smaller compared with natural images (some example images are shown in Fig. 1), thus the missing of boundary information has more severe impact. On the other hand, VDSR is very deep (20 layers), the receptive field of the last convolutional layer in VDSR is 41×41 (a very large extent) [10], which makes the boundary effect more obvious.

We propose an image padding method as a remedy to the VDSR network on small text images. After the network is trained, before inputting an image into the network, we enlarge the image by padding the boundaries with a strip width of 10 pixels. The VDSR network does not change the resolution, and thus the network output is cropped at the center to produce the final super-resolved image that has the same size as the original. Our padding method is designed in a recursive manner, in each recursion the image is padded with a strip width of 1 pixel, and there are 10 recursions. The pixel value of each padded pixel is decided as the average of pixel values of the 11 pixels that are the nearest to the padded one and are on the image boundaries. For comparison purpose, we also test a commonly used padding method, i.e. duplication padding that duplicates the boundary pixel values outwards.

IV. Experimental Results

Dataset. We use the ICDAR 2015 TextSR dataset provided by the ICDAR 2015 Competition on Text Image Super-Resolution [2]. The dataset consists of a training set and a test set. The training set consists of 567 pairs of HR-LR grayscale images and the ground-truth OCR results. Both HR and LR images are down-sampled from the high quality images extracted from the French television video flux by factors of 2 and 4, respectively. OCR results include English letters, numbers, and 14 special characters such as ",". Fig. 1 shows several examples of HR-LR image pairs in the training data. The test set consists of 141 pairs of HR-LR images as well as their ground-truth OCR results. For more details of the dataset, please refer to [2]. We selected 30 image pairs from the training set for validation purpose, so the training data used in our experiment are the remaining 537 pairs of images.

TABLE II
Results of different padding methods

Method	PSNR	MSSIM
No padding	32.18	0.9841
Duplication padding	32.31	**0.9842**
Our proposed padding	**32.42**	**0.9842**

Implementation Details. All LR images are up-sampled by a factor of 2 using bicubic interpolation to obtain the interpolated low-resolution (ILR) images. Sobel operator is applied on the HR images to calculate weight maps. The ILR images, HR images, and weight maps, are cut into 18×18 sub-images to produce training samples. In total there are 157,321 training samples. We use the deep learning framework Caffe to perform experiments. The initial learning rate of convolutional filters is 10^{-2} while the initial learning rate of biases is 10^{-3}. Learning rate is decreased to 1/10 every 49,160 iterations and the training is stopped after 196,640 iterations. The training is observed to converge quickly.

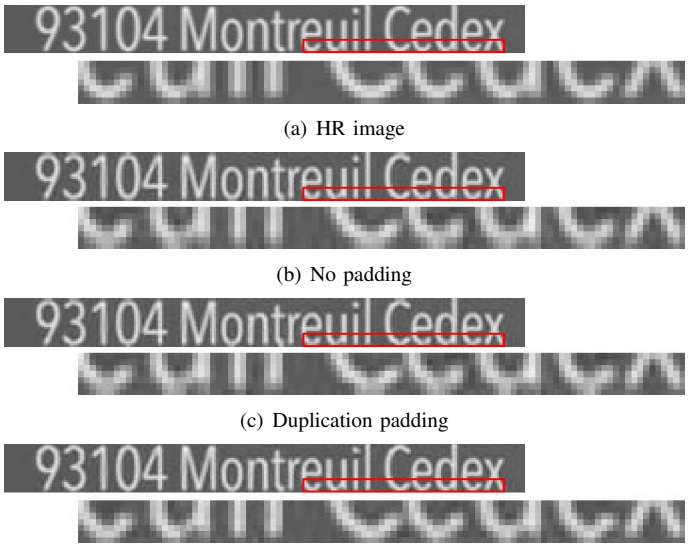
A. Image Padding Results

We first verify the effectiveness of the proposed image padding method. The VDSR trained by normal MSE is tested here. Fig. 2 shows some example results including the HR image, the result without padding, and the results with different padding methods. It can be observed that when not using padding, the resulting image contains some artifacts around the image boundaries. The artifacts are greatly reduced by adopting padding. In addition, the average PSNR and MSSIM results are summarized in Table II, which shows our proposed padding method performs the best.

B. Results of New Loss Function

We then verify the proposed gradient-based WMSE as loss function. Results are summarized in Table III, where three forms of weighting function are compared. The OCR accuracy is measured by using the Tesseract-OCR software[1]. We empirically observed the initialization of CNN has impact on the performance of the trained CNN, which was also mentioned in [1]. Therefore, for each weighting function, we perform four times of experiments and report the performance of each trained network, and the average performance of the four times in Table III. It can be observed the best results

[1]Tesseract-OCR version 3.02, http://code.google.com/p/tesseract-ocr/



(a) HR image

(b) No padding

(c) Duplication padding

(d) Our proposed padding

Fig. 2. Example results using different padding methods.

TABLE III
RESULTS OF DIFFERENT WEIGHTING FUNCTIONS AND USING DIFFERENT (RANDOM) INITIALIZATIONS WHEN TRAINING CNN

Weighting Function	Network No.	OCR (%)	MSSIM	PSNR
$f(x) = x^2$	Net1.1	76.34	0.9795	31.65
	Net1.2	76.54	0.9800	31.71
	Net1.3	76.41	0.9796	31.58
	Net1.4	75.86	0.9795	31.59
	Average	76.29	0.9797	31.63
$f(x) = x$	Net2.1	76.54	0.9815	31.88
	Net2.2	76.82	0.9812	31.80
	Net2.3	75.15	0.9813	31.91
	Net2.4	76.44	0.9811	31.75
	Average	76.24	0.9813	31.84
$f(x) = \sqrt{x}$	Net3.1	77.23	**0.9828**	**32.00**
	Net3.2	75.15	0.9821	31.95
	Net3.3	76.72	0.9826	31.94
	Net3.4	**77.98**	0.9827	31.99
	Average	76.77	0.9826	31.97

of single network, in terms of PSNR, MSSIM, and OCR accuracy, are all achieved when using $f(x) = \sqrt{x}$. In the average sense, $f(x) = \sqrt{x}$ also performs the best. It is also noticeable that all the trained networks are worse, in terms of PSNR, than that using normal MSE (shown in Table II), because optimizing PSNR is equivalent to minimizing normal MSE. However, in terms of OCR accuracy, the networks using WMSE perform better. The best single network achieves an OCR accuracy of 77.98% which exceeds the best result of Table I.

C. Model Combination Results

We perform a simple model combination experiment, where we choose the two networks, Net3.2 and Net3.4 in Table III, for combination. Both networks are trained with $f(x) = \sqrt{x}$ that is known to perform well, and we choose these two networks because they differ the most in terms of OCR accuracy. More different networks are presumed to be more complementary. We use the weights $w_1 = \alpha$ and $w_2 = 1 - \alpha$, and we search the optimal α within $[0, 1]$ with step size 0.01.

TABLE IV
RESULTS OF MODEL COMBINATION

Data	Network	OCR (%)	OCR of combined (%)
Training data	Net3.2	73.58	74.25
	Net3.4	73.93	
Test data	Net3.2	75.15	78.10
	Net3.4	77.98	

The OCR accuracy on the *training* data is used to decide the optimal α. Table IV shows that the combined model, with $\alpha = 0.67$, achieves higher OCR accuracy than either network on both training and test data. The final best performance on test data is an OCR accuracy of 78.10%, which is close to the result of using original HR images for OCR (78.80%).

V. CONCLUSION

In this paper, we summarize our attempts of improving CNN-based text image SR method to facilitate OCR. We propose a new loss function when training CNN, which assigns more weights on the edge regions to guide the CNN to focus on high-frequency image details. We also propose a simple yet effective image padding method, and conduct model combination to further improve the performance. Experimental results show that our method can achieve a high OCR accuracy from super-resolved text images, which is close to that of using original HR images. We will continue our researches in two directions, first, investigate a method to better initialize the CNN; second, study an advanced model combination method.

REFERENCES

[1] C. Dong, X. Zhu, Y. Deng, C. C. Loy, and Y. Qiao, "Boosting optical character recognition: A super-resolution approach," *arXiv preprint arXiv:1506.02211*, 2015.
[2] C. Peyrard, M. Baccouche, F. Mamalet, and C. Garcia, "ICDAR2015 competition on text image super-resolution," in *ICDAR*, 2015, pp. 1201–1205.
[3] C. Peyrard, F. Mamalet, and C. Garcia, "A comparison between multi-layer perceptrons and convolutional neural networks for text image super-resolution," in *VISAPP (1)*, 2015, pp. 84–91.
[4] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International Conference on Curves and Surfaces*, 2010, pp. 711–730.
[5] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *ACCV*, 2014, pp. 111–126.
[6] R. F. Moghaddam and M. Cheriet, "A multi-scale framework for adaptive binarization of degraded document images," *Pattern Recognition*, vol. 43, no. 6, pp. 2186–2198, 2010.
[7] R. Walha, F. Drira, F. Lebourgeois, C. Garcia, and A. M. Alimi, "Resolution enhancement of textual images via multiple coupled dictionaries and adaptive sparse representation selection," *International Journal on Document Analysis and Recognition*, vol. 18, no. 1, pp. 87–107, 2015.
[8] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014, pp. 184–199.
[9] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *ECCV*, 2014, pp. 372–386.
[10] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.
[11] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *CVPR*, 2017, pp. 4681–4690.
[12] F. Aghdasi and R. K. Ward, "Reduction of boundary artifacts in image restoration," *IEEE Transactions on Image Processing*, vol. 5, no. 4, pp. 611–618, 1996.