SVDocNet: Spatially Variant U-Net for Blind Document Deblurring

Bharat Mamidibathula[1], Prabir Kumar Biswas[2]
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur[1,2]
bharatmamidi@gmail.com[1], pkb@ece.iitkgp.ac.in[2]

Abstract

Blind document deblurring is a fundamental task in the field of document processing and restoration, having wide enhancement applications in optical character recognition systems, forensics, etc. Since this problem is highly ill-posed, supervised and unsupervised learning methods are well suited for this application. Using various techniques, extensive work has been done on natural-scene deblurring. However, these extracted features are not suitable for document images. We present SVDocNet, an end-to-end trainable U-Net based spatial recurrent neural network (RNN) for blind document deblurring where the weights of the RNNs are determined by different convolutional neural networks (CNNs). This network achieves state of the art performance in terms of both quantitative measures and qualitative results.

1 Introduction

With the advent of digitization, document and text based images have become very prominent in one's quotidian lifestyle, spanning over reports, certificates, receipts, handwritten documents, etc. During image acquisition, numerous unavoidable factors such as camera shake, focusing errors, and noise may corrupt the image, leading to loss of valuable information. Hence, image post-processing became mandatory. This step is especially vital in automated information retrieval and optical character recognition systems. The process of image degradation in single image deblurring is modelled as

$$\mathbf{y} = \mathbf{x} * \mathbf{k} + n \tag{1}$$

where **y** is the observed image, **x** is the original image, and **k** is the unknown blurring kernel, also known as the point spread function (PSF), and n denotes uncorrelated additive noise. Blind deconvolution is the method of obtaining the original image and, in some cases, the PSF, from the observed image. The problem of blind image deblurring is highly ill-posed and non-convex.

Many techniques have been used for deblurring of text-based images. Early on, statistical and learning based methods were prominent for blur kernel estimation. With the emergence of deep learning, using CNN based approaches were proposed as function approximators to predict the deblurred image.

Although these methods have proven to give admirable results, they still have certain pitfalls. We assume that the function modelled by the CNN for image restoration is a spatially invariant function, whereas this may not be true, as in the case of dynamic scenes[12]. Also, deconvolution of different types of blur kernels would inevitably increase the model parameters and computational expenses. Hence, model adjustment based on the PSF and the need for spatial variance became necessary.

We propose SVDocNet, an end to end trainable spatially variant network based on the well known U-Net encoder-decoder architecture, consisting of recurrent layers in the skip connections between the encoder-decoder blocks. Additionally, we have three auxiliary networks that do not contribute to

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

any intermediary features or outputs, but learn the internal adjustments that must be customized to each image in the form of the primary network's weights to guide the propagation of features. We evaluate the model on benchmark datasets and compare the results with state of the art solutions.

2 Previous Work

Document deblurring is an important constituent of digital document analysis and has undergone incredible advancements of late. Document restoration is done extensively to remove the effects of warping, shading distortions, noise artifacts, blurring, etc.

Early image deblurring methods utilized statistical properties of images such as sparsity priors[6], L_0-norm gradient prior[11], iterative learning procedures[10], etc.

In recent years, neural networks and deep learning have taken over the fields of data analytics, computer vision, and natural language processing. CNNs have produced astonishing results in well versed computer vision problems like image classification, object detection, image segmentation, etc. Image deblurring was no exception, with various CNN models having been proposed. Chakrabarti[1] designed a CNN to predict the Fourier coefficients of a deconvolution filter for image restoration. Nah[8] followed a coarse-to-fine CNN approach while Kupyn[5] developed a conditional adversarial network for deblurring. Zhang[12] proposed a spatially variant RNN for dynamic scene deblurring.

Similar approaches have been proposed for document deblurring, producing commensurate results. Chen[2] observed document image characteristics and integrated priors, while Pan[9] proposed an L_0-based regularized prior to deblur text-based images. Hradiš[3] trained a CNN to restore document images. Jiao[4] discretized the entire blur kernel space into multiple sub-spaces and used a two stage CNN for sub-space classification and trained a separate CNN for each sub-space.

3 Motivation

Consider an input signal x being passed through a system having an impulse response k to generate an output signal y. In the one dimensional scenario, the output of the system is calculated as

$$y[n] = \sum_{m=0}^{M} k[m]x[n-m] \quad \textit{which} \quad \textit{gives} \quad x[n] = \frac{y[n]}{k[0]} - \sum_{m=1}^{M} \frac{k[m]x[n-m]}{k[0]} \tag{2}$$

where M is the maximum range of the impulse response. The system response (PSF in our case) is assumed to be a finite impulse response (FIR), a supposition which holds true for a most instances.

By recursive expansion of the values of $x[n-m]$ in (2), we see that an infinite signal is required to model $x[n]$ from the given $y[n]$ values. Since the blur kernel is assumed to have a finite receptive field, we can infer that the receptive field of its inverse filter will be much larger than the blur kernel itself. Using conventional CNNs for deconvolution, which are nothing but convolutions followed by nonlinear activations, to approximate the original image would require a huge number of parameters and a large, complex architecture to cover the entire receptive field.

Alternatively, we would require much lesser coefficients by reproducing the IIR model for deblurring as shown in (2). We can see that fewer parameters ($k[m], m = 0, 1, ..., M$) are needed in this model provided that we are able to incorporate a sufficiently large receptive field into the restoration process.

A spatially variant RNN satisfies all of these requirements. However, a few refinements are necessary to merge information extracted from different filtering directions. This is accomplished by splitting the weights' map yielded by the secondary networks into four parts i.e. the forward and reverse directions for both the x-axis and y-axis and selecting the best direction for each element of the feature space. To account for a large receptive field, we add a convolutional layer between consecutive RNN layers to merge the extracted information. This complete model allows us to efficiently cover a large receptive field with use of minimal network parameters, significantly reducing the computation cost.

4 Network Architecture

Our network embodies one primary spatially variant U-Net for the basic deblurring task along with three additional U-Net based networks. Contrary to the standard U-Net, the skip connections between

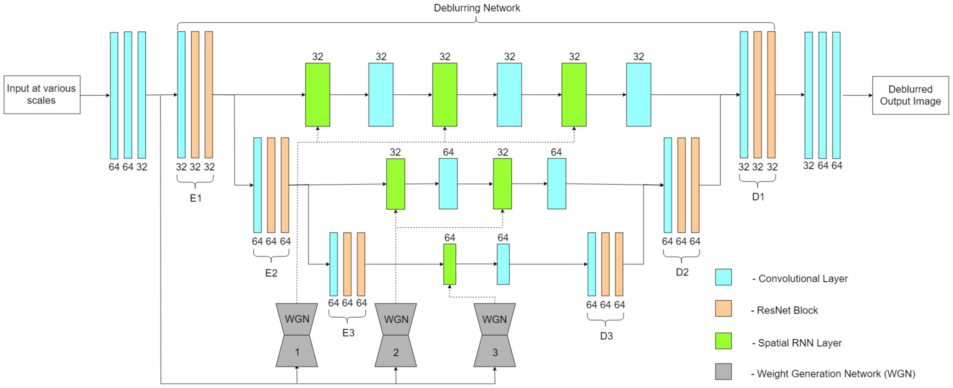

Figure 1: Detailed network architecture of SVDocNet. The values indicate the number of channels corresponding to the layer or block. The values for the spatial RNN layer indicate the number of filters for each of the four directions. E-x denotes an encoder block and D-x denotes a decoder block.

the encoder and decoder blocks consist of alternating convolutional and RNN layers. Each RNN layer has four subsections, to account for spatial information from each of the four directions. From top to bottom, the skip connections have three, two, and one RNN layers respectively. After every RNN layer, a convolutional layer is present to effectively merge the features extracted from the RNN layers. Each block has a convolutional layer and two ResNet blocks to enhance gradient propagation.

The ReLU activation is used for all layers of the primary network. Convolutional layers of stride 2 are utilized for downsampling in the encoder and bilinear upsampling is applied in the decoder. The first three and last three convolutional layers external to the deblurring network have 9x9, 7x7, and 5x5 filters to account for a larger receptive field. All of the remaining layers consist of 3x3 filters.

We used three conventional U-Net based architectures consisting of a similar ResNet and convolutional layer setup as that of the primary network for the weight generation networks (WGNs). The filter size for each layer is 3x3 and the number of filters for each of the three encoder-decoder pairs going from top to bottom is 32, 64, and 128 respectively. The last layer of each WGN yields the weights required for the RNN layers and uses the 'tanh' activation function. This operation is vital as the magnitude of the generated weights need to be less than 1 as in Liu[7] in order to stabilize the overall system. The composite network architecture with the network configurations are depicted in Figure 1.

We trained our model for 100 epochs using the Adam optimizer and a learning rate of 10^{-4} on the Keras framework with an Nvidia GTX 1080Ti GPU. We used 10000 image pairs from Hradiš'[3] training dataset which were created from random patches of documents from the CiteSeerX repository.

5 Results

We tested our proposed model on Hradiš'[3] test dataset containing unseen document images using unique kernels with different levels of noise degradation. We compared our results with the baseline network presented by Hradiš, a model which can be duly regarded as a benchmark in the problem space of blind document deblurring. We used two popular measures of image quality i.e. the peak signal to noise ratio (PSNR) and the structural similarity index (SSIM) for evaluation of our network.

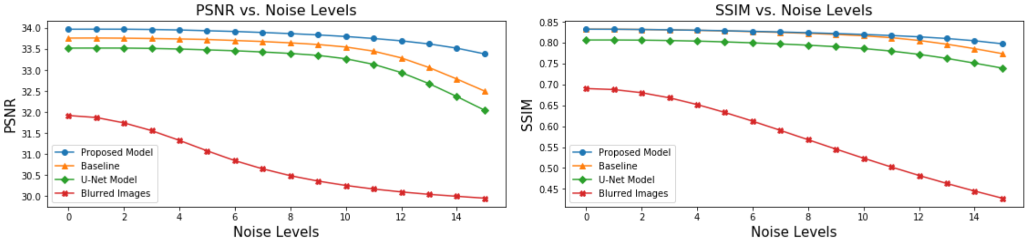

Figure 2: Quantitative results of proposed model in comparision with benchmark models.

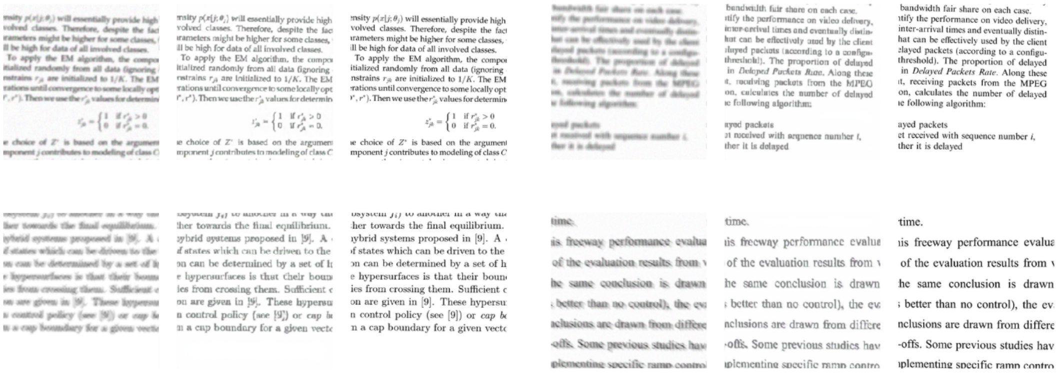

Figure 3: Qualitative results of our proposed model. In each set of three images, the first image is the blurred image, the second image is the the deblurred image yielded by our proposed network, and the third image is the original sharp image.

From Figure 2, we can clearly see that our proposed model performs better compared to both the standard U-Net architecture (assuming all the network parameters to be the same) and also the benchmark architecture proposed by Hradiš[3], especially at higher noise degradation levels.

6 Conclusion

We proposed SVDocNet, an end-to-end trainable spatially variant U-Net based architecture for blind document deblurring, replacing the skip connections between the encoder and decoder blocks with alternating convolutional and recurrent layers for efficient feature extraction. Three auxiliary U-Net networks are present to predict suitable weights for the recurrent layers by examining the input blurred image. We demonstrated the potency of this system both quantitatively and qualitatively.

References

[1] A. Chakrabarti. A neural approach to blind motion deblurring. In *European conference on computer vision*, pages 221–235. Springer, 2016.

[2] X. Chen, X. He, J. Yang, and Q. Wu. An effective document image deblurring algorithm. In *CVPR 2011*, pages 369–376. IEEE, 2011.

[3] M. Hradiš, J. Kotera, P. Zemcık, and F. Šroubek. Convolutional neural networks for direct text deblurring. In *Proceedings of BMVC*, volume 10, page 2, 2015.

[4] J. Jiao, J. Sun, and N. Satoshi. A convolutional neural network based two-stage document deblurring. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 703–707. IEEE, 2017.

[5] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

[6] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971. IEEE, 2009.

[7] S. Liu, J. Pan, and M.-H. Yang. Learning recursive filters for low-level vision via a hybrid neural network. In *European Conference on Computer Vision*, pages 560–576. Springer, 2016.

[8] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[9] J. Pan, Z. Hu, Z. Su, and M.-H. Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2908, 2014.

[10] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf. Learning to deblur. *CoRR*, abs/1406.7444, 2014.

[11] L. Xu, S. Zheng, and J. Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013.

[12] J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R. W. Lau, and M.-H. Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2521–2529, 2018.