Improving the Perceptual Quality of Document Images Using Deep Neural Network

Ram Krishna Pandey and A G Ramakrishnan

Department of Electrical Engineering,
Indian Institute of Science,
Bangalore, India
ramp@iisc.ac.in, agr@iisc.ac.in

Abstract. Given a low-resolution *binary document image*, we aim to improve its perceptual quality for *enhanced readability.* We have proposed a simple, deep learning based model, that uses convolution with transposed convolution and sub-pixel layers in the best possible way to construct the high-resolution image. The proposed architecture scales across the three different scripts tested, namely Tamil, Kannada and Roman. To show that the reconstructed output has enhanced readability, we have used the objective criterion of optical character recognizer (OCR) character level accuracy. The reported results by our CTCS architecture shows significant improvement in terms of the subjective criterion of human readability and objective criterion of OCR character level accuracy.

Keywords: Readability · Binary Document Image · Super-resolution · Deep Learning · OCR .

1 Introduction

The perceptual quality of binary document images can be ascertained in terms of the subjective criterion of human readability and the objective criterion of the character level accuracy (CLA) of an optical character recognizer (OCR). In this work, our primary goal is to improve the quality of low-resolution, binary document images for *better human readability.* OCR character level accuracy is used as a metric to objectively show that the quality of the document images has significantly improved. Thus, a secondary objective for the proposed technique can be as a preprocessing step before feeding a low resolution, document image as input to an OCR.

The performance of an OCR in terms of character and word level accuracies decreases when the documents are of very poor quality and resolution (see Table 1). Language models can be applied, as a post-processing step, on the text output by the OCR to correct the recognition errors. Alternately, a new classifier can be designed to operate on the low-resolution character images and still achieve good accuracy.

In this work, while dealing with our primary goal, we address three different tasks: (i) improving the human readability of the documents, (ii) improving the quality of the input image such that the existing OCRs perform well (the character level accuracy of the OCR should be better than that on the input image), (iii) ensuring that the method works on multiple languages and resolutions. The advantages of this approach are: (i) low-resolution images stored on digital libraries are rendered better for direct human readability. (ii) We can avoid designing a new classifier to operate on such low-resolution document images. This reduces the burden of changing the design of the existing OCRs by training the classifier for each language independently, which in turn needs huge training data from each of these languages.

We have approached to solve the above-mentioned problem using techniques based on deep learning. Given an LR document image, the challenge is to generate a HR version of it, which should be perceived by a human as better readable, than the corresponding input image. Also the OCR should achieve higher CLA on the reconstructed output. This problem was earlier attempted in [4], [5], [6] and [7]. In [4], the authors have shown that the quality of the down-sampled version of the document image can be enhanced and brought to that of the input image. In [5], the authors have proposed an efficient convolutional neural network (CNN) architecture, coupled with bicubic interpolation, to achieve better performance in terms of OCR accuracy, starting from a down-sampled version of the same image. In [6], the authors have used the traditional interpolations (bicubic, bilinear and nearest neighbor), coupled with CNN, to achieve better performance in terms of WLA, when the input image is directly fed to the model. In [7], the authors have shown language dependent quality enhancement of document images. Motivated by these works, that aim to enhance the quality of low-resolution input images for better OCR recognition, here we propose an architecture which can upscale any low-resolution document image and improve the perceptual quality so that humans find it easy to read and also the generated output should have more recognition accuracy than the input. The architecture developed here is for an upscaling factor of 2. Our contributions can be summarized as follows:

- We have performed comprehensive experiments on a huge collection of document images. We found that the OCR character level accuracies can be increased by improving the resolution of poor resolution, binary document images as shown in Table 1.
- We *define, formulate and address* (to a good extent) this problem of increasing the perceptual quality of binary document images for better human readability.
- We have used state-of-art deep learning techniques to find the optimal and judicious combination of transposed convolution [10] and sub-pixel [11] layers to obtain super-resolution of binary document images.
- We have created a *unique dataset* that captures maximum possible variations in the input image space, particularly to address this challenging problem (details in Sec. 4).
- To our knowledge, this is the first report on the enhancement of the perceptual quality of document images for better human readability. (see Table 2).
- Our proposed algorithm enhances the quality of low-resolution binary document images and improves the OCR (CLA) by a good margin.

- Our algorithm works *independent of the three languages and the resolutions* it has been tested so far.

2 Related Work

Image super-resolution is a well-known problem for natural images. There are mainly two classes of image super-resolution, namely, single image super-resolution (SISR) and multi-image super-resolution (MISR). SISR has been the focus of many important publications [15, 14, 16–19]. The main assumption in all these papers is that the image in the LR space has the same local geometry as that of the image in the HR space. The researchers have endeavored to find out the representation common for both the spaces, using which the HR image can be formed. In [15], the authors have found out the common representation in the concatenated feature space of LR and HR dictionaries.

2.1 Traditional Approaches not Based on Deep Learning

Some traditional approaches that deal with enhancing the quality of document images for better OCR recognition are mentioned below. Shi et al. [1] deal with the removal of noise (large blob or clutter noise, salt, and pepper noise) and non-text objects such as form line or rule lines from handwritten document images. They describe a region growing algorithm to fix salt and pepper noise. They also provide an approach to eliminate noisy artifacts that include multiple categories of degradation.

A non-parametric, unsupervised method is presented in [2] to deal with color or gray images e.g. camera captured or mobile document images. It Uses contrast limited, adaptive histogram equalization separately on HSV color space, an optimal conversion algorithm to transform the document image to the gray level and un-sharp masking to sharpen the useful information. The sharpened image is binarized by Otsu algorithm and fed to the OCR to obtain the final text.

Kumar et al. [3] have extended the application of sparse coding and dictionary learning techniques. Here, the basis/atoms of the dictionary for the binary document image are learnt by treating binary document images as distinct from natural images. They claim that their method restores degradations such as cuts, merge, blobs and erosion in the document. Besides these, there are reports in the literature that deal with noise removal from the document images with the sole aim of improving the OCR accuracy.

All the approaches mentioned above are used to improve the OCR accuracy for certain kinds of degradation (noise) associated with the input image. However, none of them deal with the improvement in the resolution of the input image or aim to enhance the readability of the document images, where, originally, a lot of missing pixels might have resulted in changes in the shape, structure and interpretation of the characters.

2.2 Approaches Based on Deep Learning

Since the advent of deep learning based models like [17], investigators have learned a multiple layer representation of the input images (called features) to capture maximum variations in the input image space. Once such a model is learned, they have used it to construct a HR from a single LR image. However, to our knowledge, the problem of super-resolution of binary document images has not been attempted prior to its conception in the recent papers [4–6]. In [4, 5], the authors have shown that the downsampled version of a document image can be worked upon to reconstruct an image of the original resolution. However, these techniques fail to generalize, when one inputs a LR binary document image directly, and not a downsampled version of an existing HR image. The technique of nonlinear fusion of multiple interpolations (NFMI), proposed in [6], performs direct upscaling of a Tamil document image scanned at any resolution, to result in a better word level accuracy. The NFMI method uses multiple interpolations, together with a CNN, to learn a mapping function, which takes in a LR document image and produces the corresponding HR image.

In this work, instead of using interpolations to perform convolution in the high-resolution space, we have used two recent techniques, famous in the deep learning community, namely transposed convolution [10] and sub-pixel [11] convolution to learn the mapping function. Unlike interpolations, transposed convolution layer learns weights from the training data to upscale the image. Hence, it can capture more variations in the input images and can be used for improving the quality, independent of the languages. We have created the training dataset in such a way that it covers maximum possible variations in the input image space.

3 Motivation

The issue addressed in this work is an actual industrial problem, where a huge number of documents have been scanned at a low resolution, and unfortunately, the original documents have been destroyed and hence, are no longer available for better scanning. The images are of poor quality, also because they are obtained from very old newspapers. Since the available regional language OCRs have a very poor recognition performance on these images, we would like to improve the quality of such images so that humans find it easy to read. It is very difficult and laborious to manually type and/or correct such a high volume of documents. Also, some of the images are so degraded that even native people find it uneasy to read them. Even if we are able to slightly improve the human readability, it is immensely useful.

Our studies have shown that documents scanned at a resolution of 100 dpi result in an average OCR accuracy of less than 50 %; when the same documents are scanned at 200 dpi, the same OCR performs reasonably well and gives CLAs of 80-98%. Hence, it is sufficient for our task, if we can find an optimal model to upscale the document by a factor of 2. And our proposed architecture gives a relative improvement of around 51% in terms of CLA.

4 Dataset Created

The training dataset has been created from binary scanned images of documents in three different scripts (languages), capturing the multitude of variations in the input image space. The LR patches are created in three ways: (i) by taking *alternate pixels* from the HR patches, (ii) degrading the LR patches by *multiplying them element-wise by random masks of zeros and ones*, and (iii) by selecting the patches directly from images scanned at a LR setting of the scanner. The total number of LR-HR patch pairs created for training is around fifty million. The LR patches are of size 16×16 and the corresponding ground truth patches are of size 32×32. A patch pair is removed, if the LR patch contains only background pixels.

5 Architecture Advanced

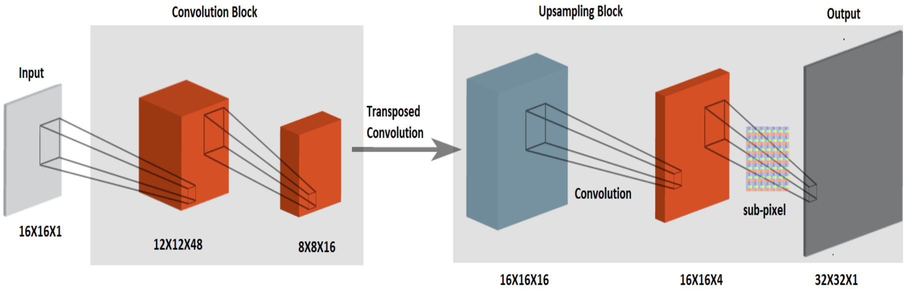

Fig. 1. The convolution-transposed convolution-subpixel (CTCS) architecture for 2X upscaling of binary document images. It is designed to deal with true LR images, scanned at a low resolution and not simulated LR images obtained by downsampling original HR images.

The convolution-transposed-convolution-subpixel (CTCS) architecture proposed by us for upscaling a document image by a factor of 2 is illustrated in Fig. 1. The architecture has a convolution block, followed by an upscaling block. The upscaling block is formed by a transposed convolution, a convolution, and a sub-pixel layer. The initial convolution block extracts the relevant features from the input low resolution, binary image. This block uses 48 filters of size 5×5, followed by another layer of 16 filters of size 5×5. The transposed convolution layer has 16 filters of size 9×9 to upscale the feature maps to double their size. The convolution layer between the transposed convolution and sub-pixel layers is to reduce the number of features to the desired size. The size of the extracted feature map is decreased because we have used convolution filters without padding, to remove the artifacts. This convolution layer provides more non-linearity in the system. These controlled feature maps are passed on to the sub-pixel layer for upscaling the features to double their size. The sub-pixel layer takes in a tensor of size 16 × 16 × 4 and gives an output of size 32 × 32 × 1.

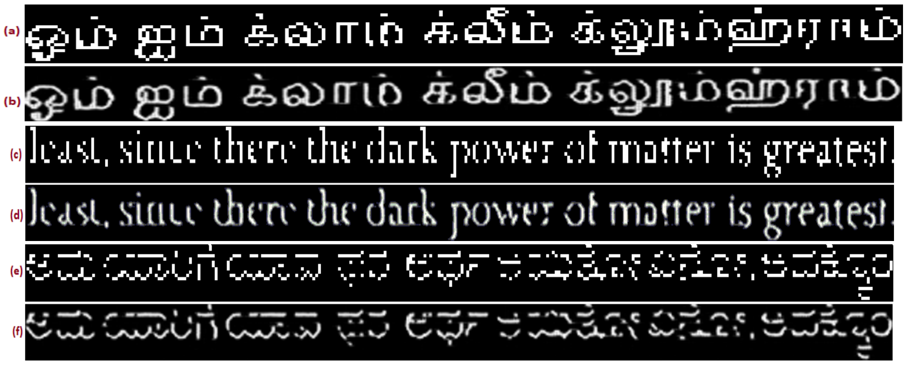

Fig. 2. (a), (c), (e): Tamil, English and Kannada input images originally scanned at 100 dpi. (b), (d), (f): the corresponding 200 dpi output images created by our CTCS architecture, with moderately enhanced human readability. It can be seen that the letter /ம்/ that appears in every Tamil word (Figs. 2 (a) and (b)) has a clearly improved perceptual quality. In the English images, the letters, "d, o, o, s, g, s" are better enhanced.

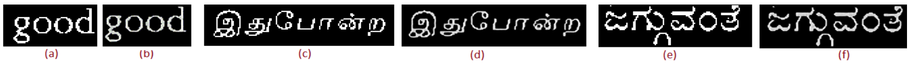

Fig. 3. (a), (c), (e): English, Tamil and Kannada input images originally scanned at 200 dpi. (b), (d), (f): the corresponding output images created by our CTCS architecture with increased perceptual quality.

5.1 Transposed Convolution (TC)

To train the model, where the input is in a low resolution space and the output is in a high resolution space, we need a mechanism to first move to the high resolution space. In an earlier work [4], interpolation techniques such as bicubic, bilinear and nearest-neighbor, or their combination have been used to take the LR images to the HR space. Unlike interpolations, transposed convolution [10] learns the weights to upscale the size of the feature map. Thus, the learning based techniques may contribute more than the interpolations in taking the image from a LR to a HR space. The size of the output obtained from the transposed convolution layer with stride = 1 and padding = 0 is given by $o = i + (f_s - 1)$, where i, o, and f_s are, respectively, the spatial sizes of the input feature map, output feature map and the filter.

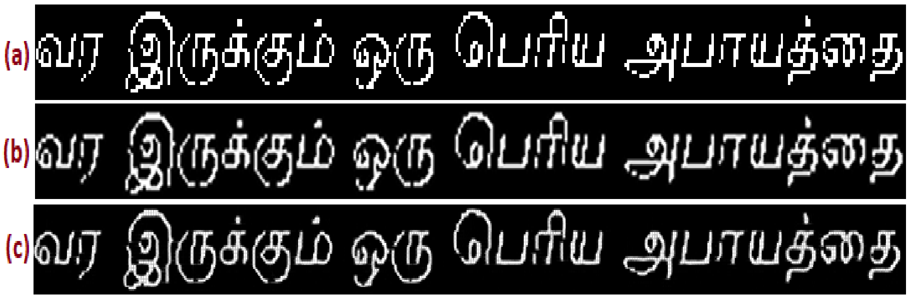
(a) வர இருக்கும் ஒரு பெரிய அபாயத்தை
(b) வர இருக்கும் ஒரு பெரிய அபாயத்தை
(c) வர இருக்கும் ஒரு பெரிய அபாயத்தை

Fig. 4. (a), (b): Input binary image and the corresponding output images (gray scale) created by our CTCS architecture. (c): image in (b) after gamma correction [20], displaying significantly improved perceptual quality and enhanced readability.

5.2 Sub-pixel Convolution (SC)

Unlike the TC, the sub-pixel convolution [11] is a technique that only rearranges the feature maps to increase its size or resolution. Hence it is faster than the TC and achieves good results, if properly positioned as a layer in the architecture. So, we avoid the use of multiple TC layers and replace the last layer with SC.

6 Training the CTCS Architecture

Given a low resolution, binary document image our goal is to construct a high resolution image with enhanced human readability and better quality in terms of CLA and WLA. Let the training set be $\{I_l^i, O_h^i\}$, $1 \leq i \leq N$, where N is the total number of patch pairs. The model weights are initialized by the technique in [13]. For each input I_l^i, the model builds a high resolution counterpart image, R_h^i. If there is an error between the produced output and the corresponding ground truth O_h^i, the weights of the model are adjusted to minimize the mean square error loss function.

The CTCS model is first trained with the patch pairs from the Tamil document image. The trained model is then further tuned on English, and then on Kannada. During the process of training, we save the weights at different stages and perform extensive experiments in order to identify the model that works best on any input image, scanned at any resolution. The model is trained for a maximum of 25 epochs. To calculate the gradients required while training, normal back-propagation is used. For adjusting the weights, Adam optimizer [9] is used with a learning rate of 0.0001, β_1 of 0.9 and β_2 of 0.99.

7 Results and Discussion

Figure 2 shows the output images for one sample input image each for three different languages, namely, Tamil, English and Kannada. In each pair of images, the top ones are the inputs scanned at a resolution of 100 dpi and the bottom one are the outputs obtained from our model. The improvement in the perceptual quality of the images is evident from observing the letter ‘ ம் ’ that occurs as the last letter of every word in Figs. 2 (a) and (b). Better enhancement is seen in the English letters, “d, o, g” and the “s” that occurs in the last two words. Figure 3 shows that our CTCS model works independent of the input resolution, where the perceptual quality of the outputs generated by our model is significantly better.

Figure 4 (a) shows a sample, Tamil binary image scanned at 150 dpi. (b) and (c) show the CTCS-reconstructed and gamma corrected images respectively. The strokes of the characters have become smoother, while preserving the structure, thus resulting in better readability.

Table 1. Mean character level accuracy of 150 images from the 3 languages, each containing around 1k characters scanned at different resolutions (100, 200 and 300 dots per inch).

	Scanned at 100 (dpi)	Scanned at 200 (dpi)	Scanned at (300 dpi)
Average CLA	31.2	90.26	95.06

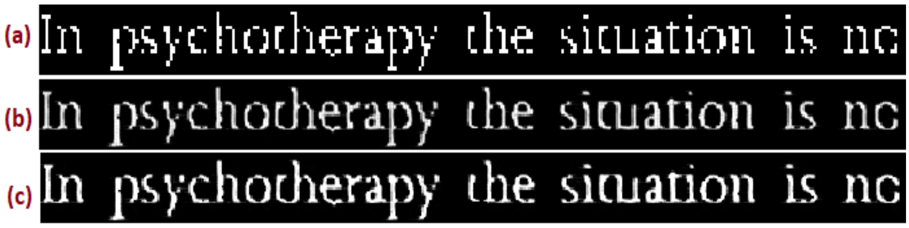
(a)
In psychotherapy the situation is nc
(b)
In psychotherapy the situation is nc
(c)
In psychotherapy the situation is nc

Fig. 5. (a): Input English image originally scanned at 150 dpi. (b): the corresponding output images created by our CTCS architecture with increase in perceived quality and ease of readability.(c): image in (b) after gamma correction [20]

Figures 5 (a) and 6(a) show a sample English and Kannada image each, scanned in binary mode at a resolution of 100 dpi. (b) and (c) show the CTCS-generated and gamma corrected respectively. In both cases, almost all the characters are better enhanced and hence the readability improves significantly.

Figure 7 illustrates how the loss of pixels changes the shapes of characters. The loss of one or two pixels from the character ‘m’ of the word ’problem’ makes it appear as two characters i.e. r and n. Similar analysis can be done for the other characters too. In the characters reconstructed by the CTCS architecture, the strokes are smooth, not pixelated. Also the gaps are filled, preserving the structure and the meaning of the characters.

We performed our initial experiments by scanning the images in binary mode at resolutions of 100, 200 and 300 dpi. Table 1 lists the OCR character level accuracy on 150 document images, each containing over 1000 characters.

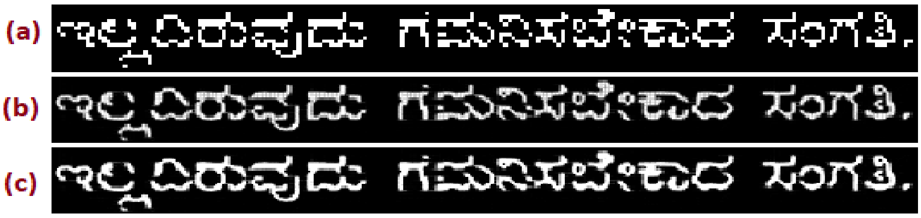

Fig. 6. (a) Input Kannada image originally scanned at 100 dpi. (b) the corresponding output images created by our CTCS architecture. (c) image in (b) after gamma correction [20]

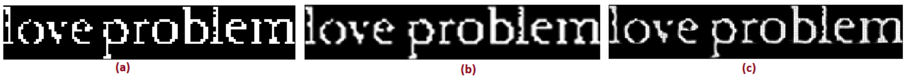

Fig. 7. (a), (b), (c): Input English, bicubic interpolated and the corresponding output images created by our CTCS architecture with increase in perceived quality and ease of readability. In the word 'love' character 'o' and in the word 'problem' character 'm' is broken in both input and the bicubic interpolated output. But the same is reconstructed in the CTCS architecture

Table 1 show that the images scanned in binary mode with less than 200 dpi can be treated as low resolution images. Also, results mentioned in the table 1 suggest that increasing the resolution help in improving the quality of the input images, that will eventually leads to better readability and OCR character level accuracy. As it can be seen that just by increasing the resolution by factor of 2 the OCR recognition accuracy has significantly improved. But these are at hardware level (means scanner skipping or taking some pixels randomly). Suppose in the situation where the documents are originally scanned at low resolution say at 100 dpi and the original document is destroyed. Can we increase the resolution such document images that the OCR perform well? The first thing that strikes our mind is the interpolation such as bicubic, bilinear or any other super-resolution algorithm can can be used. Traditional interpolation helps in increasing the OCR accuracy slightly and marginally the readability. To tackle this real challenging problem we created the dataset mentioned in sec. 4 and proposed the architecture shown in Fig. 1. We obtained the mean opinion score on the perceptual quality of the input and output images from 10 subjects each, for each of the three languages. The overall MOS of all the evaluators on all the language documents are listed in Table 2. The HR images obtained by the CTCS architecture have been evaluated at a MOS of 7.5, as compared to 4.5 for the input.

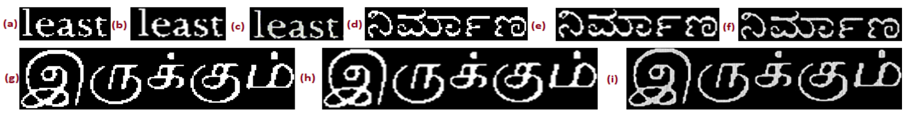

Fig. 8. (a) Input (b) bicubic (c) CTCS (d) input (e) bicubic (f) CTCS (g) Input (h) bicubic (i) CTCS: All images are scanned at 200 dpi: Shows our model is resolution independent.

Table 2. Mean opinion score (10 point scale) of the enhancement in the perceptual quality of the document images scanned at 200 dpi

	Input	CTCS
Average MOS	4.5	7.5

Table 3. Mean character level accuracies (CLA) on 15 document images, 5 each from Tamil, Kannada and English, before and after upscaling by our CTCS architecture. Input images are scanned at 100 dpi.

	Input CLA%	CTCS%
Average CLA	33.1	49.99

Table 3 shows the average character level accuracies obtained from the images scanned at 100 dpi, and the outputs created by our CTCS architecture. The results show significant improvements in CLA. The mean improvements in terms of CLA relative to input is 51%.

Our model is designed to be independent of these three languages and the resolution and hence, can be applied on document images of any one of them. Hence, we do not need to change the design of the OCR for the above mentioned performance gain. The results can further be enhanced, if we incorporate (capture) further possible variations in the input training images.

8 Conclusion

We have created a *unique and diverse dataset* that captures maximum possible variations in the input image space from documents of three different languages (key to the success of our model). After performing extensive experiments, We have obtained an effective architecture based on deep learning, that is trained on this diverse dataset. We have shown that the proposed method enhances the perceptual quality of the input document images in terms of the subjective criterion of *human readability* and the objective criterion of OCR recognition, *independent of the languages and resolutions tested.* To our knowledge, this is the first report on enhancement of document images for better readability. Our method scales across the three languages tested and works for multiple input resolutions.

References

1. Shi, Z., Setlur, S., and Govindaraju, V.: Image enhancement for degraded binary document images. Document Analysis and Recognition (ICDAR), IEEE, (2011).
2. El Harraj, A., and Raissouni, N. : OCR accuracy improvement on document images through a novel pre-processing approach. arXiv preprint arXiv:1509.03456, (2015).
3. Kumar, V., Bansal, A., Tulsiyan, G. H., Mishra, A., Namboodiri, A., and Jawahar, C. V.: Sparse document image coding for restoration. In: 12th IEEE International Conference on Document Analysis and Recognition (ICDAR), pp. 713-717, (2013).

4. Pandey, R. K., and Ramakrishnan, A. G.: Language Independent Single Document Image Super-Resolution using CNN for improved recognition. arXiv preprint arXiv:1701.08835 (2017).
5. Pandey, R. K., and Ramakrishnan, A. G.: Efficient document-image super-resolution using convolutional neural network. Sadhana **43**.2 (2018): 15.
6. Pandey, R. K., Maiya, S. R., and Ramakrishnan, A. G.: A new approach for upscaling document images for improving their quality. In: 14th IEEE India Council International Conference (INDICON). IEEE, (2017).
7. Pandey, R. K., Vignesh, K., Ramakrishnan, A. G., and Chandrahasa, B.: Binary Document Image Super Resolution for Improved Readability and OCR Performance. arXiv preprint arXiv:1812.02475, (2018).
8. LeCun, Y., Bottou, L., Orr, G., and Muller, K.: Efficient backprop in neural networks: Tricks of the trade. Lecture Notes in Computer Science, **1524**(98), 111 (1998).
9. Kingma, D. P., and Ba, J.: Adam: Amethod for stochastic optimization. In: Proc. 3rd Int. Conf. Learn. Representations, (2014).
10. Xu, L., Ren, J. S., Liu, C., and Jia, J.: Deep convolutional neural network for image deconvolution. In: Advances in Neural Information Processing Systems, pp. 1790-1798 (2014).
11. Shi, W., Caballero, J., Huszr, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016).
12. Shivakumar, H. R., and Ramakrishnan, A. G.: A tool that converted 200 Tamil books for use by blind students. In: Proceedings of the 12th International Tamil Internet Conference, Kuala Lumpur, Malaysia, (2013).
13. He, K., Zhang, X., Ren, S., and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp. 1026-1034 (2015).
14. Glasner, D., Shai, B., Michal, I.: Super-resolution from a single image. In: 12th IEEE International Conference on Computer Vision, (2009).
15. Yang, J., Wright, J., Huang, T. S., and Ma, Y. : Image super-resolution via sparse representation. IEEE transactions on image processing, **19**(11), 2861-2873 (2010).
16. Timofte, R., De Smet, V., and Van Gool, L. : Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision, (2013).
17. Dong, C., Loy, C. C., He, K., and Tang, X. : Learning a deep convolutional network for image super-resolution. In: European conference on computer vision, pp. 184-199. Springer, Cham, (2014).
18. Ledig, C., Theis, L., Huszr, F., Caballero, J., Cunningham, A., Acosta, A., Tejani, A., Totz, J., Wang, Z., and Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network.: In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105-114 (2017).
19. Lai, W. S., Huang, J. B., Ahuja, N., and Yang, M. H. : Deep laplacian pyramid networks for fast and accurate superresolution. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. **2**, no. 3, p. 5, (2017).
20. Kumar, D., and Ramakrishnan, A. G. : Power-law transformation for enhanced recognition of born-digital word images. In: International Conference on Signal Processing and Communications (SPCOM), 2012 , pp. 1-5. IEEE, (2012).