Enhancing OCR Accuracy with Super Resolution

Ankit Lat C. V. Jawahar
Centre for Visual Information Technology
IIIT Hyderabad, INDIA

***Abstract*—Accuracy of OCR is often marred by the poor quality of the input document images. Generally this performance degradation is attributed to the resolution and quality of scanning. This calls for special efforts to improve the quality of document images before passing it to the OCR engine. One compelling option is to super-resolve these low resolution document images before passing them to the OCR engine.**

In this work we address this problem by super-resolving document images using Generative Adversarial Network (GAN). We propose a super resolution based preprocessing step that can enhance the accuracies of the OCRs (including the commercial ones). Our method is specially suited for printed document images. We validate the utility in wide variety of document images (where fonts, styles, and languages vary) without any pre-processing step to adapt across situations. Our experiments show an improvement upto 21% in accuracy OCR on test images scanned at low resolution. One immediate application of this can be in enhancing the recognition of historic documents which have been scanned at low resolutions.

I. Introduction

Documents are pervasive in our everyday life and we store them on our mobile phones or personal computers either by taking pictures or by scanning on flat bed scanners. We also inherit large document image collections from digital libraries of historic documents. Often the quality of documents is not very good for automated recognition of the text. This poor quality limits their use for various applications. Any attempt to restore the resolution of these documents using interpolation, gives rise to blurry outcomes, which does not help either. In this work we use deep learning to super-resolve document images. Super-resolution (SR), unlike image zooming, involves adding details to the image which is not present in the low-resolution (LR) image. It is typically an ill-posed inverse problem. Simple interpolation (e.g. Bilinear, Bi-cubic or spline) leads to a blurry outcome. In our work, we super-resolve low-resolution document images scanned at as low as 75dpi. These super resolved images are then fed to OCR engine for the recognition. We validate the performance of our solution using quantitative measures like accuracy of the OCR and improvements in PSNR on the output image. We compare it with baselines such as Bi-Cubic interpolation. Our results show significant improvement, as high as 21.19% in the accuracy of OCR for document images in English. We also show that by virtue of our training methodology, once the network is trained using English document images, it works well across other languages and font type and sizes. Figure 1 depicts a document before and after the super resolution. Some of the words are shown by zooming in. Original document (at the top) has a character accuracy of 95.5%. While the super resolution leads to an improved accuracy of 99.5%. Both these are recognized by the same commercial OCR – ABBYY[1].

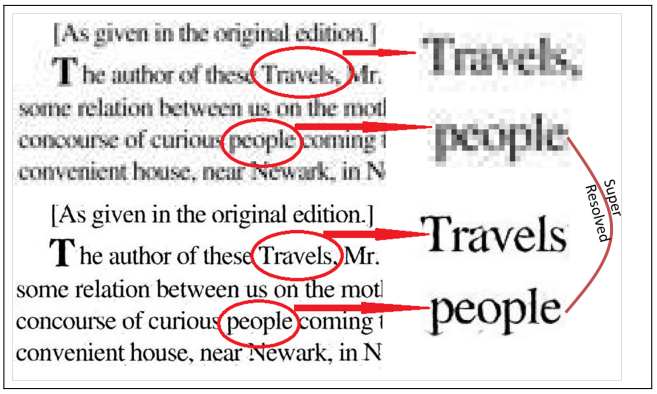

Fig. 1: Sample qualitative OCR accuracy of proposed SR method. Here we obtain a character accuracy of 99.5% on super-resolved image (bottom) while original low resolution image (top) gives 95.5% with ABBYY

Methods for super-resolution can be broadly grouped into two categories: (i) The classical sub-pixel aligned multi-image ([3], [9], [17]), and (ii) Example-Based super-resolution ([11], [10], [22]). Single image super resolution can be example based or sparse coding based. In multi-image based method, each of the low resolution images imposes a constraint on the unknown pixel in the higher resolution grid. If enough low-resolution images are available, which is impractical for document images, then these images can be used to recover high-resolution image. This may provoke a thought that if we have many sub-pixel shifted images then we can get proportional amount of super-resolution. However this has been shown ([1], [25]) that under practical and synthetic conditions an image can be super-resolved by a factor of 1.6 and 5.7 respectively. In example based methods, mapping is learned from a database of low and high resolution image pairs and then given a new low-resolution image recover its most likely high resolution version. This method is highly data hungry and language specific [5].

Apart from super-resolution based approaches many other image processing techniques have been employed to increase the accuracy of OCR. These classical approaches dwell on some form of image restoration [20] and enhancement as a preprocessing or post processing step for improving OCR accuracy. Among the techniques studied in literature most of them involve some form of character segmentation (or image

[1]ABBYY is a commercial OCR tool. https://www.abbyy.com/en-eu/

binarization [15]) in its OCR pipeline. In [16] authors have trained a set of classifiers for character recognition using segmented characters. The proposed method accuracy depends highly on the segmentation accuracy. In [12] authors have used independent principal component analysis to augment the OCR accuracy, whose performance is marred by variation in character shapes. Furthermore, suggested approach operates on word images rather than full document page.

Document images have some special interesting properties. They have repeating character patterns. They also have the strokes/glyphs that are shared across multiple characters, languages and styles. We believe, this makes the preprocessing stage that we develop to be applicable across a wide variety of situations.

Recent advances in machine learning has given a set of new tools to tackle advanced vision problems. Rather than manually designing an algorithm, a machine learns complex representations from the data. With the success of Convolutional Neural Networks in the image classification [23], it has been used for many complex vision tasks (e.g. Segmentation [26], Object Detection [29] etc.). However despite of this progress we still need to design loss function to meet our training objectives. And if we take a naive approach and minimize Euclidean distance between ground truth and predicted output then it will tend to produce blurry results [18], wherein super-resolution we like to have more details.

When it comes to natural images many works have been reported in literature with state of art performance. Dong *et al.* in [6] and [7] have used bicubic interpolation to upscale an input image and trained a deep fully convolutional network end-to-end to achieve state of-the-art super-resolution performance and also show that with deeper networks and more training data improves reconstruction. Subsequently, it was shown that enabling the network to learn the upscaling filters directly can further increase performance both in terms of accuracy and speed ([8], [30], [33]). With their deeply-recursive convolutional network (DRCN), Kim *et al.* [21] presented a highly performant architecture while keeping the number of model parameters small.

However when it comes to document image super-resolution similar techniques have not been fully extended. In [2] authors have used energy minimization framework using MRF. The proposed solution is iterative in nature and hence output quality varies with number of iterations. Because of iterative nature it is computationally slow. In [5] authors have used exemplar based approach with image prior based regularization. The problem with these kinds of methods is the training database size and the selection of patch size. Optimal patch size selection depends on language and its font-size. Despite of overcoming these limitations results remain blurry.

To the best of our knowledge, in all the works reported so far, super-resolution has been shown on images scanned at 100dpi or above. In this work we demonstrate super-resolution on documents scanned at as low as 75dpi along with improvement in the accuracy of OCR. We also demonstrate that the trained network has cross-language compatibility and super-resolves text across a range of font sizes and types along with preserving the color of the input image.

II. Method

To motivate the problem, we conducted a small experiment. We took a set of 10 high resolution English document images randomly from web with different font-types, sizes and spacing. We prepared three different set of test images. First by scanning these images with a CannonScan 5600F color scanner at 75dpi of size 620 $\times$ 876 pixels. One such patch extracted from a scanned document is shown in Figure 2. Second and third sets were prepared using mobile camera

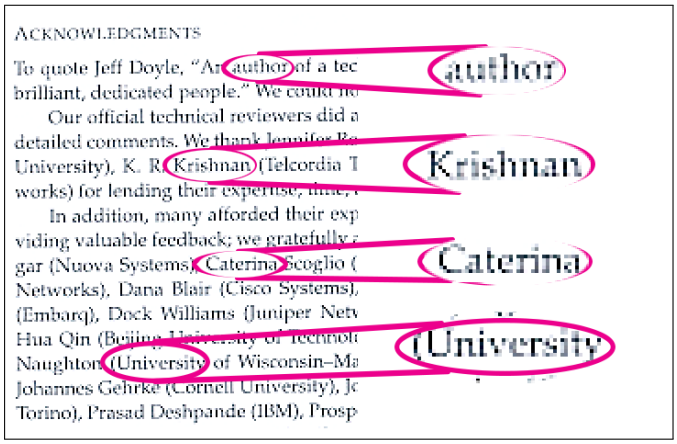

Fig. 2: Challenges posed by images scanned at low resolution. Zoomed in words show a lots of cuts and merges which are difficult to segment.

based scanner apps namely OfficeLens[2] and CamScanner[3] using their default settings. Images captured using mobile apps were of dimension 1920 $\times$ 1080 and 1280 $\times$ 960 pixels respectively. Although scanning at such low resolution may be acceptable to human perception and understanding but when it comes to machine readability accuracy goes down significantly. On all these sets the document word and character accuracy of OCR was estimated using two popular OCR tools namely ABBYY FineReader and Tesseract[4]. Table I shows the accuracy of OCR on all the three sets of document images.

Generally factors which affect accuracy of OCR can be blur, character merging and fragmentation [4], resolution (e.g. [32], [28]), language, illumination, skew and noise [19]. For more comprehensive treatise please refer to [27]. Each of the above listed problem has been addressed separately in the past; most of them being language specific.

As we captured data for the above test under controlled condition so potential factors like blur, illumination, skew etc. are less likely to contribute towards OCR inaccuracy. However one thing which is common among all the tests set is the poor resolution. Due to poor resolution strokes of the

[2]https://www.microsoft.com/en-in/store/p/office-lens/9wzdncrfj3t8
[3]https://www.camscanner.com/
[4]https://opensource.google.com/projects/tesseract

OCR Tool	Accuracy (%)	Cannon Scan 5600F	Cam Scanner	Office Lens
ABBYY	Word	91.60	91.50	97.00
	Character	96.95	93.74	96.55
Tesseract	Word	0.10	55.00	81.50
	Character	17.50	78.14	86.18

TABLE I: Benchmark accuracy of OCR on low resolution images scanned from flatbed scanner (CannonScan 5600F) and mobile apps CamScanner and OfficeLens (from Microsoft).

characters become more zigged and start merging with the document background which leads to erroneous segmentation of characters. One can easily make out this degradation from zoomed image patches shown in Figure 2. As a result of these degradations the accuracy of OCR degrades which is evident from Tesseract output shown in Table I. Thus we posit that accuracy of OCR can be increased by super-resolving the document images by enhancing its details. Though super-resolution of document images can be done using example based methods or by using a standard interpolation technique. Using either of these methods have their own limitations as discussed above in section on introduction. This motivates us to look for a method which can handle all these modalities namely language, font-types, font-size and document color.

GAN is a well known tool capable of generating high-frequency details. In [18] authors demonstrate the capability of the GAN at generating high frequency details, which is at the heart of image generation. As document images contain a lot of high frequency details owing to character strokes and boundary, we strongly believe that GANs capability can be leveraged to accomplish this task of super-resolution. We therefore train a variant of GAN, known as SRGAN [24] with desired modification to handle color documents as explained later. This is simple yet deep architecture with huge capacity. Generator network consist of 16 identical residual blocks (each residual block has two convolution layer with kernel of size 3×3 followed by batch normalization layers and Parametric-cReLU). Output from these residual blocks is fed to two sub-pixel convolution layers (as suggested by Shi et al [30]) to increase the resolution of the output image. The output image so generated is then fed to a discriminator network consisting of eight convolutional layers with increasing number of 3×3 filter kernels, increasing by a factor of 2 from 64 to 512. As the number of features are doubled in every consecutive convolutional layers, strided convolution is used to contain the size of the final output feature map to 512. These 512 features are then fed to two fully connected layers followed by a sigmoid activation function to get probability for sample classification. It follows the same training methodology to train generator and discriminator alternatively as suggested in [14].

It is hard to obtain natural training data for super-resolution. Therefore we artificially create the training data automatically. Since, typical document images have white background and black foreground, creating a dataset of LR-HR pair by down-sampling and training standard SRGAN will learn a network biased towards creating an output with white background and black foreground. In order to reap the advantage of plethora of document images publicly available for training

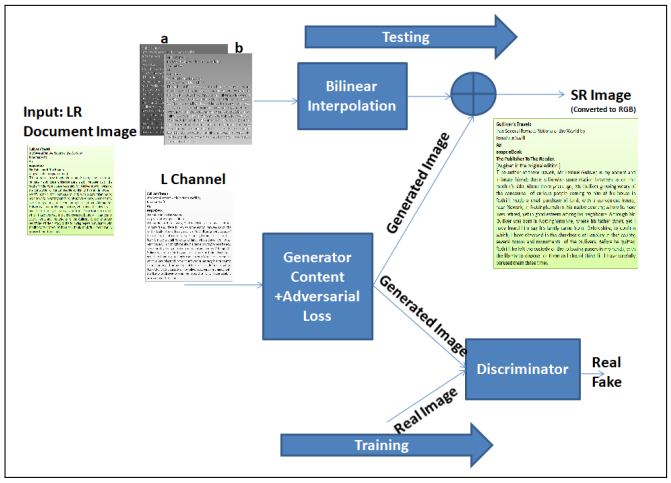

Fig. 3: Document image L component is only taken for training and 'ab' channels are upscaled using bilinear interpolation and combined with generator output post training to generate SR image

we modify the SRGAN architecture to handle color. Instead of training SRGAN on color image we first convert the RGB document image to Lab color space, wherein all the channels are mutually orthogonal. After converting, training is done only on the L component [13] while a and b components are up-scaled using standard interpolation. This keeps the look and feel of super-resolved same as the low-resolution image. The generator loss function used for training modified SRGAN has two components namely content loss [24] and adversarial loss [14]. This content loss is defined as the sum of mean square loss (between ground truth and generated image) and VGG (for VGG refer [31]) loss. The VGG loss, which is refined while training on document images, plays a very important role in reconstructing high frequency details. It is defined as mean square error between VGG feature maps of generated and ground truth image. When we generate training image pairs noise, cuts and merges are automatically induced (see Figure 6) when we down-sample the images to 60dpi. And when training is done using these image pairs, network also learns to clean noise and at the same time super-resolving the details. Training is done end-to-end on a set of 100 document image pairs for 1000 epochs. Figure 3 shows training and testing methodology. It takes roughly six hours to train the network. Typical time taken for generating a high resolution document image of size 2404×2404 from a 601×601 image is 1.56s. For more details on dataset please refer section on experiments and results. Figure 4 shows the generator and discriminator loss incurred during training. Initially image generated looks blurred but as the training progress they get sharper and sharper.

III. Experiments and Results

A. Data Sets

In order to train and test we prepared dataset using English novels. First, these pdfs were converted to images of high quality at 600 dpi of size 5100×6601. For training, 100 low-resolution and high-resolution pairs were generated from these

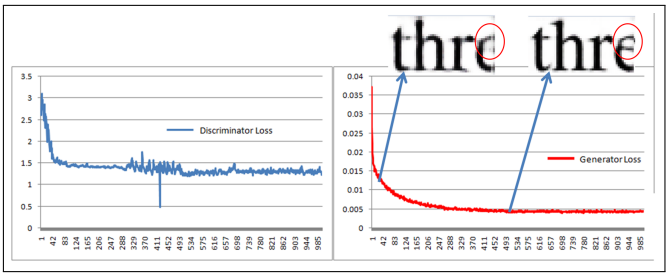

Fig. 4: (Left) Discriminator loss (Right) Generator loss. Images on top shows improvement in reconstruction as training progresses. A close look at the images show that the network learns to reconstruct high frequency details (focus on character 'r' and 'e')

high quality images by re-sampling them to 510 × 660 and 2040 × 2640 respectively. This down sampling induces noise, cuts and merges in the generated low resolution images. These modalities so induced help in robust learning. One such patch from a document pair is shown in Figure 6. For evaluating the performance of trained network we create three different types of dataset described below:

- **English Novel Dataset**: It contains synthetically generated low-resolution document images: 16 images of size 495 × 545 pixels and 13 images of size 311 × 425 taken randomly from two different English novels other than the one used for training. Here onwards it will be referred by name END.
- **Cross Language Dataset**: Its a random collection of document images of different languages. In subsequent sections it will be referred as CLD dataset.
- **Scanned Web Dataset**: It is collection of random English document images crawled from web with different fonttypes and sizes and scanned with CannonScan 5600F at 75dpi. Each image is 620 × 876 pixel. We shall refer this dataset from now on as SWD dataset.

Example patches from CLD and SWD dataset are shown in Figure 7.

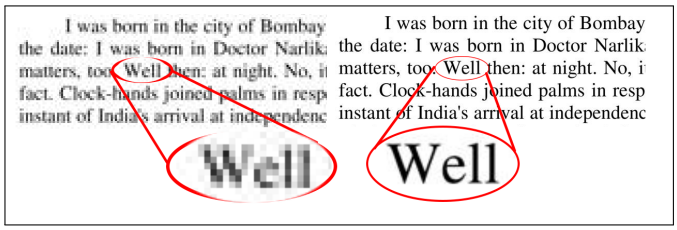

Fig. 6: Sample patch visualization from artificially created training (END) dataset: (Left) low-resolution (60dpi) showing cuts and merges and (Right) high-resolution (240dpi)

B. Evaluation Metrics

Evaluation of results on images are carried out in two ways

- **Objective**: For objective evaluation different metric measures are used such as Peak-to-Signal Noise Ratio (PSNR), Root Mean Square Error (RMSE), Structural Similarity Index (SSIM), Optical Character Recognition (OCR). These methods scale well to a large number of images but do not necessarily correspond to the perceived visible quality.
- **Subjective**: This is purely based on human perception. These types of methods are of two types. i) Reference: In this method a human is asked to rate the quality of the output image in comparison to the reference ground truth. ii) Non-Reference: Quality of the output is assessed on a preset scale. We use PSNR and accuracy of OCR for objective evaluation of the results.
  Figure 9 shows the improvement in visual quality on the image patches taken from END and SWD dataset. In the next section we discuss about evaluation metric along with the results.

a) PSNR*:* Let I be the noise free image and K be its noisy version then PSNR is defined as (for an 8-bit image):

$$\text{PSNR} = 10 \log \frac{255^2}{\text{MSE}}$$

where MSE is defined as

$$\text{MSE} = \frac{\sum_{j=1}^{N} \left(\sum_{i=1}^{M} (I_{i,j} - K_{i,j})^2 \right)}{MN}$$

In the above equation M & N refer to the height and width of the document image under consideration.

b) OCR *Accuracy:* For computing accuracy of OCR Levenshtein distance is measured between two document image, after performing OCR on it using ABBYY Fine Reader and Tesseract. Accuracy is quantified in terms of word as well as character level.

Accuracy (%)	LR	Bi-Cubic	SR
Character	98.69	98.72	99.65
Word	94.22	93.49	98.90

TABLE II: Improved accuracy of OCR on END dataset: Significant ($> 4\%$) improvement in word accuracy of OCR with ABBYY

IV. Results and Discussion

a) ***Low Resolution Images****:* For objective performance evaluation of the trained network, images from END dataset were zoomed by 4x using trained network and compared with standard Bi-cubic interpolation. Bar graph shown in Figure 8 shows PSNR improvement compared to Bi-Cubic on an average by 3.7 dB for 16 test images.

b) **OCR** ***Accuracy****:* In this experiment we evaluate the performance of SR with respect to accuracy of OCR. Accuracy is evaluated on END and SWD dataset. OCR has been done using ABBYY Fine Reader and Tesseract. Measured accuracies are tabulated in Table II and III. On SWD dataset word level accuracy improves by 2.5% for ABBYY and by 21.19% for Tesseract. This quantitative improvement in accuracy occurs as a result of enhanced details in the super-resolved images.

c) ***Cross-Language****:* We trained our network on English low quality document images. In order to explore the generalizability of this trained network we carried out cross language experiment. In this experiment we tested the trained network on CLD dataset. Our experiments reveal that the

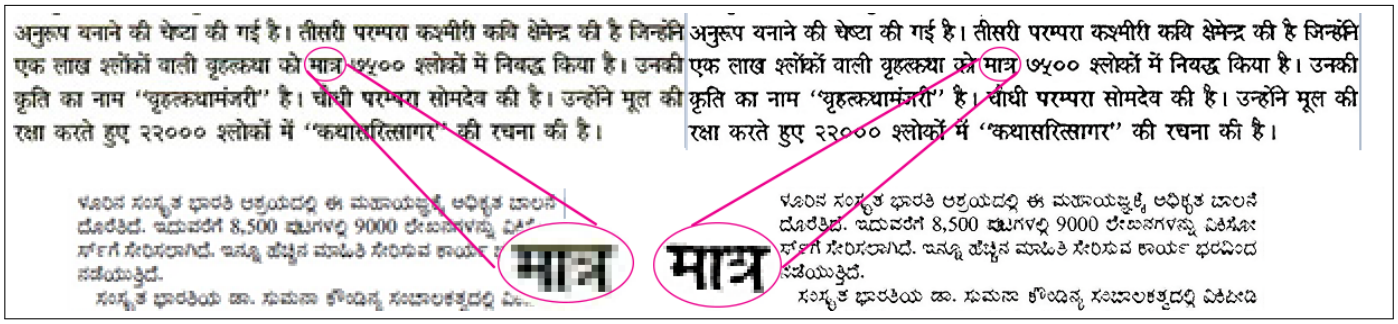

Fig. 5: Reconstructed patch of Hindi (Top) and Kannada (Bottom) documents (taken from CLD dataset) with significantly less cuts and merges. (Left) low-resolution and (Right) super-resolved.

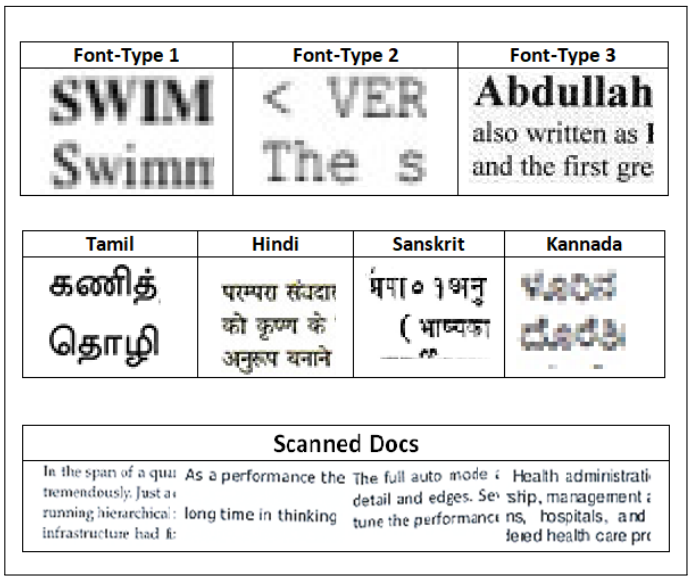

Fig. 7: Sample patches depicting: (Top) font-type and size variation in SWD dataset. (Middle) four different languages namely Tamil, Hindi, Sanskrit and Kannada in CLD dataset. (B) scanned image patches from CLD dataset.

OCR Tool	Accuracy (%)	LR	Bi-Cubic	SR
ABBYY	Word	91.60	92.70	95.20
	Character	96.95	96.95	97.88
Tesseract	Word	0.10	64.77	85.96
	Character	17.50	77.8	89.70

TABLE III: Performance evaluation of OCR on SWD dataset: Significant (21.19%) relative (compared to Bi-Cubic) improvement in word accuracy of OCR with Tesseract

same network performs well for other languages. For the purpose of illustration we present our result on Hindi, Sanskrit, Tamil and Kannada. Figure 10 shows small patches of super-resolved images. Figure 5 shows big patches for Hindi and Kannada texts. From these results one can easily make out the improvement in the quality of the zoomed image. Images super-resolved using trained network are more sharp and clear.

The proposed method using a modified version of SRGAN performs very well in super-resolving low resolution document images. As the document images are not as complex as natural

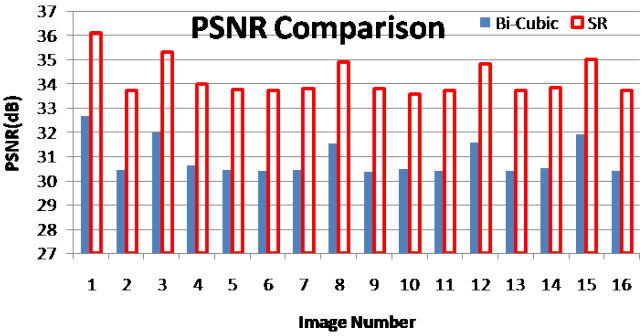

Fig. 8: PSNR average increase by 3.7dB. Image number on the horizontal axis refers to the 16 images taken for comparison

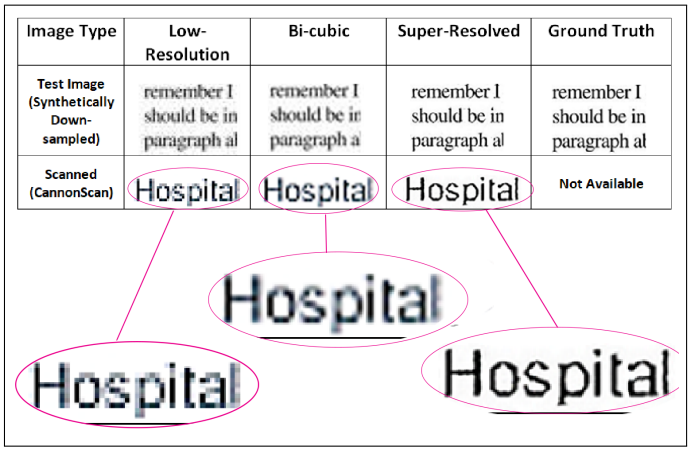

Fig. 9: Subjective quality comparison. Zoomed in image shows the superior quality of SR image patch compared to baseline Bi-Cubic interpolation and LR.

images, training of the network does not take much time as compared to time intensive training involved in image recognition or segmentation networks. Once trained on a particular language, here in English, it performs equally well on other languages. This clearly indicates that the network learns the abstract features which are necessary for super-resolving the document images across languages. Improvement in the PSNR clearly indicates the objective improvement in the quality

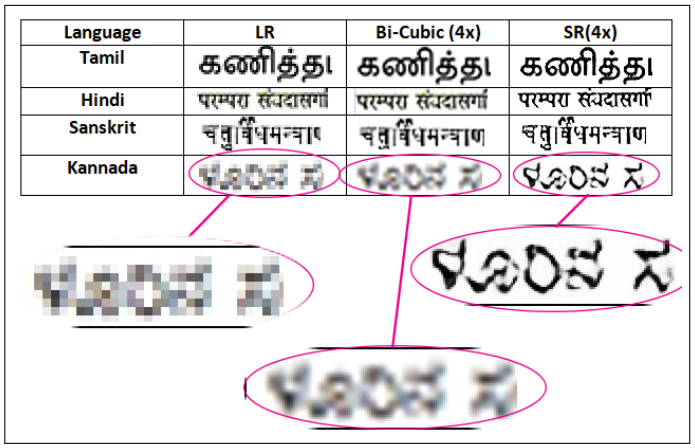

Fig. 10: Visual image quality improvement on images patches taken from languages namely Tamil, Hindi, Sanskrit and Kannada

of the document image. As this network was trained using artificially created data, it is unable to generalize well on low resolution documents with low quality paper. It is able to link edges using blur as prior to some extent however if the gap is more then it is left out as the network is not trained to in-paint. From the Table III it is evident that accuracy of OCR increases with both ABBYY and Tesseract. In case of Tesseract there is significant improvement in word accuracy by 21.19% when compared with Bi-cubic interpolation. This increase in accuracy owes not merely to the increased resolution alone. Zoomed images using SR-network are more sharper and clean, thus their recognition improves compared to that of Bi-cubic zoomed images.

V. Summary and Conclusions

In this work we demonstrate improved OCR accuracy using super-resolution on low resolution document images scanned at as low as 75dpi using Generative Adversarial Network. Our methodology improves character level accuracy by 21.19% in case of Tesseract. We also show that the trained network is invariant to color, language, font-type and size as well. One extension of this work that can be explored is to super-resolve document images with low-quality of paper and to bridge the broken characters, which will further boost OCR. Another line of work can be on document font-stylization.

Acknowledgement

This work was partly supported by IMPRINT project, Govt. of India.

References

[1] Simon Baker and Takeo Kanade. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002.
[2] Jyotirmoy Banerjee and C.V. Jawahar. Super-resolution of text images using edge-directed tangent field. *DAS*, 2008.
[3] D. P. Capel. Image mosaicing and super-resolution. *University of Oxford*, 2001.
[4] R. G. Casey and E. Lecolinet. A survey of methods and strategies in character segmentation. *PAMI*, 1996.
[5] Dmitry Datsenko and Michael Elad. Example-based single document image super-resolution: a global map approach with outlier rejection. *Springer*, 2007.
[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. *Computer Vision - ECCV*, 2014.
[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.
[8] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. *Computer Vision - ECCV*, 2016.
[9] Sina Farsiu, Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multi-frame super-resolution. *IEEE Transactions on Image Processing*, 2003.
[10] William T. Freeman, Thouis R. Jones, and Egon C. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 2002.
[11] William T. Freeman, Egon C. Pasztor, and Owen T. Carmichael. Learning low-level vision. *Intl. J. Computer Vision*, 2000.
[12] Utpal Garain, Atishay Jain, Anjan Maity, and Bhabatosh Chanda. Machine Reading of Camera-Held Low Quality Text Images: an ICA-Based image enhancement approach for improving ocr accuracy. *ICPR*, 2008.
[13] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. *ICCV*, 2009.
[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems 27*, 2014.
[15] Maya R. Gupta, Nathaniel P. Jacobson, and Eric K. Garcia. Ocr binarization and image pre-processing for searching historical documents. *Pattern Recogn.*, 2007.
[16] J. C. Handley. Improving ocr accuracy through combination: a survey. *Systems, Man, and Cybernetics, IEEE Conference*, 1998.
[17] Michal Irani and Shmuel Peleg. Improving resolution by image registration. *CVGIP: Graph. Models Image Process.*, 1991.
[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, 2016.
[19] A. Khohnzad and Y.H. Hong. Invariant image recognition by zernikc moments. *PAMI*, 1990.
[20] Van Cuong Kieu, Florence Cloppet, and Nicole Vincent. Adaptive fuzzy model for blur estimation on document images. *Pattern Recognition Letters*, 2017.
[21] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. *CoRR*, 2015.
[22] Kwang In Kim and Younghee Kwon. Example-based learning for single-image super-resolution. *Proceedings of the 30th DAGM Symposium on Pattern Recognition*, 2008.
[23] Alex Krizhevsky, I Sutskever, and G. E Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.
[24] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, 2016.
[25] Zhouchen Lin and Heung-Yeung Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.
[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.
[27] George Nagy. Twenty years of document image analysis in pami. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2000.
[28] Abderrahmane Namane and M.A. Sid-Ahmed. Character scaling by contour method. *PAMI*, 1990.
[29] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *CVPR*, 2017.
[30] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *CVPR*, 2016.
[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.
[32] D. E. Troxel and R. A. Ulichney. Scaling binary images with the telescoping template. *PAMI*, 1982.
[33] Yifan Wang, Lijun Wang, Hongyu Wang, and Peihua Li. End-to-end image super-resolution via deep and shallow convolutional networks. *CoRR*, 2016.