GRADIENT IMAGE SUPER-RESOLUTION FOR LOW-RESOLUTION IMAGE RECOGNITION

Dewan Fahim Noor[1] *Yue Li*[1] *Zhu Li*[1] *Shuvra Bhattacharyya*[2] *George York*[3]

[1]University of Missouri-Kansas City, MO, USA
[2]University of Maryland, College Park, MD, USA
[3]US Air Force Academy, USA

ABSTRACT

In visual object recognition problems essential to surveillance and navigation problems in a variety of military and civilian use cases, low-resolution and low-quality images present great challenges to this problem. Recent advancements in deep learning based methods like EDSR/VDSR have boosted pixel domain image super-resolution (SR) performances significantly in terms of signal to noise ratio(SNR)/ mean square error(MSE) metrics of the super-resolved image. However, these pixel domain signal quality metrics may not directly correlate to the machine vision tasks like key points detection and object recognition. In this work, we develop a machine vision tasks-friendly super-resolution technique which enhances the gradient images and associated features from the low-resolution images that benefit the high level machine vision tasks. Here, a residual learning deep neural network based gradient image super-resolution solution is developed with scale space adaptive network depth, and simulation results demonstrate the performance gains in both gradient image quality as well as key points repeatability.

Index Terms— Image Super-resolution, Difference of Gaussian, Gradient Image, SIFT repeatability

1. INTRODUCTION

One of the huge challenges in image recognition is dealing with low-resolution images. Especially, in military and surveillance applications, recognition is done from low quality input images. However, if the image is captured from a further distance, the quality remains very low and unrecognizable which is actually a great concern in some sectors e.g. Department of Defense (DoD) while dealing with counter Unmanned Aircraft System (UAS) . One of the popular solutions in this case would be image super-resolution. Super-resolution [1] means finding a mapping from the low-resolution (LR) image to its high-resolution (HR) version. In the case of single frame super-resolution (SISR), the number of pixels for a single image is increased so that it can visually look better as well as can be efficacious while recognition. However, in addition to super-resolving the image, the key concern is to preserve the features so that it can be recognized accurately. Nowadays, image SR is driven by the emergence of deep learning methods. Recently, numerous deep learning based super-resolution methods have been introduced. In [2], SR-CNN method is established which is an end to end system between the input low-resolution images and its interpolated high-resolution images. The results exhibit quite a good gain over the other methods. In [3], VDSR method is established which generates a very deep convolutional neural network(CNN) with stages of small filters resulting in faster convergence and much gain in PSNR. In [4], the proposed enhanced deep learning based super-resolution (EDSR) method is further replicated in stages to finally produce the deep layers of super-resolution network being inspired from residual network.

In typical super-resolution methods, the goal is to improve peak signal to noise ratio(PSNR). But, from the practical point of view, these SR methods generate more eye-soothing high-quality image by increasing PSNR which eventually contribute towards losing key features. So, while identifying those images, we need to preserve the important local and global features e.g. recognizing captured low quality images from surveillance cameras using their features or identifying an aircraft using key feature points in Air Force. There are quite a few works on low-resolution image recognition. In [5], very low-resolution recognition (VLRR) problem has been dealt with deep learning based model for demonstrating the task with face recognition, font recognition, digit recognition. In [6], another deep CNN based method is proposed to deal with face and other objects with low quality. In many recognition tasks, gradient images are important information derived from pixel images. To define, gradient image generally refers to a change in the direction of the intensity or color of an image. Numerous works regarding image recognition have been done using gradient of images. In [7] and [8], Harris Detector and Laplacian of Gaussian are used to find out the features of edges and corners and blobs of an image respectively. In [9], SIFT feature detection is used which discovers local features after computing maxima and minima from the Difference of Gaussian(DoG) image set. In recognition, key points from an object are extracted to provide a description of the features which are used for recognizing the object. So, it should be important to keep in mind that extracted features should be able to be used in case of scale, noise and illumination changes. SIFT can handle these change making SIFT an ideal method for feature extraction.

There is few research regarding the preservation of features. In [10], a visual query compression for preserving local features is introduced. Here, they go through a new method in visual key points compression which uses subspaces for optimization of preserving key point feature matching properties than the reconstruction performance. Our proposed method in this paper is not an end to end system. Rather, it is a super-resolving network which generates SIFT repeatability. So the objective is to super-resolve the images in gradient domain so that it preserves SIFT features which will eventually contribute for better recognition. Our SR network is constructed

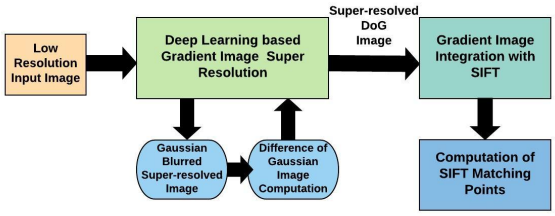


Fig. 1: Proposed Network Architecture

upon the concept of generating gradient images. The network actually consists of five SR networks. For each one, we establish deep learning method inspired from EDSR and Squeeze and Excitation Network [11] but instead of producing the super-resolved image of original input, we produce the Gaussian blurred images with different standard deviation to finally compute the DoG images. In SIFT, DoG images [12] are produced from the input image with different scale and different standard deviations. In our method, the network produces the Gaussian blurred images with similar standard deviations and compute the DoG images and integrate with SIFT method to find out the key points which are used for matching. Overall, our proposed method intends to generate super-resolved gradient images which preserve the SIFT features to produce SIFT repeatability.

2. PROPOSED METHOD

Our proposed method consists of a deep learning pipeline for image super-resolution. Our network is not an end- to end system. We wish to produce SIFT repeatability. So, we generate super-resolved Gaussian blurred image with different standard deviation instead of high-resolution image in pixel domain. The idea is to generate high-resolution gradient images from the Gaussian blurred image to finally integrate with SIFT to preserve SIFT matching points. Gradient images are generally constructed from the original image being convolved with a filter. Our image gradient method is based on the SIFT method. In SIFT method, from an input image, different Gaussian blurred images are first produced with different standard deviation. Then DoG image is computed for different scales which are called octaves. From DoG images, maxima and minima are computed to find key feature points. Let, *I(x,y)* is the original image; *G(x,y,σ)* is the Gaussian Kernel. Equation (1) and (2) show the formulation of Gaussian blurred images.

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{1}$$

$$L(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{2}$$

Where, *L(x,y,σ)* is the Gaussian blurred image with specific σ which is the standard deviation, x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis

So, the DOG will be as follows in equation (3) and (4):

$$D(x,y,\sigma_1,\sigma_2) = (G_1(x,y,\sigma_1) - G_2(x,y,\sigma_2)) * I(x,y)) \tag{3}$$

$$D(x,y,\sigma_1,\sigma_2) = L_1(x,y,\sigma_1) - L_2(x,y,\sigma_2) \tag{4}$$

Where, *D(x,y,σ1,σ2)* is the of DoG image, $\sigma_1$is the standard deviation of the first blurred image and $\sigma_2$is the standard deviation of the second blurred image. G_1,G_2 are Gaussian filters. L_1,L_2 are Gaussian blurred images.

The loss function E is the MSE loss between the DoG of the super-resolved blurred generated image and the DoG from convolution with original image which can be shown in (5):

$$E(\hat{D}, D_{original}) = \sum_{i=1}^{n}\sum_{j=1}^{m} (\hat{D}^{ij} - D^{ij}_{original})^2 \tag{5}$$

Where $\hat{D}$ is the predicted DoG image which is upscaled and $D_{original}$ is the DoG image computed from of the original one convolved with Gaussian filter. n and m are the numbers of pixels in x and y direction.

The gradient descent of the loss function is the differentiation with respect to $\hat{D}$ as followed in equation (6),(7) and (8):

$$\frac{\delta E}{\delta \hat{D}} = \frac{\delta(\sum_{i=1}^{n}\sum_{j=1}^{m} (\hat{D}^{ij} - D^{ij}_{original})^2)}{\delta \hat{D}} \tag{6}$$

$$\begin{aligned}\frac{\delta E}{\delta \hat{D}} =& 2\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{D}^{ij} - (\frac{1}{2\pi{\sigma_1}^2}P - \frac{1}{2\pi{\sigma_2}^2}Q)) \\ & (1 - (\frac{1}{2\pi{\sigma_1}^2}\frac{\delta P}{\delta \hat{D}} - \frac{1}{2\pi{\sigma_2}^2}\frac{\delta Q}{\delta \hat{D}}))\end{aligned} \tag{7}$$

$$P = e^{\frac{-(x_i^2+y_j^2)}{2{\sigma_1}^2}} * I(x_i,y_j), Q = e^{\frac{-(x_i^2+y_j^2)}{2{\sigma_2}^2}} * I(x_i,y_j) \tag{8}$$

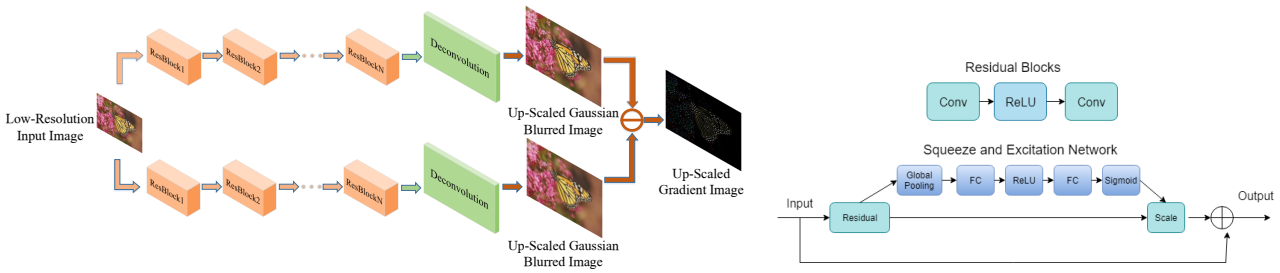
Here, equation 7 is derived from equation 6 after differentiating it with respect to $\hat{D}$. In equation 7, due to the complexity of the equation we introduce two terms *P* and *Q* [shown in equation 8]which are the exponential terms for the Gaussian filter in each image convolved with the original image $I(x_i, y_j)$ where x_i is the distance from the origin in the horizontal axis, y_j is the distance from the origin in the vertical axis.

As the loss function and its gradient descent seem to be very complex, it can be simplified if we use the MSE loss between Gaussian blurred images as our loss function and then we compute the DoG images from the Gaussian blurred image. The following equation (9) is the simplified loss function

$$E(\hat{L}, L_{original}) = \sum_{i=1}^{n}\sum_{j=1}^{m} (\hat{L}^{ij} - L^{ij}_{original})^2 \tag{9}$$

Where $\hat{L}$ is the predicted blurred image which is upscaled and $L_{original}$ is the Gaussian blurred image of the original image with same standard deviation.

There are different stages in our proposed method. From the low-resolution input images, the deep learning based Gradient Image super-resolution stage creates super-resolved Gaussian blurred images which in turns produces the DoG images. The SIFT integration stage integrates the DoG images for show-casting SIFT repeatability. Figure 1 shows the full network architecture of our proposed method. For the super-resolution network design, the residual blocks concept is taken from EDSR. Each of the five networks contains several ResBlocks followed by deconvolution layers as in Figure 2a. Each ResBlock contains a residual block which is followed by a Squeeze and Excitation network unlike EDSR. Residual blocks have a convolutional layer followed by rectified linear unit(ReLU)



(a) Deep Learning Gradient Image Super resolving network

(b) Residual Blocks with Squeeze and Excitation Network

Fig. 2: Deep Learning Gradient Image Super resolving network to compute upscaled Gradient Image

and again a convolutional layer. Each convolutional layer has filter kernel size of 3X3 with 64 number of features. In the Squeeze and Excitation network, the output from residual block is followed by a global pooling layer, fully connected layer, ReLU, a fully connected layer again and a sigmoid function followed by the scaling. The input to the residual block is added to the output of Squeeze and Excitation network for the residual learning. The Squeeze and Excitation network improves channel-wise feature responses by modelling the relationships between channels [11] as shown in Figure 2b which works as a boosting factor in our method. The deconvolutional layer [13] does the upscaling of the image. Here, stride value 2 or 4 is used for either 2X or 4X upscaling.

However, the number of ResBlocks is not fixed. We design an adaptive solution to the number of Resblocks. As we have 5 separate SR networks for generating 5 different Gaussian blurred images with different standard deviation value(σ), we adapt the number of blocks according to the sigma value unlike EDSR. For higher σ value the number of ResBlocks is reduced. We chose the σ values of 1.249, 1.545, 1.946588, 2.452527 and 3.090016 in accordance with the design of SIFT. After trial and error, we optimized the number of Resblocks as 16,12,10,8,6 respectively for lower to higher σ values. The depth of layers has been reduced as we increase the σ.

After the Gaussian blurred images are produced from the SR networks having MSE loss in DoG domain, DoG images [12] are computed simply from the subtraction between the images. It is to be noted that in our network, the ground truth is the Gaussian blurred image of the original high-resolution input whereas the input is the downsampled version of the original high-resolution image. There are five different Gaussian blurred images with different standard deviations. From each pair of blurred image, a DoG image is computed. From five blurred image, four DoG images are computed. So, finally our network gives four DoG images as the outputs.

Integration with SIFT: Once we generate the four DoG images computed from the five generated high- resolution blurred images from the network, we integrate it to the SIFT network [14]. Instead of calculating DoG images by SIFT itself which is done in SIFT method, we directly load our DoG images into the SIFT network. So, the SIFT network will find key points from our produced DoG images. The purpose of integrating with SIFT is that SIFT itself computes DoGs in different scale to find out the maxima and minima in DoG images for identifying key points. As our network already produces super-resolved DoGs, the integration will enhance the number of the SIFT matching feature points.

3. EXPERIMENTAL SETUP AND DATASET

For training, we used the CVPR DIV 2K dataset [15] with 800 training images. We first downscaled the images by both 2 and 4 times. The input images are then cropped to 32X32 patch size. The training process is conducted in Python with PyTorch [16] deep learning tool. For testing, we used the MPEG Compact Descriptors for Visual Search (CDVS) dataset [17]. CDVS is a comprehensive collection of images of various objects which consists of 186k labeled images of CDs and book covers, paintings, video frames, buildings and common objects. We experimented on all the categories of the dataset separately and chose 200 matching image pairs from each one. We used five pretrained models for generating the upscaled blurred images. The DoG images are then computed and integrated to SIFT. As it is not an end to end network, rather a network to produce SIFT repeatability, we simply did not compute any recognition scheme; rather we show the number of SIFT matching points.

4. RESULTS

For the evaluation of the performance, we basically compare our result with bi-cubic interpolation and original EDSR that generate upscaled image unlike our blurred version. We categorize the CDVS dataset into buildings,graphics(books, cards, CDs, DVDs, print), objects, videos and paintings. We collected 200 matching image pairs from each category and evaluated the performance. From the generated Gaussian blurred image, we computed the DoG images and compared the PSNR with DoG images produced from EDSR and bi-cubic interpolated images for both 2X and 4X upscaling.

Table 1: PSNR(in dB) comparison of DoG Images for 2X upscaling for CDVS full dataset.

DoG (σ_1,σ_2)	Proposed Method	EDSR	Bi-cubic
σ_1=1.24, σ_2 =1.54	33.30	31.24	30.2
σ_1=1.54, σ_2=1.94	37.60	35.58	34.75
σ_1=1.94, σ_2=2.45	44.75	42.48	41.5
σ_1=2.45, σ_2=3.09	48.38	46.12	45.55

Table 1 and Table 2 show the result for PSNR in dB for four DoG images generated from difference of Gaussian blurred images, blurred at σ_1 and σ_2 using our proposed method, DoG images generated from EDSR images convolved with Gaussian filters and DoG images generated from bi-cubic interpolated images convolved with Gaussian filters for 2X and 4X upscaling. It is crystal clear that DoG

Table 2: PSNR(in dB) comparison of DoG Images for 4X upscaling for CDVS full dataset.

DoG (σ_1,σ_2)	Proposed Method	EDSR	Bi-cubic
σ_1=1.24, σ_2=1.54	31.20	29.15	28.65
σ_1=1.54, σ_2=1.94)	35.68	33.53	33.05
σ_1=1.94, σ_2=2.45	40.6	38.15	37.68
σ_1=2.45, σ_2=3.09	45.9	43.60	43.15

images from our SR network have acquired around 2 -2.3 dB gain for 2X and 2-2.4 dB gain for 4X upscaling over the DoG images generated from original EDSR convolved with Gaussian filter and 2.7-3 dB gain for 2X and 2.5-2.9 dB gain for 4X upscaling over bi-cubic interpolation.

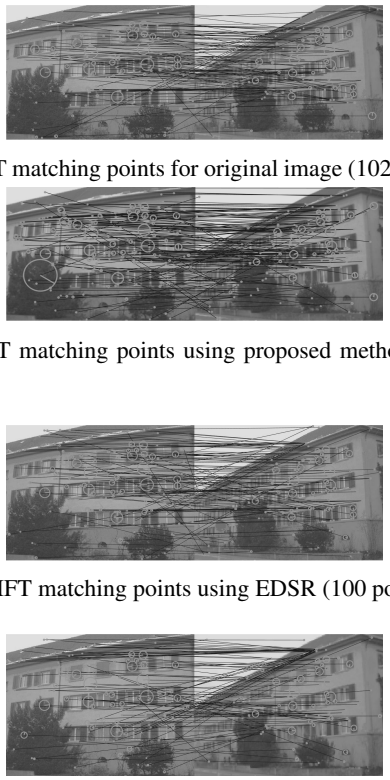
Table 3: Average number of SIFT matching points for 200 matching image pairs from each category of the CDVS dataset.

Category	Factor	Original	Proposed	EDSR	Bi-cubic
Buildings	2X	125.8	130.4	116.3	112.4
Buildings	4X	125.8	115.4	105.6	100.4
Graphics	2X	101.6	102.8	94.5	92.8
Graphics	4X	101.6	90.4	86.7	85.4
Objects	2X	115.3	118.5	106.9	102.6
Objects	4X	115.3	108.8	99.1	96.2
Paintings	2X	114.4	120.5	105.9	100.7
Paintings	4X	114.4	109.8	101.5	96.1
Video	2X	94.3	94.4	87.2	85.2
Video	4X	94.3	85.5	80.1	79.2

As we generate DoG images, we integrate them into SIFT to generate SIFT repeatability. In Table 3, the result is shown for five different categories for the average number of SIFT matching points from generated super-resolved images using our proposed method , EDSR method and bi-cubic interpolation methods and also the original images which are already high-resolution images for 200 matching image pairs. For 2X upscaling, our method is having a gain of around 7-14 points over EDSR and 9-18 points over bi-cubic interpolation. For 4X upscaling, the gain is around 4-10 points over EDSR and 5-15 points over bi-cubic interpolation. The best result is achieved in the buildings category with 10-14 points and 15-18 points gain over EDSR and bi-cubic respectively. The worst result is achieved in the graphics and video categories with a gain of around 5-10 points gain. The goal of our proposed method is to produce SIFT repeatability rather than constructing an end to end system for full recognition. The SIFT repeatability bears the testimony that the produced images have more matching feature points which contribute for recognition.

In comparison with the original image, our proposed method achieves approximately even 0.1 to 4 more matching points than the original image for 2X upscaling factor as shown in Table 3. The reason is that while super-resolving from lower resolution image, the Gaussian blurred image stored the information of the features more rigorously. So after computing the DoG, SIFT feature extraction method finds more maxima and minima while discovering key points.

Figure 3 shows the images of SIFT matching points for image, image generated using our method,original EDSR and bi-cubic interpolation for 2X upscaling. It is seen that our proposed method shows more gain over the other methods.



(a) SIFT matching points for original image (102 points)

(b) SIFT matching points using proposed method (112 points)

(c) SIFT matching points using EDSR (100 points)

(d) SIFT matching points using bi-cubic interpolation(96 points)

Fig. 3: SIFT Matching Points Comparison

5. CONCLUSION

Low-resolution images present great challenges to a variety of visual recognition problems in real world navigation and surveillance applications. In this work, we developed a novel gradient image super-resolution solution that opens up more degree of freedom (DoF) in the SR network design by allowing scale space adaptation in both network architecture and depth. Simulation results demonstrated that the SR performance in both gradient image quality and subsequent machine vision tasks like key point repeatability are improved compared with the state of art solutions in pixel domain super-resolution. In the future, we will develop task-specific deep neural network integration with triplet loss and softmax loss networks to drive better task level performances.

6. ACKNOWLEDGEMENT

This research was supported in part by the U.S. Air Force Office of Scientific Research under the Dynamic Data Driven Applications Systems (DDDAS) Program.

7. REFERENCES

[1] W. Siu and K. Hung, "Review of image interpolation and super-resolution," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, Dec 2012, pp. 1–10.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb 2016.

[3] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 1646–1654.

[4] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1132–1140.

[5] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4792–4800.

[6] Sivaram Prasad Mudunuri, Soubhik Sanyal, and Soma Biswas, "Genlr-net: Deep framework for very low resolution face and object recognition with generalization to unseen categories," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[7] Z. Ye, Y. Pei, and J. Shi, "An adaptive algorithm for harris corner detection," in *2009 International Conference on Computational Intelligence and Software Engineering*, Dec 2009, pp. 1–4.

[8] H. Kong, H. C. Akakin, and S. E. Sarma, "A generalized laplacian of gaussian filter for blob detection and its applications," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1719–1733, Dec 2013.

[9] Cong Geng and X. Jiang, "Sift features for face recognition," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, Aug 2009, pp. 598–602.

[10] Z. Zhang, L. Li, Z. Li, and H. Li, "Visual query compression with locality preserving projection on grassmann manifold," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sept 2017, pp. 3026–3030.

[11] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," 2018.

[12] Y. Lv, G. Jiang, M. Yu, H. Xu, F. Shao, and S. Liu, "Difference of gaussian statistical features based blind image quality assessment: A deep learning approach," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sept 2015, pp. 2344–2348.

[13] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1520–1528.

[14] Xing Di, *SIFT Feature Extraction. Online Available:https://www.mathworks.com/matlabcentral/fileexchange/50319-sift-feature-extreaction*, 2015.

[15] Eirikur Agustsson and Radu Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[16] PyTorch, *PyTorch Documentation. Online available: https://pytorch.org/docs/stable/index.html*, 2016.

[17] S. S. Tsai N. M. Cheung H. Chen G. Takacs Y. Reznik R. Vedantham R. Grzeszczuk J. Bach V. Chandrasekhar, D. Chen and B. Girod, "The stanford mobile visual search dataset," in *Proceedings of ACM Multimedia Systems Conference (MMSys), San Jose, California*, February 2011.