A Novel Multi-Frame Color Images Super-Resolution Framework based on Deep Convolutional Neural Network

Zhe Li, Shu Li, Jianmin Wang and Hongyang Wang

Department of Electronics Science and Technology, Harbin University of Science and Technology, Harbin, Heilongjiang, 150080, China

histarfish@sina.com

Keywords: Super-resolution, Deep Convolutional Neural Networks, Multi-frame Color Images.

Abstract. With the extensive application of machine learning. Deep convolution neural network (DCNN) learning method is developed on the basis of a multi-layer neural network for image classification and identification of specially designed. It has been improved and applied for single image super-resolution problem and demonstrated state-of-the-art quality. In this paper, we presents a novel framework based on deep convolutional neural network to realize the multi-frame color images super-resolution. The system contains two parts, multi-frame Image pixel processing and structure design of DCNN. The prior information could be utilized during the image pixel processing. Experimental results prove its effectiveness and confirm out framework can be effectively applied to multi-frame color images super-resolution. The generated super-resolution image achieves a better restoration image quality compared to state-of-the-art methods.

Introduction

In many military, medical and civilian applications, high-resolution (HR) images are desirable and required. Image super-resolution (SR) is to overcome the resolution limitation of sensor and restore a high resolution images from single or multiple low resolution (LR) images. Super-resolution can be used in many areas like Medical imaging, Satellite imaging, Remote imaging, video surveillance. The key issue is to solve the inherent ill-possed problem.

Various techniques have been proposed in multi-frame images super resolution to handle this problem. The reconstruction based method is a kind of common methods. They are classified into two major parts: frequency domain algorithms and spatial domain algorithms. Foremost research in frequency domain algorithms by Tsai and Huang [1]. These methods are simple and computationally cheap. They are extremely sensitive to the image noise, limiting their use in the spatially invariant noise model. For spatial domain algorithms, seems to be the most widely and popular method in recent years which can perform directly on pixel. The representative research are as follows. The non-uniform interpolation-based methods: these methods have low computational cost. However, degradation models are not applicable in these methods if the blur and the noise are different for LR images. POCS (Projections onto convex set) method have advantage of simplicity, but they are non-uniqueness of solution. Besides, low convergence rate limit the speed of the iteration, and the computational load is heavy. IBP-based methods can restore HR image in a straightforward way and it can be used in complex motion models. The disadvantage is no unique solution in IBP methods. MAP [2] (Maximum a posterior) methods are kind of effective robust statistical methods. They are used for complex SR which the scenes contain multiple independently moving objects. Regularization-based methods [3] can solve the ill-posed regularization problem with prior information by the Bayesian approach. Spatial domain algorithms can lead to better SR reconstruction results than the frequency domain algorithms.

Learning-based SR algorithms [4,5] are popular in single image super-resolution. They suppose the LR image is lost high frequency details from HR image. The lost details are estimated by learning the training dataset contains LR and HR images. Example-based method: Yang [6] proposed a sparse-coding-based method. Dong [7,8] proposed a deep convolutional neural network for image

super-resolution. This method learned end-to-end mapping from LR to HR images patches pairs, and achieves good restoration quality.

This paper is organized as follows. In section II the CNN background is described. Mathematical backgrounds are given in order to define how the CNN can be used for resolution. In section III. The CNN based multi-frame image super-resolution framework is proposed formally. Also, the methods of multi-frame images pixel processing and CNN super-resolution process are introduced. Before concluding, experiments are performed in order to discuss the quality and speed of the framework we proposed and compared with other state-of-the-art methods. Dataset used comes from Set5 (5 images) and Set 14(14 images).The conclusions and future works are discussed in section V.

Backgrounds of CNN

Convolutional Neural Network (CNN) is a feed-forward back-propagation multilayer perceptron model. CNN works on a supervised back-propagation learning technique during the training phase of the network. CNN generally consists of input layer, convolutional layer, hidden layer and output layer. Each layer has multiple numbers of neurons, which vary according to the complexity of the network. The learning process takes place in the perceptron by altering the weight factors after each training epoch. Weight factors are adjusted accordingly by calculating the mean error of the expected result in contrary with the output result. The aim of training a neural network is to search for a set of weight factors which links the provided input with the expected output. The process of optimizing the number of hidden layers and amount of neurons used in each layer greatly affects the performance of the entire network.

Methods

Super-resolution system

In order to restore a HR image from multi-frame color images. An effective super-resolution framework based on deep convolutional neural network is proposed. The framework is shown in Fig.1. In the process of image pixel processing, each LR image is simply a linear combination of the target (TR) image pixel values. The TR image is generated by the process of the cost function we will discuss in next section. In the process of CNN super-resolution, the training phase is added to the super-resolution. We learn the end-to-end mapping between the HR and TR image patches pairs in the training process.

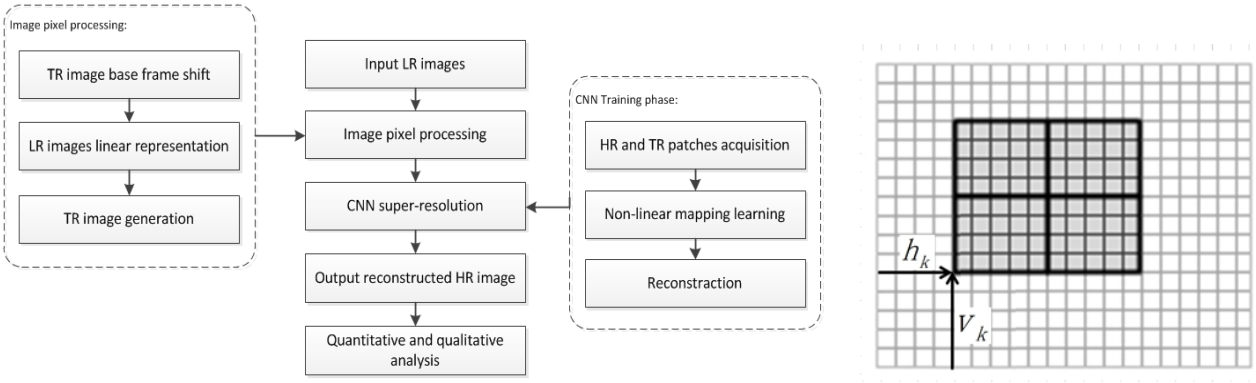

Fig. 1. Basic structure of the color images SR framework. **Fig. 2.** Pixel frame in the TR image frame

Multi-frame images pixel processing

In this work, we fused multiple color images to one target color image which contains all the measured frames $x_k(n1, n2)$. First, let us denote the multiple low resolution (LR) images by

$x_k(n1,n2)\ k=1,\mathbf{L},p$, and the target image by $t(n1,n2)$. The size of the LR image is $S1=N_1\times N_2$, and size of the target image is $S2=LN_1\times LN_2$, L is upscale factor and it is a positive integer. Here, the target image can be viewed as an underling reference pixel grid (marked by gray borders), and the physical pixels acquired from single LR image marked by bold borders (see Fig.2). This frame is shifted a distance of h_k and v_k in horizontal and vertical directions respectively.

Then we put all the captured frame in a single vector $X_k=[x_{k,1},x_{k,2},\mathbf{L},x_{k,S1}]$, and put the target image pixels in a vector $T=[t_1,t_2,\mathbf{L},t_{S2}]$. Here the LR image pixel value can be determined by a linear combination of target image pixels and single LR image frame shift h_k and v_k. Mathematical formulation as follows:

$$\hat{X}_{k,m}=\sum_{n=1,\mathbf{L},S2}W_{k,m,n}(h_k,v_k)\cdot t_n+h_k \tag{1}$$

where $m=1,\mathbf{L},S1$, $\hat{X}_{k,m}$ donates the calculated LR pixel value, $W_{k,m,n}$ is a weighting coefficient under different parameters which shows the contribution of target pixels to LR frame pixels. $W_{k,m,n}$ can be approximated by a Gaussian distribution over the target image pixels. h_k represents additive noise samples and it is assumed to be independent and identically distributed (i.i.d). The target image can be produced by the following function:

$$C(T)=\frac{1}{3}\sum_{\substack{k=1,\mathbf{L},p\\ m=1,\mathbf{L},S1}}W_{k,m,n}(X_{k,m}-\hat{X}_{k,m})^2+\frac{a}{3}(Y_{fil}^T\bullet Y_{fil})+\frac{b}{3}g(X_{k,m}) \tag{2}$$

where the function contains three terms, the first term calculates squared error between the LR pixels and target image pixels. Minimize this term can optimize the target image. The second term restores the frequency components of the fused process. Y_{fil} is a high-pass filter, α is a weight to the high frequency filtration. Here, we make a = 1 and a Laplacian kernel for high-pass filtration process. The last term is a smoothing prior term, a and b can balance the proportion of the high-pass filtration process prior and the smoothing prior. From the analysis, large α and small weight b can get smoother but more blurred images. On the contrary, the restored image has better details because of the prior learned increased.

DCNN structure design

In recent learning-based image super-resolution algorithm research, feature choose and feature representation is an outstanding issues. Different features have been extracted in many algorithms such as raw image data, high-frequency information and image primitives. In convolutional neural networks, the end to end mapping must be well trained with sufficient number of training samples. As in Fig.3, the first three layers of DCNN is used to restore HR which has the same structure as Dong's CNN network. Our operation is to recover HR image from the target image. In the learning of mapping F .The image patches extracted from target image are represented as high-dimensional vector. These vectors comprise a set of feature maps which are convolved by a set of filters.

$$Y_i=W*T_i+B_i \tag{3}$$

where i is the ith convolutional layer, W is the filters , Bi represents the ith convolutional layer bias. The first two convolutional layers are followed by an rectified linear units (RELU). Thus the patches become:

$$Y_i=\max(0,\ W*T_i+B_i) \tag{4}$$

We use the Mean Squared Error (MSE) between the ground truth images and the HR images as the reconstruction loss function. We use stochastic gradient descent with standard back propagation minimizing the loss function. Fig.3 shows the structure design of the framework we proposed.

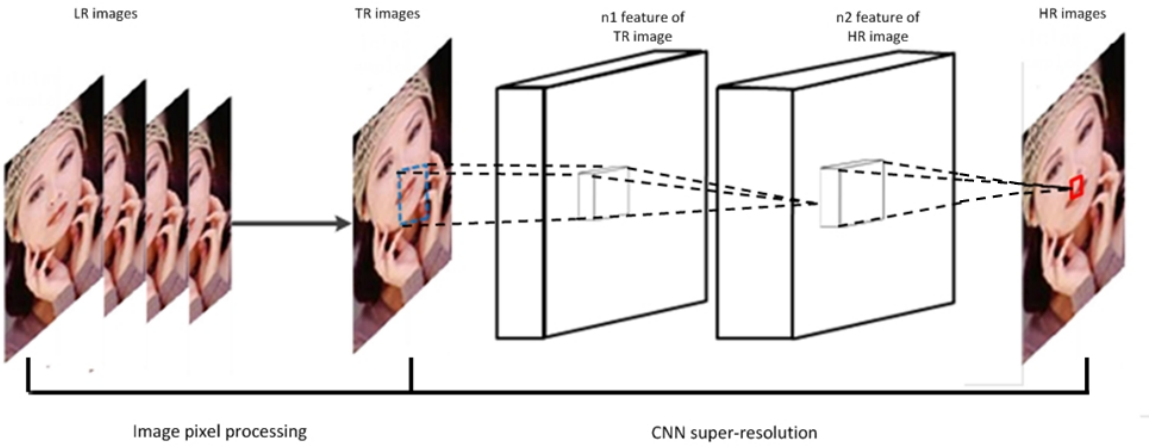

Fig. 3. Structure design of the framework

Experiments and analysis

Data set and parameters design

For a fair comparison of these state-of-the-art quality methods such as SRCNN, K-SVD [9], ANR [10], NE+LLE [11] etc. We use the same training set (91 images) and test sets. The Set5 (5 images) and Set 14 (14 images) are used to evaluate the SR results. Besides, the size of the three convolutional layers are $f1\times f1\times n1$, $n1\times1\times1\times n2$ and $n2\times f3\times f3\times1$. We set three convolutional layers parameters to be $f1 = 9$, $f2 = 1$, $f3 = 5$, $n1 = 64$, $n2 = 32$. In the training part, we use the 91 images dataset as our training set. They can be decomposed into 24800 sub-images which are extracted from raw images with the same stride of 14 as Dong's work. Here, we first transform the color images into the $YCbCr$ space. According to the CNN training experience, the training strategy in the framework we proposed are only applied on the Y channel, and Cb , Cr channels are upscaled by bicubic interpolation. We use the Caffe training package to implement the training process. As the framework we designed above. The testing multi-frame colorful images are obtained from Set5 and Set14 by a manual image degradation process. We use coordinate transformation and interpolation methods to rotate different angles of the image, and add low-pass filter and downsampling to acquire a set of LR images to verify the methods we proposed above.

Quality evaluation

We use the peak signal-to-noise ratio (PSNR) and the visual effect of these images constitute the criteria. The formulates are defined as follow:

$$MSE(A,B)=\frac{1}{M}\sum_{i=1}^{M}(a_i-b_i)^2 \tag{5}$$

where image A is the restored HR image $A=a_1,a_2,\mathbf{L},a_M$, and image B is the ground truth image $B=b_1,b_2,\mathbf{L},b_M$. M is the number of pixels, a and b are pixels of their images. PSNR (peak-signal-to-noise ratio) is calculated from all three color channels in luminance, contrast and structure of an image. Luminance is the average pixel intensity. The contrast is the variance between the reference and distorted image, while structure is obtained by calculating the cross correlation of the two images.

$$PSNR(A,B)=10\log_{10}(\frac{MAX^2}{MSE(A,B)}) \tag{6}$$

Results and Discussion

We compared our result of PSNR (dB) and running time(s) on Set5 and Set14 with other methods. Table 1 shows the details of the each comparison result of Set5 with the factor 3. Here, we only show the Set5 detail result and give the average result of Set 14 in Table 2. We got the average PSNR 32.89 (Set5) and 30.63(Set14) in our method. In addition to quantitative evaluation, we also present some qualitative results in Fig.4. The results demonstrate our SR method can lead to a good quality of the restored HR image.

Fig. 4. Qualitative comparison among original and resolved results by Bicubic, NE+LLE, SRCNN and PAPER, respectively

Table 1. The results of PSNR (dB) and running time(s) on Set5 with the factor 3

Set5	Bicubic		K-SVD		ANR		SRCNN		NE+LLE		PAPER	
	PSNR	TIME	PSNR	TIME	PSNR	TIME	PSNR	TIME	PSNR	TIME	PSNR	TIME
baby	33.86	-	35.50	105.6	36.58	24.31	36.32	13.3	36.30	135.7	36.54	15.14
bird	31.97	-	33.87	41.24	34.10	16.01	35.69	5.66	34.92	62.30	33.58	5.73
butterfly	25.67	-	25.90	34.02	29.12	14.31	29.14	4.86	27.88	52.73	29.67	4.86
head	30.38	-	32.44	40.26	32.90	14.50	32.94	5.03	31.54	55.40	32.63	5.04
woman	28.45	-	28.71	40.60	30.84	14.69	30.06	5.05	30.06	56.81	32.03	5.05
average	30.06	-	31.28	52.34	32.70	16.76	32.83	6.78	32.14	72.58	**32.89**	**7.14**

Table 2. Average results of PSNR (dB) and running time(s) on Set14 with the factor3

Set14	Bicubic		K-SVD		ANR		SRCNN		NE+LLE		PAPER	
	PSNR	TIME	PSNR	TIME	PSNR	TIME	PSNR	TIME	PSNR	TIME	PSNR	TIME
average	28.06	-	29.28	58.13	29.70	16.60	30.47	7.98	30.02	79.93	**30.63**	**8.84**

Conclusions

To sum it all, we have presented a multi-frame colorful images super-resolution method using a deeply-recursive convolutional network. We got the average PSNR 32.89 (Set5) and 30.63(Set14) in our method. The experimental result demonstrates that our CNN framework can be modeled to perform image super resolution, and it outperforms other super resolution methods. Besides, our framework can be applied to nature images of different kinds. Future work involves framework optimization and speed. More CNN training methods and prior acquisition optimization methods will be tried to apply in our framework to improve the quality of restoration. In addition, the running speed of the SR process is an important problem to be solved. We will improve the framework to be expanded to tackle real-time multi-frame images SR reconstruction problems in future.

Acknowledgements

This work was supported by Natural Science Foundation of Heilongjiang Province of China (E201446, F201113).

References

[1] H. J. L., “Diffraction and resolving power,” J.Opt. Society of America A, vol. 54, no. 7, pp. 931-933, 1964.

[2] B. H. H. Shen, L. Zhang and P. Li, “A map approach for joint motion estimation segmentation and super resolution,” IEEE Transactions on Image Processing, vol. 16, no. 2, pp. 479-490, 2007.

[3] X. Z. Yang, “A robust multi-frame super-resolution algorithm based on half quadratic estimation with modified btv regularization,” Digital Signal Processing, vol. 23, no. 1, pp. 98-109, 2013.

[4] S. M. Z. J. Y. Qiu, “Multi-frame super-resolution reconstruction based on self-learning method,” Mathematical Problems in Engineering, vol. 2015, no. 17, pp. 1-12, 2015.

[5] T. G. K. S. T. Sakurai, “Learning-based super-resolution image reconstruction on multi-core processor,” IEEE Transactions on Consumer Electronics, vol. 58, no. 3, pp. 941-946, 2012.

[6] H. T. Yang J, Wright J and Y. Ma, “Image super-resolution via sparse representation,” IEEE Transaction on Image Processing, vol. 19, no. 11, pp. 2861-2873, 2010.

[7] C. L. C. Dong and K. He, “Learning a deep convolutional network for image super-resolution,” European Conference on Computer Vision, vol. 8692, pp. 184-199, 2014.

[8] K. H. C. Dong, C.C. Loy and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE transactions on instrumentation and measurement, vol. 38, no. 2, pp. 295-307, 2016.

[9] M. E. R. Zeyde and M. Protte, “A technique for estimating the state of health of lithium batteries through a dual-sliding-mode observer,” International Conference on Curves and Surfaces, vol. 6920, pp. 711-730, 2010.

[10] V. D. R. Timofte and L. V. Gool, “Anchored neighborhood regression for fast example-based super-resolution,” Computer Vision, pp. 1920-1927, 2013.

[11] D. Y. H. Chang and Y. Xiong, “Super-resolution through neighbor embedding.” IEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 275-282, 2010.