VIDEO RESTORATION USING CONVOLUTION NEURAL NETWORKS

[1]Sushant Deshmukh, [2]Rajesh Patil

[1]M.Tech (Electronics and Telecommunication Engineering), [2]Associate Professor
[1,2]Department of Electrical Engineering,
[1,2]Veermata Jijabai Technological Institute, Mumbai, India

Abstract*—— Neural networks have shown very promising results in a large number of research areas. With the introduction of convolution neural networks, they have been widely used in image processing. In this paper we implement Convolution Neural Network for video restoration. This is achieved by introducing higher frequency details using pre trained networks. Most of the research aims at improving video quality by increasing PSNR, but sometimes due to this the videos may become aesthetically less satisfying. While large image databases are available to train deep neural networks, it is more challenging to create a large video database of sufficient quality to train neural nets for video restoration. The dataset used for training the model is from DIV2K - bicubic downscaling x4 competition.Video restoration remains a challenging problem despite being a very active area of research. Even with huge strides made with single-image super-resolution, multi-frame techniques, which utilize multiple frames in improving the quality of a given frame, we have yet to fully take advantage of the power of deep learning.

***IndexTerms*— Neural Networks, Video Restoration**

I. Introduction

Image spatial resolution refers to the capability of the sensor to observe or measure the smallest object, which depends upon the pixel size. As two-dimensional signal records, digital images with a higher resolution are always desirable in most applications. Imaging techniques have been rapidly developed in the last decades, and the resolution has reached a new level. The question is therefore: are resolution enhancement techniques still required?

The application of restoration techniques has entered our daily life. Low resolution video images can be converted to high-definition images using restoration techniques. Hitachi Ltd. achieved the conversion of standard definition TV (SDTV) to high-definition television (HDTV) using restoration technology for videos, which makes it a particularly interesting and relevant research topic The fact is, although the high-definition displays in recent years have reached a new level (e.g., 1080*1920 for HDTV, 3840*2160 for some ultra HDTV, and 2048*1536 for some mobile devices), the need for resolution enhancement cannot be ignored in many applications. For instance, to guarantee the long-term stable operation of the recording devices, as well as the appropriate frame rate for dynamic scenes, digital surveillance products tend to sacrifice resolution to some degree. A similar situation exists in the remote sensing field: there is always a tradeoff between the spatial, spectral, and temporal resolutions. As for medical imaging, within each imaging modality, specific physical laws are in control, defining the meaning of noise and the sensitivity of the imaging process. How to extract 3D models of the human structure with high-resolution images while reducing the level of radiation still remains a challenge.

Based on these facts, the current techniques cannot yet satisfy the demands. Resolution enhancement is therefore still necessary, especially in fields such as video surveillance, medical diagnosis, and remote sensing applications. Considering the high cost and the limitations of resolution enhancement through "hardware" techniques, especially for large-scale imaging devices, signal processing methods, have become a potential way to obtain high-resolution (HR) images.

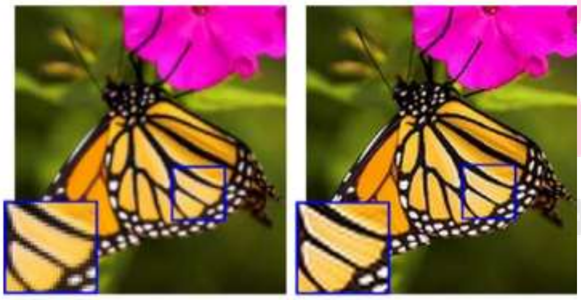

Fig.1 Restored image (left) and original image (right)

II. Literature survey

Most algorithms can be divided into two categories, model-based and learning-based algorithms. Model-based approaches [1]–[5] model the Low Resolution (LR) image as a blurred, subsampled version of the High Resolution (HR) image with additive noise. The reconstruction of the HR image from the LR image is an ill-posed problem and therefore needs to be regularized. In a Bayesian framework, priors controlling the smoothness or the total variation of the image are introduced in order to obtain the reconstructed HR image. For example, Babacan et al. [1] utilize the Bayesian framework to reconstruct an HR image from multiple LR observations, subject to rotation and translation amongst them. Belekos et al. [2] and later Liu and Sun [3] also use the Bayesian framework to derive an algorithm that is able to deal with complex motion and real world video sequences. With all these algorithms, the motion field and the HR reconstructed image, along with additionally required model parameters are estimated simultaneously from the observed data. Ma et al. [5] presented an algorithm that extended the same idea to handle motion blur..

III. ARCHITECTURE

Images and videos come in various resolutions hence more suitable approach is to design an architecture which can scale for the desired resolution. However designing a flexible architecture is a very complicated process hence as suggested by [6] we scale the input to desired size before feeding it to the input. Bicubic interpolation is used for scaling. Keeping the above restrictions in mind we use architecture which has only convolution layer as suggested by [6]. It has only convolutional layers which have the advantage that the input images can be of any size and the algorithm is not patch-based. The setup is shown in Figure 2. In it Y represents the input LR image and X the output HR image. It consists of three convolutional layers, where the two hidden layers H1 and H2 are followed by a Rectified Linear Unit (ReLU) . The first convolutional layer consists of 1 × f 1 × f 1 × C 1 filter coefficients, where f 1 × f 1 is the kernel size and C 1 the number of kernels in the first layer. We use this notation to indicate that the first dimension is defined by the number of input images, which is 1 for the image SR case. The filter dimensions of the second and third layers are C 1 × f 2 × f 2 × C 2 and C 2 × f 3 × f 3 × 1, respectively. The last layer can only have one kernel in order to obtain an image as output. Otherwise an additional layer with one kernel otherwise a post processing or aggregation step is required. The input image Y is bicubically upsampled so that the input (LR) and output (HR) images have the same resolution. This is necessary because upsampling with standard convolutional layers is not possible. Typical image classification architecture often contains pooling and normalization layers, which help to create compressed layer outputs that are invariant to small shifts and distortions of the input image. In this task, we are interested in creating more image details rather than compressing them. Hence the introduction of pooling and normalization layers would be counterproductive.

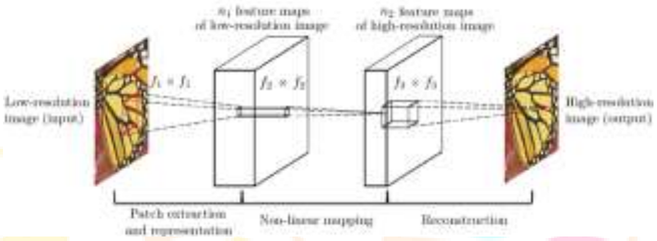
n_1 feature maps of low-resolution image
n_2 feature maps of high-resolution image
$f_1 \times f_1$
$f_2 \times f_2$
$f_3 \times f_3$
Low-resolution image (input)
High-resolution image (output)
Patch extraction and representation
Non-linear mapping
Reconstruction

Fig.2 The architecture of SRCNN as presented in [6].)

In addition, we performed L2 regularization for each weight matrix W, which limits the size of the weights at each layer by adding a term to the loss equal to some hyper parameter lambda times the sum of the squares of each weight in the weight matrix.

IV. Initialization and weight update

A. Xavier Initialization

The main reason for using Xavier Initializations is that it allows for signal to reach deep into the network. If the weights in a network start too small, then the signal shrinks as it passes through each layer until it's too tiny to be useful. If the weights in a network start too large, then the signal grows as it passes through each layer until it's too massive to be useful. Xavier initialization makes sure the weights are just right, keeping the signal in a reasonable range of values through many layers. It initializes the weights in the network by drawing them from a distribution with zero mean and a specific variance, the distribution used is typically Gaussian or uniform.

B. Adam Update

Adam is an optimization algorithm that can be used instead of the classical stochastic gradient descent procedure to update network weights iterative based in training data. Adaptive Moment Estimation is another method that computes adaptive learning rates for each parameter. In addition to storing an exponentially decaying average of past squared gradients like Adadelta and RMSprop, Adam also keeps an exponentially decaying average of past gradients similar to momentum.

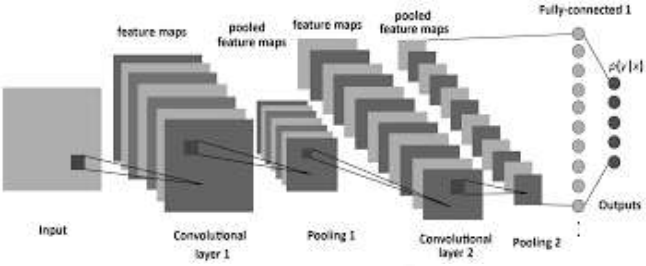

Fig.3 General architecture of a convolution neural network

V. EVALUATION METRICS

The most commonly used metric to test the quality of an image/video is Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measurement (SSIM), which are described in detail below. These two serve as quantitative evaluation metrics that are more easily interpretable.

A. PSNR:

Peak signal-to-noise ratio, often abbreviated PSNR, is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is usually expressed in terms of the logarithmic decibel scale. For color images with three RGB values per pixel, the definition of PSNR is the same except the MSE is the sum over all squared value differences divided by image size and by three. Alternately, for color images the image is converted to a different color space and PSNR is reported against each channel of that color space

B. SSIM:

SSIM is used for measuring the similarity between two images. The SSIM index is a full reference metric; in other words, the measurement or prediction of image quality is based on an initial uncompressed or distortion-free image as reference. SSIM is designed to improve on traditional methods such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE). The difference with respect to other techniques mentioned previously such as MSE or PSNR is that these approaches estimate absolute errors; on the other hand, SSIM is a perception-based model that considers image degradation as perceived change in structural information, while also incorporating important perceptual phenomena, including both luminance masking and contrast masking terms. Structural information is the idea that the pixels have strong inter-dependencies especially when they are spatially close. These dependencies carry important information about the structure of the objects in the visual scene. Luminance masking is a phenomenon whereby image distortions (in this context) tend to be less visible in bright regions, while contrast masking is a phenomenon whereby distortions become less visible where there is significant activity or texture in the image.

VI. RESULT AND DISCUSSION

We have trained the model and with fine tuning the hyperparametres and have obtained the results as shown in the table below. In order to get a decent comparison to other algorithms, we compute the mean PSNR and SSIM of Bicubic interpolation on our data set (VidSet12) and normalize our results such that both data sets have the same Bicubic interpolation mean PSNR and SSIM. As an example, suppose our test set PSNR is S. Then, letting B_{p0} be the mean PSNR of our data set and B_{p1} the mean PSNR of theirs, we report that our PSNR is equal to $S\frac{B_{p1}}{B_{p0}}$ (likewise for SSIM). As it happens, the ratio $\frac{B_{p1}}{B_{p0}}$ is close to one so this normalization does not have a profound effect.

Table 1 Results of proposed method with respect to other method

Method	**PSNR(db)**	**SSIM**
BiCubic	31.72	0.9483
BayesSR	29.57	0.9451
Proposed Method	31.52	0.9477

VII. CONCLUSION AND FUTURE ENHANCEMENT

In this paper we present a video restoration method based on convolution neural networks. Although our technique provides on par performance with respect to other methods it is yet inconsistent for general use due to its high dependency on the data used for training it. The main advantage of our method is that most other methods result estimates have high peak signal-to-noise ratios, but they are often lacking high-frequency details and are perceptually unsatisfying in the sense that they fail to match the fidelity expected at the higher resolution.

REFERENCES

[1] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Variational Bayesian Super Resolution.” IEEE Transactions on Image Processing, vol. 20, no. 4, pp. 984–999, 2011..

[2] S. P. Belekos, N. P. Galatsanos, and A. K. Katsaggelos, “Maximum aposteriori video super-resolution using a new multichannel image prior,” IEEE Transactions on Image Processing, vol. 19, no. 6, pp. 1451–1464, 2010.

[3] C. Liu and D. Sun, “On Bayesian Adaptive Video Super Resolution,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 2, pp. 346–360, 2014.

[4] A. K. Katsaggelos, R. Molina, and J. Mateos, Super Resolution of Images and Video, san rafael ed., 2007.

[5] Z. Ma, J. Jia, and E. Wu, “Handling Motion Blur in Multi-Frame Super-Resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. d, 2015.

[6] 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, 2015.

[7] a. A. C. B. Zhou Wang, "A Universal Image Quality Index," IEEE SIGNAL PROCESSING LETTERS, vol. 9, pp. 81-84, 2002.