# VIDEO SUPER RESOLUTION USING DEEP NEURAL NETWORKS

**[1]Anantha Sree Skanda S N, [2]Abhilash B S, [3]Akash Deep, [4]Amrit Kumar Bhuwania**
[1]Student, [2]Student, [3]Student & [4]Student
Computer Science and Engineering
BMS College of Engineering, Bangalore, India

***Abstract:*** ***Convolutional neural networks (CNN) are a special type of deep neural networks (DNN). They have so far been successfully applied to image super-resolution (SR) as well as other image restoration tasks. Consecutive frames are motion compensated and used as input to a CNN that provides super-resolved video frames as output. We investigate different options for combining the video frames within one CNN architecture. 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. We show that by using images to pre-train our model, a relatively small video database is sufficient for the training of our model to achieve and even improve upon the current state-of-the-art. We compare our proposed approach to current video as well as image SR algorithms.***

***IndexTerms - Bi-cubic Interpolation, Convolutional neural net, Motion Compensation, FlowNet, Multi-Frame systems, Spatio-Temporal networks.***

## I. INTRODUCTION

Super-resolution is the process of obtaining high-resolution images from relatively lower resolution images at a certain upscaling factor. Extensive research has been done in the field of Single image super-resolution. Technological advancements in the last four decades have seen a drastic improvement in the levels of accuracy of super-resolution images. With the recent advancement in the field of neural networks, several learning-based algorithms have come into the scene to solve this classical problem of super-resolution. Modern learning-based algorithms aim at using these existing image super-resolution models and applying them for the purposes of super-resolution videos. The main obstacle that has kept this area of study still very much active is the concept of optical flow which is not catered by the existing learning based image super-resolution techniques. Many variations have been devised through the years which exploit the fields of deep learning and multi-frame super resolution to improve upon the existing methods and getting a higher level of accuracy and optical flow.

The following paper describes in brief detail about the various single image and video super-resolution techniques and how they evolved over the years and how they differ from each other. It also explains what specific problem a technique aims to solve and what advantages one technique holds over the other.

In this method, the nearest value is copied for interpolation and this technique has less computational complexity. Nearest neighbor interpolation is recommended for categorical data such as land use classification.

## II. LITERATURE SURVEY

The existing approaches for super-resolution of images can be mainly classified into two categories, Frequency Domain Approach and Spatial Domain Approach. Frequency Domain Approach takes the low-resolution[LR] image and transforms it into its Discrete Fourier Transform(DFT) domain and combines them according to the relationship between the aliased DFT coefficients of the observed LR images and that of the unknown high-resolution image. While this approach cannot handle real-world applications, it is an efficient way as it is a spontaneous way to enhance the details as it has lower computational complexity. In Spatial Domain Approach the frequency domain approach has a major disadvantage that it does not support any use of prior knowledge. The main techniques used under this domain are Interpolation, Iterative Back Projection, Classical Multi-Image Super-resolution and Example-based Super-resolution.

### 2.1 Interpolation

Interpolation is the process of transferring the image from one resolution to another without losing image quality. When we increase the resolution of the image from low to high, it is called up-sampling or up-scaling while the reverse is called down-sampling or downscaling. There exist several ways to interpolate an image:

#### 2.1.1 Bilinear interpolation

In Bilinear interpolation, the interpolated point is filled with four closest pixel's weighted average. Bilinear interpolation is recommended for continuous data like elevation and raw slope values.

#### 2.1.2 Bicubic interpolation

Looks at the sixteen nearest cells and fits a smooth curve through the points to find the output value. Bicubic interpolation is recommended for smoothing continuous data, but this incurs a processing performance overhead.

#### 2.1.3 Nearest Neighbor interpolation

In this method, the nearest value is copied for interpolation and this technique has less computational complexity. Nearest neighbor interpolation is recommended for categorical data such as land use classification.

**2.2 Iterative Back Projection**

This is an iterative approach for super-resolution images. In this approach, the high-resolution image is estimated by projecting back the difference between the estimated and the actual LR image on the interpolated image. This iterative process goes on till the cost function minimization is achieved.

**2.3 Classical Multi Image Super Resolution**

This approach works on the assumption that two or more LR images must have distinguishable features. An SR image is reconstructed if there are enough images available. This is inefficient in practical situations if multiple images with distinguishable features are not available.

**2.4 Example-Based Super-Resolution**

Contrary to the Classical Multi Image approach, this method is useful when only single LR image is available. Each LR patch in an image is replaced by its corresponding High-Resolution patch to generate the SR image.
There exist several approaches under Example-Based Super-Resolution:

**2.4.1 Neighbor Embedding**

This approach selects several LR candidate patches in the dictionary by using the nearest neighbor search and employs their HR version for the reconstruction of HR output patches.

**2.4.2 Super-Resolution Forests**

This approach relies on direct mapping from LR to HR patches using random forests (RFs). A relation of contemporary SISR to locally linear regression and try to fit in RFs into this approach has been demonstrated.

**2.4.3 Naive-Bayes Super Resolution forest**

This is a probabilistic approach for example-based SR using an external learning strategy, that provides a fast local linearization search, followed by a fast Local Naive Bayes strategy for patch-wise estimation.

**2.4.4 Learning-Based Super Resolution**

This method heavily depends on the training data as to what High-Resolution patches are retrieved for corresponding Low-Resolution patches. The classification is done for LR and HR patches in the early stages which reduces immensely the number of comparisons to be done during super-resolution.

These existing techniques for single image super-resolution (SISR) can be extended further for super-resolution Low-resolution videos to High-Resolution. This is possible by taking each frame as a separate Single Image and using the previously mentioned techniques and applying them to super resolute individual frames. This approach doesn't work in practical applications as it doesn't account for motion compensation or optical flow. With the advent of deep-learning in recent years, several techniques have come up. This is further helped by huge datasets available for training a neural network and fast processors. The modern state-of-the-art techniques that exploit deep-learning use the following sub-systems to tackle the problem of video super-resolution:

Bicubic Interpolation: This is the first step of most of the techniques used. This step is used to interpolate individual frames by resizing them to the resolution of desired High-resolution video frame. Bicubic interpolation looks at the sixteen nearest cells and fits a smooth curve through the points to find the output value. Bicubic interpolation is recommended for smoothing continuous data, but this incurs a processing performance overhead.

Motion Compensation: This step is done to compensate for the motion blur and optical flow which is not accounted for by the systems for single image super-resolution. Networks/Algorithms like FlowNet etc. are used which output a motion compensated frame depending on the consecutive frames existing before and after it.

Optical Flow/Image Distortion Fixing: Image warping is the process of digitally manipulating an image such that any shapes portrayed in the image have been significantly distorted. Warping is used in this step for fixing the distortions and noises produced in the previous steps. It uses a process of forwarding and reverses mapping of consecutive frames for optical flow estimation.

Convolutional Neural Networks: This is the step where actual super-resolution of individual frames occur. The output retrieved from previous steps after motion compensation and interpolation is fed into the convolutional neural network for actual super-resolution. The process of super-resolution is done by using already-trained networks that are trained using LR and HR images.

## III. PROPOSED WORK

There have been numerous algorithms/techniques which may or may not use the above-specified sub-systems, relevant ones of which are mentioned below:

**3.1 Multi-Frame Super-Resolution**

Multi-Frame super-resolution (MFSR) is the process of taking multiple low-resolution (LR) video frames and constructing a single high-resolution (HR) video frame. Generic image SR techniques utilize image priors to generate an HR version of a single LR image input. Example-based methods such as [1], [2], [3] seemed to produce the most promising results. such as [3], [4] train dictionaries to learn a mapping from LR to HR image patches. The nearest neighbor (NN) of the input image patch is found among the LR patches in the dictionary and its corresponding HR image patch is used to reconstruct the input. [4] improves the above technique by using a sparse coding

formulation to replace the above NN strategy. Experimentation with better mapping functions kernel regression [2] and anchored neighborhood regression [3] has increased the speed and accuracy of the traditional example based techniques.

There are trained two different classes of models, one using only single images or frames (SICNN) and one utilizing information from multiple frames in a video (MFCNN). For the MFCNN we take a different approach which allows us to take in a frame and its adjacent frames and output the high-quality version of the middle frame. In order to allow the adjacent input frames to retain all of their spatial information, we form a single training example by concatenating all of the input frames along the channel dimension.

**3.2 Spatio-Temporal Networks**

This technique enabled accurate image super-resolution in real-time. Spatio-temporal sub-pixel convolution networks effectively exploit temporal redundancies and improve reconstruction accuracy while maintaining real-time speed.
Video SR methods have mainly emerged as adaptations of image SR techniques. Kernel regression methods [5] have been shown to be applicable to videos using 3D kernels instead of 2-D ones [6]. Dictionary learning approaches, which define LR images as a sparse linear combination of dictionary atoms coupled to an HR dictionary, have also been adapted from images [7] to videos [8]. Spatial transformer networks [9] provide a means to infer parameters for a spatial mapping between two images. These are differentiable networks that can be seamlessly combined and jointly trained with networks targeting other objectives to enhance their performance. Recently, it has been shown how spatial transformers can encode optical flow features with unsupervised training [10, 11, 12, 13]. Some of the merging techniques at the sub-pixel convolutional network are as follows:

**3.2.1 Early fusion**

One of the most straightforward approaches for a CNN to process videos is to match the temporal depth of the input layer to the number of frames. This will collapse all temporal information in the first layer and the remaining operations are identical to those in a single image SR network.

**3.2.2 Slow Fusion**

In this case, the temporal depth of network layers is configured to be and therefore some layers also have a temporal extent until all information has been merged and the depth of the network reduces to 1. This architecture, termed slow fusion, has shown better performance than early fusion for video classification [14].

**3.2.3 3D-convolutions**

Another variation of slow fusion is to force layer weights to be shared across the temporal dimension, which has computational advantages. Assuming an online processing of frames, when a new frame becomes available the result of some layers for the previous frame can be reused.

**3.3 End to End Learning Technique**

The task of providing a good estimation of a high-resolution (HR) image from low-resolution(LR) minimum up-sampling effects, such as ringing, noise, and blurring studied extensively [15,16,17,18]. Motion estimation is a longstanding research topic in computer vision, and a survey is given in [19]. In this work, we aim to perform video super-resolution with a CNN-only approach. The FlowNet2 by Ilg et al. [20] yields state-of-the-art accuracy but is orders of magnitudes faster than traditional methods.
For motion estimation, we used the FlowNet2-SD variant from [20]. For the warping operation, implementation from [20] is used, which also allows a backward pass while training. The combined network is trained on complete images instead of patches. Motion compensation with FlowNet2 [20] seems to be marginally sharper than motion compensation with Drulea [21]. The joint training reduces ringing artifacts.

**3.4 Bi-Directional Recurrent Convolutional Neural Networks**

Super-resolving a low-resolution video is usually handled by either single-image super-resolution (SR) or multi-frame SR. Single-Image SR deals with each video frame independently and ignores intrinsic temporal dependency of video frames which actually plays a very important role in video super-resolution. Existing multi-frame SR methods generally model the temporal dependency by extracting sub-pixel motions of video frames, e.g., estimating optical flow based on sparse prior integration or variation regularity [22, 23, 24]. Recurrent neural networks (RNNs) can well model long-term contextual information for the video sequence.

Bidirectional recurrent convolutional network (BRCN) is used to efficiently learn the temporal dependency for multi-frame SR. The proposed network exploits three convolutions. Feedforward convolution models visual-spatial dependency between a low-resolution frame and its high-resolution result. Recurrent convolution connects the hidden layers of successive frames to learn temporal dependency. Different from the commonly-used full recurrent connection in vanilla RNNs, it is a weight-sharing convolutional connection here. Conditional convolution connects input layers at the previous time-step to the current hidden layer, to further enhance visual-temporal dependency modeling. In each sub-network, there are four layers including the input layer, the first hidden layer, the second hidden layer and the output layer, which are connected by three convolutional operations.

**IV CONCLUSION**

Utilizing profound learning methods for SISR, we have proposed a novel multi-outline VSR method that effectively exploits pixel data from nearby outlines. While the field of VSR is excessively youthful, making it impossible to have a plainly cutting-edge technique, our model essentially beats the strategies that we could think about it to be. Our outcomes likewise affirm that VSR is a less demanding and hence more encouraging issue to unravel than SISR, as we were ready to enhance the nature of the video considering different outlines more than if we had just access to one casing at a period. Be that as it may, while our strategy has incredible execution on normal, it is awfully conflicting starting at yet to be utilized for the most part. Future work incorporates adjusting the models altogether to create a heartier calculation for VSR that isn't subject to the same number of anomalies and sudden or unforeseen drops in execution. One plan to battle this is

to include higher regularization to the models, which may decrease the mean change be that as it may, altogether diminish its fluctuation. Another is to outfit numerous models and take the most prevalent decision among these for every pixel to diminish the likelihood that a solitary model on a subset of pixels.

**V REFERENCES**

[1] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. IEEE International
[2] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. IEEE International
[3] R. Timofte, V. D. Smet, and L. V. Gool. Anchored neighborhood regression for fast example-based super-resolution.IEEE International Conference on Computer Vision pages 1920–1927, 2013.
[4] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. IEEE Transactions on Image Processing, 19(11):2861–2873, 2010.
[5] H. Takeda, S. Farsiu, and P. Milanfar. Kernel regression for image processing and reconstruction. IEEE Transactions on Image Processing, 16(2):349–366, 2007.
[6] H. Takeda, P. Milanfar, M. Protter, and M. Elad. Superresolution without explicit subpixel motion estimation. IEEE Transactions on Image Processing, 18(9):1958–1975, 2009.
[7] J. Yang, S. Member, and Z. Wang. Coupled dictionary training for image super-resolution. IEEE Transactions on Image Processing, 21(8):3467–3478, 2012.
[8] Y. Ganin, D. Kononenko, D. Sungatullina, and V. Lempitsky. DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation. European Conference on Computer Vision(ECCV), pages 311–326, 2016.
[9] A. Ahmadi and I. Patras. Unsupervised convolutional neural networks for motion estimation. IEEE International Conference on Image Processing (ICIP), pages 1629–1633, 2016.
[10] V. Patraucean, A. Handa, and R. Cipolla. Spatio-temporal
video autoencoder with differentiable memory. International
Conference On Learning Representations (ICLR) Workshop,
2016.
[11] A. Handa, M. Bloesch, V. Patraucean, S. Stent, J. McCormac, and A. Davison. gvnn: Neural Network Library for Geometric Computer Vision. European Conference on computer vision (ECCV) Workshop on Deep Geometry, 2016.
[12]. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. IEEE Computer Graphics and Applications (ICGA) 22(2), 56–65 (Mar 2002)
[13]. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June
2004)
[14]. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution as the sparse representation of raw image patches. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2008)
[15]. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 19(11), 2861–2873 (Nov 2010)
[16]. Drulea, M., Nedevschi, S.: Total variation regularization of local-global optical flow.In: IEEE Conference on Intelligent Transportation Systems (ITSC). pp. 318–323 (Oct 2011)
[17] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
[18] S. Baker and T. Kanade. Super-resolution optical flow. Technical report, CMU, 1999.
[19] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L. A. Morel. Low-complexity single-image superresolution
based on nonnegative neighbor embedding. British Machine Vision Conference, 2012.
[20] C. Liu and D. Sun. On Bayesian adaptive video super-resolution. IEEE Transactions on Pattern Analysis
and Machine Intelligence pages 346–360, 2014.
[21] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. European Conference on Computer Vision, pages 184–199, 2014.
[22] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. IEEE International Conference on Computer Vision, pages 633–640, 2013.
[23] V. Jain and S. Seung. Natural image denoising with convolutional networks. Advances in Neural Information Processing Systems, pages 769–776, 2008.
[24] Y. A. Y. Al-Najjar and D. D. C. Soong. Comparison of image quality assessment: Psnr, hvs, ssim, uiqi. International Journal of Scientific and Engineering Research, 3(8), 2012.
[25] C. Dong, C. C. Loy, K. He, and X. Tang. Image superresolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015.
[26] C. Liu and D. Sun. On Bayesian adaptive video super-resolution. IEEE TPAMI, 2013.
[27] Alex Greaves, Hanna Winter. Multi-Frame Video Super-Resolution Using Convolutional Neural Networks. Stanford University 450 Serra Mall, Stanford, CA 94305.
[28] Shuai YUAN, Masahide ABE, Akira TAGUCHI* and Masayuki KAWAMATA. High Accuracy Bicubic Interpolation Using Image Local Features. Department of Electronic Engineering, Tohoku University Department of Electrical and Electronic Engineering, Musashi Institute of Technology.
[29] Osama Makansi, Eddy Ilg, and Thomas Brox. End-to-End Learning of Video Super-Resolution with Motion Compensation. Department of Computer Science, University of Freiburg.