Video Super-Resolution using Multi-Frames Fusion and Perceptual Loss

Xiaonan Zhu, Yucong Zhao, Ping Wang, Yuanchen Wang and Jiquan Ma

Computer Science and Technology, Heilongjiang University, Harbin, Heilongjiang, 150080, China

*Corresponding author’s e-mail: majiquan@hlju.edu.cn

Abstract. Video super-resolution (VSR) has become more and more important in vision perception. Currently, most of video super-resolution methods used optical flow estimation by explicit motion compensation. We proposed a new network using multi-frames as input without motion compensation. In order to improve the performance of the reconstructed high-resolution (HR) frame, perceptual loss function is used to evaluate the deviation between the prediction and ground truth. Perceptual loss characteristic is able to improve the quality of image reconstruction on the vision, but with a low Peak Signal to Noise Ratio (PSNR). Experimental results show that compared with other methods, our model can recover more details in high-resolution image, although it does not get higher PSNR value.

1. Introduction

Super-resolution is the process of transforming the low-resolution (LR) image to the high-resolution (HR) image. Traditional super-resolution methods have got better performance, such as sparse representation and interpolation, but some details in low-resolution image can not be reconstructed using these methods. With the development of deep learning, super-resolution based on learning has got exciting results. In singe image super-resolution (SISR), SRCNN [2] was proposed by Dong et al. In this method, CNN was applied to reconstruct high-resolution image. In [7], residual learning was applied by the VDSR network. DRRN was proposed in [10], DRRN was a deep recursive residual network, it has deep and concise features. All above methods used pixel-based MSE as the loss function to optimize the parameters of network. Furthermore, in order to improve the performance, perceptual loss function was introduced in [5], the experimental results have revealed that the quality of reconstruction based on perceptual loss can get art of the state performance under the human vision perception and have proved that the quality of reconstruction based on pixel measurement is not the only standard.

The difference between VSR and SISR is the correlation and continuity in sequenced frames. Some VSR methods ignored the characteristics of consecutive frames, such as ESPCN [4]. So the idea of optical flow was applied to [1,3,8]. [8] used Spatial Transformer Networks (STN) [6] in order to make full use of consecutive frames. [9] achieve more reasonable frame alignment through sub-pixel motion compensation (SPMC) layer. So then the Long Short Term Memory (LSTM) framework was used to predict the HR frame by multi-frames input. [12] used recursive network, the HR prediction generated in the previous time was used as the input of the next prediction, which can not only produce time-continuous results but also reduce the computational complexity.

Our model is divided into two parts: 1. Multi-frames fusion sub-network. In this stage, continuous video frames is combined as the real input for the reconstructed network. 2. Reconstruction sub-
network. It is used to predict the final HR. A pre-trained VGG16 model is introduced to optimize our model and evaluate the perceptual semantic bias of prediction and ground truth.

2. Method
Our goal is to convert the input LR frames \( \{ X_i \} \) to output HR frames \( \{ Y_i \} \). The \( W \) is VSR Network, \( \Theta \) is the network parameters, \( i \) is time steps. The formula can be expressed

\[
Y'_i = W_{\Theta}(X_{i-N} : X_{i+N})
\]  

(1)

Where \( N \) is the temporal radius. \( \{ X_i \} \) is LR frames. The frame size is \( C \times H \times W \). \( C \) is the number of channels. Our model architecture is illustrated in Figure 1.

![Figure 1. Our model architecture. Multi-frames fusion sub-network and reconstruction sub-network achieve super-resolution through end-to-end learning. The perceptual loss is applied to adjust semantic bias.](image)

2.1. Multi-frames fusion sub-network

![Figure 2. Multi-frames fusion sub-network. The network is divided into two parts: convolution process and concatenation process. The function of convolution process is not only to extract features, but also to transform video frames, it is conducive to better fuse (concatenation). The \( F \) is the number of input consecutive frames (LR).](image)

Multi-frames fusion sub-network (in Figure 2) uses different convolution blocks to transform the adjacent frames so that they are aligned with the reference frames in the convolution process automatically. The network is helpful to provide more effective detail features for reconstructing reference frames.
For the input LR frames (F=3, down-sampling=4), the convolution block that processes adjacent frames of reference frame is composed of four filters, the size are 9*9, 7*7, 3*3 and 1*1. The convolution block for processing reference frame consists of two filters, the size are 3*3 and 1*1.

2.2. Reconstruction sub-network
The reconstruction sub-network is to get the final HR prediction. Each convolution layer except the last layer is followed by Instance Normalization (IN) and Rectified Linear Unit (ReLU) activation. The reconstruction process is illustrated in Figure 3.

In the network, Conv block consists of a 9*9, a 7*7, a 3*3 and a 1*1 filter. In the Conv block, large filters are mainly used to extract features, 1*1 filter is used to smooth the features.

Residual network can solve the side effects (degradation problems) and improve network performance when increasing network depth. So we added five residual blocks in our network, each convolution layer is followed by an ReLU activation.

Before deconvolution, the size of feature maps remains unchanged. Mapping LR to HR space via two deconvolutions. Before we get the final HR prediction, we use a 1*1 filter for smoothing.

2.3. Loss function
In this experiment, we used two loss functions to optimize our model. MSE loss and perceptual loss.

2.3.1. MSE loss
\[ L_{sr} = \| Y_i^r - Y_i \|^2_2 \] (2)

\( Y_i^r \) is the prediction, \( Y_i \) is the ground truth.

2.3.2. Perceptual loss network
Since the perceptual loss is based on semantic information, what we pursue is similar to the original image visually, rather than increasing the PSNR value of the input image simply. In this experiment, we used pre-trained VGG16 model on the ImageNet dataset, it contains a large amount of feature information, so we send the prediction and the ground truth to the VGG16 model and calculate the loss by extracting the feature map from the model relu2_2. The reason for extracting the feature of relu2_2 is that the deeper network extraction features are more abstract and have more semantic information, which is beneficial to the reconstruction of features. The perceptual loss is as

\[ L_{per}(Y,Y') = \frac{1}{C_iH_jW_j} \| \Phi(Y') - \Phi(Y) \|^2_2 \] (3)

Figure 3. Reconstruction sub-network. The input is the output of the multi-frame fusion sub-network.
Where \( j \) represents the \( j \)-layer of the VGG16 network. \( \Phi(Y') \) and \( \Phi(Y) \) are the feature maps of \( j \)-layer. \( C, H, W \) represents the size of feature map from \( j \)-layer.

In the last stage, we jointly tune the whole system using the total loss as

\[
L_{\text{total}} = L_{\text{sr}} + \lambda L_{\text{feat}}
\]

(4)

Where \( \lambda \) is the weight balancing two losses.

3. Experiments

3.1. Dataset

For super-resolution, the quality and quantity of the dataset are very important. They require a wealth of scenes, including animals, people, landscapes and so on. But video super-resolution does not have public and abundant train dataset, so the dataset we trained and tested is 1920*1080 HD video downloaded from YouTube.

For the downloaded HD video, we process it into 30,000 consecutive video frames and ignore the scene change problem. In addition, our model can be used to test video of any size.

3.2. Implementation details

For consecutive frames entered into the model, we cut it to patch of size 288*288, and then down-sampling = 4 to obtain the input by bicubic interpolation. We used frames of the three channels for training and testing. We set the batch size to 4 and use multi-threading for training. Our experiment runs on the Nvidia Geforce GTX 1080 Ti, it takes about a day to converge the model.

We compared with Li’s [5] and VESPCN [8], because they are symbolic methods to introduce perceptual loss and motion compensation into super-resolution.

3.3. Results

Table 1. Comparison with SR methods. down-sampling = 4.

| Method      | PSNR | SSIM |
|-------------|------|------|
| Bicubic     | 21.69| 0.58 |
| Li’s method | 24.62| 0.66 |
| VESPCN-F3   | 30.58| 0.89 |
| Ours-F3     | 29.16| 0.81 |

Our result values is illustrated in Table 1. The F3 means that the number of input consecutive frames is 3. Because perceptual loss is based on semantic information optimization model, while PSNR and SSIM are based on pixel evaluation. So the quality of reconstruction based on pixel measurement is not the only standard. When perceptual loss is added to our model, the result values are not as high as VESPCN at the pixel level, but the details are better restored visually. Figure 4 shows the performance of our results and other methods.
4. Discussion

In the course of experiment, we try to change the number of residual blocks, the comparison results show that five residual blocks can make the model converge better. Since deep feature maps contain a large number of abstract semantics, which is conducive to detail recovery, but further deepening will lose more feature information. So we also try to extract feature maps from different positions of VGG16 model to optimize the loss. Experiments show that relu2_2 layer is most conducive to optimization model and feature reconstruction.
5. Conclusion
Compared with [5], our results show significant improvements in values and visual perception. Meanwhile, compared with [8], our full convolution network has the following advantages: 1. More concise. It can make the consecutive frames learn the relationship between video frames in the convolution process and align with the reference frame automatically. 2. More efficient. Our network can accept any size of input frames and does not need separate training for different size of videos. In the next work, we will continue to optimize the model and improve the quality of reconstruction.

References
[1] C. Liu, D. Sun. (2011) A bayesian approach to adaptive video super resolution. In CVPR, pp. 209–216.
[2] C. Dong, C. Loy, K. He, and X. Tang. (2014) Learning a deep convolutional network for image super-resolution. In ECCV, pp. 184–199.
[3] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos. (2016) Video super-resolution with convolutional neural networks. In IEEE, pp. 109–122.
[4] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. (2016) Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In CVPR, pp. 1874–1883.
[5] Justin Johnson, Alexandre Alahi, Li Fei-Fei. (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In ICCV, pp. 694-711.
[6] Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu. (2016) Spatial Transformer Networks. Advances in neural information processing systems, pp. 2017-2025.
[7] Jiwon Kim, Jung Kwon Lee and Kyoung Mu LeeDepartment of ECE. (2016) Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In CVPR, pp. 1646–1654.
[8] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. (2017) Real-time video super-resolution with spatio-temporal networks and motion compensation. In CVPR, pp. 4778-4787.
[9] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, Jiaya Jia, (2017) Detail-revealing Deep Video Super-resolution. In ICCV, pp. 4472-4480.
[10] Ying Tai, Jian Yang, Xiaoming Liu. (2017) Image Super-Resolution via Deep Recursive Residual Network. In CVPR, pp. 2790-1798.
[11] Sajjadi, M. S., Vemulapalli, R., Brown, M. (2018) Frame-recurrent video super-resolution. In CVPR, pp. 6626-6634.