Enhanced Video Super Resolution System using Group-based Optimized Filter-set via Shallow Convolutional Neural Network for Super-Resolution

Sangchul Kim
Dept. of Computer Science and Engineering
Sogang University
Seoul, Korea, Republic of
sckim@sogang.ac.kr

Jongho Nang
Dept. of Computer Science and Engineering
Sogang University
Seoul, Korea, Republic of
jhnang@sogang.ac.kr

***Abstract*—Scaling up video resolution has conventionally been achieved via linear interpolation, however this method occasionally introduces blurring to the output. Super-resolution (SR), an approach to preserve image quality in enlarged still images, has been exploited as a substitute for linear interpolation, however, the output at times exhibits image qualities worse than what linear interpolation produces primarily because the initial goal of SR is preservation of image quality when a still image is enlarged. In this context, this paper proposes a fast-performance adaptive system for scaling-up other resolutions like X2→X3 or X3→X2 by (1) first grouping frames that would use similar filter sets (2) then conducting fine-tuning of shallow CNN for SR on each frame group. Filter sets fine-tuned for each group resulted in significantly improved PSNR over either linear interpolation or conventional SR in our experiment. In the fine-tuning stage for each group, 0.5K to 2.5K iterations were sufficient to improve PSNR by 10%. By fine-tuning instead of performing full training, the number of sufficient iterations was reduced from 300K to mere 0.5K to 2.5K.**

Keywords-video resolution Scaling up; convolutional neural network; shot change detection; gradual transition detection; deep learning; fine tuning; super resolution; CNN;

I. INTRODUCTION

The latest standard for high-resolution videos is Ultra HD but most videos are being produced in Full HD format. When a Full HD video is played back on an Ultra HD screen, the resolution is scaled up usually by means of linear interpolation. One side effect of linear interpolation is blurring, which negatively affects video quality.

To address this issue, Super-resolution (SR) can be adopted, the initial goal of which is preserving the image quality of enlarged still shots. Characteristics of motion pictures and still images are inherently different, so SR occasionally degrades video quality more than linear interpolation does in effect.

This paper proposes video resolution scaling-up system that (1) first puts frames into groups that are expected to have similar filter sets (2) then conducts fine-tuning for each group to obtain filter sets tailored for each group. This system noticeably improves video quality. We defined two group types: shots and gradual shot transitions. In a shot frame group, frames share very similar objects and background, and the frames would use similar filter sets. We handle gradual shot transition frame groups separately because alpha-blending and blurring effects are dominant in the frames.

Experimentally we verified the proposed system increases the overall PSNR by more than 10% compared to the shallow CNN-based SR because each trained model is used as an optimized filter set for each group.

Since the proposed system uses pre-trained models, only 0.5K to 2.5K training iterations were needed to produce the results rather than 300K iterations as in regular model training. Section 2 introduces related works. Section 3 explains group-based optimized filter training for SRCNN system. In Section 4, experiment analysis is presented. This paper concludes in Section 5 with suggestions for future work.

II. RELATED WORKS

A. Convolutional Neural Network (CNN)

The Convolutional Neural Network Model consists of convolutional layers and sub-sampling layers. In a CNN model, the feature map has a larger target area in upper layers [1]. A feature map robust for feature position is trained through the link structure of receptive fields. A feature map trained in this way is used for categorization, image reconstruction, and other similar determination problems using deep neural network (DNN) structure. DNN structure need tremendous time to train a model, fine-tuning[2], which use pre-trained model and adopt network to the model, is usually conducted to reduce training time and achieve their goal.

B. Shot Change Detection

Video shots are taken and later merged in a contiguous form during editing. A shot is the smallest unit with semantic meaning, and shot change detection is receiving much attention for research. In [3][4], adjacent frames are compared to detect abrupt changes in brightness or optical flow and attempts to identify shot boundaries.

Gradual shot transitions are difficult to detect for this approach, hence cumulative change in brightness and motion vectors has been studied to identify gradual shot transition boundaries [5].

C. *Super Resolution*

Super-resolution scales up image resolution while preventing image quality degradation. When the resolution of an image is increased, the image is blurred and noise is introduced.

So, de-blur and de-noise filters are trained using one image in single image-based SR [6] or using multiple images in multi-frame based SR [7]. A large number of images are used to train filter sets in advance before applying in example-based SR [8].

These approaches come with the issues of machine learning methods, hence CNN-based SR [9][10][11], which train filter sets with deep learning, have been researched with significant improvement in output quality. However, these method has limitation to change variable scaling resolution because these train result is fixed size scaling model

III. VIDEO RESOLUTION SCALING UP SYSTEM USING GROUP-BASED OPTIMIZED FILTER SET TRAINING VIA SHALLOW CNN

A. *Proposed System Architecture*

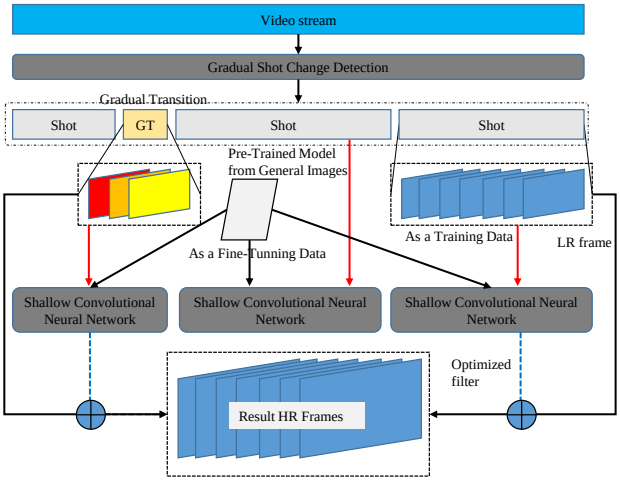

Figure 1. Proposed Video Resolution Scaling up System Architecture.

Fig. 1 illustrates the overall architecture of the proposed system. In the system, shot frames and gradual transition frames are distinguished in the Gradual Shot Transition Detection module. In a gradual shot transition, blurring and alpha-blending effects are predominant and a specific SR filter exists specific to this type of frames. Frame groups are designated as CNN-based SR filter set training, thus, these trained CNN-based SR filters sets optimized for each group. Since models are fine-tuned rather than trained in full, CNN-based SR filter set training iterations are significantly reduced per group to quickly find a model for increasing resolution.

B. *Gradual Transition Detection*

The goal of SRCNN is forming a filter that reducing noises while improving on edges. Edges are weak in gradual shot transition frames. In frames for gradual shot transitions, the image quality of SRCNN output is worse than what can be obtained with linear interpolation. In this paper, shots and gradual transitions are distinguished using [5], which relies on motion vectors and cumulative brightness change.

C. *Shallow CNN-based SR with training group-based optimized filter set to change variable scale resolution*

Fig. 2 illustrates the SR network used in our system.

We adopted [8] for shallow CNN-based SR to get optimized filter sets for frame groups. CNN-based SR method [10][11] showed significantly improved output quality by adopting very deep neural network, but large deep networks take longer training time per iteration and training also requires a large amount of data.

However, the average shot length is around 3 seconds with approximately 90 frames, and a shallow CNN is adequate for this size of training sets, hence we opted for SRCNN [9].

The conventional CNN-based SR finds still image frames of good image quality for training data sets, label sets, and test sets tailored for general cases. However, video frames are usually shaky and blurry as shots are taken, and frames are distorted after data compression. Moreover it has limitation to change variable scaling their resolution. Because its method train a model considering only difference between input scale and output scale.

Within a single frame group, objects, background, camera walk, and image effects tend to be identical. This implies that filters used in SR for each frame group are all similar. Therefore, we hypothesize that a filter has to be optimized for each frame group in order for SR to perform better than filter sets for generic SR.

Our approach is to start by fine-tuning existing pre-trained models to finish training quickly. This reduces training time cost despite that SRCNN training method is used per frame group.

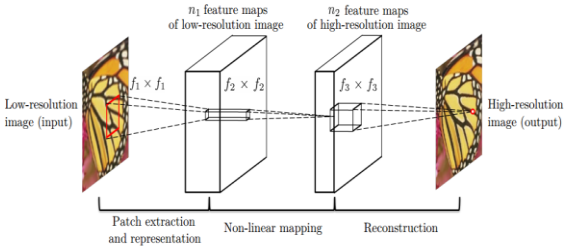

Figure 2. SRCNN Network structure presented in [9]. We use this network to train optimized filter set and conduct scaling up resolution of frames in our system

IV. EXPERIMENT AND EVALUATION

A. *Dataset and CNN setting*

For dataset, we obtained 823 frames from randomly chosen 10 frame groups (7 general shots and 3 gradual transitions) from Korean documentary films.

Learning rate was increased by 5 fold, and weight filler by 10 fold for fast convergence and overfitting in the same network as [9]. And we use [9]'s x2 model and x3 model to conduct fine-tuning to variable scale size. We test only x2, x3 scale output of fine-tuned version by cross checking.

B. *Quality of Resolution Scaled-up Result*

Table 1 shows the average PSNR of scaled-up images after applying different SRCNN models.

Case (b) has iterated 2,500 trainings with shot frames as train/test label set, and the output was much blurred, which is worse than bicubic output.

Case (c) used SRCNN train set[9], and the performance is slightly worse than bicubic method.

It went through generous filter training with a dataset of clean still images, hence motions of objects, blurs due to camera walking, and ghosting were not factored in for filter training.

In Case (d), each frame group is designated with a more suitable general filter set and trained, which produces a better output than what bicubic conduct.

We analyze this result in the next subsection.

TABLE I. PSNR (PEAK SIGNAL-TO-NOISE RATIO) OF RESOLUTION SCALED-UP RESULT FRAMES FOR EACH CNN MODEL

Average PSNR	**(a)**	**(b)**	**(c)**	**(d) (Ours)**
	Bicubic (Linear Interpolation)	**Frames in a shot (2500 iteration)**	**SRCNN trainset[9] (state-of the-art)**	**Fine tuning (2500 iteration)**
	31.04045 dB	28.75475 dB	29.9929 dB	33.5176 dB

C. *Correlation between Trained Filter set and CNN training iteration count*

SRCNN has a simple structure, which causes the loss rate of the loss layer to converge at the 500th iteration.

Fig. 3 charts the relation between the number of training iterations and Average PSNR of results of SRCNN fine-tuning models trained from both SRCNN x3 to group data for scaling x2 and SRCNN x2 to group data for scaling x3. PSNR increases by more than 10% at the 500th iteration, and more training iterations afterwards yields diminishing returns. On average, 2,500 iterations or more gave the impression that the image quality has been improved.

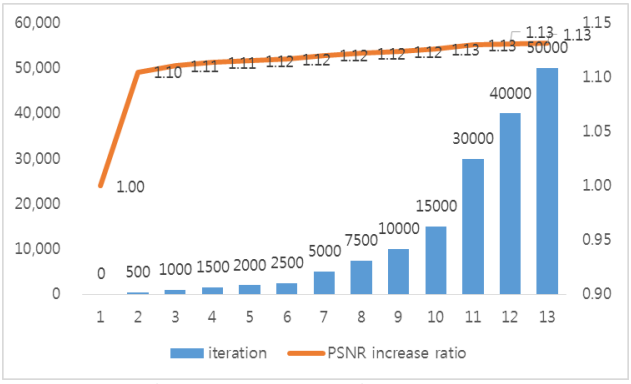

Figure 3. Graph of correlations between fine-tuning training iterations and PSNR that different resolution scale

Average PSNR of results of SRCNN fine-tuning models trained from both SRCNN x3 to group data for scaling x2 and SRCNN x2 to group data for scaling x3

Fig. 4 charts the relation between Average PSNR and the number of training iterations of both grouping data and SRCNN training data. Fine tuning of equivalent scaling model has slightly improve its PSNR in contrast to SRCNN training set. Since SRCNN model's weights are already converged, otherwise, frames in a group to be conducted super resolution has similar characteristic such as edge, brightness and has are different from SRCNN training set. It is able to change their weights of CNN. Fig. 5 is a demonstration of SR output at different training iterations. Fewer iterations as in (b) has lower loss rate since blur filter has been trained, but the subjective quality is worse. As training progresses, edge filters start to be effective and image appears more sharpened. This implies that models pre-trained with a large number of iterations can be exploited for our system to fine-tune with few additional iterations for adding adequate denoise filters and removing unnecessary edge filters.

Fig. 6 shows that the results by conducting SR for each filter set with fine-tuning model. low training iterations of fine-tuning model remains edge noise because pre-trained SRCNN model has sharpen filters and blur filters for this group is yet insufficiently trained, however, edge noise is eliminated as filter set for this group is sufficiently trained by increasing training iterations.

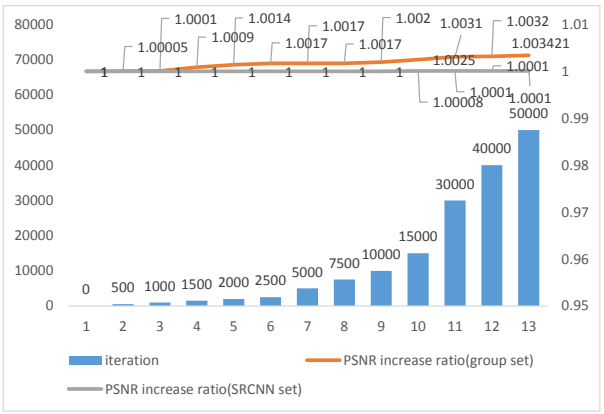

Figure 4. Graph of correlations between fine-tuning training iterations and Average PSNR of results of SRCNN fine-tuning models trained from both SRCNN x2 to SRCNN data for scaling x2 and SRCNN x2 to group data for scaling x2.

TABLE II. QUALITY OF CNN MODEL TRAINED WITH FRAMES IN GRADUAL TRANSITION ONLY. AVERAGE PSNR OF SRCNN(X2)→FINETUNED(X3), SRCNN(X3)→FINETUNED(X2) AND SRCNN(X2)→FINETUNED(X2).

Method	**(a)**	**(b)**	**(c)**
	BiCubic	**trained (Gradual Transition) GT Frames only**	**Trained frames include shot and GT frames**
PSNR (dB)	33.113752 (dB)	34.121934 (dB)	33.585275 (dB)

Table 2 lists the PSNR values when gradual shot transition frames were processed with separate SR effects. Case (c) combines gradual transition shots with the adjacent shots, and its output is worse than when training was performed only with gradual shot transition frames as in Case (b).

The cause is inclusion of gradual shot transition frames. The average length of gradual shot transitions is 20 frames, and the adjacent shots are mostly disparate.

In a gradual shot transition, blurring and alpha blending effects are predominant but edge filters intended for non-transition shots remain effective.

If training is limited only to gradual transition shots, filters suitable for blurring and alpha blending take effects and improve output frame quality.

(a) : Original

(b) : 2500 iteration

(c) : 300K iteration

Figure 5. Examples of SR output for Different Training Iterations.

I. Conclusion and Future Works

We propose video resolution scaling up system conducting super resolution with optimized filter set trained by deep-learning in a shallow convolutional neural network. In our work we observed that segregating regular shots and gradual transition shots for SR makes significant output improvement over existing pre-trained models.

The number of required training iteration was only 2.5K rather than 300K as we only needed to fine-tune existing pre-trained models, and separate application of CNN to each frame groups was feasible in resolution scale-up even for streamed videos.

In our future work, we will simplify SR CNN by considering the temporal redundancies of frame groups in order to optimize with fewer iterations.

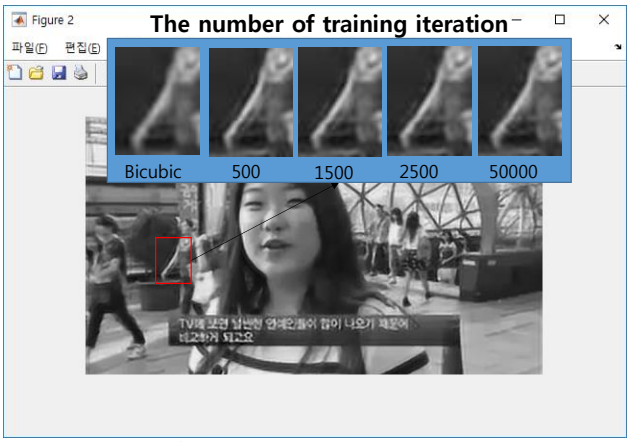

Figure 6. Examples of image output applied trained shallow CNN-based filter-set for different training iterations with fine-tuning model.

Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP. [R0126-16-1112, Development of Media Application Framework based on Multi-modality which enables Personal Media Reconstruction]

References

[1] C. Garcia, M. Delakis, “Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection, ” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 11, 2004, pp.1408-1423.

[2] G. E. Hinton, R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” Science, Vol. 313, No. 5786, 2006, pp.504-507.

[3] J. Yuan, H. Wang and L. Xiao, "A Formal Study of Shot Boundary Detection", IEEE Transactions on Circuits and Systems for Video Technology, vol.17, no.2, 2007, pp.168-186.

[4] A. Amel, B. Abdessalem and M. Abdellatif, "Video Shot Boundary Detection Using Motion Activity Descriptor," Journal of Telecommunications, vol.2, no.1, 2010, pp.54-59.

[5] S. Kim, H. Hong, J. Nang, “A Gradual Shot Change Detection using Combination of Luminance and Motion Features for Frame Rate Up Conversion,” Proc. of Signal-Image Technology & Internet-Based Systems (SITIS 15), 2015, pp.295-299.

[6] D. Glasner, S. Bagon, M. Irani, “Super-resolution from a single image,” in Proceedings of IEEE International Conference on Computer Vision, 2009, pp. 349-356.

[7] S. Farsiu, D. Robinson, M. Elad, P. Milanfar, “Fast and Robust Multiframe Super Resolution, “ IEEE Transactions on Image Processing, “Vol. 13, No. 10, , 2004, pp. 1327-1344.

[8] W. T. Freeman, T. R. Jones, E. C. Pasztor, “Example-based super-resolution,” IEEE Transactions on Computer Graphics and Applications, Vol.22, No.2, 2002, pp. 56-65.

[9] C. Dong, C. C. Loy, K. He, X. Tang, “Learning a Deep Convolutional Network for Image Super-Resolution,” Proc. of European Conference on Computer Vision (ECCV 14) , 2014, pp.1-16.

[10] J. Kim, J. Kwon Lee, and K. Mu Lee, “Accurate image super-resolution using very deep convolutional networks,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition(CVPR 16), 2016.

[11] J. Kim, J. Kwon Lee, and K. Mu Lee, “Deeply-Recursive Convolutional Network for Image Super-Resolution,” Proc. of IEEE Conference on Computer Vision and Pattern Recognition(CVPR 16), 2016.