PERCEPTUAL VIDEO CODING USING DEEP NEURAL NETWORK BASED JND MODEL

Jongho Kim, Dae Yeol Lee, Seyoon Jeong and Seunghyun Cho

ETRI

Contents

Introduction

- The **perceptual video coding or compression (PVC)** is one important branch of studies to **maximize the compression efficiency** or the perceptual quality **by applying human perceptual mechanisms to video compression.**

- **Just notice difference (JND)** is the one of the effective PVC methods which **can achieve the additional compression gain by reducing perceptual redundancies in video up to unnoticeable level**.

- In this paper, we proposed PVC method employs the pre-processing approach, where we **suppress the visual redundancy of the input video by applying the deep neural network (DNN) based JND model** before the encoding.

Proposed Algorithm

❑ Proposed DNN based JND Suppression Model

A. Main Goal

- The proposed JND model's goal is **to reduce the perceptual redundancy of the input video through a deep neural network (DNN)** and further improve the compression efficiency while minimally affecting the perceptual quality.

B. Network Structure of the Proposed JND Model (shown in Figure. 2&3)

- One **convolution layer consists of 64 filter of the size 3 × 3 × 64**, where a filter operates on **3 × 3 spatial regions across 64 channels**.
- Each ResBlock is a combination of **three pairs of convolution layer and rectified linear unit (ReLU)** and contains a **skip connection for local residual learning**.

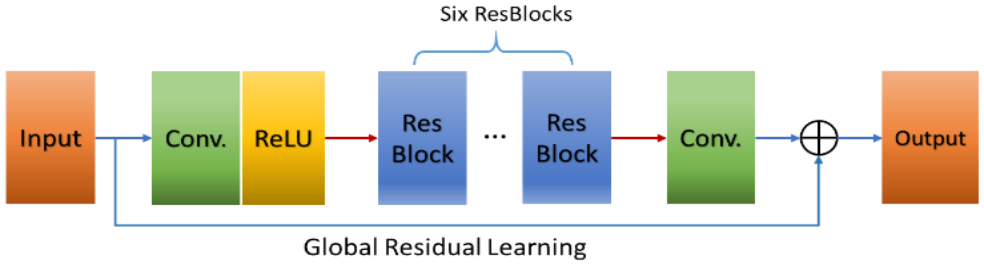


Figure 1: The architecture of the proposed DNN based JND model.

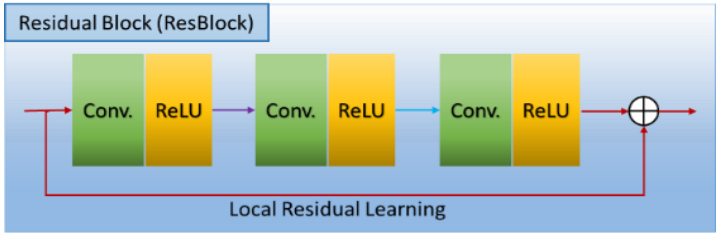


Figure 2: Residual Block (ResBlock) architecture**.**

Proposed Algorithm

C. Dataset: VideoSet (Video Subject Evaluation Test)

- The **'VideoSet' database** was used for training and verifying the performance of the proposed DNN based JND model**.**
- The VideoSet is a **large-scale compressed video quality dataset provided by USC.**
  - 220 video sequences which has 5s duration in four resolutions (i.e., 1920 × 1080, 1280 × 720, 960 × 540 and 640× 360) → total 880 sequences
  - Each of the 880 video clips are encoded using the H.264 codec with QP ranging from 0 to 51. → total 45,769 bitstream.
  - The JND point labels contain the QP values for the first, second, and third JND points of each sequence, which was acquired through large scale human study.

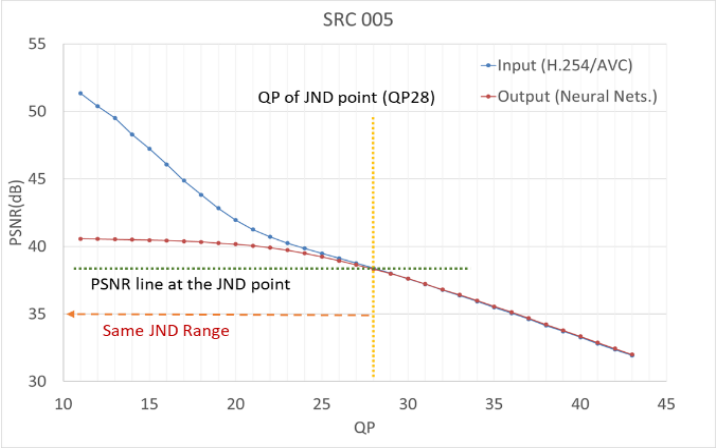
D. Training

- The purpose of the proposed JND model is **to remove the perceptual redundancy of the input image or video up to the point where it reaches the JND point**, we designed the cost function of the learning model to **minimize the L2 norm between the JND point video and the input video with smaller QPs than that of JND point.**
- Learning proceeds in **32x32 block patches** and it is carried out by optimizing the regression objective using **mini-batch gradient descent based on backpropagation**.

Proposed Algorithm

E. Verification

- To evaluate whether the proposed model properly exploits the JND characteristics of humans, we considered the following two points.
  - 1) If the input video has a smaller QP than that of the JND point, the proposed model can further throw away the negligible visual information to become compression-friendly, until it reaches the JND point, where the distortions start becoming noticeable.
  - 2) If the input image has a higher QP than the JND point, the proposed method should retain the input image as is, since the image already contains noticeable distortions.
- If we input the **video with QPs smaller than the JND point**, we can see that **the PSNR of the input video has fell to that of the JND point,** which means that **the JND network can suppresses the perceptual redundancy of input video to the unnoticeable range.**
- If we input the **video with QPs higher than the JND point**, we can see that the **output video mostly maintains its PSNR of the input video**, which means that **the network does not manipulate the input video once the image is outside of its JND range of interest**.



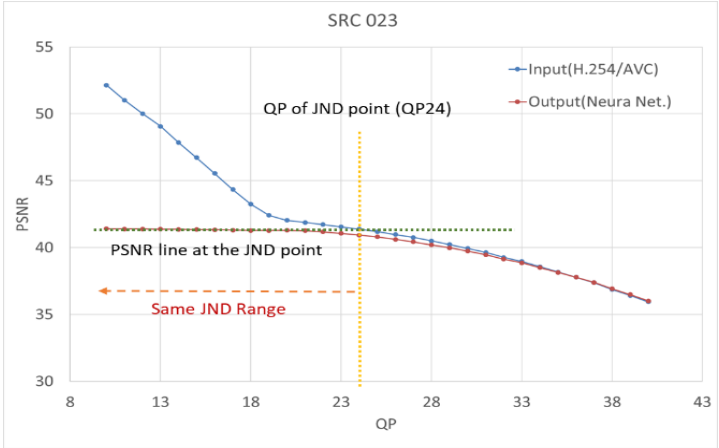


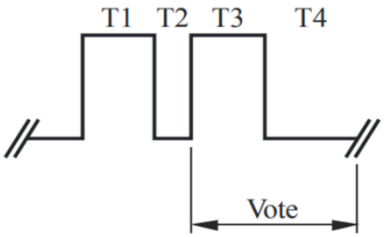
Figure 3: PSNR relationship between the input video and the output video when the videos with various QPs enter the network input.

Experimental Results

❑ Video Coding Experiment using JND Suppression Model

A. Test Condition

- The proposed JND model is a **pre-processing approach and suppresses the perceptually redundant information of the input video** using the DNN before encoding.
- Applied our JND model as a pre-processor for HM 16.0
- Experiments were conducted under **Random Access (RA) condition in JCT-VC common test condition.**
- For subjective quality experiment, we use the single stimulus method for adjectival categorical judgement in ITU-R BT.500. (shown in Figure. 4)



–3	Much worse
–2	Worse
–1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

Figure 4: Configuration of presentation and score rating for subjective quality assessments.

B. Test Results

- Average 19.74% of coding bits can be reduced with negligible loss in perceptual quality.
- Average 3.91% of encoding time can be reduced by eliminating the perceptual redundancy.

Table 1: Performance comparison for the HM16.0.

Sequence	Bit Reduction (%)	DMOS	Δ Time Saving (%)
VVC Class A (UHD)	18.17	-0.01	3.74
VVC Class B (FHD)	21.62	-0.29	4.12
Average	19.74	-0.14	3.91

Thank you !