|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding by Artificial Intelligencewww.mpai.community |

**Public document**

|  |  |
| --- | --- |
| N606 | 2022/03/23 |
| Source | Requirements (EVC) |
| Title | MPAI-EVC Evidence Project report and plan |
| Target | MPAI Members |

The goal of the group is to enhance EVC (Essential Video Coding) using AI-tools to reach at least 25% improvement over the baseline profile. The group is currently working on three coding tools: Intra prediction, Super Resolution, and in-loop filtering. For each tool, in the following we describe the proposed approach and the steps of database building, learning phase and inference.

**BVI dataset preparation**

MPAI-EVC has decided to use: **BVI-DVC Part 1 (University of Bristol)**

Data from<https://data.bris.ac.uk/data/dataset/3h0hduxrq4awq2ffvhabjzbzi1>

The sequences need a pre-processing to be used in the MPAI-EVC experiments. In the following the steps adopted by the group to prepare the sequences for future training.

Figure 1 illustrates the processing workflow.



Figure 1 processing workflow

We have started the coding at fixed QP (as per the Common Test Conditions) to provide references and the downsampled videos to be processed by the super-resolution network.

This activity is still ongoing and results will be available soon.

**Further datasets**

We added to the BVI-DVC the following datasets for a total of 350 4K sequences:

* Ultravideo dataset, containing 16 4K 10-bit raw sequences, available here <http://ultravideo.fi/>
* The opensource SVT datasets (7 new plus 5 old 4K sequences), available here <https://www.svt.se/opensource/content>
* The Tencent video dataset (85 4K sequences), available here <https://multimedia.tencent.com/resources/tvd>

Other datasets, such as the Youtube UGC dataset are being investigated.

**Intra prediction tool**

We address the challenge of predicting an intra-coded block given its context (Intra prediction) as an image inpainting problem, i.e. recovering pixels of an image that are unavailable due to, e.g. occlusions or information loss. Masked convolutional neural networks have been recently proposed for image inpainting exploiting the apriori information from the context to recover the missing image area. The method we propose relies on masked convolutions to generate the block predictor starting from a decoded context of 64 × 64 pixels (Figure 2). For example, for each 32x32 coding unit a 64x64 context is sent to the autoencoder. The autoencoder returns to the EVC encoder a 32x32 predictor that is considered as a 6th EVC Intra predictor mode that is put into competition with the other 5 predictors. The generated bitstream is fully decodable under the assumption that the autoencoder network is also available at the decoder side.



Figure 2: context con the left and the predictor on the right

The masked autoencoder (Figure 3) is trained in a supervised manner for 1000 Epochs over randomly sampled patches from about 800 images representing various types of contents by minimising absolute error (ABS) between its output and the original picture block.



Figure 3: Procedure for training the convolutional autoencoder used to generate the Intra predictor.

Table 1 shows the results of the proposed method over the standard JVET test sequences for the 22-37 and 22-47 QP ranges. The experiments show best results for, in the order, 720p, HD, 4K, 480p and 240p sequences, with BDRate gains from above 10% to just slightly below 5%. Our analysis of the training material showed that these gains distribution match the picture height distribution of the train sequences. About screen contents (class F), our training set does not include computer-generated sequences, and that may explain the comparatively low performance.



Table 1 Proposed Intra prediction method based on a convolutional autoencoder.

Future plans include:

* Enriching the training set with more diverse content including computer-generated screens
* Minimization of a better proxy of the coding rate
* Reducing the number of learnable parameters of the autoencoder

**Super-resolution tool**

The super-resolution step is added as a post-processing tool. The picture before encoding with EVC baseline profile is downscaled and then the super-resolution network is applied to the decoded picture to get the native resolution.

We have carried out extensive training of the selected deep-learning approach for super-resolution on 4 QPs (15,30,37 and 45). We have tested its performances on 8 test sequences for the case of SD to HD, and on 3 test sequences for the case of HD to 4K.

The group has worked on the computation of the BD-rate SD to HD and this has shown an improvement of -4.701% when compared with the ground truth EVC.

For the HD to 4K we are in the phase of calculating the BD-rate.

Because the training step for the case HD24K was carried out by different partners and on machines configurations that were substantially different among them. We have been carrying out substantial work for verification purposes. First, we have verified that usage of dual GPUs setting was not changing the quality performances when compared with the single GPU setting. Second, we have verified that among the two partners involved in this activity, the patches extraction to generate the training dataset, was producing the same dataset. Since the two above activities have provided an outcome where the results were showing no discrepancies among them, the only component of the working framework that could have affected the quality performances was training. After a careful examination, we have noted that the two partners were applying a slightly different approach for the training of the used deep learning approach. One option was training the deep learning approach from scratch. The other methodology was performing a so-called transfer learning approach where the original weights of the deep-learning approach, obtained from a different domain dataset, were used as a starting point. This has clearly provided different results among the two partners, and it has shown that training the neural network from scratch gives better performance compared to the fine-tuning starting from pre-trained weights. This activity is still ongoing on all QPs and the results will be available soon.

**In-loop filter**

Starting from the paper A Deep Learning Approach for Multi-Frame In-Loop Filter of HEVC we have started to go in-depth in the available code. This approach is implemented in HEVC and the plan is to port it into the EVC codec.

We are in the study phase to verify feasibility.

The next steps are:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **Date** | **Topic** | **Who** |
| Intra prediction | 1 meeting cycle | More experiments to improve the BD-rate | Attilio, Alessandra, Roberto |
| 2 meeting cycles | Adding 6-th Intra mode  | Attilio |
|  | 2 meeting cycles | Reduce ing number of parameters | Attilio |
|  | 2 meeting cycles | Find a proxy for the encoding rate | Attilio, Alessandra, Roberto |
| Super Resolution | 2 meeting cycles | More experiments to improve the BD-rate | Francesco, Antonio, Mattia and Alessandro |
|  | 1 meeting cycles | Verify patch extraction among partners | Francesco, Antonio, Mattia and Alessandro |
|  | 1 meeting cycles | Visual evaluation of the compressed test sequences | All |
| In-loop | 1 meeting cycle | Ask collaboration to the author of the available code (Ren Yang) | All |
|  | 1 meeting cycle | Run the Ren Yang software to review the performance on our test sequences  | Roberto |
|  | 1 meeting cycle | Have a look into the python Neural Network code  | All |
|  | 1 meeting cycle | Have a look at the deblocking filter inside EVC  | Alessandra |
|  | 1 meeting cycle | Have a look at the dataset for training inloop available on the git  | Alessandra |

**Future Plan**

* motion compensation: improve the motion compensation using NN architecture
* inter prediction: use NN architectures to refine the quality of inter-predicted blocks; introduce new inter prediction mode which tries to predict a frame directly without the use of side information; leverage on Optical Flow algorithm for the motion estimation.
* quantization: uniform scalar quantization used in classical video codec standards does not conform to the characteristics of the human visual system. It is possible to use a quantization strategy based on neural networks.
* arithmetic encoder: improve the CABAC performance by leveraging NN to directly predict the probability distribution of intra modes instead of the handcraft context models