Enhanced Video Compression based on Effective Bit depth Adaptation

Fan (Aaron) Zhang, Mariana Afonso and David Bull

University of Bristol

September 2019 @ IEEE ICIP, Taipei

Outline

Outline

Introduction

Context

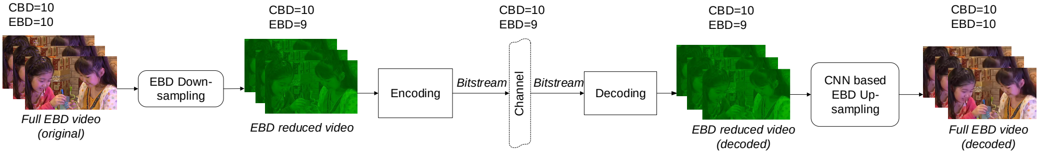
- Video compression is at the key position in ensuring **visual quality** and maintaining compatibility with the **transmission bandwidth**.
- New **standardised video coding** algorithms have been initiated including MPEG 's VVC [Bross *et al.*, 2019] and AOM's AV1 [AOM, 2019].
- **Machine learning** has started to play a more important role in video coding [Yeh *et al.*, 2018; Liu *et al.*, 2018], although it still remains an underdeveloped research area.
- **Resolution re-sampling** has also been employed in video coding but only for spatial and temporal resolutions [Ma *et al.*, 2012; Afonso *et al.*, 2019; Li *et al.*, 2018].

The Summary of Contributions

- The first CNN based **effective bit depth adaptation** approach (EBDA-CNN) for video compression.
- It achieves **consistent bit rate savings** over HEVC HM for content at various resolutions.
- An early version of this work was contributed by the University of Bristol (JVET-J0031) [Bull *et al.*, 2018] to the JVET "Call for proposals" for Versatile Video Coding (VVC) [Segall *et al.*, 2017].

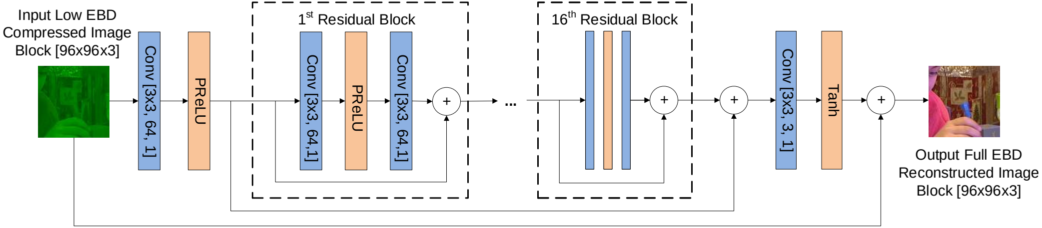
Outline

The Proposed EBDA Video Coding Framework



- **Coding bit depth (CBD)** is conventionally used to describe pixel bit depth in video coding and is defined as *InternalBitDepth* in HEVC HM codecs (10 bit in this paper).
- **Effective bit depth (EBD)**: this is the actual bit depth used to represent the video content, which is lower than or equals CBD in the proposed approach (9 bit or 10 bit in this paper).
- It reduces the effective bit depth of an input video before encoding, and reconstructs the original bit depth at the decoder.
- A fixed **QP offset** of -6 is applied on the initial base QP value when adaptation is enabled to obtain similar bit rates with full EBD coding (without adaptation).
- A modified CNN was employed at the decoder for full bit depth reconstruction.

The Employed CNN Architecture



- The proposed CNN architecture is a modified from **SRResNet** [Ledig and , 2017] – the generator of SRGAN that was developed for super-resolution.
- **Batch normalisation** (BN) layers have not been used here, as they were reported to decrease image feature variability and influence overall performance [Lim *et al.*, 2017].
- The **loss function** employed for training the network is $\ell 1$ loss rather than $\ell 2$ based on the results reported in [Johnson *et al.*, 2016].

The CNN Training and Evaluation

- **Eighty 10 bit video sequences** at various spatial resolutions were used for training the CNN, each of which was converted to 9 bit and compressed by HEVC HM 16.20 using **four initial base QP values** (22, 27, 32 and 37).
- Compressed and corresponding original frames from each QP group were **randomly selected** and split into 96$\times$96 pixel blocks. This results in **15,000 pairs** of input/target image blocks for each group.
- The CNN was built and trained using **Tensorflow** (1.8.0) [Martín , 2015] using the following parameters: Adam optimisation [Kingma and Ba, 2015], batch size of 16, learning rate of 10^{-4}, weight decay of 0.1 and 200 epochs.
- This results in four CNN models for different QP values in evaluation:

$$\mathrm{CNNs} = \begin{cases} \mathrm{model}_1, \text{ if} & \mathrm{QP}_{\mathrm{base}} \leq 24.5 \\ \mathrm{model}_2, \text{ if} & 24.5 < \mathrm{QP}_{\mathrm{base}} \leq 29.5 \\ \mathrm{model}_3, \text{ if} & 29.5 < \mathrm{QP}_{\mathrm{base}} \leq 34.5 \\ \mathrm{model}_4, \text{ if} & \mathrm{QP}_{\mathrm{base}} \geq 34.5 \end{cases} \tag{1}$$

- In the evaluation phase, each frame is also split into 96×96 overlapping blocks, with an **overlap size** of 4.
- The full EBD blocks produced by the CNN (output) are then **aggregated in the same way** to form a final reconstruction frame.

Outline

Algorithm Evaluation Configuration

- The proposed approach was **integrated into HEVC HM 16.20**, and was evaluated on JVET Common Test Conditions [Bossen *et al.*, 2019] using the Random Access configuration (Main10 profile).
- All the test sequences are from **JVET CTC SDR video class A1, A2, B, C and D**, different from those used in CNN training.
- Results for EBD up-sampling with **bit shifting** were also generated for additional benchmarking.
- All the results are based on **Bjøntegaard Delta (BD) measurements** [Bjøntegaard, 2001] over all frames using PSNR (Y channel only).
- Both encoding and decoding were executed on a **shared cluster, BlueCrystal Phase 4** based in the University of Bristol.
- The encoding jobs were run on nodes with 14 core 2.4 GHz Intel E5-2680 v4 (Broadwell) CPUs with 128GB of RAM.
- The decoding jobs were run on GPU nodes with an **additional graphic card** NVIDIA P100.

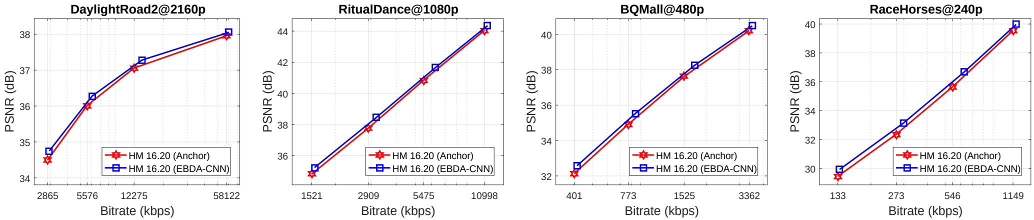
Compression Results and Complexity

Class-Sequence	EBDA-CNN		EBDA-w/o CNN	
	BD-Rate	BD-PSNR	BD-Rate	BD-PSNR
A1-Campfire	-11.4%	+0.18dB	-9.8%	+0.17dB
A1-FoodMarket4	-4.2%	+0.13dB	+1.4%	-0.04dB
A1-Tango2	-6.1%	+0.09dB	-0.0%	+0.01dB
A2-CatRobot1	-7.9%	+0.14dB	-0.1%	+0.01dB
A2-DaylightRoad2	-8.5%	+0.11dB	+0.6%	+0.00dB
A2-ParkRunning3	-12.0%	+0.50dB	-7.8%	+0.32dB
Class A	**-8.4%**	**+0.19dB**	**-2.6%**	**+0.08dB**
B-BQTerrace	-7.5%	+0.12dB	-0.2%	+0.02dB
B-BasketballDrive	-7.4%	+0.17dB	-1.7%	+0.04dB
B-Cactus	-1.4%	+0.03dB	-0.3%	+0.00dB
B-MarketPlace	-2.7%	+0.07dB	+2.3%	-0.06dB
B-RitualDance	-4.4%	+0.21dB	+1.2%	-0.05dB
Class B	**-4.7%**	**+0.12dB**	**+0.3%**	**-0.01dB**

Class-Sequence	EBDA-CNN		EBDA-w/o CNN	
	BD-Rate	BD-PSNR	BD-Rate	BD-PSNR
C-BQMall	-5.6%	+0.21dB	+1.6%	-0.06dB
C-BasketballDrill	-6.0%	+0.26dB	+1.2%	-0.05dB
C-PartyScene	-3.7%	+0.16dB	+1.6%	-0.06dB
C-RaceHorses	-6.3%	+0.24dB	-2.0%	+0.08dB
Class C	**-5.4%**	**+0.22dB**	**+0.6%**	**-0.02dB**
D-BQSquare	-7.0%	+0.25dB	+2.1%	-0.07dB
D-BasketballPass	-7.2%	+0.36dB	-0.8%	+0.05dB
D-BlowingBubbles	-4.5%	+0.18dB	+1.3%	-0.05dB
D-RaceHorses	-6.9%	+0.34dB	-1.2%	+0.06dB
Class D	**-6.4%**	**+0.28dB**	**-0.1%**	**+0.01dB**
Overall	**-6.4%**	**+0.20dB**	**-0.6%**	**+0.02dB**

The average encoding time is **1.02**$\times$ that of the original HM 16.20, while the average decoding time is **69.3 times** that of HM.

Example Rate-PSNR Curves



(a) DaylightRoad2 (b) RitualDance (c) BQMall (d) RaceHorses

- Without using CNN-based up-sampling, EBDA-w/o CNN does **not provide any significant improvement** in coding efficiency.
- EBDA-CNN has achieved (although different) **coding gains for all test sequences**.
- The improvement is **consistent across the whole tested QP range** for all test content.

Outline

Conclusions and Future Work

Conclusions

- An **effective bit depth adaptation** (EBDA-CNN) approach has been presented for video coding.
- It reduces effective bit depth (EBD) by 1 bit before encoding, and reconstructs full bit depth at the decoder using a deep CNN based up-sampling method.
- This approach has been integrated into HEVC reference codec **HM 16.20** and fully evaluated on JVET CTC test sequences.
- The results show that **consistent coding gains** can be achieved for all tested sequences, with an average BD-rate of -6.4%.

Future work

- To reduce the **complexity** of the CNN for EBD up-sampling.
- To extend its application on **higher dynamic range** (bit depth) content.
- To assess and investigate **the subjective quality** of the EBDA reconstructed content.

Acknowledgements

EPSRC
Investing in research for discovery and innovation
Platform Grant: Vision for the Future:
EP/M000885/1

NVIDIA GPU Seeding Grants

When the Proposed EBDA is integrated into VTM 4.01

Class-Sequence	BD-Rate
A-Campfire	-11.3%
A-FoodMarket4	-0.6%
A-Tango2	-1.9%
A-CatRobot1	-3.6%
A-DaylightRoad2	-5.8%
A-ParkRunning3	-17.0%
Class A	**-6.7%**

Class-Sequence	BD-Rate
B-BQTerrace	+0.1%
B-BasketballDrive	-3.9%
B-Cactus	-4.0%
B-MarketPlace	+4.2%
B-RitualDance	-2.2%
Class B	**-1.2%**

Class-Sequence	BD-Rate
C-BQMall	-2.1%
C-BasketballDrill	+0.7%
C-PartyScene	-3.2%
C-RaceHorses	-3.3%
Class C	**-2.0%**

Class-Sequence	BD-Rate
D-BQSquare	-1.8%
D-BasketballPass	-4.6%
D-BlowingBubbles	-2.6%
D-RaceHorses	-5.6%
Class D	**-3.7%**
Overall	**-3.6%**

Here the CNN model was re-trained with VTM 4.01 compressed content.

References I

M. Afonso, F. Zhang, and D. R. Bull. Video compression based on spatio-temporal resolution adaptation. *IEEE Trans. on Circuits and Systems for Video Technology*, 29(1):275–280, January 2019.

AOM. AOMedia Video 1 (AV1), 2019.

G. Bjøntegaard. Calculation of average PSNR differences between RD-curves. In *13th VCEG Meeting*, number VCEG-M33, Austin, Texas, USA, April 2001. ITU-T.

F. Bossen, J. Boyce, X. Li, V. Seregin, and K. Sühring. Jvet common test conditions and software reference configurations for sdr video. In *the JVET meeting*, number JVET-M1001. ITU-T and ISO/IEC, 2019.

B. Bross, J. Chen, and S. Liu. Versatile video coding (draft 6). In *the JVET meeting*, number JVET-O2001. ITU-T and ISO/IEC, 2019.

D. Bull, F. Zhang, and M. Afonso. Description of SDR video coding technology proposal by University of Bristol (JVET-J0031). In *the JVET meeting*, number JVET-J0031, San Diego, US, April 2018. ITU-T and ISO/IEC.

J. Johnson, A. Alahi, and F.-F. Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

C. Ledig and *et al.* Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114. IEEE, 2017.

Yue Li, Dong Liu, Houqiang Li, Li Li, Feng Wu, Hong Zhang, and Haitao Yang. Convolutional neural network-based block up-sampling for intra frame coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2316–2330, 2018.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

J. Liu, S. Xia, W. Yang, M. Li, and D. Liu. One-for-all: Grouped variation network based fractional interpolation in video coding. *IEEE Transactions on Image Processing*, 2018.

References II

Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang. Modeling of rate and perceptual quality of compressed video as functions of frame rate and quantization stepsize and its applications. *IEEE Trans. on Circuits and Systems for Video Technology*, 22(5):671–682, 2012.

A. Martín *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

A. Segall, V. Baroncini, J. Boyce, J. Chen, and T. Suzuki. Joint call for proposals on video compression with capability beyond hevc. In *the JVET meeting*, number JVET-H1002, Macao, CN, October 2017. ITU-T and ISO/IEC.

C.-H. Yeh, Z.-T. Zhang, M.-J. Chen, and C.-Y. Lin. HEVC intra frame coding based on convolutional neural network. *IEEE Access*, 6:50087–50095, 2018.