Next Generation Video coding in 8K era - Versatile Video Coding and AI

Tomohiro Ikai, Eiichi Sasaki, Yukinobu Yasugi, Tomonori Hashimoto, Tianyang Zhou, Takeshi Chujoh, Tomoko Aono, Norio Itoh

[1]Sharp Corporation

Keywords: Versatile Video Coding, CNN, Video Super Resolution

ABSTRACT

Displays and video compression are key drivers in emerging 4K/8K and VR/AR video market. Versatile Video Coding (VVC), which is under development as the next generation video coding, inevitably changes our society in the 2020s. This paper shows VVC key components including simplification and improvement aspects and shows neural network's difficulty and significance in compressed video.

1 INTRODUCTION

Displays are key devices for video consumptions, where image quality is one of the most important issues. Video codec as well plays a crucial role for this purpose where traffic bandwidth is also concerned. In July 2020, Versatile Video Coding (VVC) is to be standardized, which promises to be more powerful than Advanced Video Coding (AVC) and High Efficiency Video Coding (HEVC).

After standardization of HEVC ratified in January 2013 and a success of joint exploration work [1] since October 2015, a call for proposals was issued in October 2017 jointly by ITU-T SG16 Q6 (VCEG) and ISO/IEC JTC 1/SC 29/W G11 (MPEG) [2]. Those call has three categories of SDR, HDR and 360 and 33 organization applied for the call [3]. Currently, the VVC working draft 6 has been developed and VVC test model (VTM) can be found in [4].

AI technologies are rapidly evolving thanks to a large amount of data and learning skill's improvement, which is also becoming relevant for image quality.

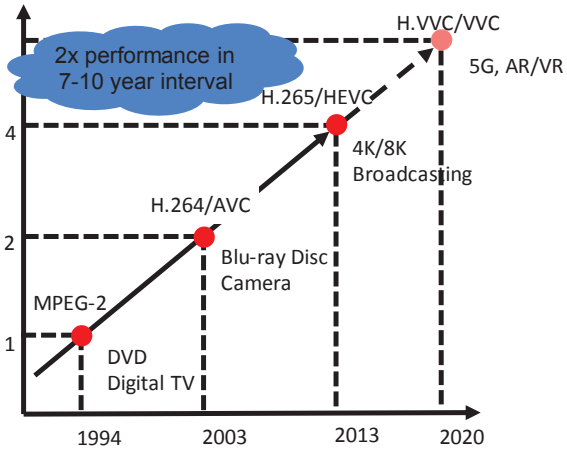

Fig. 1 Development of standard video codec

2 Versatile Video Coding

Versatile Video Coding Test Model (VTM), the reference software has been developing and the version 6 (VTM6) shows the coding gain of 39 % compared to HEVC for 4K videos as shown in Fig. 2. The gain is based on PSNR and subjective improvement is thought to be more than that (it can be up to 65 %).

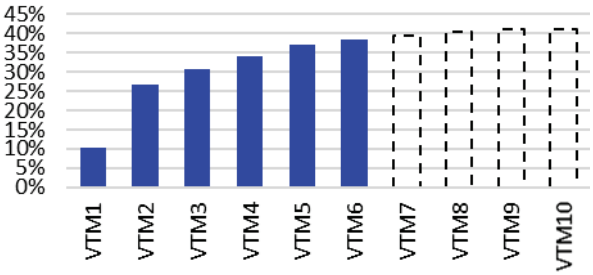

Fig. 2 Coding performance progress in VTM

2.1 VVC coding tools and its characteristics

VVC has a lot of coding tools with coding performance in Table 1 [5]. Here we'd like to discuss what is the point of these new tools. In our understanding, the computation power and better modelling is going on behind the scene. Also interesting if we categories coding tools into the following two modelling concepts.

- **Human centric model** (parametric model)
- **Data centric model** (learned model)

Fig. 3 and Fig. 4 shows its positioning.

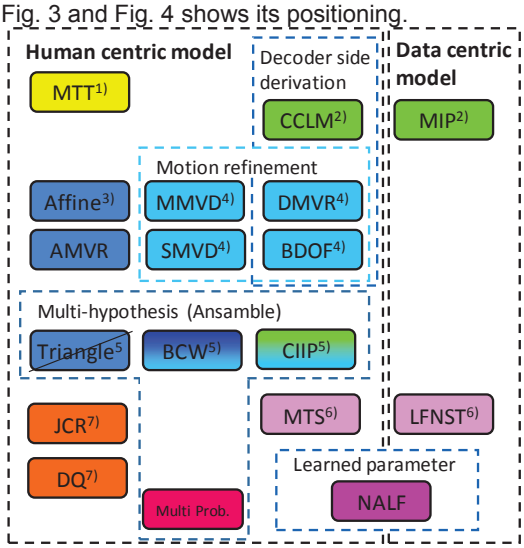

Fig. 3 Categorize of coding tools on modelling

Table 1 VVC coding tools and its performance

	Y	U	V
Partition[1]			
Multi Tree Type Partitioning (**MTT**)			
Chroma Separate Tree	0.14%	6.05%	7.57%
Intra Prediction[2]			
CCLM	**0.90%**	**15.85%**	**16.95%**
MIP	0.27%	0.26%	0.32%
MRLP	0.20%	0.03%	-0.02%
ISP	0.13%	0.04%	0.08%
Inter Prediction[3-5]			
Affine [3]	**2.53%**	**1.83%**	**1.75%**
AMVR	**1.05%**	**1.51%**	**1.48%**
TMVP	**1.19%**	**0.93%**	**1.03%**
SbTMVP	0.43%	0.29%	0.31%
DMVR [4]	0.82%	1.05%	1.07%
BDOF [4]	0.78%	0.24%	0.15%
MMVD [4]	0.58%	0.64%	0.64%
SMVD [4]	0.25%	0.25%	0.23%
CIIP [5]	0.38%	0.08%	0.02%
Triangle [5]	0.35%	0.63%	0.67%
BCW [5]	0.43%	0.53%	0.58%
SBT	0.41%	-0.02%	0.07%
Transform[6]			
Multiple Transform Selection (**MTS**)	0.33%	0.51%	0.42%
Low Freq.Non-Separable Transform (**LFNST**)	0.79%	0.39%	0.81%
Residual Coding[7]			
Dependent Quantization (**DQ**)	**1.71%**	**-0.51%**	**-0.72%**
Joint Chroma Residual (**JCR**)	0.25%	0.65%	0.72%
Loop filter and Loop process[8]			
Non-linear Adaptive Loop Filter (**NALF**)	**4.91%**	**4.56%**	**4.07%**
Luma mapping with chroma scaling (LMCS)	**1.39%**	**-2.83%**	**-2.55%**
Entropy Coding [10]			
Multi. Prob. CABAC	0.94%	1.00%	0.79%

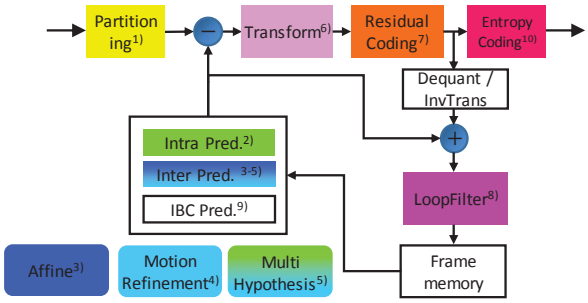

Fig. 4 Core coding tools in encoding structure

The **Human centric model** has been used for standards with an advantage of less computing resources but evolves with the following aspects.

- **Multi parametric motion**: Affine prediction
- **Decoder-side derivation** in Chroma Component Linear Model (CCLM), Decoder-side Motion Vector Refinement (DMVR), Bi-Directional Optical Flow (BDOF)
- **Motion refinement** in Merge Mode with motion Vector Difference (MMVD), Symmetric Motion Vector Difference (SMVD), DMVR, and BDOF
- **Multi hypothesis** in Combined Intra Inter Prediction (CIIP), Triangle partitioning, Bi-prediction with CU Weights (BCW), Multi. Prob. CABAC
- **State machine** in Dependent Quantization (DQ)

Typical equations are:

$$CCLMSamples[x][y] = (w * pred[x][y]) >> k + b$$

$$CIIPSamples[x][y] = (w * predIntra[x][y] + (4 - w) * predInter[x][y])>>2$$

where the number of parameters (i.e. w) is 1 or 2.

On the other hand, **Data centric model** is included in the standard for the first time:

- Prediction with **learned weights** in Matrix based Intra Prediction (MIP)
- Transform with **learned kernel** in Low Frequency Non Separable Transform (LFNST)
- Filter with **learned weights and clipping values** in Non-linear Adaptive Loop Filter (NALF)

Typical equations are:

$$MIPSamples[x][y]=((\Sigma w[i][x][y] * p[i] + oW) >> sW) + b$$

$$LFNSTSamples[x][y] = (\Sigma w[j][i] * coeff[i]+64) >> 7$$

$$ALFSamples[x][y] = (\Sigma w[i] * ref[i] + 64) >> 7$$

where i = 0..N-1, N is 4 (MIP), 8 or 16 (LFNST), 12 (ALF)

It is obvious that more params. and computations, which become doable in the 2020s would be this basis but in details all are with linear equations and the number of operations is under roughly 16 multiplies per pixels for each specific tool. In this sense, video codec evolution is linear one. It seems more than 3 weight params. in equation needs **Data centric model** while 1-3 params. can be specified by **Human centric model**.

VTM1 started with the most powerful tool MTT where recursive multi-type (quad, ternary, binary) tree partitioning provides super compact and sufficient modeling for particular area, VTM 2 adopted basic tools CCLM, Adaptive Motion Vector Resolution (AMVR), Sub-block Temporal motion vector prediction (SbTMVP), and ALF, VTM3 / VTM4 adopted multi. param. Affine, Decoder-side motion derivation (BDOF and DMVR), Multi-hypothesis and IntraBlockCopy (IBC) which are all powerful but needs special consideration for complexity, and VTM5 adopted data centric tools of MIP and LFNST where we human needs to believe and confirm the learned models . VTM6 adopted a scalable functionality.

2.2 VVC development in terms of simplification and improvements

Let's see a set of interesting adopted contributions, which significantly reduces the complexity or improvement. First a CCLM simplification [6] replaces two big tables with just 48 bits. Specifically,

A original 16 bit 512 entry table

{65536, 32768, 21845, 16384, 13107, 10922, 9362, 8192, 7281, 6553, 5957, 5461, 5041, 4681, 4369, 4096, 3855, 3640…, 128, 128, 128} was replaced with 3 bit 8 entry one

{0, 7, 6, 5, 5, 4, 4, 3, 3, 2, 2, 1, 1, 1, 1, 0}

It's 99.7 % reduction.

Affine motion (multi-parametric motion) extended conventional 2 translation params, of *e*, *f* to 4 params. [7] of *a*=*d*, *b*=*c*, *e*, *f* and 6 params. [8] of *a*, *b*, *c*, *d*, *e*, *f*.

$$vx = a * dx + b * dy + e \quad \text{(eq.1)}$$

$$vy = c * dx + d * dy + f$$

where 6 params. are useful to model true motion with shape changes (different scaling factor in horizontal and vertical directions) in contrast to 4 params. for zoom and rotation. Another interesting thing is BDOF [9], where optical flow (gradient based pixel level motion vector information) is utilized to improve prediction in decoder side. More than 10 bit support is improved in [10]. At the latest Gothenburg meeting, the optical flow based pixel level motion refinement is further used in Affine as predicted refined optical flow (PROF) [11].

Regarding functionality. In VVC, reference picture resampling (RPR) for adaptive resolution change and scalable coding / sub-picture, flexible tile[12] and wavefront CTU line [13] segmentation for parallel processing and region controls / gradual image refresh for low delay [14] / wrap-around prediction in Equirectangular and loop filter control in arbitrary horizontal/vertical lines [15] for 360 video projection are introduced.

3 AI era video coding

3.1 Progress of Neural network based video process

In the area of SISR (single image super resolution), EDSR [16] and R-CAN [17] shows impressive image quality by employing depth to resolution technique [18] and local / global residuals network with Dense or SE-NET For video processing, NN based optical flow [19] is included as end-to-end learning [20] based on spatial transformer network [21], where image generation and optical flow is simultaneously learned, which is then used in super-resolution [22] and frame rate up conversion [23]. Subjective quality conscious loss functions, such as perceptual loss [24] and GAN loss [25] becomes common. Recently video super resolution employs a re-current network, which reuses generated HR images with optical flow network [26][27]. Degradation process, e.g. blurring kernels, motion-blur etc. is addressed in [28] but the process has not been fully tested in compressed video except for video loop filtering [29][30].

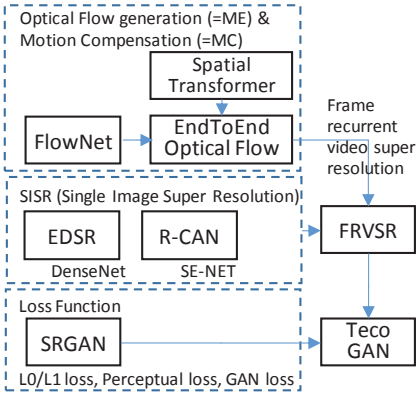

Fig. 5 Progress of Neural network based video process

3.2 Loop filtering and super resolution for video compression

The NN based technologies are promising for its power of generating realistic image. There're two difficulties to realize that technology for compressed video.

- Computational complexity
- Compression artifacts

At NN based image recognition, the computation complexity can be about 224 * 224 * number of weights / 100 [31] while at NN based filtering [32], it becomes about 3840 * 2160 * number of weights. Thus the ratio is 224*224/100 : 3840*2160 = 1 : 16500. In other words, we need **1/10000** network size for image processing to achieve the same level real time operation. One possible solution is to use really small size network learned for each specific bitstream, where only 1440 weight params. is used and signalled [33]. This is an incredibly small network considering hundreds of millions is ok in super-resolution research. We've also proposed a relatively small 7k params. loop filter [35] to JVET with spatial separable convolution (3x1 and 1x3 kernels instead of 3x3 kernels) and SE-Net (Attention) [34] is used.

Current super-resolution techniques might be not so effective where known or estimated degradation process is a key factor of detailed generation. In video compression, degradation process is very different in block by blocks and frames by frames.

We employed a frame recurrent video super resolution for video compressed image with 3x3 kernels with 64 channels, which shows progress compared to single image super resolution in Fig 6. The benefit of multi-frame based super-resolution (FRVSR [26]) is confirmed if the compression ratio is not large.

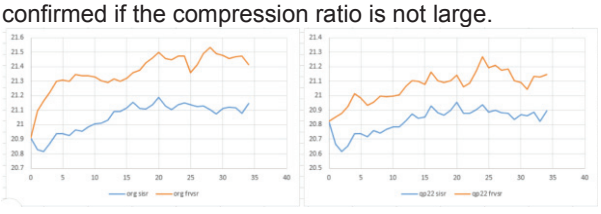

Fig. 6 Video based super resolution (Left: raw video, Right: compressed video with qp 22)

4 CONCLUSIONS

We present core technologies and tool development in Versatile Video Coding and discuss the state-of-the-art neural network based video processing with our experiment in compressed video.

REFERENCES

[1] "Algorithm description of Joint Exploration Test Model 7 (JEM7)," JVET-G1001 (2017).

[2] "Joint Call for Proposals on Video Compression with Capability beyond HEVC", N17195, MPEG (2017).

[3] Report of results from the Call for Proposals on Video Compression with Capability beyond HEVC, JVET-J1003 (2018).

[4] "Versatile Video Coding (Draft 6)", JVET-O2001

[5] JVET AHG report: "Tool reporting procedure (AHG13), " JVET-N0013 (2019).

[6] Y. Yasugi, F. Bossen, E. Sasaki, "Non-CE3: CCLM table reduction and bit range control," JVET-M0064 (2019).

[7] L. Li, H. Li, D. Liu, Z. Li, H. Yang, S. Lin, H. Chen, F. Wu, "An Efficient Four-Parameter Affine Motion Model for Video Coding," IEEE TCSVT, Vol. 28 , Issue: 8 , Aug. (2018).

[8] F. Zou, J. Chen, M. Karczewicz, X. Li, H.-C. Chuang, W.-J. Chien, "Improved affine motion prediction", JVET-C0062 (2016).

[9] A. Alshin ; E. Alshina, "Bi-directional Optical Flow for Future Video Codec," IEEE DCC (2016).

[10] T. Chujoh, T. Ikai , Non-CE9: An improvement of BDOF, JVET-M0063 (2019).

[11] J. Luo, Y. He,"CE2-related: Prediction refinement with optical flow for affine mode, "JVET-N0236 (2019).

[12] S. Deshpande, Hendry, Y.-K. Wang, M. M. Hannuksela, Y. He, L. Chen, W. I. Choi, B. D. Choi, R. Sjöberg, R. Skupin, "AHG12: On Tile Grouping", JVET-M0853 (2019).

[13] T. Ikai, "AHG12: One CTU delay wavefront parallel processing," JVET-N0150 (2019).

[14] S. Deshpande, Y.-K. Wang, Hendry, R. Sjöberg, M. Pettersson, L. Chen, "Gradual Random Access," JVET-N0865 (2019).

[15] S.-Y. Lin, L. Liu, J.-L. Lin, Y.-C. Chang, C.-C. Ju, P. Hanhart, Y. He, "AHG12: Loop filter disabled across virtual boundaries," JVET-N0438 (2019).

[16] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," IEEE Conf. CVPR (2017).

[17] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, "Image Super-Resolution Using Very Deep Residual Channel Attention Networks, " ECCV (2018).

[18] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken,R. Bishop, D. Rueckert, Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," IEEE Conf. CVPR (2016).

[19] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, T. Brox, "FlowNet: Learning Optical Flow with Convolutional Networks," IEEE ICCP (2015).

[20] Jason J. Yu, Adam W. Harley, K. G. Derpanis, "Back to Basics: Unsupervised Learning of Optical Flow via Brightness Constancy and Motion Smoothness," arXiv:1608.05842 (2016).

[21] M. Jaderberg K. Simonyan A. Zisserman K. Kavukcuoglu, "Spatial Transformer Networks, "Conf. on NIPS (2015).

[22] J. Caballero, C. Ledig, A. Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, Wenzhe Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," arXiv:1611.05250 (2016).

[23] H. Jiang, Deqing Sun, V. Jampani, M.-H. Yang, E. L.-Miller, Jan Kautz, "Super SloMo: High Quality Estimation of Multiple Intermediate Frames for Video Interpolation," arXiv:1712.00080 (2017).

[24] J. Johnson, A. Alahi, Li Fei-Fei, Perceptual Losses for Real-Time Style Transfer and Super-Resolution, ECCV (2016).

[25] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network", IEEE ICIP (2018).

[26] M. S. M. Sajjadi, R. Vemulapalli, M. Brown, "Frame-Recurrent Video Super-Resolution," IEEE ICIP (2018).

[27] M. Chu, Y. Xie, L. Leal-Taixé, N. Thuerey, "Temporally Coherent GANs for Video Super-Resolution (TecoGAN)", arXiv:1811.09393 (2018).

[28] K. Zhang, W. Zuo; L. Zhang, "Deep Plug-and-Play Super-Resolution for Arbitrary Blur Kernels", arXiv: 1903.12529 (2019).

[29] L. Zhou, X. Song, J. Yao, L. Wang, and F. Chen, "Convolutional Neural Network Filter (CNNF) for intra frame," JVET-I0022, Gwangju, KR, January (2018).

[30] J. Yao, X. Song. S. Fang, and L. Wang, "AHG9:Convolutional Neural Network Filter for inter frame," JVET-J0043, San Diego, US, April (2018).

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," In: Proceedings of the IEEE Conf. CVPR. pp. 4510–4520 (2018).

[32] Y.-L. Hsiao, O. Chubach, C.-Y. Chen, T.-D. Chuang, C.-W. Hsu, Y.-W. Huang, S.-M. Lei, "CE10-1.2: Convolutional neural network loop filter,"JVET-O0056 (2019).

[35] E. Sasaki, T. Hashimoto, T. Chojoh, T. Ikai, "SENet-CNN filter using spatial separable convolution network" (in Japanese), P-4-3, PCSJ (2018).