INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG04
MPEG VIDEO CODING

ISO/IEC JTC 1/SC 29/WG 04 m58048
October 2021, Online

Title: Decoder-side depth estimation with input depth assistance
Source: PUT: Dominika Klóska, Dawid Mieloch, Adrian Dziembowski, Marek Domański
ETRI: Gwangsoon Lee, Jun Young Jeong

Abstract

This document presents a description of the extension of the MIV DSDE, where we send depth maps for the subset of views. These depth maps help the IVDE to obtain better quality, and to decrease the computational time of depth estimation. The results show, that the proposed approach performs better than the G17 anchor in terms of quality and depth estimation time.

The recommendation is to open an EE, to test the possibility of using the proposed scheme as one of the MIV anchors.

1 Proposed approach

In the proposal, which is the combination of V17 and G17 configuration, the geometry is sent only for views packed into the first of three texture atlases.

V17:

G17:

Proposal:

For basic views from the 2nd, and 3rd atlas, the geometry information is not being sent at all. For these views, depth maps are estimated at the decoder side, basing on textures and depth maps already available in the decoder.

IVDE was updated to allow using input depth maps during the estimation process. This way, the depth estimation becomes much faster and more robust, as the presence of input depth maps for some views decreases the number of depth candidates for other views (because of the inter-view consistency term in the IVDE optimization step).

In order to provide good quality depth information for the large part of the scene, the basic views are reshuffled (when compared to the G17 and V17 anchor), and the first atlas contains the most distant views (chosen by the TMIV view selector/labeler launched for the 2nd time, only for basic views).

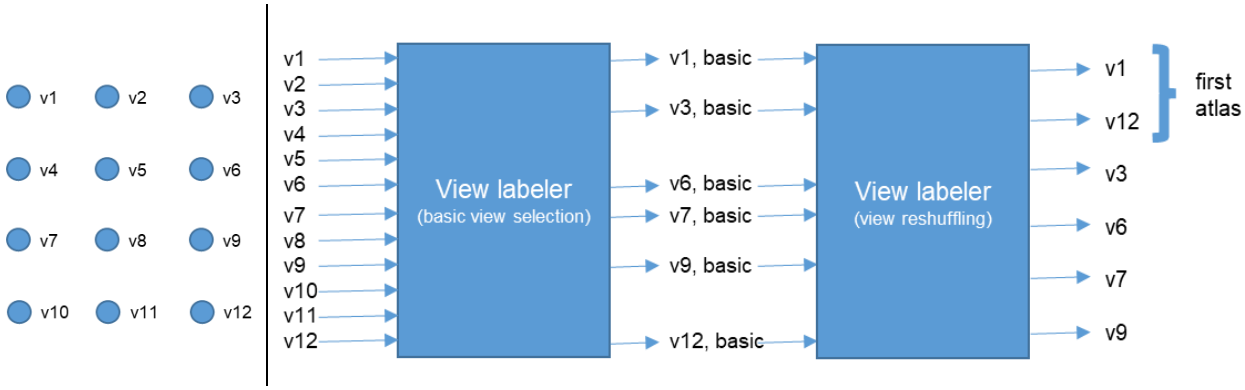

Fig. Basic view reshuffling in case, where an atlas contains 2 views. Left: camera arrangement, right: view selection process (basic views are additionally processed to find most distant ones).

The proposal utilizes syntax elements already available in the MIV Extended profile – the `vps_geometry_video_present_flag[ atlasID ]` is set to 0 for `atlasID` > 0 (in MIV Main it has to be set to 1).

2 Results

The proposal was tested on all perspective content (mandatory + optional) and compared to the A17, V17, and G17 anchors.

A17 vs. the proposal:

Mandatory content - Proposal vs. Low/High-bitrate Anchors

Sequence		High-BR BD rate Y-PSNR	Low-BR BD rate Y-PSNR	Max delta Y-PSNR	High-BR BD rate IV-PSNR	Low-BR BD rate IV-PSNR
Fan	O	---	-75.9%	9.17	-60.6%	-56.5%
Kitchen	J	-2.9%	-11.5%	13.54	156.0%	50.7%
Painter	D	-61.0%	-57.3%	6.67	-46.9%	-48.6%
Frog	E	-61.3%	-54.7%	6.83	-47.4%	-46.4%
Carpark	P	7.0%	-9.5%	9.81	12.9%	-7.4%
Group	R	-30.8%	-42.6%	19.49	44.5%	-7.2%
MIV		---	---	**13.00**	---	---

Runtime ratio (%)

Atlas encoding	Video encoding	Decoding & Rendering
3.2%	86.4%	2304.0%
4.9%	81.3%	1621.0%
2.6%	70.2%	2315.8%
2.7%	74.7%	3344.0%
4.1%	106.5%	977.2%
2.3%	82.5%	1719.5%
3.6%	84.5%	1608.7%

Optional content - Proposal vs. Low/High-bitrate Anchors

Fencing	L	0.9%	-33.3%	12.62	-9.7%	-34.1%
Hall	T	---	43.7%	16.00	71.4%	5.9%
Street	U	-46.7%	-29.7%	7.01	-14.1%	-9.5%
Mirror	I	-19.3%	-33.6%	14.27	-0.8%	-22.0%
Cadillac	G	-60.3%	-67.5%	14.59	-31.7%	-54.7%
MIV		---	---	**17.47**	---	---

3.1%	71.8%	1015.0%
3.4%	81.4%	821.9%
3.9%	85.8%	879.2%
2.5%	74.7%	1641.7%
4.4%	56.6%	1926.6%
3.4%	73.5%	1105.5%

For almost all perspective sequences, the proposal provides better BD rates than A17, especially for low bitrates. The results show also that the proposal provides a reduction of the complexity of the encoder (30 times faster). The rendering process is longer, as it now includes depth estimation. The experiment was also done in configuration with faster depth estimation (smaller number of superpixels and optimization cycles in IVDE).

A17 vs. the proposal with faster depth estimation:

Mandatory content - Proposal vs. Low/High-bitrate Anchors

Sequence		High-BR BD rate Y-PSNR	Low-BR BD rate Y-PSNR	Max delta Y-PSNR	High-BR BD rate IV-PSNR	Low-BR BD rate IV-PSNR
Fan	O	---	-75.8%	9.51	-59.5%	-55.8%
Kitchen	J	43.0%	3.8%	14.34	258.9%	59.9%
Painter	D	-60.1%	-56.7%	7.07	-45.1%	-47.7%
Frog	E	-61.1%	-54.6%	6.86	-47.2%	-46.3%
Carpark	P	11.6%	-8.1%	9.85	16.8%	-5.6%
Group	R	276.1%	-31.0%	19.82	######	56.0%
MIV		---	---	**13.21**	---	---

Runtime ratio (%)

Atlas encoding	Video encoding	Decoding & Rendering
3.2%	86.4%	697.4%
4.9%	81.3%	469.0%
2.6%	70.2%	671.3%
2.7%	74.7%	977.3%
4.1%	106.5%	327.6%
2.3%	82.5%	525.2%
3.6%	84.5%	493.5%

Optional content - Proposal vs. Low/High-bitrate Anchors

Fencing	L	3.2%	-31.8%	12.68	-8.9%	-33.7%
Hall	T	941.0%	45.1%	15.98	163.7%	16.6%
Street	U	-43.6%	-27.6%	7.00	-12.9%	-8.8%
Mirror	I	-19.6%	-32.6%	14.35	-0.9%	-20.6%
Cadillac	G	-58.4%	-67.0%	14.81	-34.2%	-55.5%
MIV		---	---	**17.59**	---	---

3.1%	71.8%	360.0%
3.4%	81.4%	259.5%
3.9%	85.8%	322.6%
2.5%	74.7%	556.5%
4.4%	56.6%	588.5%
3.4%	73.5%	376.2%

When simplified depth estimation is performed, the decoding and rendering become just 5 times longer than in A17, while the objective quality is still better in most cases.

V17 vs. the proposal:

Mandatory content - Proposal vs. Low/High-bitrate Anchors

Sequence		High-BR BD rate Y-PSNR	Low-BR BD rate Y-PSNR	Max delta Y-PSNR	High-BR BD rate IV-PSNR	Low-BR BD rate IV-PSNR
Fan	O	-76.7%	-72.8%	9.17	-55.8%	-53.1%
Kitchen	J	-39.9%	-31.9%	13.54	-9.9%	-5.3%
Painter	D	-44.4%	-38.3%	6.67	-34.3%	-31.9%
Frog	E	-44.1%	-37.4%	6.83	-41.9%	-35.9%
Carpark	P	-16.6%	-19.5%	9.81	-21.1%	-22.0%
Group	R	32.1%	3.1%	19.49	118.8%	30.4%
MIV		---	---	**13.00**	---	---

Runtime ratio (%)

Atlas encoding	Video encoding	Decoding & Rendering
21.4%	121.4%	7234.1%
29.7%	127.8%	6698.4%
11.4%	105.1%	7794.5%
15.1%	108.0%	########
19.0%	129.0%	3138.9%
20.8%	152.6%	6874.1%
21.3%	128.5%	5935.1%

Optional content - Proposal vs. Low/High-bitrate Anchors

Fencing	L	-17.9%	-24.5%	12.62	-24.8%	-26.9%
Hall	T	-3.9%	-8.2%	16.00	4.3%	-9.8%
Street	U	-8.6%	-13.7%	7.01	-19.8%	-19.2%
Mirror	I	-66.2%	-48.1%	14.27	-64.6%	-44.0%
Cadillac	G	-51.2%	-49.7%	14.59	-41.1%	-37.5%
MIV		---	---	**17.47**	---	---

14.7%	125.9%	3456.9%
13.8%	98.2%	2784.2%
13.3%	114.7%	2482.7%
18.3%	122.2%	6195.7%
21.0%	102.2%	7051.7%
17.1%	117.4%	3990.6%

G17 vs. the proposal:

Mandatory content - Proposal vs. Low/High-bitrate Anchors

Sequence		High-BR BD rate Y-PSNR	Low-BR BD rate Y-PSNR	Max delta Y-PSNR	High-BR BD rate IV-PSNR	Low-BR BD rate IV-PSNR
Fan	O	0.5%	37.1%	9.17	1.6%	40.2%
Kitchen	J	6.9%	-6.6%	13.54	15.4%	-0.8%
Painter	D	-12.5%	-1.0%	6.67	-19.6%	-6.1%
Frog	E	-11.2%	-1.8%	6.83	-1.3%	4.4%
Carpark	P	50.8%	55.9%	9.81	15.0%	30.7%
Group	R	---	---	19.49	---	---
MIV		---	---	**10.92**	---	---

Runtime ratio (%)

Atlas encoding	Video encoding	Decoding & Rendering
466.8%	65.8%	64.9%
769.5%	88.1%	74.5%
236.3%	87.1%	72.9%
264.6%	91.2%	58.7%
342.3%	130.4%	71.7%
557.2%	82.3%	80.4%
439.5%	90.8%	70.5%

Optional content - Proposal vs. Low/High-bitrate Anchors

Fencing	L	-47.3%	-20.7%	12.62	-5.3%	13.2%
Hall	T	-71.6%	-71.4%	16.00	-71.1%	-63.3%
Street	U	8.1%	18.1%	7.01	19.2%	28.4%
Mirror	I	-10.6%	-6.9%	14.27	-16.2%	-10.4%
Cadillac	G	5.3%	7.5%	14.59	16.7%	14.3%
MIV		---	---	**17.38**	---	---

264.3%	87.5%	60.3%
305.9%	75.0%	74.4%
306.9%	106.0%	68.1%
311.3%	107.2%	83.5%
482.2%	103.2%	70.2%
613.3%	96.3%	69.1%

When compared with G17, the proposal provides a similar objective quality, but the depth estimation is faster, as the number of atlases with textures was decreased. Moreover, the views from the 1st atlas are being estimated much faster, as they utilize input depth maps available in the decoder.

3 Recommendation

We recommend opening an EE, to test the possibility of using the proposed scheme as one of the MIV anchors.

4 Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00207, Immersive Media Research Laboratory).