OPTIMIZATION AND COMPARISON OF CODING ALGORITHMS FOR MOBILE 3DTV

G. Tech$^{(1)}$*, A. Smolic*$^{(1)}$*, H. Brust*$^{(1)}$*, P. Merkle*$^{(1)}$*, K. Dix*$^{(1)}$*, Y. Wang*$^{(1)}$*, K. Müller*$^{(1)}$*, and T. Wiegand*$^{(1)(2)}$

(1) Fraunhofer Institute for Telecommunications
Heinrich-Hertz-Institut
Image Processing Department
Einsteinufer 37, 10587 Berlin, Germany
gerhard.tech@hhi.fraunhofer.de

(2) Technical University of Berlin
Image Communication, Faculty of EE & CS
Einsteinufer 17, 10587 Berlin, Germany

ABSTRACT

Different methods for coding of stereo video content for mobile 3DTV are examined and compared. These methods are H.264/MPEG-4 AVC simulcast transmission, H.264/MPEG-4 AVC Stereo SEI message, mixed resolution coding, and video plus depth coding using MPEG-C Part 3. The first two methods are based on a full left and right video (V+V) representation, the third method uses a full and a subsampled view and the fourth method is based on a one video plus associated depth (V+D) representation. Each method was optimized and tested using professional 3D video content. Subjective tests were carried out on a small size autostereoscopic display that is used in mobile devices. A comparison of the four methods at two different bitrates is presented. Results are provided in average subjective scoring, PSNR and VSSIM (Video Structure Similarity).

***Index Terms*—** mobile 3DTV, stereo video, mixed resolution, video plus depth, H.264/AVC, MPEG-C Part 3, subjective assessment

1. INTRODUCTION

3D video is a new upcoming trend in entertainment industry. This technology will also be applied to mobile 3DTV. The requirements of mobile technology impose certain constraints to transmission and coding. Limited bandwidth and receiver complexity have to be considered as well as special properties of the small, portable viewing devices. The subjective impression of video content differs on small and large screens. Furthermore, screen size on mobile devices limits the viewing angle and the number of concurrent viewers is one. Hence transmission of two views is sufficient. Various 3D video representation formats and coding algorithms have been proposed for this purpose.

In [1] and [2] the optimization of video plus video and video plus depth approaches is examined. The mixed resolution approach is discussed in [3]. This paper reviews and compares four of the methods for 3D video on mobile devices using objective video quality metrics and subjective testing.

In sections two and three the 3D video representation and coding methods and the test data are introduced. The objective metrics and the optimization of the methods are discussed in sections four and five. Subjective tests are described in section six. Sections seven and eight provide the results and the conclusion.

2. CODING METHODS

2.1. H.264/AVC Simulcast

The left and right view are transmitted independently, each coded using H.264/MPEG-4 AVC. Hence this method does not need any pre- or post processing before coding and after decoding, the complexity on sender and receiver side is low. Redundancy between channels is not reduced, thus coding efficiency is not optimized.

2.2. H.264/AVC Stereo SEI Message

H.264/MPEG-4 AVC enables inter-view prediction through the Stereo SEI syntax. Practically it is based on interlacing the left and the right view prior to coding and exploring interlaced coding mechanisms. It has been shown that the principle and efficiency of this approach is very similar to Multiview Video Coding (MVC), which is a H.264/MPEG-4 AVC extension to code two or more related video signals [1].

2.3. Mixed Resolution coding

Binocular suppression theory states that perceived image quality is dominated by the view with higher spatial resolution [3]. The mixed resolution approach utilizes this attribute of human perception by decimating one view before transmission and up-scaling at the receiver side. This enables a tradeoff between spatial subsampling and amplitude quantization. For experiments in this scope the right view was decimated by a factor of about two in horizontal and vertical direction.

	Snail	Car	Horse	Hands
Low birate	60	120	177	578
High bitrate	188	639	1023	2918

Table 1. Low and high bitrates in kbit/s

2.4. Video plus Depth

MPEG-C Part 3 defines a video plus depth representation of the stereo video content. Depth is generated at the sender side for instance by estimation from an original left and right view. One view is transmitted simultaneously with the depth signal. At the receiver the other view is synthesized by depth image based rendering [2]. Compared to video a depth signal can in most cases be coded at a fraction of the bitrate at sufficient quality for view synthesis. Nevertheless errors in depth estimation and problems with disocclusions introduce artifacts to the rendered view.

3. TEST DATA

Four sequences were used for objective and subjective tests. The sequences were produced by the professional 3D-content production company KUK Filmproduktion GmbH. The sequences are named: *Snail*, *Hands*, *Car* and *Horse*. The content of the sequences has different complexity. In sequence *Snail* motion and complexity of the scene is low. The sequence *Hands* provides a fast moving and complex scene that is challenging for coding. *Car* contains fast motion and the *Horse* sequence contains a lot of fine structures.

To determine a high and a low bitrate for the test sequences the quantization parameter (QP) for simulcast was set to 30 and 40. Resulting bitrates for the sequences are shown in Table 1. The other three methods were optimized to the same bitrates. This leads to a total of $4 \times 4 \times 2 = 32$ sequences. The H.264/MPEG-4 AVC reference software JM 15.0 was used for all experiments.

4. VIDEO QUALITY METRICS

Within this paper Peak Signal-to-Noise Ratio (PSNR) and Video Structure Similarity (VSSIM) [4] are used as objective metrics for optimization. Although PSNR is averaged over individual frames traditionally, the mean PSNR as the average from the two individual PSNR values is not useful in case of different resolutions to evaluate total stereo video quality. A more suitable metric used in this scope and already utilized in [3] is the PSNR calculated from average MSE of both views. VSSIM is a structure similarity based metric using luminance and motion weighting to simulate the properties of the human visual system.

5. OPTIMIZATION

Transmission of TV signals requires random access points to provide transmission error robustness. This also applies to mobile applications, where burst errors can also occur. To address this requirement, an I picture period of 16 was used.

5.1. Simulcast and Stereo SEI Message

A detailed optimization of video plus video approaches has been carried out in [1]. Here, it is shown that hierarchical B pictures significantly increase coding efficiency. Nevertheless hierarchical B pictures require increased complexity of the decoder and the encoder, which limits application in mobile devices. In contrast to [1] nonhierarchical B pictures are examined in this paper. A GOP structure of IBBP was chosen.

5.2. Mixed Resolution

To determine optimal bitrate distribution between the views of the mixed resolution method, the approach suggested in [3] is used: To take binocular suppression theory into account VSSIM and PSNR are calculated using the blurred original view as reference for the low-resolution view. Resulting RD characteristics for different bitrates of the full view are shown in Figure 1 for the sequence Snail. It can be seen that the optimal distribution of bitrate assigns 2/3 for the full view and 1/3 for the down-sampled view as reported in [3]. VSSIM shows a similar characteristic.

5.3. Video plus depth

For the video plus depth approach an optimal distribution of bitrate between depth map and video signal had to be found, and for that a reference view had to be defined. However, comparing a rendered view to an original view in terms of PSNR or VSSIM is not appropriate since irrelevant imperfections like slight shifts of the content, which are visually not noticeable, will results in large errors. Hence it is not useful to use the original view as reference. For the optimization of the video plus depth approach VSSIM and PSNR are calculated using the view rendered from the uncoded video and depth as reference for the view rendered from coded data. Resulting PSNR and VSSIM characteristics for the sequence *Hands* are depicted in Figure 2. With respect to PSNR optimal distributions of bitrate were found to range from 10% to 30% for depth depending on sequence content and total bitrate. As can be seen in Figure 2 the optimization of the VSSIM results in a lower bitrate for depth.

6. SUBJECTIVE TESTS

Subjective tests were carried out using a 3.5” autostereoscopic display used in mobile devices with a resolution of 640x480 pixels and barrier-technology. 16 experts in video and image

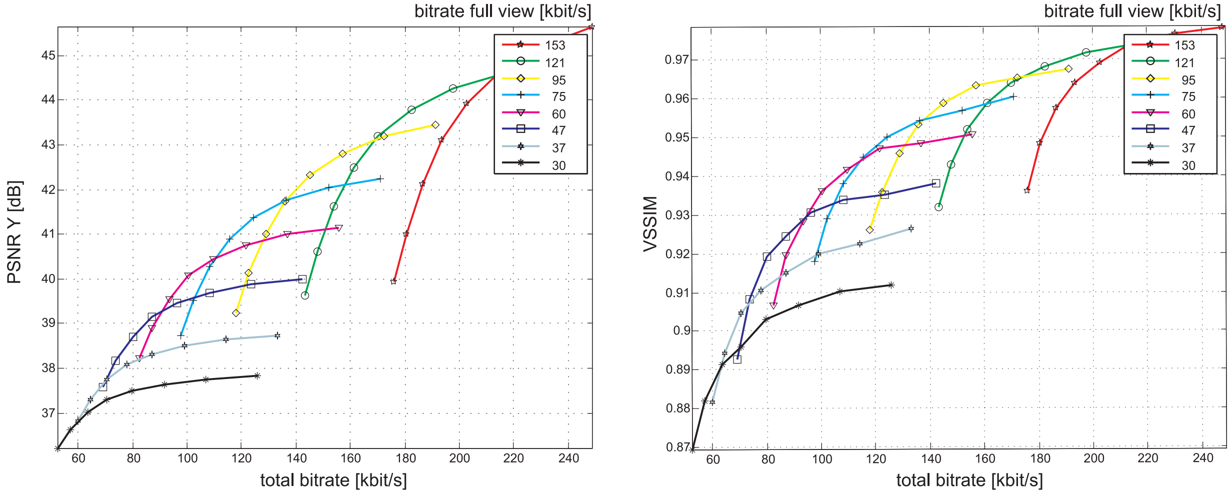

Fig. 1. Mixed resolution coding: PSNR Y and VSSIM vs. total bitrate for varying bitrate of the full view, sequence *Snail*

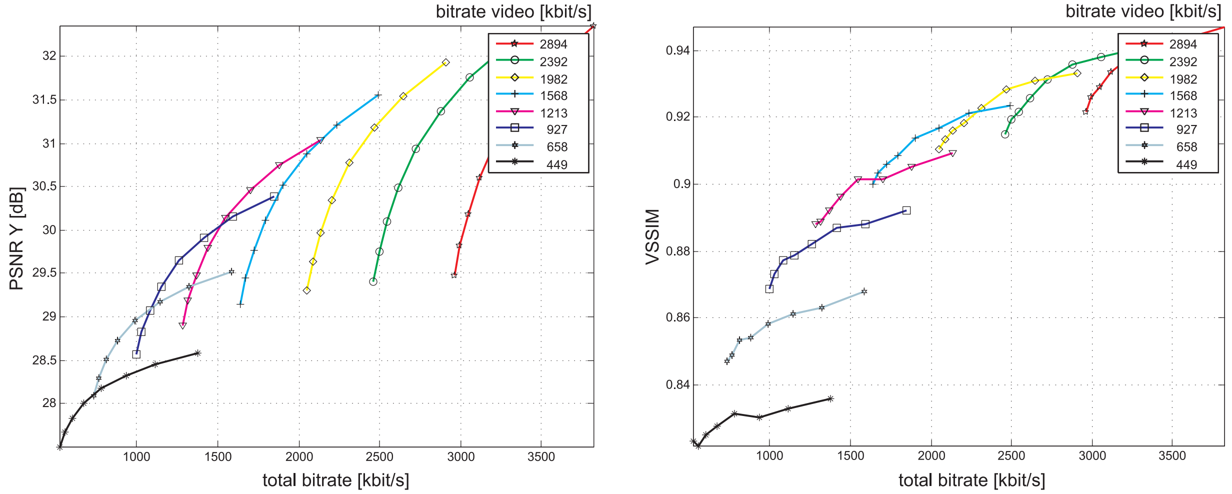

Fig. 2. Video plus depth coding: PSNR Y and VSSIM vs. total bitrate for varying bitrate of the video data, sequence *Hands*

processing participated in the test. Initially the four uncoded original sequences were presented as reference. Furthermore participants were informed of the influence of their head position on perceived depth impression on the autostereoscopic device. After that viewers rated all randomized 32 sequences interactively on a scale including the quality levels from 5 (excellent) to 1 (bad). Additionally participants were asked after the test for their overall impression. Most participants had problems in rating the *Snail* sequence.

7. RESULTS

The resulting MOS scores are shown in Figure 3 for each sequence individually and for the average over all sequences. At the high rate all methods perform very similar. This proves the binocular suppression theory since test persons did not notice the difference between the mixed resolution sequences and the other methods. It also proves the suitability of the video plus depth concept, since rendering artifacts were not noticeable. On the other hand, none of the advanced approaches outperforms simple simulcast in this case of sufficient bit budget (as given for the QP30-case).

Differences become evident at the lower bitrate. All other methods outperform simple simulcast for every sequence. The results vary for different sequences: *Hands* is a very complex sequence that requires significant bitrate for decent quality. In comparison, the corresponding depth data can be compressed at relatively low bitrate. In case of the *Horse* sequence the depth maps are very smooth with low structure thus also resulting in a relatively low bitrate compared to the corresponding color video. Thus for these two sequences the video plus depth approach is most efficient and outperforms the other approaches. In contrast to that *Snail* is very simple (low motion and structure) in general and *Car* has a complex depth struc-

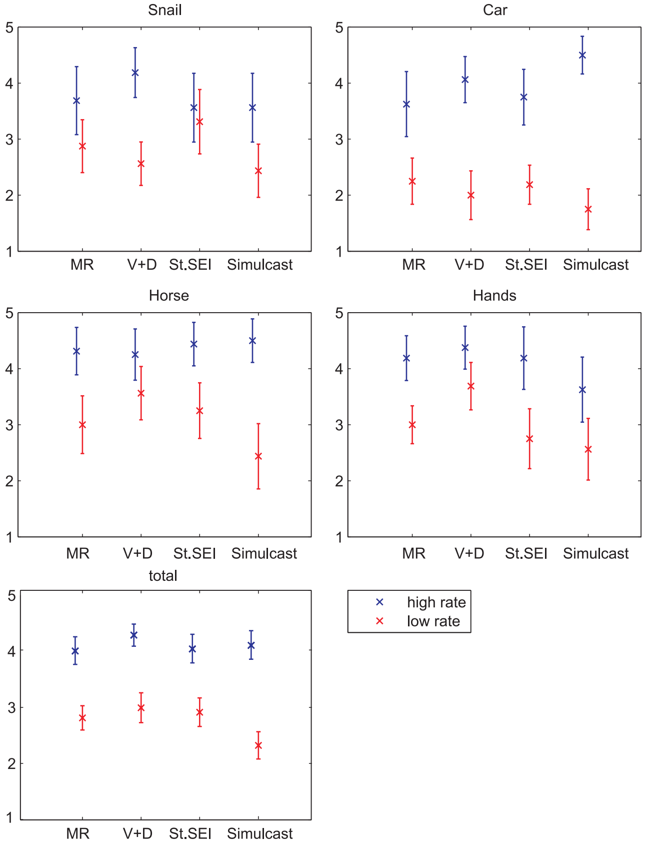

Fig. 3. MOS of each sequence and of all sequences, with 95% confidence intervals

	High rate			Low rate		
	MOS	PSNR	VSSIM	MOS	PSNR	VSSIM
MR	4.0	40.2*	0.97*	2.8	32.7*	0.89*
V+D	4.2	35.7+	0.95+	2.9	31.0+	0.84+
Interl.	4.0	37.9	0.95	2.9	31.4	0.83
Simul.	4.0	37.0	0.94	2.3	30.6	0.80

Table 2. Average results at high and low rate; *blurred view from uncoded data as reference for the low resolution view; + rendered view from uncoded data as reference for the rendered view

ture. Therefore for these sequences the video plus depth approach does not have an advantage. In total, the 3 approaches perform very similar but all outperform simple simulcast.

Further findings can be drawn by comparing the objective quality metrics with the MOS in Table 2. In contrast to the MOS values, PSNR and VSSIM of the mixed resolution approach are higher than the values of the other methods. Hence, PSNR and VSSIM (which were obtained using the blurred original view as reference for the view with low resolution) overestimate objective quality. A contrary effect can be observed for PSNR calculated using the rendered view from uncoded data as reference for the rendered view: Here, video quality is underestimated although artifacts not related to coding are not covered by PSNR.

8. CONCLUSION AND FUTURE WORK

Four methods for transmission and coding of stereo video content have been optimized and analyzed. Subjective ratings show that the mixed resolution approach and the video plus depth approach do not impair video quality at high bitrates. At low bitrates simulcast transmission it outperformed by the other methods. Objective quality metrics, utilizing the blurred or rendered view from uncoded data as reference, can be used for optimization of single methods. They cannot be used for comparison of methods since they have a positive or negative bias.

For further evaluation, the subjective tests will be broadened. Further development of individual methods will include combinations like inter-view prediction for mixed resolution coding and depth representation at reduced resolution.

9. ACKNOWLEDGEMENTS

This work was supported by EC within FP7 (Grant 216503 with the acronym MOBILE3DTV). We would like to thank KUK Filmproduktion GmbH (http://www.kuk-film.de/), Munich, Germany, for providing the *Hands*, *Car*, *Horse* and *Snail* stereo video data.

10. REFERENCES

[1] P. Merkle, H. Brust, K. Dix, A. Smolic, and T. Wiegand, "Stereo video compression for mobile 3D services," *3DTV Conference*, 2009.

[2] P. Merkle, Y. Wang, K. Müller, A. Smolic, and T. Wiegand, "Video plus depth compression for mobile 3D services," *3DTV Conference*, 2009.

[3] H. Brust, A. Smolic, K. Müller, G. Tech, and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices," *3DTV Conference*, 2009.

[4] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.