# STEREO VIDEO ENCODER OPTIMIZATION FOR MOBILE APPLICATIONS

*Philipp Merkle[1], Jordi Bayo Singla[1], Karsten Müller[1], and Thomas Wiegand[1,2]*

[1]Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Berlin, Germany;
[2]Berlin Institute of Technology, Germany

## ABSTRACT

This paper presents a stereo video encoder optimization for mobile applications. While video coding applications mostly concentrate on finding a good trade-off between rate and distortion, the additional constraint of limited processing power has to be considered for mobile applications. Realizing mobile video coding applications for stereo instead of 2D video is challenging, as twice the amount of video data has to be processed. Therefore, we investigate how the trade-off between rate, distortion, and complexity can be optimized for the multiview video coding (MVC) extension of the H.264/AVC standard. In the paper, we focus on the encoder, as its complexity is much higher and more configuration dependent than at the decoder. By enabling and disabling certain tools and setting different parameter values, the encoder complexity can be adapted for mobile applications. The presented results show that an optimized MVC encoder configuration performs significantly faster without impairing the rate-distortion performance.

***Index Terms***— 3D video, multi-view video coding, mobile applications, stereoscopic representation

## 1. INTRODUCTION

Currently, 3D video is emerging from 3D cinemas to home entertainment and mobile device applications. This requires efficient technologies for the whole 3D video processing chain, including content production, coding, transmission and display [1]. Considering a next-generation smartphone that is equipped with a stereo camera as well as an auto-stereoscopic display, 3D video applications like playback of live-streams or real-time communication can be realized. However, the implementation of such applications for mobile devices is challenging. Essential restricting factors are the limited processing power of the hardware as well as the limited bandwidth of mobile radio channels.

Coding is the core component of these 3D video applications and in the case of stereo video MVC [2] is used. Here, the multi-frame motion compensation method of H.264/AVC is extended in a way that a picture can have temporal as well as inter-view reference pictures for motion- and disparity-compensated prediction, respectively. Advantages of MVC are the improved coding efficiency and backward compatibility to H.264/AVC. For that, the first view is compressed using a profile conforming to H.264/AVC without multi-view capability. The remaining views are coded using MVC and typically require 30-50% less bit rate than H.264/AVC simulcast coding at the same quality.

The rate-distortion (R-D) efficiency of MVC is achieved at the expense of a high complexity, especially for the encoding process. Therefore, a study on the encoder performance is presented in the following sections, optimizing the trade-off between rate, distortion and complexity (R-D-C) for mobile applications by modifying the configuration in terms of enabling/disabling certain tools and varying the value of certain parameters.

## 2. PROCESSING SETUP

Figure 1 illustrates the processing setup of the study experiments. In order to classify the MVC complexity results, frame-compatible stereo in side-by-side format (SbS) and AVC compression serves as a reference.

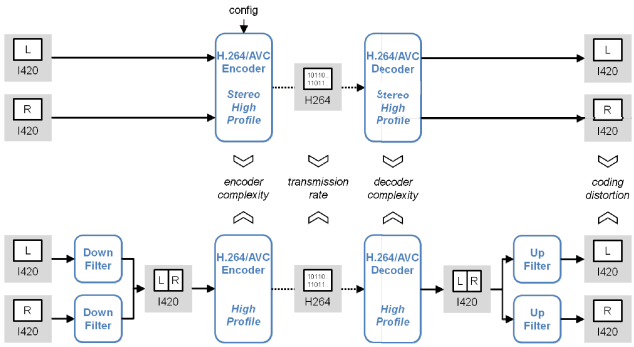


**Figure 1.** Processing and evaluation setup for MVC (top) and SbS reference (bottom). Blue outlined boxes indicate processing steps and grey boxes the according data formats (with *FourCC* identifiers).

For both configurations, the full resolution left and right view YUV 4:2:0 (I420) video files of a stereo sequence are used as encoder input and decoder output, while one H.264/AVC compliant bit-stream (H264) is transmitted between encoder output and decoder input. Consequently, MVC processing simply consists of running the encoder and decoder in 2-view mode. For SbS the processing consists of

978-1-61284-162-5/11/$26.00 2011 © IEEE

pre- and post-processing of the frame-compatible format plus running the encoder and decoder in single-view mode. Here, preprocessing means down-sampling the left and right view horizontally and merging them into a single frame-compatible stream, while post-processing means splitting the left and right view and up-sampling each of them to the original size. For up- and down-sampling the SVC *DownConvert* tool [3] is used.

## 3. CODING CONDITIONS

All coding experiments are carried out with the H.264/AVC Reference Software JM 17.2 [4]. For supporting MVC stereo coding the *Stereo High profile* has to be selected, while SbS reference coding uses the *High profile*.

One of the most important influencing factors regarding the coding complexity is the temporal prediction structure. This is addressed by selecting the following four typical temporal prediction structures for the coding experiments: intra only (IIII), inter predictive (IPPP), bi-predictive (IBBP), and hierarchical bi-predictive (IbBb) with a GOP (group of pictures) size of 8 [5].

| Parameter | Values | | | |
|---|---|---|---|---|
| Profile | Stereo High profile | | | |
| Number views | 2 | | | |
| Temporal pred. | IIII | IPPP | IBBP | IbBb |
| Intra period | 1 | 16 | | |
| GOP size | - | | 3 | 8 |
| QPs | 38/44/50 | 28/34/40 | | |
| Symbol mode | CAVLC/CABAC | | | |
| Search range | 0/16/32 | | | |
| Search mode | full/EPZ search | | | |
| RD optimization | RD-off/RD-on | | | |
| Subpel ME | on/off | | | |
| I modes | on/off | | | |
| P modes | - | on/off | | |
| B modes | - | | on/off | |

**Table 1.** Encoder configuration for MVC stereo, combining different prediction structures and values of complexity relevant settings and parameters.

For the optimization study we focus on testing those settings and parameters, which are expected to have a major influence on both the rate-distortion and the complexity performance. Typically, a better R-D performance is achieved by more complex tools and vice versa [6]. Due to interdependencies between the selected settings and parameters, all possible combinations need to be tested for achieving an efficient configuration. Table 1 summarizes the relevant encoder configuration settings and parameters. For each combination of values three QPs are tested, corresponding to low, medium, and high bit rate.

In order to avoid an excessive number of coder runs, the MVC stereo coding tests are divided into two groups. According to Table 1 the first group consists of the following parameters:

- Symbol mode: The two supported entropy coding methods, namely CAVLC and CABAC, are tested.
- Search range: The parameter determines the neighborhood area for motion estimation.
- Search mode: Different search strategies for motion estimation are supported. Tested methods are full search and enhanced predictive zonal (EPZ) search.
- RD optimization: Different algorithms for Lagrangian based rate-distortion optimized mode decision are supported. Tested methods are RD-off (low complexity) and RD-on (high complexity).
- Subpel ME: The parameter enables/disables the support of sub-pixel precision for motion estimation.

Combining all given variations, results in a total of 576 coder runs for each test sequence. After evaluating the results an optimum configuration for the first group of parameters can be derived, which is also used as the fixed configuration for SbS reference coding. Based on this configuration, the following parameter variations are tested in the second group:

- I modes: Enabling/disabling intra prediction modes (4x4, 16x16, chroma, IPCM).
- P modes: Enabling/disabling P slice modes (intra pred., 4x4…16x16 inter pred. and motion comp.).
- B modes: Enabling/disabling B slice modes (4x4…16x16 inter pred. and motion comp., 8x8…16x16 bi-predictive motion estimation).

Combining all variations of the second group, results in 222 coder runs for each test sequence. Please note that all other parameters are set to default or typical configuration.

## 4. EVALUATION PROCEDURE

As illustrated in Figure 1, the results of the experiments are evaluated with respect to rate, distortion, and complexity.

Measuring rate and distortion is well known from normal video coding and can easily be applied to stereo coding. For the latter, the total bit rate in *kbps* is derived from the size of the encoded bit stream (containing both views) and the frame rate of the sequence. Regarding distortion, the mean squared error (MSE) is measured between the frames of the decoded and the original video sequences for the left and the right view as the PSNR of the luma component in *dB*. Note, that the full resolution sequences are evaluated for both MVC and SbS setup. The total distortion is derived by averaging the two PSNR values of the left and right view.

Regarding complexity or processing speed, the execution time for processing a certain number of frames is measured. The processing speed in *fps* is derived by

dividing the number of frames by the execution time. Note, that a frame includes the left and right view picture in the case of stereo video. According to Figure 1, for MVC stereo coding only the encoder process execution time needs to be measured. For SbS coding either only the encoder process execution time or the complete sender- and receiver-side processing time is measured, including the up-/down-sampling and merging/splitting processes. In contrast to determining rate and distortion values, measuring the processing speed requires considering distorting side effects that are caused by hardware-software interaction, such as buffering, virtual memory, parallel processing, etc. Therefore, the process execution time on a single CPU rather than the system start and stop time has to be measured.

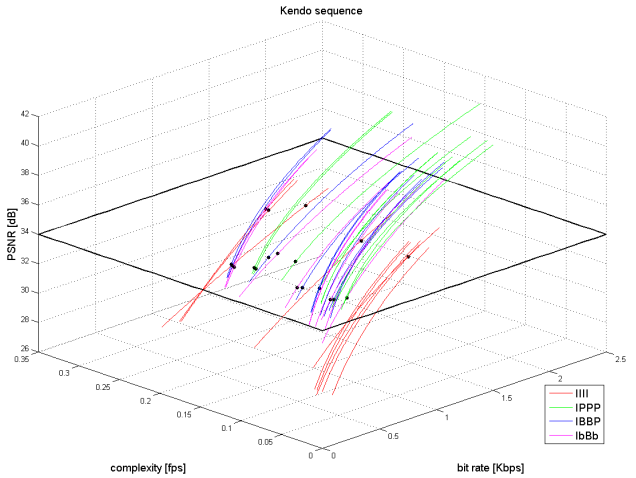


**Figure 2.** Sample rate-distortion-complexity diagram with constant quality cutting plane and piercing points for Kendo sequence.

Finally, one R-D-C triplet of values is achieved for each configuration and each QP. Here, comparison and evaluation of the combined R-D-C performance results for different prediction structures and encoder configurations is not as easy as evaluating only R-D performance without complexity. Two-dimensional R-D diagrams are very well known in the field of video coding and each curve of such a diagram represents the R-D performance of a certain configuration or method at different QPs. The evaluation of such a diagram is very intuitive: One method has a better R-D performance, if it has a lower bit rate and/or higher PSNR. Including the additional value for the complexity or processing speed to the results leads to three-dimensional R-D-C diagrams. The example in Figure 2 highlights that neither evaluation nor illustration is as self-evident as for R-D diagrams. Therefore, we analyze the results at a fixed, constant quality, as indicated by the black outlined cutting plane at $PSNR_Y = 34\,dB$ in Figure 2. The three QPs are first interpolated, using a piecewise cubic interpolation function. By doing so, the intermediate R-C value for the selected distortion can be easily derived (see black markers in Figure 2). The resulting two-dimensional R-C diagrams allow for evaluating the different configurations according to the trade-off between bandwidth and processing power at a certain quality – the two essential restricting factors for stereo video applications on mobile devices. Similar to R-D diagrams, a point that lies left-above another point shows a better performance, due to lower bit rate and/or higher processing speed.

## 5. RESULTS

As the scope of the presented study is mobile applications, the hardware for implementing the experiments is an ultra-mobile PC (UMPC, Sony Vaio VGN-UX1XN).

Regarding test data, three stereo sequences with complex content (i.e. fine structures, complex motion and depth structure) are analyzed for the experiments: “Kendo”, “Newspaper”, and “Hulahoop”. The first two are test data sets used in the MPEG 3DV activity [7], while Hulahoop was captured for the 3DPhone project, using a consumer market digital stereo camera. All sequences have a frame rate of 30 fps and in order to simulate realistic conditions for mobile applications, the resolution of the sequences is down-sampled to VGA (640×480).

The R-C results for encoding the two groups of MVC stereo as well as the SbS reference are shown in Figure 3. Here, diamond markers represent the first group and square markers the second group MVC results. The filled markers highlight efficient configurations. Regarding the MVC results for the first group of parameters, the evaluation of the results shows that the most efficient overall configuration (filled diamond markers) for all prediction structures is as follows: CABAC, a search range of 16, EPZ search, low complexity RD optimization, and sub-pixel motion estimation. This configuration is also used for SbS coding (filled circle markers). Comparing the MVC and SbS results shows that they achieve an equal performance. The only exception is the SbS result for IIII, with very high processing speed (~2.3 fps), but also a very high rate (~1.5×MVC_IIII). For all SbS results the pre-processing time is not included. Pre-processing runs at about 4.5 fps and in combination with encoding the resulting processing speed would be lower. The most efficient configuration for the second group of parameters (filled square markers) is: all I, P, and B modes disabled. For IPPP, IBBP, and IbBb prediction structures this configuration achieves a significantly higher processing speed at almost the same rate. Especially the B modes have a significant influence on the processing speed.

## 6. SUMMARY & CONCLUSION

We have presented a rate-distortion-complexity optimization study, which includes the processing power as an important delimiting factor for realizing video coding applications on mobile devices. This is especially true for stereo video in general and here for the encoder in

particular. Accordingly, the study focused on encoding stereo video with MVC. By evaluating the results for a constant distortion, the trade-off between bit rate and processing speed can be analyzed for all configurations. Altogether, the results show that the optimized configuration performs significantly faster than a default configuration without impairing the rate-distortion performance. However, the encoding speed is still far from real-time. In contrast, additional evaluation of the decoder complexity showed that the decoding speed is close to real-time for most configurations. Therefore, real-time stereo video applications with MVC encoding on mobile devices would require combining the presented optimized configuration with an encoder implementation that is highly adapted to the special hardware capabilities and restrictions of mobile devices. A significant complexity reduction can be expected from optimizing the JM reference encoder, as the purpose of this implementation is rather supporting all features of the standard than providing maximum processing speed. Summarizing the overall results shows that the configuration with the highest processing speed is about 20 times faster than the most complex configuration at the same bit rate.

## ACKNOWLEDGMENT

This work was supported in part by EC within FP7 under Grant No. 213349 (with the acronym 3DPHONE).
We would like to thank the Nagoya University (Japan) for providing the *Kendo* data set and the ETRI (Korea) for providing the *Newspaper* data set.

## REFERENCES

[1] P. Merkle, K. Müller, and T. Wiegand, "3D Video: Acquisition, Coding, and Display," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, pp. 946-950, May 2010.
[2] Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services", EURASIP Journal on Advances in Signal Processing, Vol. 2009, No. 1, January 2009.
[3] H. Schwarz, J. Vieron, T. Wiegand, M. Wien, A. Eleftheriadis, and V. Bottreau, "JSVM software, text, and conformance status," Doc. JVT-AF013, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, Geneva, Switzerland, Nov. 2009.
[4] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services," ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), Version 10, March 2010.
[5] H. Schwarz, D. Marpe, and T. Wiegand: Analysis of hierarchical B pictures and MCTF, IEEE International Conference on Multimedia and Expo (ICME'06), Toronto, Ontario, Canada, July 2006.
[6]T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. on Circuits and Systems for Video Tech., vol. 13, no. 7, pp. 560-576, July 2003.
[7] "Description of Exploration Experiments in 3D Video Coding", ISO/IEC JTC1/SC29/WG11, Doc. N11831, Daegu, Korea, January 2011.

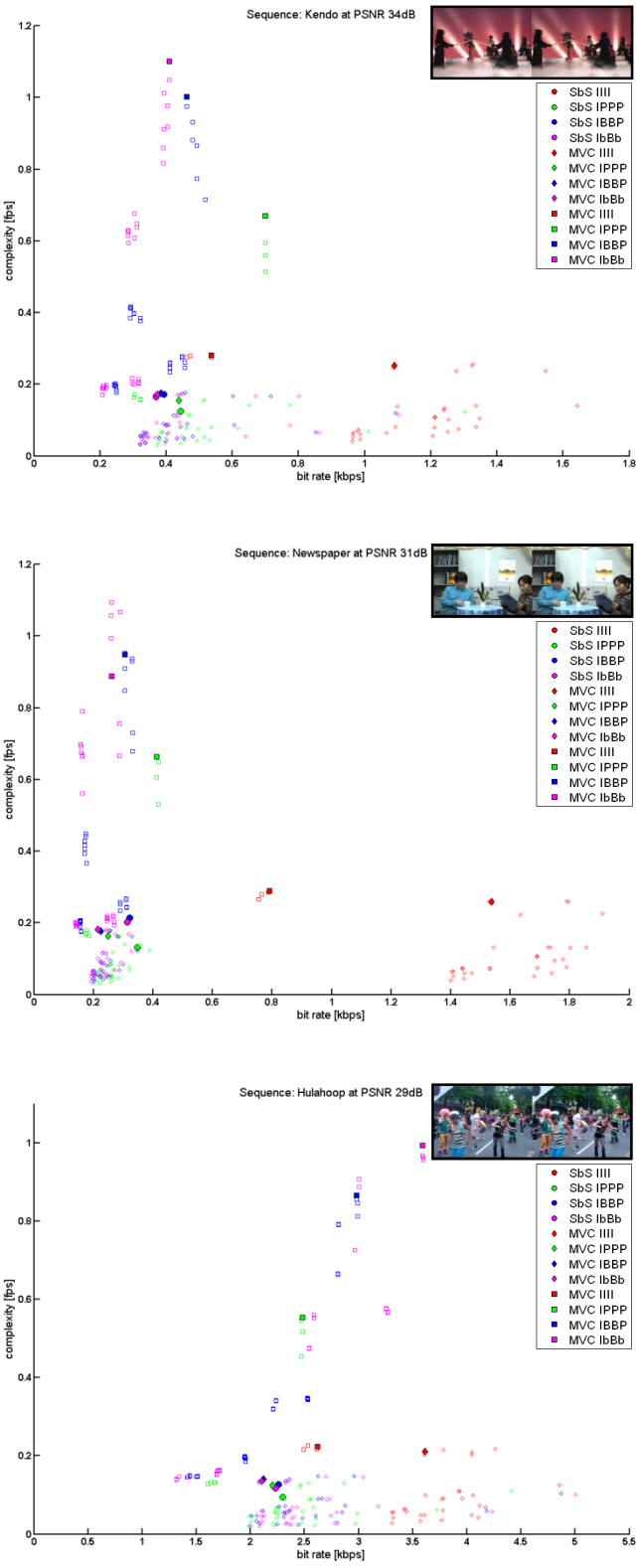


**Figure 3.** Rate-Complexity diagrams with MVC and SbS reference results for Kendo (top), Newspaper (middle), and Hulahoop (bottom). Filled markers highlight most efficient configurations.