MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Analysis of 3D and Multiview Extensions of the Emerging HEVC Standard

Vetro, A.; Tian, D.

TR2012-068   August 2012

Abstract

Standardization of a new set of 3D formats has been initiated with the goal of improving the coding of stereo and multiview video, and also facilitating the generation of multiview output needed for auto-stereoscopic displays. Part of this effort will develop 3D and multiview extensions of the emerging standard for High Efficiency Video Coding (HEVC). This paper outlines some of the key technologies and architectures being considered for standardization, and analyzes the viability, benefits and drawbacks of different codec designs.

SPIE Applications of Digitnal Image Processing

This work may not be copied or reproduced in whole or in part for any commercial purpose. Permission to copy in whole or in part without payment of fee is granted for nonprofit educational and research purposes provided that all such whole or partial copies include the following: a notice that such copying is by permission of Mitsubishi Electric Research Laboratories, Inc.; an acknowledgment of the authors and individual contributions to the work; and all applicable portions of the copyright notice. Copying, reproduction, or republishing for any other purpose shall require a license with payment of fee to Mitsubishi Electric Research Laboratories, Inc. All rights reserved.

Copyright © Mitsubishi Electric Research Laboratories, Inc., 2012
201 Broadway, Cambridge, Massachusetts 02139

Analysis of 3D and multiview extensions of the emerging HEVC standard

Anthony Vetro, Dong Tian
Mitsubishi Electric Research Laboratories, 201 Broadway, 8th Floor, Cambridge, MA, USA 02139

ABSTRACT

Standardization of a new set of 3D formats has been initiated with the goal of improving the coding of stereo and multiview video, and also facilitating the generation of multiview output needed for auto-stereoscopic displays. Part of this effort will develop 3D and multiview extensions of the emerging standard for High Efficiency Video Coding (HEVC). This paper outlines some of the key technologies and architectures being considered for standardization, and analyzes the viability, benefits and drawbacks of different codec designs.

Keywords: video coding, 3D, multiview, HEVC, codec

1. INTRODUCTION

3D and multiview video formats are able to provide depth perception of a visual scene through the appropriate 3D display system. The types of 3D displays include stereoscopic displays that require glasses to view the depths of a scene, and auto-stereoscopic displays that emit view-dependent pixels and do not require glasses for viewing. A more comprehensive review of 3D display technologies has been given by Urey, et al.[1]. In addition to enhancing the viewing experience through depth, these 3D and multiview video formats also enable free-viewpoint video, which may be useful in surveillance or immersive teleconference applications. In this scenario, the viewpoint and view direction can be interactively changed and the system allows viewers to freely navigate through the different viewpoints of the scene.

The H.264/MPEG-4 Advanced Video Coding (AVC) standard[2,3] is the basis for current stereo and multiview video coding formats. AVC has been extensively deployed for a wide range of video products and services, including stereoscopic services. There exist two primary categories of stereoscopic formats based on AVC: frame compatible, and Multiview Video Coding (MVC). Frame compatible formats refer to a class of stereo video formats in which the two stereo views are filtered, sub-sampled and arranged into a single coded frame or sequence of frames, i.e., the left and right views are packed together in the samples of a single video frame[4]. Popular arrangements include the side-by-side and top-bottom formats. The primary benefit of frame-compatible formats is that they facilitate the introduction of stereoscopic services through existing infrastructure and equipment. Frame compatible formats have been embraced for the first phase of broadcast services. Further details on these formats and their signaling can be found in[2,4]. In contrast to frame-compatible video, the MVC extension of AVC provides a direct encoding of the stereo views at their full-resolution and leverages inter-view prediction to improve compression capability, in addition to ordinary intra and inter-prediction modes. Another important aspect of the MVC design is the inherent support for 2D/backwards compatibility with existing legacy systems. In other words, the compressed multiview stream includes a base view bit stream that is coded independently from all other views in a manner compatible with decoders for single-view profile of the standard. The MVC format was selected by the Blu-Ray Disc Association as the coding format for stereo video with high-definition resolution, and is now being considered for stereo broadcast as well.

A new video coding standard for High Efficiency Video Coding (HEVC) is now being finalized with a primary focus on efficient compression of monoscopic video. Preliminary results have already demonstrated that this new standard will provide the same subjective quality at half the bit rate compared to AVC High Profile. A primary usage of HEVC is to support the delivery of ultra-high definition (UHD) video. It is believed that many UHD displays will also be capable of decoding stereo video as well. The first version of the standard will be approved by January 2013. Recently, a new Joint Collaborative Team on 3D Video Coding Extensions Development (JCT-3V) has been formed between ISO/IEC and ITU-T for the development of new 3D standards, including extensions of HEVC. This paper presents the architectures under consideration and some corresponding tools, and provides an analysis of the different schemes in terms of compression performance, implementation and deployment potential.

2. CODING ARCHITECTURES

There are several different coding architectures that can be considered in the development of 3D and multiview extensions of HEVC, which are briefly outlined and reviewed in this section.

2.1 Multiview HEVC

The most straightforward architecture is a multiview extension of HEVC that utilizes the same design principles of MVC in the MPEG-4/H.264 AVC framework[4]. This scheme would provide backwards compatibility for monoscopic decoding and utilize inter-view prediction between the texture views, where inter-view prediction is enabled through modifications to the reference picture management that enable inclusion of inter-view reference pictures in the reference picture lists that are maintained for prediction. To achieve this, high-level syntax must also signal the dependencies between different views.

A key feature of this architecture is that the basic block-level decoding process would remain unchanged. This design allows for existing single layer codec designs that have been initially designed for 2D applications to be extended without major implementation changes to support stereo and multiview applications. According to current plans, this extension of HEVC is expected to be finalized by early 2014.

2.2 Multiview HEVC with Block-Level Tools

To achieve higher compression efficiency, yet still maintain backwards compatibility with monoscopic video coded by HEVC, an alternative coding architecture would leverage the benefits of block-level coding tools. In this architecture, and similar to the architecture described in section 2.1, the base view is fully compatible with HEVC in order to extract monoscopic video, and only the dependent views would utilize additional tools, such as those described in section 3. As an example, it has been recognized that there is significant correlation between motion and mode parameters between the base and dependent views. Exploiting this correlation could lead to notable bit rate savings.

2.3 Multiview HEVC with Depth

Depth-based representations are another important and emerging class of 3D formats. Such formats are unique in that they enable the generation of virtual views through depth-based image rendering techniques, which may be required by auto-stereoscopic or multiview displays[5]. Depth-based 3D formats can also allow for advanced stereoscopic processing, such as adjusting the level of depth perception with stereo displays according to viewing characteristics such as display size, viewing distance or user preference. The depth information itself may be extracted from a stereo pair by solving for stereo correspondences or obtained directly through special range cameras; it may also be an inherent part of the content, such as with computer generated imagery

In terms of compression formats, it is also anticipated that extension of the HEVC standard would be developed that support the efficient inclusion of depth information. One desirable characteristic of this format is for stereo video to be easily extracted to support existing stereoscopic displays; in such cases, the dependency between the video data and depth data may be limited. However, allowing for a greater degree of dependency between the different components may provide more significant benefits in terms of compression capability and rendering performance.

2.4 Hybrid AVC and HEVC

From a pure compression efficiency point of view, it is always best to use the most advanced codec. However, when introducing new services, providers must also consider capabilities of existing receivers and an appropriate transition plan. Considering that most terrestrial broadcast systems are based on MPEG-2 or AVC, it may not be easy to simply switch codecs in the near-term.

One solution to this problem is to transmit the 2D program in the legacy format, while transmitting an additional view to support stereo services in an advanced coding format, e.g. HEVC. The obvious advantage is that backward compatibility with the existing system is provided with significant bandwidth savings relative to simulcast in the legacy format. One drawback of this approach is that there is a strong dependency between the 3D program and the 2D program, which does not allow for independent 2D and 3D content programs that may be desirable for production. Also, this approach requires legacy and advanced codecs to operate synchronously, which may pose implementation challenges for certain receiver designs. Nevertheless, broadcasting trials of hybrid MPEG-2 and AVC based systems are underway in Korea, and there are plans to standardize the transmission of such a hybrid format in ATSC.

In the context of depth-based 3D formats, there are clearly many variations that could be considered. In an AVC-compatible framework, the base view would be coded with AVC, while additional texture views and supplemental depth videos could be encoded with HEVC. A slight variation on this would be for the stereo pair of the texture to be coded with MVC, and only the depth videos are coded with HEVC.

3. COMPRESSION TECHNOLOGY

For the efficient compression of 3D video data with multiple video and depth components, a number of coding tools are used to exploit the different dependencies among the components. It is assumed that one video component is independently coded by a conventional block-based 2D video coding method, such as AVC or HEVC without additional tools in order to provide compatibility with existing 2D video services. For each additional 3D video component, i.e. the video component of the dependent views as well as the depth maps, additional coding tools are added on top of a 2D coding structure. Thus, a 3D video encoder can select the best coding method for each block from a set of conventional 2D coding tools and additional new coding tools, some of which are described in the following subsections.

3.1 Inter-View Prediction

The basic concept of inter-view prediction, which is employed in all standardized designs for efficient multiview video coding, is to exploit both inter-view and temporal redundancy for compression. Since the cameras of a multiview scenario typically capture the same scene from nearby viewpoints, substantial inter-view redundancy is present. This holds for both texture views and the corresponding depth map images associated with each view, thus inter-prediction can be applied to both types of data independently.

A sample prediction structure is shown in Fig. 1. In modern video coding standards such as AVC and HEVC, inter-view prediction is enabled through the flexible reference picture management capabilities of those standards. Essentially, the decoded pictures from other views are made available in the reference picture lists for use by the inter-picture prediction processing. As a result, the reference picture lists include the temporal reference pictures that may be used to predict the current picture along with the inter-view reference pictures from neighboring views. With this design, block-level decoding modules remain unchanged and only small changes to the high-level syntax are required, e.g., indication of the prediction dependency across views and corresponding view identifiers. The prediction is adaptive, so the best predictor among temporal and inter-view references can be selected on a block basis in terms of rate-distortion cost.

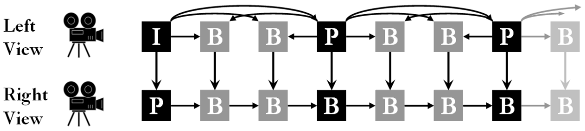


Figure 1. Illustration of inter-view prediction.

In the following subsections, coding tools that go beyond picture-based inter-view prediction are described. Many require changes to block-level syntax and decoding processes, with the benefit of additional gains in coding efficiency.

3.2 Motion/Mode Parameter Prediction

For the joint coding of multiview video, as well as multiview video with associated depth data, dependencies between the different components could be identified and exploited. For instance, scene objects projected to different viewpoints have similar motion and texture characteristics. Additionally, the edge information that is present in the depth components, which correspond to depth discontinuities in the scene, are typically a subset of the edges that could be extracted from the corresponding texture component.

In the context of multiview video coding, it is possible to infer side information used in the decoding process, e.g., motion vectors for a particular block, based on other available data, e.g., motion vectors from other blocks (see Fig. 2). Such inference of coded block data between views could be considered an extension of the basic principle of direct mode prediction in AVC for 2D video coding. Specific extensions to the conventional skip and direct coding modes for multiview video coding were proposed by Koo, et al.[6,7]. Specifically, this method infers side information from inter-view references rather than temporal references. A global disparity vector is determined for each neighboring reference view. The motion vector of a corresponding block in the neighboring view may then be used for prediction of the current block in a different view. This signaling is very minimal and this method has the potential to offer notable reduction in bit rate.

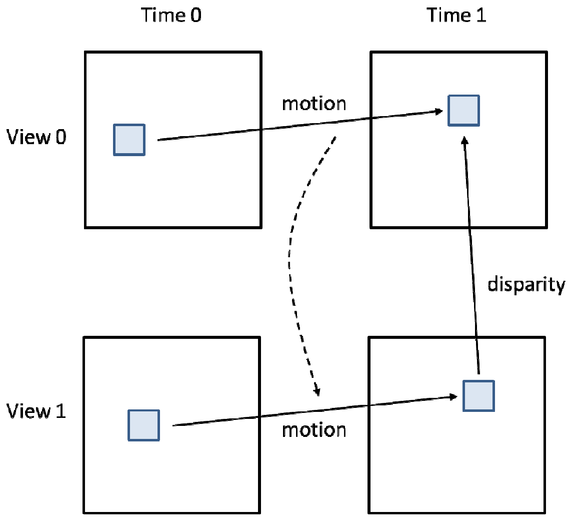


Figure 2. Illustration of motion prediction between views, where the motion vector of view 1 is inferred from the motion vector of view 0 from corresponding blocks at time 1 based on the disparity between those blocks.

More recently, it has been proposed to estimate the disparity for sample locations instead of using a global disparity vector[8]. In this way, motion parameters of a block in the reference view can be used as a motion candidate for the current view. Advanced schemes based on the concept of motion vector competition have also been show to provide very competitive performance and further gains[9].

For more efficient coding of depth maps, an additional coding mode that infers the block partitioning of sub-blocks and associated motion parameters from the co-located block in the associated video picture has been described by Winken, et al.[10]. This technique adaptively decides whether partitioning and motion information are inherited from the co-located region of the video picture for each depth block, or whether new motion data should be transmitted. If such information is inherited, no additional bits for partitioning and motion information are required.

3.3 Depth Coding

As discussed earlier, depth information could be used at the receiver for view generation or used at the encoder to realize more efficient compression with view synthesis prediction schemes. Although the depth data is not directly output to a display and viewed, maintaining the fidelity of depth information is important since the quality of the view synthesis result is highly dependent on the accuracy of the geometric information provided by depth. A depth sample represents a shift value in texture samples from the original views. Thus, coding errors in depth maps result in wrong pixel shifts in synthesized views. This may lead to annoying artifacts, especially along visible object boundaries. Therefore, a depth compression algorithm needs to preserve depth edges much better than traditional coding methods. It is also crucial to strike a good balance between the fidelity of depth data and the overall bandwidth requirement.

A depth signal mainly consists of larger homogeneous areas inside scene objects and sharp transitions along boundaries between objects at different depth values. Therefore, in the frequency spectrum of a depth map, low and very high frequencies are dominant. Video compression algorithms are typically designed to preserve low frequencies and image blurring occurs in the reconstructed video at high compression rates. The need for compression techniques that are adapted to these special characteristics of the depth signal and the requirement on maintain the fidelity of edge information in the depth maps has motivated research in this area.

Approaches have recently been developed to code the depth based on geometric representation of the data. Morvan, et al.[11] model depth images using a piece-wise linear function; they referred to this representation as platelets. Example functions are shown in Fig. 3. Given these functions, the image is subdivided using a quadtree decomposition and an appropriate modeling function is selected for each region of the image in order to optimize the overall rate-distortion cost. This concept has been further refined and a complete set of depth modeling modes, which aim to represent wedge and contour-based patterns of the depth signal, have been introduced[12]. During encoding, each depth block is analyzed for significant edges. If such an edge is present, a block is subdivided into two non-rectangular partitions. The partitions can be separated by a straight line as an approximation of the depth edge, with each partition represented by a constant value. To encode the line, an explicit signaling can be used or the position information can be derived from neighboring blocks. Alternatively, the line position can be derived from the corresponding texture block. When the depth block contains a more complex pattern, its contour can also be derived from the corresponding texture block. The best mode would be determined as part of an optimal rate-distortion process.

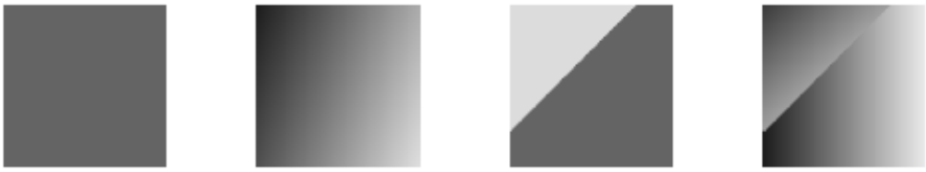


Figure 3. Illustration of geometric representations of depth blocks.

4. EVALUATION

To evaluate the compression efficiency of the different architectures and coding techniques, simulations are conducted using reference software and experimental evaluation methodology that has been developed and is being used by the standardization community[13]. In the experimental framework, multiview video and corresponding depth are provided as input, while the decoded views and views synthesized at select positions are generated as output.

For HEVC simulcast, coding results are based on HM 6.0[14]. The simple multiview extension of HEVC described in section 2.1 is referred to as MV-HEVC (multiview HEVC), while the extension with block-level tools described in section 2.2 is referred to as 3D-HTM (3D HEVC Test Model). The software used for both MV-HEVC and 3D-HTM architectures was HTM 3.1[15], which is based on HM 6.0. Encoder configurations follow those specified in the common test conditions for 3D video coding[16].

Table 1 provides a detailed comparison of performance for the various architectures including Simulcast versus MV-HEVC, Simulcast versus 3D-HTM, and MV-HEVC versus 3D-HTM. For each architecture comparison, three Bjøntegaard delta bit rates are provided:

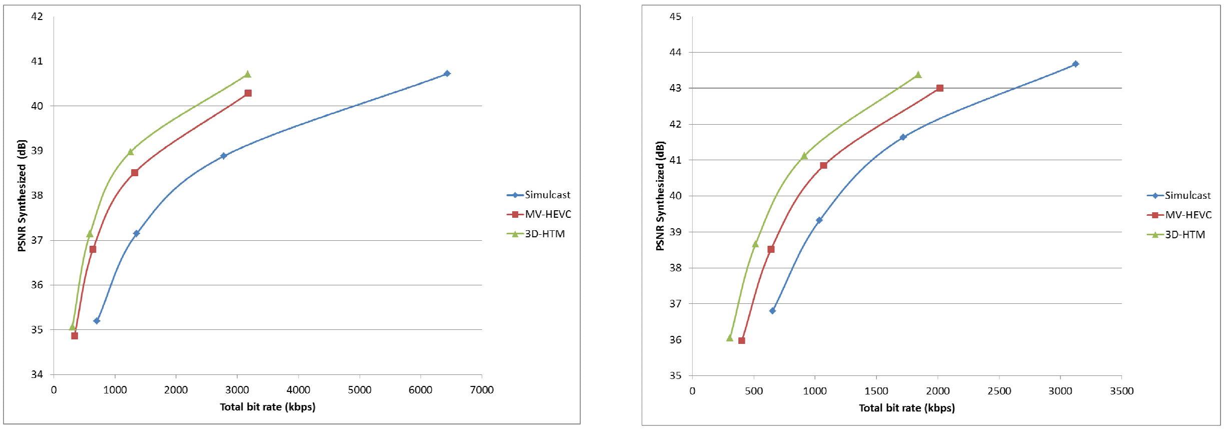
- video only: PSNR of decoded videos, and overall bit rate of texture and depth
- synthesized only: PSNR of synthesized videos, and overall bit rate of texture and depth
- coded & synthesized: PSNR of decoded and synthesized videos, and overall bit rate of texture and depth

Additionally, sample plots for GT_Fly and Kendo sequences are shown in Fig. 4, with the horizontal axis representing overall rate of texture and depth and the vertical axis representing the PSNR of synthesized videos.

Relative to simulcast, which does not utilize inter-view prediction, it is shown through these experiments that inter-view prediction is responsible for the majority of the coding efficiency gains. This leads to a simplified design for efficient multiview video coding (both texture and depth) with good compression capability. Examining the performance between MV-HEVC and 3D-HTM, it is evident that gain measured only on the decoded video is relatively modest. However, the block-based tools offer more substantial gains when also accounting for synthesis quality.

Table 1. Summary of performance comparison of different 3D video coding architectures.

Sequences	**MVHEVC vs Simulcast**			**HTM vs Simulcast**			**HTM vs MVHEVC**		
	video only	synthesized only	coded & synth.	video only	synthesized only	coded & synth.	video only	synthesized only	coded & synth.
Balloons	-28.7%	-24.0%	-24.6%	-33.4%	-39.8%	-37.6%	-6.6%	-20.5%	-17.6%
Kendo	-29.5%	-25.6%	-25.8%	-35.4%	-42.2%	-41.3%	-8.5%	-22.8%	-21.7%
Newspaper	-33.0%	-28.2%	-28.8%	-35.4%	-42.7%	-38.3%	-3.7%	-19.6%	-14.1%
GT_Fly	-47.0%	-44.0%	-44.8%	-53.8%	-57.8%	-55.7%	-13.2%	-24.8%	-20.3%
Poznan_Hall2	-26.6%	-23.0%	-23.4%	-31.4%	-44.3%	-40.4%	-6.8%	-26.5%	-21.5%
Poznan_Street	-41.2%	-37.4%	-38.2%	-43.5%	-45.6%	-44.3%	-4.0%	-13.7%	-10.5%
Undo_Dancer	-46.6%	-42.9%	-44.1%	-49.5%	-60.8%	-56.4%	-5.7%	-31.0%	-21.9%
Average	**-36.1%**	**-32.1%**	**-32.8%**	**-40.3%**	**-47.6%**	**-44.9%**	**-6.9%**	**-22.7%**	**-18.2%**



(a) (b)

Figure 4. Performance comparison of different 3D video coding architectures for (a) GT_Fly and (b) Kendo. The rate includes texture and depth bit rate, while the PSNR is computed on the synthesized videos.

5. CONCLUDING REMARKS

This paper presents several architectures to enable multiview and 3D extensions within the HEVC framework. Associated compression techniques including inter-view prediction, methods for motion and mode parameter prediction and depth coding methods have been reviewed. An evaluation of performance in terms of coded data and synthesized views has been provided.

From the experiments that have been conducted, it is shown that the enabling of inter-view prediction for multiview video signals provides an attractive operating point in terms of coding efficiency and complexity. As demonstrated in the multiview extension of AVC, the architecture is relatively straightforward to implement and requires only minimal changes to high-level syntax and reference picture management. This design enables relatively fast deployment of stereo and multiview implementations based on 2D codecs. An extension of HEVC based on this design principle is expected to be finalized by early 2014.

The inclusion of depth information is another key target of the current 3D video coding extensions development activity. Building on the extensive research that has been done on efficient representation of depth itself, as well as the utilization of depth for texture coding, the benefits of various block-level coding tools in the context of the HEVC design are currently being evaluated through the core experiment process. Experiments have shown that the current set of tools can provide only modest gains when considering the quality of the decoded video components, but substantially higher gains can be achieved when considering the synthesized video quality.

When considering the standardization of depth-based formats, an important consideration in the design is the inter-component dependencies. When inter-view prediction is enabled, there exists a dependency between the different views; this would exist for both texture and depth components. Additionally, decoded information from texture components may be used in the decoding of depth, e.g., the motion prediction techniques discussed in section 3.2. With such a dependency, the texture components could still be extracted independent of the depth, which may be desirable to maintain compatibility with stereo decoders that do not recognize or support the decoding of depth components. On the other hand, there are certain tools that require the decoded depth information to decode the texture, e.g., view synthesis prediction[17]. Such tools have the potential to provide further compression gains but at the cost of stereo compatibility. All of these dependencies will ultimately need to be evaluated in terms of their compression and rendering performance as well as desired level of compatibility and implementation complexity in the standardization development process.

Finally, hybrid solutions have been discussed as possible architecture when compatibility with AVC is desired, for either monoscopic of stereoscopic video. As one would expect, the compression performance would be between that of full-compatible AVC and HEVC solutions[18]. Standardized formats that support mixed codec designs are expected in the near future. However, at this stage, the market needs require further study and the possible inter-component dependencies must be considered as well.

REFERENCES

[1] Urey H., Chellephan K. V., Erden E., and Surman P., "State of the Art in Stereoscopic and Autostereoscopic Displays," *Proceedings of the IEEE*, vol. 99, no. 4, 540-555 (2011).

[2] ITU-T and ISO/IEC JTC 1, "Advanced video coding for generic audiovisual services", ITU-T Recommendation H.264 and ISO/IEC 14496-10 (2012).

[3] Wiegand T., Sullivan G. J., Bjøntegaard G., and Luthra A., "Overview of the H.264/AVC video coding standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576 (2003).

[4] Vetro A., Wiegand T., and Sullivan G. J., "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/AVC Standard", *Proc. of the IEEE*, vol. 99, no. 4, 626 - 642 (2011).

[5] Müller K., Merkle P., and Wiegand T., "3D Video Representation Using Depth Maps", *Proc. of the IEEE*, vol. 99, no. 4, 643 - 656 (2011).

[6] Koo H.S.; Jeon Y.J. and Jeon B.M., "MVC motion skip mode", Joint Video Team (JVT) Doc. JVT-W081, San Jose, CA (2007).

[7] Koo H.S.; Jeon Y.J. and Jeon B.M., "Motion information inferring scheme for multi-view video coding," *IEICE Transactions on Communications*, E91-B(4), 1247-1250 (2008).

[8] Schwarz H. and Wiegand, T. "Inter-view prediction of motion data in multiview video coding" *Proc. Picture Coding Symposium*, Krakow, Poland (2012).

[9] Lin J.L., Chen Y.W., Chang Y.L, Tsai Y.P, Huang Y.W, and Lei S., "3D-CE5.a results on the motion vector competition-based Skip/Direct mode with explicit signaling," Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) Doc. JCT3V-A0045, Stockholm, SE (2012).

[10] Winken M., Schwarz H. and Wiegand T., "Motion Vector Inheritance for High Efficiency 3D Video plus Depth Coding," *Proc. Picture Coding Symposium*, Krakow, Poland (2012).

[11] Morvan Y., Farin D., and de With P.H.N., "Depth-image compression based on an R-D optimized quadtree decomposition for the transmission of multiview images," *Proc. IEEE International Conference on Image Processing*, San Antonio, TX (2007).

[12] Schwarz H., et al., "3D video coding using advanced prediction, depth modeling, and encoder control methods," *Proc. Picture Coding Symposium*, Krakow, Poland (2012).

[13] Video Group, "Report on Experimental Framework for 3D Video Coding," ISO/IEC JTC1/SC29/WG11 MPEG Doc. N11631, Guangzhou, CN (2010).

[14] SVN repository for HM 6.0, https://hevc.hhi.fauhofer.de/svn/svn_HEVCSoftware/tags/HM-6.0

[15] SVN repository for HTM 3.1, https://hevc.hhi.frauhofer.de/svn/svn_3DVCSofware/tags/HTM-3.1

[16] Rusanovskyy D., Müller K., Vetro A., "Common Test Conditions of 3DV Core Experiments," Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) Doc. JCT3V-A1100, Stockholm, SE (2012).

[17] Yea S., and Vetro A., "View Synthesis Prediction for Multiview Video Coding", *Signal Processing: Image Communication*, vol. 24, no. 1+2, pp. 89-100 (2009).

[18] Van Leuven, S., et al. "Overview of the coding performance of 3D video architectures," ISO/IEC JTC1/SC29/WG11 MPEG Doc. m24968, Geneva, CH (2012).