Texture-Aware Depth Prediction in 3D Video Coding

Ce Zhu, *Senior Member, IEEE*, Shuai Li, Jianhua Zheng, Yanbo Gao, and Lu Yu

***Abstract*—3D video has raised great interest in the last decade and currently a new 3D video coding standard, known as 3D video coding extension of High Efficiency Video Coding (3D-HEVC), has been developed. The standard investigates the coding of multiview video plus depth, which consists of texture videos and depth videos of multiple views. Depth video, as a description of geometry information of a scene, is generally composed of large flat regions separated by sharp edges. The conventional video coding may fail to generate an accurate prediction for units with sharp edges due to its block-based prediction which cannot compensate (minor) boundary changes well. In order to attack the problem, a new texture-aware depth inter-prediction method is proposed, which incorporates pixel-oriented weighting in the bi-prediction process by exploiting motion and structure similarities between texture and depth videos. Furthermore, such pixel-oriented weighting scheme can be extended to the uni-prediction process by considering more prediction blocks with small motion vector displacements. Experimental results demonstrate that the adapted 3D-HEVC codec with the proposed method can achieve better rate-distortion performance compared to the original 3D-HEVC standard codec.**

***Index Terms*—Depth prediction, multiview video plus depth, 3D-HEVC, weighted prediction, 3D video.**

I. Introduction

MULTIVIEW video plus depth [1] has been widely recognized as the promising candidate for 3D video representation [2]. It is composed of texture videos and depth videos of multiple views. The Moving Picture Experts Group (MPEG) has established the Joint Collaborative Team on 3D Video Coding (JCT-3V) in order to standardize the 3D video coding, and the standard, namely the 3D Video Coding Extension of High Efficiency Video Coding (3D-HEVC) [3]–[5], was finalized in 2015. While there are a lot of proposals targeting at the part of multiview video coding, great efforts have been made on depth video coding. Compared to texture video, depth video, which describes the geometry of the 3D scene, exhibits distinct characteristics. It is usually composed of large portions of flat regions separated by sharp edges. The conventional video coding methods such as the latest video coding standard - High Efficiency Video Coding (HEVC) [6], are designed for coding the general texture video which may not be efficient in coding the depth video of different characteristics.

On one hand, depth video, which depicts the geometry of a 3D scene, is associated with the corresponding texture video representing the color information of the scene. Strong similarities [7], especially in motion and structure of the scene, can be observed between the depth video and the corresponding texture video. In other words, there is redundancy which can be exploited to enhance the coding efficiency. To be backward compatible with the 2D video, texture video is generally coded before the depth video. Thus the coded color information is available to further enhance the coding of the depth video. On the other hand, depth video is generally used to synthesize virtual views with its corresponding texture video in the decoder side. Therefore, the distortion metric for the depth video coding is supposed to assess quality of the synthesized view rather than that of the depth video itself.

Recent efforts have been made to address the aforementioned problems, such as the depth modeling modes, motion parameter inheritance and synthesized view distortion change models [8]–[11]. Among them, the motion parameter inheritance technique is used to exploit the motion similarity between depth and texture videos. Each unit in the depth image is first split into several sub-units (such as 4*4 unit) and then each sub-unit inherits the motion vector and reference of the corresponding texture sub-unit in the coding. In this way, the redundancy in motion information between the depth video and the corresponding texture video appears to be removed. However, the characteristics of the depth video significantly differ from that of the texture video and the block-based motion vectors used for the texture video may not be good in depth prediction, in view of sharp edges in depth maps and incompetency of block-based prediction in handling an even minor change of depth boundary.

In our previous work [12], we proposed a pixel-wise inter prediction scheme for depth coding on the platform of H.264/AVC, which employs a pixel-wise motion estimation process based on the coded texture information. Since the method is restricted to the edge depth block, the complexity may be usually controlled to a certain extent. However, the worst-case complexity, which requires the full motion estimation process in the decoder side, is high. More efficient prediction techniques are in need.

In particular, bi-directional motion estimation is employed to facilitate more efficient coding of the smooth-prone texture video, where the transition between regions in the texture video is relatively smooth. In contrast, the transition between different regions in the depth video tends to be much sharper, and the averaging of two prediction units selected in the bi-directional motion estimation may produce blurred in-between pixels, thus leading to a poor prediction in the transition areas in the bi-directional prediction of depth video.

To cope with the problem of poor inter-prediction for edge depth blocks, we further develop texture-aware weighted prediction of depth information by incorporating a pixel-wise weighting process in the depth prediction. The weight for each pixel is determined based on a simple comparison of the corresponding texture pixel, thus incurring a minimum amount of computations. The proposed method can be applied in both bidirectional and unidirectional prediction. Experimental results demonstrate that the proposed method can further improve the state-of-the-art 3D-HEVC codec with better rate-distortion performance.

Manuscript received January 4, 2016; accepted March 7, 2016. Date of publication April 20, 2016; date of current version June 3, 2016. This work was supported in part by the National High Technology Research and Development Program of China under Grant 2015AA015903, in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2014Z003, in part by the Ph.D. Programs Foundation of Ministry of Education of China under Grant 20130182110010, and in part by the National Natural Science Foundation of China under Grant 61571102.

C. Zhu and S. Li are with the School of Electronic Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China, and also with the Center for Robotics, UESTC, Chengdu 611731, China (e-mail: eczhu@uestc.edu.cn; shuailichn@gmail.com).

J. Zheng is with Hisilicon Technologies Company Ltd., Shenzhen 518129, China.

Y. Gao is with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

L. Yu is with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2016.2550762

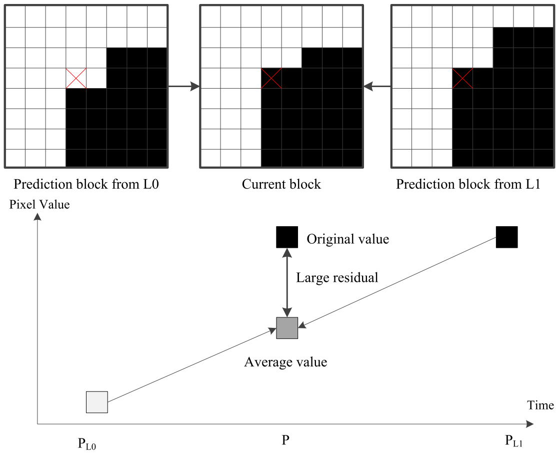


Fig. 1. Illustration of bi-prediction process for a depth block with sharp edge.

The remainder of the paper is organized as follows. In Section II, the proposed texture-aware weighting prediction of depth is presented in detail. Section III describes experimental results and conclusion is drawn in Section IV.

II. Texture-Aware Depth Prediction

Two types of inter-prediction techniques, namely the uni-prediction with one motion vector indicating one prediction (reference) unit and the bi-prediction with two motion vectors indicating two prediction (reference) units, are extensively applied in the video coding. In the prediction process, weighted prediction is introduced primarily for compensating the overall illumination change among different frames. Different methods [13]–[16] have been proposed for estimation of the weight parameters. However, such methods focus on compensating the global illumination change, which is inappropriate or ineffective for the prediction of the depth video with sharp edges.

Depth video depicts geometric characteristics of a scene, which generally presents sharp transitions between different regions (objects) in the depth maps. As shown in Fig. 1, a depth unit comprising parts of two different objects (foreground and background) is coded with bi-prediction, where either prediction unit from prediction list L0 or L1 fails to accurately predict the current unit around edge. Taking the pixel marked by "X" in Fig. 1 as an instance, the average of the prediction pixels from the two prediction blocks, indicated by "P_{L0}" and "P_{L1}" as shown in the lower part in Fig. 1, clearly deviates from the original pixel "P", leading to a large prediction error. As a matter of fact, even with the conventional weighted prediction in texture video coding, the prediction values still cannot match those around the transition area well. To deal with the problem, we develop a texture-aware depth prediction method by appropriately selecting the weight assigned to each prediction pixel, thus greatly increasing the prediction accuracy.

A. Texture-Aware Depth Prediction in Bi-Prediction

The proposed method is to improve the prediction accuracy for those depth pixels of which the two reference pixels differ significantly, as that shown in the abovementioned example in Fig. 1. Those pixels are generally located in the transition area between different objects/regions. We may identify such a pixel (noted as a hard-to-be-predicted pixel) in the current coding unit by comparing the difference of its two reference pixels against a threshold *Diff*. Here we set the *Diff* to be 10 (a wide range of values above 10 appear to work well according to our experiments).

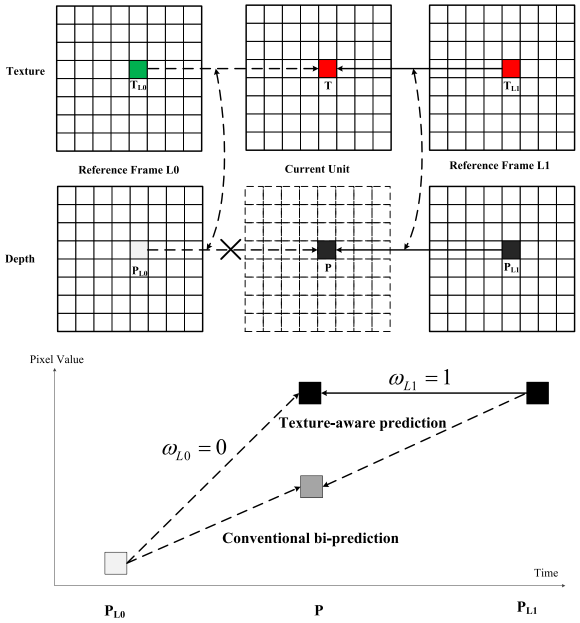


Fig. 2. Illustration of the proposed texture-aware depth prediction method in the bi-prediction process.

In order to obtain better prediction values for the identified hard-to-be-predicted depth pixels, we take the corresponding texture information into consideration. Since texture video and its associated depth video are the color and geometrical projections of a same scene, respectively, they share similarity in both structure and motion. Consequently, the prediction depth pixel that better describes the current depth one may be determined by the corresponding texture counterpart. That is to say, we may identify the better prediction pixel in the texture domain instead and then use the pixel in the depth prediction. In short, the prediction depth value can be obtained as

$$P = P_{L0} \cdot w_{L0} + P_{L1} \cdot w_{L1} \tag{1}$$

where

$$w_{L0} + w_{L1} = 1; \\ \begin{cases} w_{L0} > w_{L1}, & if\ |T_{L0} - T| < |T_{L1} - T| \\ w_{L0} < w_{L1}, & otherwise \end{cases} \tag{2}$$

where T, T_{L0} and T_{L1} are the corresponding texture values of the current pixel and the two prediction pixels, respectively. Considering that the depth transition between different regions is generally sharp, the weight factor is set to 0 or 1 in this work for simplicity. In such a way, the wrong prediction value may be fully excluded and thus the final prediction value becomes better, as shown in Fig. 2.

From the above it can be seen that the proposed method only needs to calculate the prediction values for the hard-to-be-predicted pixels, which is just a small portion of the unit. Compared to our previous work [12], the complexity can be further reduced. Moreover, it is worth noting that the proposed method can be easily generalized to the case of using multiple prediction units.

B. Texture-Aware Depth Prediction in Uni-Prediction

In the uni-prediction case of video coding, only one reference (prediction) unit is considered with one motion vector. In order to identify the hard-to-be-predicted pixels in the current depth block, additional prediction units that are also good matched ones to the current unit are needed. In view that the depth image is generally composed of large flat regions and motion estimation is performed on the block

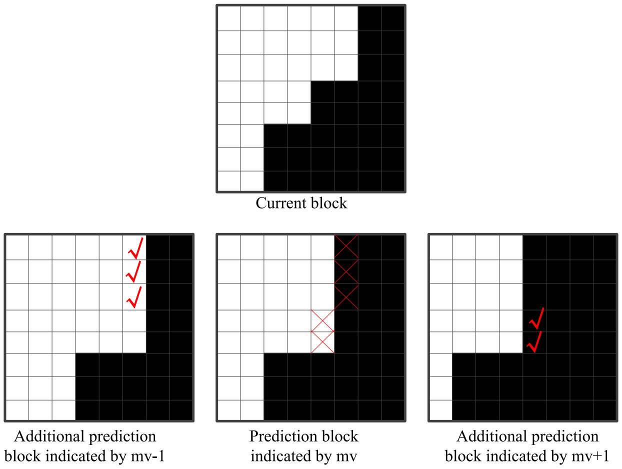

Current block
Additional prediction block indicated by mv-1
Prediction block indicated by mv
Additional prediction block indicated by mv+1

Fig. 3. Illustration of two additional prediction blocks with a horizontal shift of one pixel.

basis, a small variation to the located motion vector may also produce a good prediction block. Especially when the prediction unit does not match the current depth unit exactly, e.g., for the depth block with sharp edges, such variations may further provide better predictions for some pixels around the edges. Therefore, we consider adding/subtracting a small displacement (Δmv) to/from the current motion vector (mv), and then take the corresponding blocks as additional prediction units. For simplicity, only a horizontal motion vector displacement is used in our current work shown in (3), which generates two additional prediction units accordingly.

$$mv_1 = mv + \Delta mv, mv_2 = mv - \Delta mv \quad (3)$$

As shown in Fig. 3, some of the pixels which cannot be predicted well by the prediction block (pixels marked with "X") may now be well approximated by pixels in the additional prediction units with small horizontal displacements of ±1. The determination of the small motion vector displacement will be discussed in the following.

With the three prediction blocks available in the uni-prediction case, it is natural and straightforward to extend the proposed texture-aware depth prediction method shown in (1) and (2), where there will be three weights. In view of the sharp depth edges and simplifying the prediction process, the weight values are also taken as 0 or 1. That is to say, among the three reference pixels, the pixel whose texture value is the closest to that of the current pixel will be selected and its depth value is then used to predict the current depth pixel.

Determination of the small motion vector displacement

The motion vector displacement (Δmv) can be set as a small integer between 1 and 5 empirically. In order to determine the best displacement for a unit, a rate-distortion optimization (RDO) process can be performed where the rate-distortion cost of using each Δmv needs to be compared. Since the bits used to signal the varying Δmv are close to each other, to simplify the computation we only compare the residuals after prediction to select the best one. According to our experiments with the standard test sequences suggested by the JCT-3V, the values of 1 and 3 appear to work best. Therefore in order to further reduce the bits to signal the selected Δmv, only the two values are considered in the current depth coding and a flag is written in the bit stream to indicate which one is used.

C. Complexity Analysis

The proposed depth prediction scheme is integrated into the 3D-HEVC codec as one more mode of depth prediction added. As the merge mode is employed in the current 3D-HEVC implementation and there are multiple merge candidates available for inter prediction, the proposed method can be applied to each of these prediction processes with a given motion vector. To reduce the complexity, we only consider the proposed method for the motion vectors selected from the conventional motion estimation and the best merge candidate (plus the motion parameters inheritance if it is not the best candidate). In this way, the added mode based on the proposed method only introduces a small complexity increase with at most three groups of block comparisons added. To further enhance the prediction accuracy, the proposed method may also be readily incorporated into the motion estimation process to select the best motion vectors in the depth prediction, which is expected to yield the best coding performance, at the cost of increasing complexity substantially.

Moreover the proposed texture-aware depth prediction can be implemented at different levels, e.g., at the pixel level as described above or at sub-unit level to further reduce the prediction complexity. For the sub-unit level implementation, the average of pixel values in a sub-unit is used to replace the pixel values in the prediction process.

TABLE I
TEST SEQUENCES SUGGESTED IN COMMON TEST CONDITIONS (CTC) BY JCT-3V

Sequence	Resolution	Frames to be encoded	Input views
Balloons	1024x768	300	1-3-5
Kendo	1024x768	300	1-3-5
Newspaper_CC	1024x768	300	2-4-6
GT_Fly	1920x1088	250	9-5-1
Poznan_Hall2	1920x1088	200	7-6-5
Poznan_Street	1920x1088	250	5-4-3
Undo_Dancer	1920x1088	250	1-5-9
Shark	1920x1088	300	1-5-9

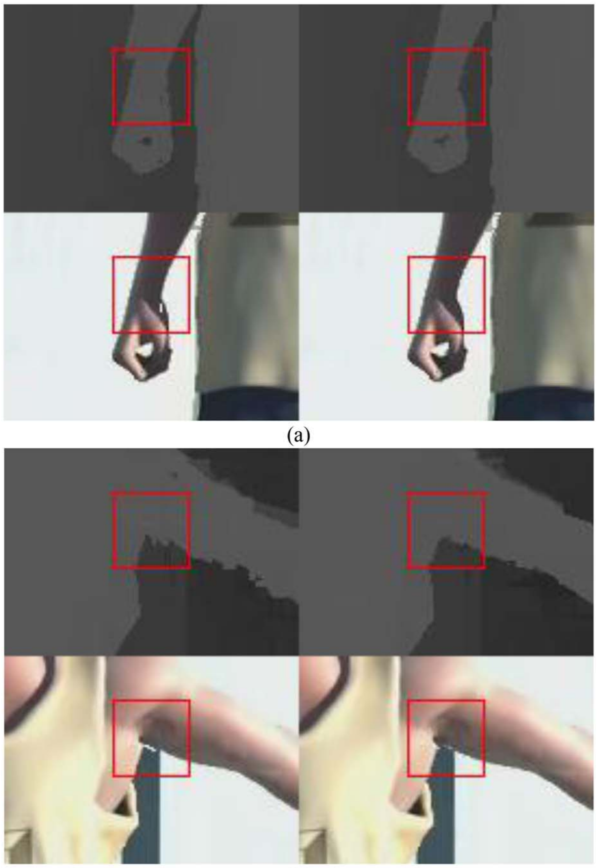
III. EXPERIMENTAL RESULTS

Experiments were performed on the platform of the 3D-HEVC test model (HTM 11.0) to evaluate efficiency of the proposed texture-aware depth prediction method. The simulation environment is set as suggested in the common test conditions (CTC) [17] specified by the JCT-3V. As the proposed method just changes the inter prediction process, only the random access (RA) configuration is needed to be tested. Eight standard test sequences are used, including three sequences of resolution 1024*768 and five sequences of resolution 1920*1088 as tabulated in Table I. Three views of the texture video and the depth video for each sequence are used. Six more views are rendered using the view synthesis method adopted in the HTM 11.0, while more advanced methods [18]–[20] may also be employed for better quality of view synthesis. BD-Rate saving [21] in terms of total bitrate of the texture videos and depth videos and the quality of the original and synthesized texture views (9 views in total) is used to measure the performance. Table II tabulates all the experimental results of the proposed method at the pixel level.

As can be seen from the table, 0.15% BD-Rate saving in average can be achieved for the synthesis view by the proposed texture-aware depth prediction method over the original state-of-the-art HTM codec, with the highest coding gain up to 0.31% for the video of "Shark". Since the performance is measured in terms of the quality of the 9 views (3 original plus 6 synthesized views) and the total bitrate (3 texture and 3 depth videos), over 0.1% BD-Rate saving is always regarded as a relatively significant coding gain in the JCT-3V community when compared with the well-developed 3D-HEVC anchor.

TABLE II
CODING RESULTS OF THE PROPOSED METHOD AGAINST HTM 11.0 UNDER RANDOM ACCESS CONFIGURATION

	video 0	video 1	video 2	synth PSNR / total bitrate	enc time	dec time
Balloons	0.00%	-0.02%	-0.03%	-0.13%	120.7%	105.6%
Kendo	0.00%	-0.07%	0.05%	-0.15%	118.2%	86.3%
Newspaper_CC	0.00%	-0.01%	0.09%	-0.13%	117.6%	93.5%
GT_Fly	0.00%	-0.02%	0.01%	-0.29%	119.6%	104.6%
Poznan_Hall2	0.00%	0.22%	-0.08%	0.04%	123.0%	119.6%
Poznan_Street	0.00%	0.03%	-0.15%	-0.01%	120.4%	92.0%
Undo_Dancer	0.00%	-0.04%	0.05%	-0.20%	119.3%	106.6%
Shark	0.00%	-0.09%	0.06%	-0.31%	120.7%	92.2%
1024x768	0.00%	-0.03%	0.04%	-0.14%	118.8%	95.2%
1920x1088	0.00%	0.02%	-0.02%	-0.15%	120.6%	103.0%
Average	0.00%	0.00%	0.00%	-0.15%	119.9%	100.1%



(a)

(b)

Fig. 4. Sample snapshots of the reconstructed depth maps and synthesized images in "Undo_Dancer" coded by: (left) HTM11.0, (right) Proposed.

Note that we only consider depth coding here without changing anything about texture coding, while slight changes in the coding results of the texture video (video 1 and 2 in the table) is due to the inter-component prediction in the HTM. Although encoding time increases due to the addition of the proposed depth prediction mode, the decoding time is almost the same as that in HTM 11.0.

Fig. 4 shows some snapshots of the reconstructed depth maps and the synthesized images by the HTM11.0 with and without the proposed coding scheme using the same QP of 25, respectively, by coding the video of "Undo_Dancer". It can be seen that our proposed depth coding scheme can obtain better results than HTM11.0 in depth quality, especially around the edges. For example, there is a small strip in the lower part of the arm showing noticeable and annoying artifacts in the synthesized image as indicated in the red box of Fig. 4(a) due to the depth value change from foreground to background value. Similar results can be observed in Fig. 4(b). From the part of the arm at the top of the red box in Fig. 4(a), we can also see that HTM11.0 may introduce severe coding (prediction) distortion due to that the distortion metric used in the coding is to measure the distortion of the synthesized view rather than the depth map itself. If all the modes available in HTM11.0 cannot present an adequately good prediction for a block, the block may be coded with a relatively large distortion in terms of depth value as long as the quality of synthesized views for a given set of target viewpoints is acceptable based on the RDO selection. However, it can be seen that with a better prediction by our proposed approach, such a block can be reconstructed with much better quality.

IV. CONCLUSION

In this paper, an effective yet efficient texture-aware depth prediction method has been proposed for depth coding. The proposed method attempts to produce good predictions of the pixels around the sharp transition areas in the depth video in the cases of bi-prediction or uni-prediction by incorporating a pixel-wise weighted prediction scheme. Those hard-to-be-predicted depth pixels are first identified and the pixel-wise prediction can be performed by checking the corresponding texture information, producing better depth prediction results for those "hard" pixels. The experimental results show that the state-of-the-art 3D-HEVC codec integrated with the proposed method can further improve coding performance in terms of BD-Rate saving.

REFERENCES

[1] K. Müller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proc. IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.

[2] C. Fehn, R. de la Barre, and S. Pastoor, "Interactive 3-DTV-concepts and key technologies," *Proc. IEEE*, vol. 94, no. 3, pp. 524–538, Mar. 2006.

[3] K. Muller *et al.*, "3D high-efficiency video coding for multi-view video and depth data," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3366–3378, Sep. 2013.

[4] G. Tech, K. Wegner, Y. Chen, and S. Yea, *3D-HEVC Draft Text 5*, document JCT-3V-I1001, Joint Collaborative Team, Sapporo, Japan, Jul. 2014.

[5] Y. Chen, G. Tech, K. Wegner, and S. Yea, *Test Model 9 of 3D-HEVC and MV-HEVC*, document JCT-3V-I1003, Joint Collaborative Team, Sapporo, Japan, Jul. 2014.

[6] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[7] J. Lei, S. Li, C. Zhu, M.-T. Sun, and C. Hou, "Depth coding based on depth-texture motion and structure similarities," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 2, pp. 275–286, Feb. 2015.

[8] M. Domanski *et al.*, "High efficiency 3D video coding using new tools based on view synthesis," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3517–3527, Sep. 2013.

[9] H. Schwarz *et al.*, "3D video coding using advanced prediction, depth modeling, and encoder control methods," in *Proc. IEEE Pict. Coding Symp. (PCS)*, Kraków, Poland, May 2012, pp. 1–4.

[10] M. Winken, H. Schwarz, and T. Wiegand, "Motion vector inheritance for high efficiency 3D video plus depth coding," in *Proc. IEEE Pict. Coding Symp. (PCS)*, Kraków, Poland, May 2012, pp. 53–56.

[11] G. Tech, H. Schwarz, K. Muller, and T. Wiegand, "3D video coding using the synthesized view distortion change," in *Proc. Pict. Coding Symp. (PCS)*, Kraków, Poland, May 2012, pp. 25–28.

[12] S. Li, J. Lei, C. Zhu, L. Yu, and C. Hou, "Pixel-based inter prediction in coded texture assisted depth coding," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 74–78, Jan. 2014.

[13] P. Bordes, *Weighted Prediction*, document JCTVC-F265, Joint Collaborative Team on Video Coding, Turin, Italy, Jul. 2011.

[14] A. Tanizawa, T. Chujoh, and T. Yamakage, *Explicit Weighted Prediction With Simple WP Parameter Estimation*, document JCTVC-F326, Joint Collaborative Team on Video Coding, Turin, Italy, Jul. 2011.

[15] Y. Ye and E.-S. Ryu, *Improved Weighted Prediction*, document JCTVC-G065, Joint Collaborative Team on Video Coding, Geneva, Switzerland, Nov. 2011.

[16] A. Tanizawa, T. Chujoh, and T. Yamakage, "Multi-directional implicit weighted prediction based on image characteristics of reference pictures for inter coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Orlando, FL, USA, Oct. 2012, pp. 1545–1548.

[17] K. Müller and A. Vetro, *Common Test Conditions of 3DV Core Experiments*, document JCT-3V-G1100, Joint Collaborative Team, San Jose, CA, USA, Jan. 2014.

[18] Y. Zhao, C. Zhu, Z. Chen, D. Tian, and L. Yu, "Boundary artifact reduction in view synthesis of 3D video: From perspective of texture-depth alignment," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 510–522, Jun. 2011.

[19] C. Zhu and S. Li, "Depth image based view synthesis: New insights and perspectives on hole generation and filling," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 82–93, Mar. 2016.

[20] C. Zhu and S. Li, "Multiple reference views for hole reduction in DIBR view synthesis," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Beijing, China, Jun. 2014, pp. 1–5.

[21] *An Excel Add-In for Computing Bjontegaard Metric and Its Evolution*, document VCEG-AE07, ITU-T SG16 Q.6, Video Coding Experts Group, Marrakesh, Morocco, Jan. 2007.