TRANSMISSION DISTORTION MODELING FOR VIEW SYNTHESIS PREDICTION BASED 3-D VIDEO STREAMING

Pan Gao, Wei Xiang, and Lijuan Zhang

School of Mechanical and Electrical Engineering, University of Southern Queensland,
Toowoomba, QLD 4350, Australia

ABSTRACT

View synthesis prediction (VSP) is an important tool for improving the coding efficiency in the next generation three-dimensional (3-D) video systems. However, VSP will result in a new type of inter-view error propagation when the multi-view video plus depth (MVD) data are transmitted over the lossy networks. In this paper, this new type of error propagation is characterized and modeled. Firstly, a new analytic model is formulated to estimate the expected transmission distortion caused by error propagation from the synthesized reference view. Then, the compound impact of the transmission distortions of both the texture video and the depth map on the quality of the synthetic reference view is mathematically analysed. Our extensive simulation results demonstrate that the proposed transmission distortion model is very accurate.

Index Terms— View synthesis prediction, 3-D video transmission, transmission distortion modeling

1. INTRODUCTION

Recently, three dimensional (3-D) video has become increasingly prevalent as a new multimedia technique, which evolves towards new systems that include glassless displays and provide an immersive sense of realism [1]. In such 3-D video systems, the multi-view video plus depth (MVD) representation, which is based on depth image-based rendering (DIBR) [2], has been generally considered to be the best format for representing 3-D scenes [3]. The MVD format consists of texture videos and depth maps for a limited number of original camera views of the same natural scene. By using the MVD representation, an arbitrary number of virtual views can be generated from the transmitted multi-view videos and their corresponding depth maps via the DIBR technique.

Although the MVD representation could greatly reduce the data volume of 3-D video having to be transmitted, the presence of multiple cameras as well as additional depth information creates new challenges for compression. Generally speaking, there are two major coding issues for the 3-D video codec. One is depth map coding, while the other is texture video coding with enhanced inter-view prediction. Since a depth map has considerably varying statistics and characteristics compared with the texture video, many advanced coding tools are incorporated into the 3-D video codec to improve the performance of depth map compression, such as platelet-based coding [5], the shape-adaptive wavelet transform [6], and the view synthesis guided rate-distortion optimization mechanism [7], etc. With regards to the texture video coding, in order to further improve coding efficiency, one of the key technologies for reducing the inter-view redundancy is view synthesis prediction (VSP). Its basic idea involves warping the reference view to the target viewpoint, whereby its available depth map helps facilitate the warping process. The synthesized view frames are then utilized as complementary reference frames for non-translational disparity compensated prediction. This idea was first proposed in [8]. Later, based on VSP, Yea *et al.* [9] devised a rate-distortion optimized Multi-view Video Coding (MVC) framework to improve the coding performance. Shimizu *et al.* [10] designed a related VSP scheme, in which the original video of base views and the residue of enhancement views are scalably encoded by a traditional video coding process. For low decoder complexity purpose, Tian *et al.* [11] proposed a backward VSP design using the depth of the current view to perform a pixel-based warping. Due to its superior performance, VSP has been adopted by both the upcoming H.264/AVC-based and high efficiency video coding (HEVC)-based 3-D video coding standards [12].

In unreliable underlying networks, transmission of compressed video is highly susceptible to channel errors, which will cause decoding failure at the receiver side. Moreover, the use of motion compensation prediction causes these errors to propagate to subsequent frames, and significantly degrade the picture presentation quality. This type of video distortion is usually called transmission or channel distortion. To date, in order to optimize performance and resource allocation in video communications, there exist a large number of reported studies in the literature on the analysis and modeling of the influence of channel losses on monoscopic video systems. These studies could be roughly categorized based on whether the distortion is estimated and tracked at the pixel level [13], the macroblock (MB) level [14], or the frame level [15]. However, to the best knowledge of the authors, only a limited number of publications have been reported on distortion modeling for multi-view 3-D video coding and transmission. Zhou *et al.* [16] developed a recursive mathematical model for MVC-based video coding systems to estimate the expected channel-induced distortion at the frame and sequence levels without the use of depth information. Machiavello *et al.* [17] introduced a synthesized view distortion model for reference frame selection in loss-resilient depth map coding, where the synthesized view distortion due to errors in the reconstructed depth map is approximated by a per-pixel quadratic weighting function. Thereafter, this idea was extended to the encoding of both texture and depth map [18]. In these two algorithms, inter-view error propagation is not considered and only the distortion in the synthesized views is modeled. In [19], in order to fully improve the overall quality of reconstructed 3-D video, a summative transmission distortion model was presented for loss-aware rate-distortion optimized mode switching, in which both the channel-induced distortions in the rendered view and the coded texture video are characterized. Although the above models can achieve some modest performance improvements on the accuracy of channel distortion estimation, they all are built upon the conventional MVD-based 3-D video coding frame-

The work was supported in part by a Smart Futures Fellowship funded by the Queensland Government of Commonwealth Australia.

work, without consideration of VSP.

In this work, we will concentrate our efforts on developing an analytic transmission distortion model for an improved 3-D video coding framework. The major contributions of this paper are two-fold. Firstly, we construct a statistical model to optimally estimate the expected distortion in the dependent view, in which the VSP based inter-view error propagation is explicitly and accurately accounted for. Secondly, through taking into account the texture image characteristics and mimicking the rendering process, the distortion of the synthetic reference view due to packet losses is mathematically decomposed into two components, i.e., the transmission distortion induced by texture errors and the transmission distortion induced by depth errors. This proposed model can provide helpful insights into the behavior of channel distortion caused by VSP. To the best of the authors' knowledge, this work is the first of its kind to theoretically model the expected distortion in VSP based 3-D video streaming.

The rest of this paper is organized as follows. Section 2 briefly describes the VSP based encoding prediction structure. In Section 3, based on the propagating behaviour of transmission errors, a new channel distortion estimation model for 3-D video transmission is developed. The experimental results are presented and discussed in Section 4. Finally, concluding remarks are drawn in Section 5.

2. VIEW SYNTHESIS PREDICTION

Disparity-compensated prediction is a well-known technique for exploiting the redundancy between different views, which can provide gains when temporal correlation is lower in comparison to spatial correlation. However, it does not utilize some essential features of multi-view video. While block translation is good for predicting temporally adjacent frames, it is less accurate for predicting spatially adjacent ones because the disparity of an object in one frame relative to another frame depends on the distance of the object to the camera, as well as the camera setup and scene geometry.

To exploit these new features of multi-view video, view synthesis has been proposed for enhanced prediction in multi-view 3-D video coding. Following the 3-D video coding standard specification in [3], this work is based on a two-view coding configuration with 1D parallel camera setting, in which view synthesis is employed as an alternative means of prediction at the encoder. Specifically, at each time instance, there are two views, i.e., a left view and a right view, with each view being composed of a texture video and a depth map. The left view is firstly encoded by traditional motion-compensated prediction, which can be compatible with the H.264/AVC or HEVC standard bit stream. Then, a virtual version of the right target viewpoint is synthesized from the already encoded left view according to the reconstructed depth information and the camera parameters. This virtual view will exhibit a object structure more similar to the original right view. Finally, based on the synthesized reference view, disparity-compensated prediction is employed to encode the texture video of the right view in addition to the existing temporal prediction. Note that in this study, in order to focus on the analysis of the error propagation behaviour with VSP, we disable the translational disparity compensation prediction directly from the left view.

3. PROPOSED TRANSMISSION DISTORTION ESTIMATION MODEL

On the basis of the improved MVD-based 3-D video coding framework, it is evident that the transmission distortion for MVD-based 3-D video consists of the expected reconstruction distortions in the texture video as well as the depth map. Since the depth map is encoded by traditional joint motion/disparity-estimation-based MVC, its transmission distortion model is actually the same as that of the MVC-based video transmission in [16]. As for the texture video transmission, the left and right views exhibit different characteristics facing transmission errors. For the left view, the texture video is encoded by temporal motion-compensated prediction without the use of the depth map. As a result, the expected distortion model of the texture video of the left view is also similar to that of single view video transmission. On the other hand, due to view synthesis based inter-view prediction, the transmission errors of the texture video in the right view come from not only itself, but also the synthetic reference frame. In particular, the channel errors of the depth map in the left view will cause incorrect projection of texture pixels, which may lead to unexpected holes or overlaps in the synthesized reference view. Moreover, this kind of geometry errors will further propagate to the right view via disparity compensation step. Therefore, this work, unlike other reported distortion estimation methods, focuses primarily on analytic transmission distortion modeling in the right view, where the inter-view error propagation often exhibits a very complicated and irregular behavior.

To better describe the proposed distortion model, in the subsequent derivation, the symbols T, D and V are used to indicate the texture, depth information, and the virtual reference view, respectively. Subscripts L and R represent the left and right views, respectively.

Above all, let $T_{R,t}^{x,y}$ be the original value of pixel (x,y) in texture frame t of the right view, $\hat{T}_{R,t}^{x,y}$ be the reconstructed signal at the encoder, and $\tilde{T}_{R,t}^{x,y}$ be the corresponding reconstructed signal at the decoder. Assume that (x,y) is predicted from a pixel (u,v) of the frame t of the virtual view with the prediction residual signal $\hat{e}_{R,t}^{x,y}$. Thus, denoted by $\hat{V}_t$ the virtual view image synthesized from the compressed texture video and depth map of the left view, the encoder prediction of the pixel (x,y) is $\hat{V}_t^{u,v}$, and we can have $\hat{T}_{R,t}^{x,y} = \hat{V}_t^{u,v} + \hat{e}_{R,t}^{x,y}$. At the receiver, if the current pixel is correctly received, the decoder reconstructs pixel (x,y) by $\tilde{T}_{R,t}^{x,y} = \tilde{V}_t^{u,v} + \hat{e}_{R,t}^{x,y}$. $\tilde{V}_t^{u,v}$ represents the decoder reconstruction of pixel (u,v) in the virtual reference view image, which is synthesized from the decoded texture video and depth map of the left view. If the current pixel is lost during transmission, we assume that the decoder performs a pixel copy from the previous texture frame $t-1$ within the right view. Denoting by $\rho(x,y)$ the estimated matching pixel in frame $t-1$ for pixel (x,y) based on the estimated motion vector, the concealed value for pixel (x,y) is then $\tilde{T}_{R,t-1}^{\rho(x,y)}$.

Suppose the packet loss rate is known as p, which is equivalent to the slice loss rate. Then the transmission distortion in terms of the mean squared error (MSE) for the texture pixel (x,y) of the right view can be derived as follows

$$\begin{aligned} d(\hat{T}_{R,t}^{x,y}) &= E\left\{(\hat{T}_{R,t}^{x,y} - \tilde{T}_{R,t}^{x,y})^2\right\} \\ &= (1-p)E\left\{(\hat{V}_t^{u,v} - \tilde{V}_t^{u,v})^2\right\} \\ &\quad + pE\left\{(\hat{T}_{R,t}^{x,y} - \tilde{T}_{R,t-1}^{\rho(x,y)})^2\right\} \\ &= (1-p)E\left\{(\hat{V}_t^{u,v} - \tilde{V}_t^{u,v})^2\right\} \\ &\quad + p\left\{(\hat{T}_{R,t}^{x,y} - \hat{T}_{R,t-1}^{\rho(x,y)})^2\right\} + p\left\{(\hat{T}_{R,t-1}^{\rho(x,y)} - \tilde{T}_{R,t-1}^{\rho(x,y)})^2\right\} \\ &= (1-p)d_{ep}(\hat{V}_t^{u,v}) + pd_{ec}(\hat{T}_{R,t}^{x,y}) + pd_{ep}(\hat{T}_{R,t-1}^{\rho(x,y)}) \end{aligned} \tag{1}$$

where $d_{ep}(\hat{V}_t^{u,v})$ denotes the error propagation distortion introduced by the reference pixel (u,v) of the rendered view, $d_{ec}(\hat{T}_{R,t}^{x,y})$ represents the so-called error concealment distortion, and $d_{ep}(\hat{T}_{R,t-1}^{\rho(x,y)})$ refers to the error propagation distortion of the concealed pixel $\rho(x,y)$. While deriving (1), $d_{ec}(\hat{T}_{R,t}^{x,y})$ is assumed to be uncorrelated with $d_{ep}(\hat{T}_{R,t-1}^{\rho(x,y)})$ [20]. $d_{ec}(\hat{T}_{R,t}^{x,y})$ can be readily measured by simulating packet losses at the encoder with the knowledge of the packet loss rate, whereas $d_{ep}(\hat{T}_{R,t-1}^{\rho(x,y)})$ can be recursively calculated under the given inter dependencies established during error concealment process. Therefore, the only thing left in (1) is how to compute the $d_{ep}(\hat{V}_t^{u,v})$. Since the synthetic reference view is rendered from the texture video and depth map of the left view by the pre-defined warping function, the local video characteristics of the synthesized view video would be very similar to those of the left view video. Thus, the reconstruction errors in the rendered reference view can reflect on the source left view.

In the 3-D warping procedure, as shown in Fig. 1, when using the compressed depth map, it is assumed that the pixel (x',y') in the left view could be projected to "Point M" in the 3-D world coordinate by the depth value Z; then "Point M" could be projected to (u,v) in the synthetic reference view. When using the decoded depth map, due to the reconstructed depth map errors caused by packet losses, the pixel (x'',y') in the left view may be projected to "Point M′" in the 3-D world coordinate by the distorted depth value Z′; then "Point M′" could also be projected to (u,v) in the synthesized reference view. The horizontal position difference between (x',y') and (x'',y') in the left view is represented as $\Delta = x'' - x'$. When the cameras are in parallel positions, Δ is already proven to be approximately proportional to the depth map error as in [7], i.e.,

$$\Delta = \alpha(\hat{D}_{L,t}^{x',y'} - \tilde{D}_{L,t}^{x',y'}) \tag{2}$$

where $\hat{D}_{L,t}^{x',y'}$ and $\tilde{D}_{L,t}^{x',y'}$ indicate the encoder and decoder reconstructed pixel values of (x',y') in the depth image of the left view, respectively, and α is the proportional coefficient determined by the following equation

$$\alpha = \frac{fL}{255}\left(\frac{1}{Z_{\text{near}}} - \frac{1}{Z_{\text{far}}}\right) \tag{3}$$

where f is the common focal length, L is the baseline distance between the left view and rendered reference view, and Z_{near} and Z_{far} are the physical values of the nearest and farthest depth of the scene, respectively.

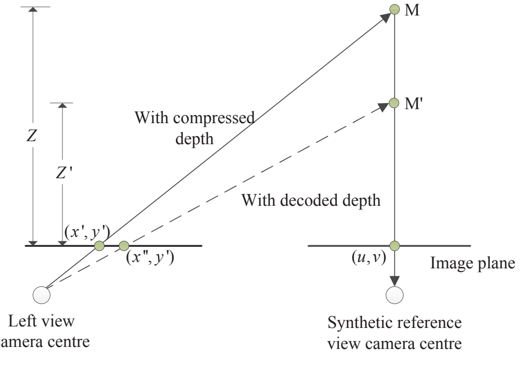

Fig. 1. 3-D warping illustration with distorted depth (parallel camera setup).

Let $\hat{T}_{L,t}^{x',y'}$ and $\tilde{T}_{L,t}^{x',y'}$ denote the reconstructed values of the pixel (x',y') in texture frame t of the left view at the encoder and decoder, respectively. The decoder reconstructed value of pixel (x'',y') in texture frame t of the left view is denoted by $\tilde{T}_{L,t}^{x'',y'}$. Based on the above rendering error analysis with the distorted depth for the synthetic reference view, $d_{ep}(\hat{V}_t^{u,v})$ can be further derived as follows

$$\begin{aligned} d_{ep}(\hat{V}_t^{u,v}) &= E\left\{(\hat{V}_t^{u,v} - \tilde{V}_t^{u,v})^2\right\} \\ &= E\left\{(\hat{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})^2\right\} \\ &= E\left\{(\hat{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x',y'} + \tilde{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})^2\right\} \\ &= E\left\{(\hat{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x',y'})^2\right\} + E\left\{(\tilde{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})^2\right\} \\ &\quad +2E\left\{(\hat{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x',y'})(\tilde{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})\right\} \end{aligned} \tag{4}$$

where $E\left\{(\hat{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x',y'})^2\right\}$ represents the average view rendering distortion induced by texture errors of the left view, i.e., the channel-induced distortion occurring at pixel (x',y') of the left view, $E\left\{(\tilde{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})^2\right\}$ represents the view rendering distortion induced by depth errors, and $E\left\{(\hat{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x',y'})(\tilde{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})\right\}$ approximates to zero [22]. From the above derivation, it can be seen that the view rendering distortion due to depth errors can be represented by the MSE between the corresponding pixel (x',y') in the reconstructed texture image and the pixel (x',y') derived by translating with a geometry displacement Δ.

For a particular rendered reference view, $E\left\{(\tilde{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})^2\right\}$ can be characterized by a linear model and expressed as follows [21]

$$\begin{aligned} E\left\{(\tilde{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x'',y'})^2\right\} &= ||\Delta||^2 \times \psi_r \\ &= \psi_r \alpha^2 E\left\{(\hat{D}_{L,t}^{x',y'} - \tilde{D}_{L,t}^{x',y'})^2\right\} \end{aligned} \tag{5}$$

where ψ_r is a linear parameter associated with the image contents, which can be readily computed from the energy density of the input texture video of the left view.

Thus, substituting (5) into (4), we can model the linear relationship between the error propagation distortion $d_{ep}(\hat{V}_t^{u,v})$ and the channel-induced distortion of the coded texture video and depth map of the left view, which can be represented as

$$\begin{aligned} d_{ep}(\hat{V}_t^{u,v}) &= E\left\{(\hat{T}_{L,t}^{x',y'} - \tilde{T}_{L,t}^{x',y'})^2\right\} \\ &\quad +\psi_r \alpha^2 E\left\{(\hat{D}_{L,t}^{x',y'} - \tilde{D}_{L,t}^{x',y'})^2\right\} \\ &= d_{ep}(\hat{T}_{L,t}^{x',y'}) + \psi_r \alpha^2 d_{ep}(\hat{D}_{L,t}^{x',y'}) \end{aligned} \tag{6}$$

where $d_{ep}(\hat{T}_{L,t}^{x',y'})$ and $d_{ep}(\hat{D}_{L,t}^{x',y'})$ represent the transmission distortions of the pixel (x',y') in the compressed texture video and depth map of the left view, respectively. Recall that the texture and depth of the left view are encoded by employing traditional motion-compensated prediction, and both $d_{ep}(\hat{T}_{L,t}^{x',y'})$ and $d_{ep}(\hat{D}_{L,t}^{x',y'})$ can be directly and recursively computed using the modeling algorithms introduced in [13] and [15].

Note that, if the current pixel of the right view is inter-coded, i.e., temporally predicted by the previous pixel within the same view, the channel distortion of this pixel can also be estimated using (1) except that the error propagation distortion comes from the preceding intra-view frame. For a intra-coded pixel, no transmission errors will be propagated from the synthesized reference frame, and the transmission distortion for intra-coded pixels is due only to packet drops, i.e., the sum of $d_{ec}(\hat{T}^{x,y}_{R,t})$ and $d_{ep}(\hat{T}^{\rho(x,y)}_{R,t-1})$.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, the performance of the proposed scheme is evaluated. The Joint Multi-view Video Coding (JMVC) version 8.0 [23] of the H.264/MVC reference software is appropriately modified to encode both the texture videos and the depth maps, and View Synthesis Reference Software (VSRS) 3.5 [24] is used to render the synthetic reference view at the encoder and the virtual intermediate views at the decoder. The standard multi-view video plus depth sequences “BookArrival”, “Newspaper”, and “Undo_Dancer” are chosen for our simulations. Note that these test sequences represent a wide range of video motion complexity and depth fidelity. Among these sequences, for “BookArrival”, views 8 and 10 are used as the left and right views, respectively. For “Newspaper”, views 4 and 6 are served as the left and right views, respectively. For “Undo_Dancer”, views 2 and 5 are employed as the left and right views. The first two sequences have a resolution of 1024 × 768 samples, while the remaining one has a resolution of 1920 × 1088 samples. For both texture video and depth map coding, context-adaptive binary arithmetic coding (CABAC) is used as the entropy coding scheme, and the functions of the variable prediction size and the loop filter are enabled. The search range for disparity and motion estimation is set to 64. The quantization parameter is fixed to 32 for texture video and depth. For each multi-view video sequence, each view is encoded with a group of pictures size of 32 frames, where the first frame in the left view is coded as an I-frame followed by all P-frames.

Each coded frame is partitioned into slices, where each depth slice contains four horizontal rows of MBs, and each texture slice contains a horizontal row of MBs due to higher associated bit rates. Each coded slice is then carried in a separate packet. It should be noted that the packet length of all the frames in our simulations is within the limit of the maximum transmission unit (MTU) for the Ethernet. The random packet loss pattern is employed to simulate packet losses [25]. Different packet loss rates of 5%, 10% and 20% are tested on both the compressed texture video and depth stream. To simulate the channel, at each packet loss rate, 100 packet loss patterns are randomly generated. The transmission distortion for each MB or frame is determined by averaging the distortions resulting from all the loss realizations. In our experiments, the error concealment method where each damaged block either in the texture or depth map is directly replaced by its co-located counterpart in the previous frame is employed at the multi-view video decoder.

Table 1. Correlation coefficients between the estimated and measured transmission distortions for each video sequence.

Sequence	Level	Correlation coefficients		
		5%	10%	20%
BookArrival	MB	0.89	0.87	0.86
	Frame	0.92	0.93	0.92
Newspaper	MB	0.85	0.86	0.84
	Frame	0.90	0.91	0.89
Undo_Dancer	MB	0.90	0.88	0.85
	Frame	0.93	0.92	0.91

In order to validate the estimation accuracy of the proposed transmission distortion model, the correlation coefficient between the actual distortion and the estimated one has been measured at the MB and frame levels. Since the proposed method focuses on modeling the reconstructed distortion of the right view induced by VSP, only the comparative results for the texture of the right view are given. Table 1 shows the values of the correlation coefficient computed for all the test sequences at various packet loss rates. As can be observed, at the MB level, the correlation coefficients are greater than 0.84, which clearly proves that the expected transmission distortion can be precisely estimated by the proposed model. That is, the effect of inter-view error propagation caused by VSP can be accurately accounted for. The same observation holds at the frame level, where the proposed distortion model yields a correlation coefficient greater than 0.89.

As a visual comparison, we also report that the frame-by-frame distortion comparison between the estimated and measured transmission distortions. Fig. 2 shows the distortion versus frame number for the BookArrival sequence when the packet loss rate is equal to 10%. It can be observed that the estimated transmission distortion still matches quite well with the actual frame-level distortion.

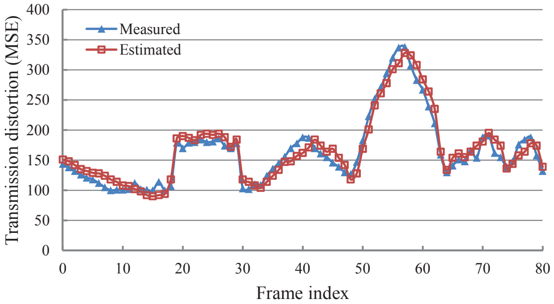

Fig. 2. Frame-by-frame evolution track of the transmission distortion estimation for the “BookArrival” sequence.

5. CONCLUSIONS

In this paper, we have proposed a new transmission distortion model for compressed MVD-based 3-D video streaming. Based on the study on the characteristics of the propagating behaviour of transmission errors due to packet losses, a recursive distortion model is derived to capture the effect of transmission errors of the synthetic reference view on the dependent view. Unlike other existing distortion models, the proposed mathematical model explicitly takes into account view synthesis based inter-view error propagation. Extensive evaluation results demonstrate the provided estimate has a good accuracy at both the MB and frame levels.

6. REFERENCES

[1] K. Muller, P. Merkle, and T. Wiegand, “3-D video representation using depth maps,” *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643-656, Apr. 2011.

[2] C. Fehn, “Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3-D-TV,” in *Proc. 11th SPIE Stereoscopic Displays Virtual Reality Syst.*, Jan. 2004, pp. 93-104.

[3] "Call for Proposal on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11, MPEG, Doc. N12036, Geneva, Switzerland, March, 2011.

[4] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1461-1473, Nov. 2007.

[5] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, P.H.N. de With, and T. Wiegand, "The effect of multiview depth video compression on multiview rendering," *Signal Process.: Image Commun.*, vol. 24, no. 1-2, pp. 73-88, Jan. 2009.

[6] M. Maitre and M. N. Do, "Joint encoding of the depth image based representation using shape-adaptive wavelets," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 1768-1771.

[7] B. T. Oh, J. Lee, and D.-S. Park, "Depth map coding based on synthesized view distortion function," *IEEE Journal of Selected Topics in Signal Process.*, vol. 5, no. 7, pp. 1344-1352, Nov. 2011.

[8] E. Martinian, A. Behrens, J. Xin, and A. Vetro, "View synthesis for multiview video compression," in *Proc. Picture Coding Symposium (PCS)*, Beijing, China, Apr. 2006.

[9] S. Yea and A. Vetro, "View synthesis prediction for multiview video coding," *Signal Process: Image Commun.*, vol. 24, no. 1-2, pp. 89-100, Jan. 2009.

[10] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multiview video coding using 3-D warping with depth map," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1485-1495, Nov. 2007.

[11] D. Tian, F. Zhou, and A. Vetro, "CE1.h: Backward View Synthesis Prediction using Neighboring Blocks," *ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JCT3V-C0152,* Geneva, CH, Jan. 2013.

[12] M. M. Hannuksela, D. Rusanovskyy, W. Su, L. Chen, R. Li, P. Aflaki, D. Lan, M. Joachimiak, H. Li, and M. Gabbouj, "Multiview-video-plus-depth coding based on the advanced video coding standard," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3449-3458, Sep. 2013.

[13] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 966-976, Jun. 2000.

[14] G. Cote, S. Shirani, and F. Kossentini, "Optimal mode selection and synchronization for robust video communications over error-prone networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 6, pp. 952-965, Jun. 2000.

[15] Z. He, J. Cai, and C. W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 511-523, Jun. 2002.

[16] Y. Zhou, C. Hou, W. Xiang, and F. Wu, "Channel distortion modeling for multi-view video transmissmino over packet-switched networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 11, pp. 1697-1692, Nov. 2011.

[17] B. Macchiavello, C. Dorea, E. M. Hung, G. Cheung, and W. Tan, "Reference frame selection for loss-resilient depth map coding in multiview video conferencing," in *Proc. SPIE Visual Inf. Process. Commun.*, 2012.

[18] B. Macchiavello, C. Dorea, E. M. Hung, G. Cheung, and W. Tan, "Loss-resilient coding of texture and depth for free-viewpoint video conferencing," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 711-725, Apr. 2014.

[19] P. Gao and W. Xiang, "Rate-distortion optimized mode switching for error-resilient multi-view video plus depth based 3-D video coding," *IEEE Trans. Multimedia*, vol. 16, no. 7, pp. 1797-1808, Nov. 2014.

[20] Y. Wang, Z. Wu, and J. M. Boyce, "Modeling of transmission-loss-induced distortion in decoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 716-732, Jun. 2006.

[21] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao, "Joint video/dpeth rate allocation for 3-D video coding based on view synthesis ditortion model," *Signal Process.: Image Commun.*, vol. 24, no. 8, pp. 666-681, Jun. 2009.

[22] H. Yuan, Y. Chang, J. Huo, F. Yang, and Z. Lu, "Model-based joint bit allocation between texture videos and depth maps for 3-D video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 485-497, Apr. 2011.

[23] ISO/IEC JTC1/SC29/WG11, WD 3 Reference Software for MVC, Doc. JVT-AC207, Busan, Korea, 2008.

[24] ISO/IEC JTC1/SC29/WG11, 3DV/FTV EE2: Report on VSRS Extrapolation, Doc. M18356, Guangzhou, China, 2010.

[25] S. Wenger, "Proposed Error Patterns for Internet Experiments," ITU-T VCEG document Q 15-I-16r1 Oct. 1999.