Temporal Synchronization Scheme in Live 3D Video Streaming over IEEE 802.11 Wireless Networks

Yohaan Yoon, Myungchul Kim, Ben Lee† and Kyungmin Go
Department of Computer Science, KAIST, Daejeon, Republic of Korea, Email: {straightfor, mck, kyungmingo}@kaist.ac.kr
†School of Electrical Engineering and Computer Science, Oregon State University, Oregon, USA, Email: {benl}@eecs.orst.edu

***Abstract*—Although 3D video has become popular, streaming over wireless network faces a number of challenges. Due to frequent frame losses in wireless networks, temporal asynchrony occurs and results in serious visual fatigue for viewers. In order to provide better quality of 3D video, this paper proposes a new scheme called the Temporal Synchronization Scheme (TSS) for live 3D video streaming over wireless networks. TSS delivers video frames for the left and right views in the same frame order with the same transmission priority and compensates for frame damage and loss during the decoding phase. In addition, a new metric called the Stereoscopic Temporal Variation Index (STVI) is proposed to measure the degree of temporal asynchrony in 3D video. Subjective assessments demonstrate that STVI is an objective metric for measuring subjective quality. Moreover, our study shows that the proposed scheme results in better 3D video quality than the conventional method in terms of STVI and MOS.**

Keywords—3D video steaming; temporal synchronization in 3D video; 3D video quality assessment; IEEE 802.11 wireless networks

I. INTRODUCTION

After Avatar's phenomenal success in 2010, three-dimensional (3D) video has become one of the most popular multimedia content formats. This popularity has also led to wide acceptance of 3D TVs and smart-phones by users. As 3D-enabled devices become more readily available, 3D video streaming over wireless networks will also become an important technology. Although 3D video streaming over wireless networks is currently available, its visual quality is not guaranteed due to error-prone medium, network congestion, and interference caused by carrier sense and hidden nodes [1]. Therefore, multimedia streaming over wireless network, especially the live 3D video streaming, is still a challenging issue.

There have been some research efforts to provide Quality of Service (QoS) for 3D video streaming over wireless networks. In [2-4], the authors proposed various schemes to improve the QoS for 3D video streaming. However, these prior studies do not consider video frame damage and loss that occur during wireless transmission. In live 3D video streaming, frame damage and loss cause the two separate views for left and right eyes that create depth perception to be out of synchronization [5].

The limitations of these prior studies are that they only considered the quality of each view independently and did not consider video frame losses due to packet losses during transmission. In particular, the frame loss cause temporal asynchrony, which leads to misalignment between the left and right views, and serious visual fatigue for viewers [6]. Thus, an important requirement of live 3D video streaming is the accurate temporal synchronization between the left and right views.

Therefore, this paper proposes a new scheme for live 3D video streaming over wireless networks called the *Temporal Synchronization Scheme* (TSS), which consists of a QoS Mapping Module and a Compensation Module. The QoS Mapping Module delivers the frames for the left and right views in the same frame order with the same transmission priority, while the Compensation Module restores damaged and lost frames to resynchronize between the left and right views. Differentiated delivery in the QoS Mapping Module is implemented using IEEE 802.11e Enhanced Distributed Channel Access (EDCA) Access Categories (ACs) [7], and restoration of damaged and lost frames in the Compensation Module is performed by copying, deleting, and replicating received frames for the damaged and lost frames with respect to temporal synchronization.

The effectiveness of the proposed TSS is studied using a new metric called *Stereoscopic Temporal Variation Index* (STVI) that measures the degree of temporal asynchrony. Our experimental study shows that TSS results in better temporal quality than existing methods in terms of Peak Signal-to-Noise Ratio (PSNR) [8], STVI, and subjective Mean Opinion Score (MOS) [9].

The important and unique contributions of the paper are the following:

- The proposed TSS is the first scheme to address and solve the temporal asynchrony issue in live 3D video streaming over wireless networks.
- TSS only requires slight modification of H.264/AVC decoder.
- A new metric STVI is proposed and applied to measure the degree of temporal asynchrony in 3D video.

The rest of this paper is organized as follows. Section II provides an overview of 3D video. Section III discusses the temporal asynchrony problem during transmission and the related studies. Section IV presents the proposed TSS scheme. Experimental results and analysis are discussed in Section V. Finally, Section VI concludes the paper and discusses future work.

II. BACKGROUND

This section presents the background on the 3D video processing chain and 3D video Quality of Experience (QoE).

A. 3D Video Processing Chain

The 3D video processing chain involves the following three steps: production, transport and display. Table I shows the three types of formats that are defined in the 3D video processing chain [10, 11].

The Production Format used in 3D video acquisition consists of two types of format called *Video Only Format* and *Depth Enhance Format*. The Video Only Format, which synthesizes two (left and right) or more views together, is commonly used in theaters, TVs, etc. On the other hand, the Depth Enhanced Format can be rendered by the video and its depth information. However, it has not been standardized or commercialized because of the algorithm's complexity.

Examples of the Transport Format used in the coding and transmission of Video Only Format are H.264/AVC standards, such as H.264/AVC Simulcast, H.264/AVC Supplemental Enhancement Information (SEI) message, and H.264/AVC Stereo High Profile. Among them, Stereo High Profile is the most popular encoding method. The Stereo High Profile is more efficient than the other methods because it uses new techniques, such as interview prediction, to encode the two videos for left and right eyes into one encoded video.

Fig. 1 shows the various types of frame order for the Stereo High Profile [12]. As it seen in the figure, a 3D video consists of a sequence of frames for the left and right views that are encoded. The *Access Unit Order* indicates the sequence of encoded frames for each view. The *Transmission Order* (also called the Decoding Order) is the order in which frames are transmitted or decoded from the encoded video. For example, frames in a Group Of Picture (GOP) are transmitted alternatively from left to right views according to the Transmission Order shown in Fig. 1. Finally, the Display Order is the order in which frames are displayed which is based on *Type 0* in H.264/AVC [12]. These orders are determined during the encoding phase and labeled into frame headers.

In the last step of the 3D processing chain, various Display Formats are available and all the formats require viewers to wear special glasses and/or a customized display in order to view 3D video.

B. QoE of 3D Video

The three elements of QoE in 3D video are classified as visual quality, depth quality/perception, and visual comfort/discomfort [13, 14].

Visual quality depends on the quality of 2D scenes. In order to produce 3D video, at least two views are needed using multi-2D cameras. Therefore, the visual quality of 2D scenes before synthesis of the stereo views should be included in 3D video quality evaluation. The 2D video quality can be evaluated using MOS, but it is expensive and time consuming. For these reasons, PSNR has been used for 2D video quality evaluation.

The two views in a 3D video give a viewer depth perception since the left and right views are projected onto a viewer's left and right eyes, respectively. The evaluation of *depth quality/perception* is related with the naturalness [15] of 3D perception, i.e., how a viewer perceives depth in real world.

Finally, *visual discomfort* is caused by the difficulty of fusing the left and right views due to excessive binocular parallax [14]. This causes some viewers to experience visual fatigue with symptoms such as eye strain, headache, and nausea. The level of visual comfort/discomfort is often measured using questionnaires. In particular, the visual comfort/discomfort implies difficulty in watching 3D video. Thus, an important element that affects the visual comfort/discomfort is the accuracy of temporal synchronization between left and right views [14, 16].

TABLE I
FORMATS ON 3D VIDEO PROCESSING CHAIN

	Production Format	**Transport Format**	**Display Format**
Objective	Acquisition	Coding and transmission	Output on a display device
Method	Capturing and post-processing	Encoding and decoding	Rendering
Examples	Video Only Format and Depth Enhanced Format	Simulcast, SEI, and Stereo High Profile in H.264/AVC	Anaglyph, Shutter glasses, Polarized glasses, and Auto-stereoscopic

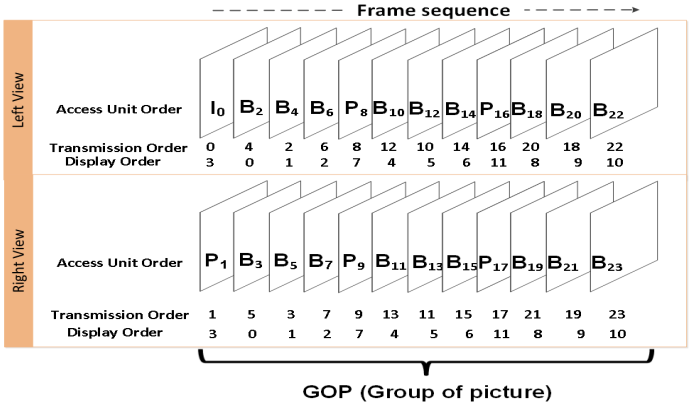


Fig. 1. Access Unit Order, Transmission Order, and Display Order in GOP

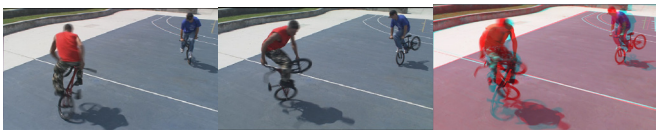
III. RELATED WORK

This section first surveys existing metrics for evaluating the temporal quality of video. Thereafter, previous work on guaranteeing the quality of 3D video streaming over networks is discussed.

A. Evaluation Metric

PSNR is a well-known objective metric for measuring the quality of 2D video and images [8]. However, it only evaluates the spatial quality of video and images. Since PSNR is a full reference method requiring both the original and received frames for calculation, it can measure damaged frames but not lost frames in 2D video.

Chan *et al.* proposed the Temporal Variation Metric (TVM) and Temporal Variation Index (TVI) to measure the temporal information of a 2D video [17]. TVM measures the temporal information of consecutive frames in a video. A large TVM value infers fast scene-to-scene transition or frame loss. However, this information alone is not enough to measure temporal quality. Therefore, temporal quality is measured by comparing the temporal information of the received video with that of the original video. The change in temporal information between the original and received video is caused by a degradation of temporal quality.



(a) Left view (b) Right view (c) Synthesized view

Fig. 2. An example of 3D video on congested networks

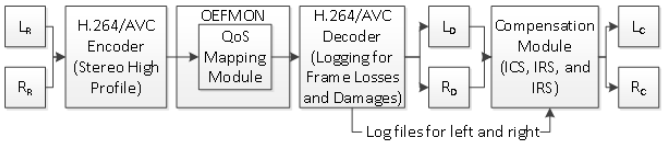


Fig. 3. The overall structure of TSS

TVI measures the temporal quality difference between TVMs of the original and the received video, and is given by

$$TVI(t) = \frac{\left|TVM_s(t) - TVM_r(t)\right|}{TVM_s(t)}, \quad (1)$$

where $TVM_s(t)$ and $TVM_r(t)$ are the TVM values of the original and the received frames, respectively, at display order t.

Compared with PSNR, TVI is a reduced reference method, which means that temporal quality is measured using only TVM values; thus, the original video is not required. However, both TVM and TVI are difficult to apply to a 3D video because of the existence of two different views and video frame losses occurring at different frame orders for the left and right views. Since PSNR and TVI are not applicable to the evaluation of the temporal quality of 3D video, a new metric is needed.

B. Video Streaming Methods

There are several cross-layer mapping approaches to assign high priority to important video frames or slices to improve the quality of 2D video over IEEE 802.11e. In addition to the Distributed Coordination Function (DCF), which is the basic access mechanism using CSMA/CA in IEEE 802.11, Extended EDCA provides QoS differentiation with four ACs according to traffic types [7]. Choudhry *et al.* proposed a method where I-frame is mapped to AC(3), which has the highest priority, and P- and B-frames are mapped to AC(2) and AC(1), respectively [18]. Ksentini *et al.* presented an approach that maps between the EDCA ACs and data partition types according to their priorities [19]. However, these two cross-layer mapping approaches apply to 2D video only. Even though Hewage *et al.* introduced a cross-layer mapping approach for 3D Video [20], the approach only apply to Depth Enhanced Format which do not use for commercial 3D-enabled devices.

A number of prior studies exist to guarantee the QoS of 3D video over networks [2-4]. However, these prior studies do not consider video frame loss during transmissions of multimedia over network, which causes 3D video to become unsynchronized.

Fig. 2 shows an example of a 3D video frame consisting of left (189th frame) and right (197th frame) views, which will be synthesized to generate the 132nd frame. The 189th and 197th frames are located at the position of 132nd frame due to frame losses in congested networks. The temporal asynchrony due to frame loss may cause visual fatigue [6]. Therefore, a method to minimize the temporal asynchrony and a new metric for measuring the degree temporal asynchrony are required.

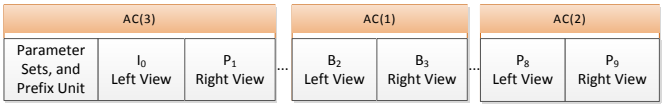


Fig. 4. Transmission Order of NALUs

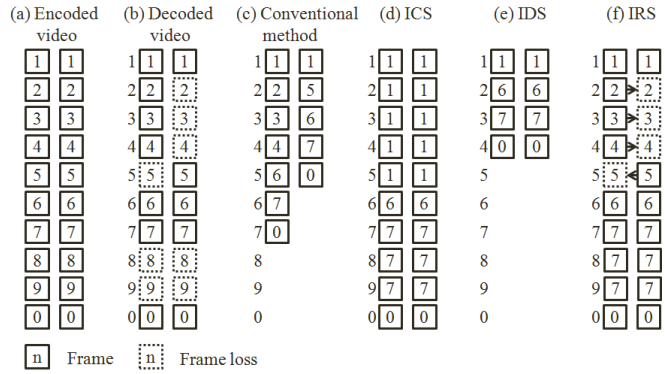


Fig. 5. Examples of a conventional method and CM given encoded and decoded videos

Moreover, QoE for 3D video is considered mostly during acquisition, but rarely during transmission. Therefore, the proposed TSS addresses an important requirement of live 3D video streaming over wireless networks, which is to provide accurate temporal synchronization between the left and right views in terms of QoE.

IV. Proposed Scheme for 3D Video Syncrhonization

The proposed TSS consists of two modules: 1) the *QoS Mapping Module* (QMM), operating at the transmission side and maintaining transmission priority at routers, being responsible for delivering the frames of both views in the same frame order and with the same transmission priority; 2) the *Compensation Module* (CM), operating at the receiver side, being responsible for restoring the frame synchronization when frame damage or loss occur.

Fig. 3 shows the overall structure of the proposed scheme. The left and right raw video files (L_R and R_R) are encoded into one video file using Stereo High Profile of the H.264/AVC encoder. The encoded video file has I-, P-, and B-frames, and each frame is encapsulated in Network Abstraction Layer Units (NALUs). The NALUs are transmitted using Real-time Transport Protocol (RTP) over IEEE 802.11e by QMM running on Open Evaluation Framework Multimedia Over Networks (OEFMON) [21]. OEFMON integrates the DirectShow as a multimedia module and the QualNet network simulator to evaluate the quality of multimedia transmissions over networks. A modified version of the H.264/AVC decoder which outputs the left and right decoded views (L_D and R_D) is used to log damaged or lost frames. Finally, the CM produces compensated views (L_C and R_C) of the 3D video sequence using several techniques. The discussion of these techniques will be provided in Section IV. B.

A. QoS Mapping Module

Based on the H.264/AVC Stereo High Profile as mentioned in Section II, the Display Order is the same for a pair of frames of synchronized left and right views in 3D video. If they are different, temporal asynchrony occurs. In order to reduce the temporal asynchrony, QMM assigns a pair of frames of

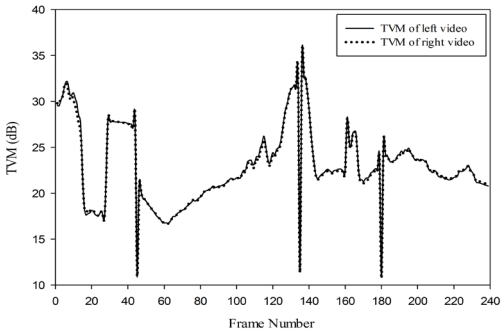


Fig. 6. TVM of the left and right views for BMX

synchronized left and right views in the same Display Order to the same ACs in IEEE 802.11e.

Fig. 4 shows the Transmission Order of NALUs in detail using Stereo High Profile of H.264/AVC. Parameter Sets (PSs), Prefix Unit, I-frame for the left view, P-frame for the right view, B-frame for the left view, B-frame for the right view, P-frame for the left view, and P-frame for the right view are transmitted in sequential order. The mapping between ACs and NALUs is shown in Fig. 4. The QMM maps PSs, Prefix Units, I-frame for the left view and P-frame for the right view in the same Display Order to AC(3). P-frames for the left and right views in the same Display Order are mapped to AC(2). All other pairs of frames are mapped to AC(1).

B. Compensation Module

Since QMM does not always guarantee temporal synchronization in 3D video due to frame damage and loss, CM recovers the synchronized Display Order for damaged and lost frames. As shown in Fig. 3, CM consists of two processes: logging the Display Order of damaged and lost frames and compensating damaged and lost frames to reduce the effect of temporal asynchrony.

Fig. 5 shows an example frame sequence for the conventional method and CM for given encoded and decoded videos. The numbers inside the boxes represent the Display Order of encoded video frames. Fig. 5(a) shows the encoded and transmitted video. Meanwhile, Fig. 5(b) shows decoded video, where some frames are lost during transmission. These frame losses would cause temporal asynchrony in the conventional method as shown in Fig. 5(c).

There are three compensation techniques in CM to reduce temporal synchrony: *Image Copy* (IC), *Image Delete* (ID), and *Image Replication* (IR). The IC replaces lost frames with the latest synchronized frames before temporal asynchrony occurs. The replacement occurs whether one of the frames or both frames from the left and right views at the same Display Order are lost. Fig. 5(d) shows the effect of applying IC, where the second, third, fourth, and fifth frames are synchronized with the first frame. The ID deletes frames that are in temporal asynchrony. As shown in Fig. 5(e), the second, third, and fourth frames on the left and the fifth frames on the right are deleted, even though these frames are successfully decoded. Finally, the IR replaces the lost frame with a successfully decoded frame in the *same* Display Order. As shown in Fig. 5(f), the second, third, and fourth frame on the left and the fifth frame on the right are copied from the second, third, and fourth frame on the right and the fifth frame on the left, respectively.

C. The Proposed Metric for Determining Temporal Asynchrony on 3D Video

Since various metrics such as PSNR [8], TVM, and TVI [18], can only measure the quality of 2D video, a new metric for measuring the temporal asynchrony in 3D video is needed. Thus, the *Stereoscopic Temporal Variation Index* (STVI) is proposed in this paper to measure the temporal asynchrony in 3D video. The *Binocular-disparity Variation Metric* (BVM) and *Stereoscopic Temporal Variation Metric* (STVM) are also proposed to calculate STVI since STVI is a reduced reference method.

First, BVM measures the binocular-disparity information of the left and right views to consider the depth quality in 3D video. In other words, BVM only uses the left and right frames in the same Display Order for measuring binocular-disparity variation and is defined as

$$BVM_p = 10\log_{10}\frac{k^2}{dLR_p}, \tag{2}$$

where k represents the maximum color depth and dLR_p is the mean squared error value of the corresponding pixels in the stereo frames given by

$$dLR_p = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(FLeft_p(i,j) - FRight_p(i,j))^2,$$

where MN is the resolution of the video, and $FLeft_p(i, j)$ and $FRight_p(i, j)$ represent the (i, j)-pixel of the p^{th} frame on left and right views, respectively. Our study uses 8-bit YUV420 format; thus, the value of k is 255.

Second, STVM represents the temporal and binocular-disparity information of 3D video. The STVM is calculated with three values, namely, the TVMs of the left and right views and BVM. While TVM represents the temporal information of a 3D video, BVM represents the binocular-disparity information of the 3D video. In [18], TVM is defined as

$$TVM_p = 10\log_{10}\frac{k^2}{d_p}, \tag{3}$$

where k represents the maximum color depth and d is the mean squared error value of the corresponding pixels in the consecutive frames given by

$$d_p = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(F_{p-1}(i,j) - F_p(i,j))^2,$$

where MN is the resolution of the video and $F_p(i, j)$ represents the (i, j)-pixel of the p^{th} frame.

Note that dLR and d are variance since the formula of mean square error is derived from the formula of variance. Based on the basic properties of the variance [22], the variance of sum of two random variables is given by

$$Var(X+Y) = Var(X) + Var(Y), \text{ if } X \text{ and } Y \text{ are independent} \tag{4}$$

Before calculating STVM, Fig. 6 shows the TVM values of the left and right views for videos such as BMX. As can be seen in Fig. 6, the TVM values of the left view is equal approximately to those of the right view. Since we assume that these two TVM values are identical, the mean value for TVMs of both views is calculated as the representative variable for TVM in Eq. (5). Thus, these results lead to STVM of TVM and BVM as follows:

$$STVM_p = 10\log_{10}\frac{k^2}{Var(d_p + dLR_p)}$$
$$= 10\log_{10}\frac{k^2}{Var(d_p) + Var(dLR_p)}$$
$$= 10\log_{10}\frac{k^2}{\frac{k^2}{10^{\frac{TVM_p}{10}}} + \frac{k^2}{10^{\frac{BVM_p}{10}}}} = 10\log_{10}\frac{10^{\frac{TVM_p + BVM_p}{10}}}{10^{\frac{TVM_p}{10}} + 10^{\frac{BVM_p}{10}}} \quad (5)$$
$$= TVM_p + BVM_p - 10\log_{10}(10^{\frac{TVM_p}{10}} + 10^{\frac{BVM_p}{10}}),$$

where

$$TVM_p = 10\log_{10}\frac{k^2}{Var(d_p)},\ Var(d_p) = \frac{k^2}{10^{\frac{TVM_p}{10}}},$$

$$BVM_p = 10\log_{10}\frac{k^2}{Var(dLR_p)},\ Var(dLR_p) = \frac{k^2}{10^{\frac{BVM_p}{10}}}$$

Finally, STVI measures the temporal asynchrony using the STVM of the encoded and received videos. STVI of 3D video at Display Order p, $STVI_p$, is defined as

$$STVI_p = \frac{\left|STVM_p^{\ S} - STVM_p^{\ R}\right|}{STVM_p^{\ S}}, \quad (6)$$

where $STVM_p^S$ and $STVM_p^R$ are the STVM values for the encoded and received frames at Display Order p.

After receiving the encoded video and its $STVM_p^S$, the video client calculates the STVI value using the $STVM_p^R$ of the received video and its log files. Note that STVI values are calculated using the left view as the base in our study. In the conventional method, if either the left or right frame in the same Display Order does not exist, the value of STVM is assigned to zero. For IR, the replicated frames are excluded from the calculation of STVI because there is no binocular-disparity information due to 2D video replication.

STVI is a reduced-reference method since it only needs the STVM values of encoded video. Therefore, STVI can be used to measure the temporal quality of 3D videos in real-time, and thus save more bandwidth and computational resources than the metric such as PSNR discussed in [6].

V. Experimental Evaluation and Analysis

This section evaluates and analyzes our proposed TSS scheme.

A. Experimental Environment

The performance of our proposed scheme is evaluated in terms of PSNR, STVI, and subjective MOS rating using the OEFMON simulator (see Section IV).

The network topology and traffic are illustrated in Fig. 7, which comprises wired and wireless networks with infrastructure mode. The topology has three servers, a router, an Access Point (AP), and three IEEE 802.11g nodes with the distance of 10 m between the AP and nodes. In detail, the wired and wireless networks provide 100 Mbps and 54 Mbps, respectively. Table II gives the details of the network parameters used in our simulation. Node B transmits the video stream to node E. Meanwhile, nodes A and C transmit background data as Constant Bit Rate (CBR) of 1-20 Mbps to nodes D and F, respectively, in order to control Packet Loss Ratio (PLR).

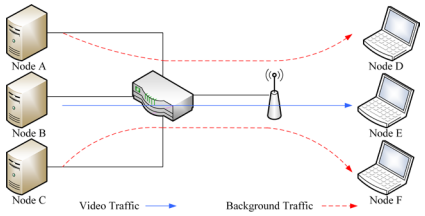


Fig. 7. Network topology and traffic

TABLE II
SIMULATION PARAMETERS

Parameters	Values
Radio type	802.11g (Infrastructure mode)
Data link rate	100 Mbps (wired networks), 54 Mbps (wireless networks)
Resolution	1024x768, 1280 × 720
Frame per second	30 and 24 fps
Video length	5 ~ 10 s
Jitter buffer	1,000 ms
Video frame type	I-, P-, and B-frames
I-frame interval	24 Frame
Video flow	One way, 1 flow

The video sources are 1024 × 768 @ 30 fps Newspaper [23] and 1280 × 720 @ 24 fps BMX [24]. The scene changes in BMX are fast, whereas the scene changes in Newspaper are slow. The sizes of the encoded Newspaper and BMX are 14.3 MB and 24.6 MB, respectively.

The PLR varies from 1% to 5% for high-resolution videos (BMX), and it varies from 10% to 20% for low-resolution videos (Newspaper). Low PLR causes more significant distortions in high-resolution videos than in low-resolution ones. Note that the PLR is calculated from all network traffic including video and background traffic, and all background traffic are mapped to AC(0) of IEEE 802.11e.

The Production and Transmission Formats of the 3D videos are Video Only Format and Stereo High Profile of H.264/AVC, respectively. The encoder used for the Stereo High Profile of H.264/AVC is Joint Model (JM) software 18.4 [25].

The JM software decoder does not decode and show the received video frames on the fly. Instead, the decoder generates a raw video file in YUV format and verifies the integrity of the decoded frames using their size information. If a frame is damaged or lost, the decoder skips the frame and jumps to the next frame. In addition, the decoder suddenly terminates when a bunch of data in a frame is lost. In order to handle this case, the JM decoder was modified to record the Display Order of the frame into log files associated with the left and right views. Therefore, the modified decoder generates not only the left and right decoded video files but also log files for the left and right views.

For subjective quality, we first streamed the two videos over the wireless network using the OEFMON simulator with

TABLE III
THE MEAN PSNR VALUES FOR TWO VIDEOS

			Conventional method		IC		ID		IR	
			DCF	*QMM*	*DCF*	*QMM*	*DCF*	*QMM*	*DCF*	*QMM*
News paper	PLR of 10%	Left	51	52	38	52	53	51	38	52
		Right	50	51	37	51	51	52	36	51
	PLR of 20%	Left	52	52	27	52	52	51	28	52
		Right	51	51	29	51	50	51	27	52
BMX	PLR of 1%	Left	32	32	49	49	28	28	49	49
		Right	28	29	48	50	27	31	48	49
	PLR of 3%	Left	27	30	47	48	26	28	47	40
		Right	29	27	41	42	27	27	40	42
	PLR of 5%	Left	30	28	45	48	29	39	44	32
		Right	26	26	30	43	29	37	32	43

various conditions. We collected a total of 40 samples of these videos with different quality. There are 16 samples of Newspaper and 24 samples of BMX. We then made YUV format files of left and right videos and synthesized a 3D video from the YUV format files using the Stereoscopic Player [26]. We engaged 15 volunteers as subjects to watch the videos on a PC. In order to calculate subjective rate, the MOS is calculated as the mean of the numerical values that were assigned to the attributes of the Absolute Category Rating (ACR) scale [9].

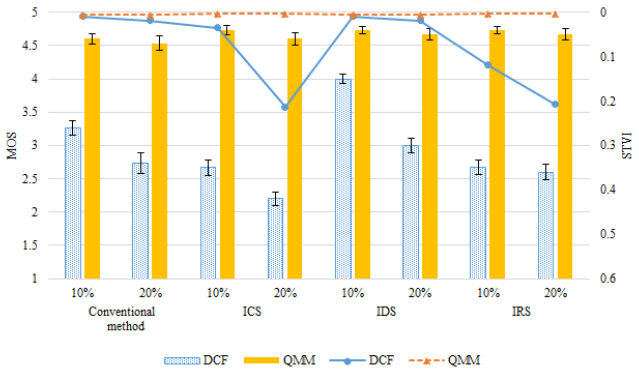
B. Experimental Result and Analysis

Our proposed TSS scheme was experimentally evaluated using the network simulation part of the OEFMON simulator and the decoding part of JM software. The experimental results of the proposed scheme will be explained for each part.

Table III shows the mean PSNR values of the left and right videos for the conventional method, IC, ID, and IR based on DCF and QMM with various PLRs. If a frame loss occurs, calculating PSNR for the frame is excluded. The mean PSNR values with IC and IR for low-resolution videos are lower than the mean PSNR values with the conventional method and ID. In contrast, the mean PSNR values with IC and IR for high-resolution videos are greater than the mean PSNR values with the conventional method and ID. The reason is that low-resolution videos have lots of empty frames which are not calculated with PSNR due to lost and deleted frames.

According to ITU-R BT.500-11 [27], a PSNR value greater than 31 dB represents a score of 3 in MOS, or Fair quality. Even though most of the videos in Table III have mean PSNR values greater than 31 dB, their subjective MOS scores are lower than 3. For instance, the mean PSNR values of all videos in Newspaper on the conventional method with DCF are greater than 31 dB, but the subjective MOS scores are lower than 3. Therefore, the mean PSNR values cannot represent subjective quality for a 3D video.

Fig. 8 shows the mean values of STVI and subjective MOS rating for the conventional method, IC, ID, and IR based on DCF and QMM. Figs. 8(a) and 8(b) show the graphs for Newspaper and BMX, respectively. Whilst each method in Fig. 8(a) has 10% and 20% of PLRs, each method in Fig. 8(b) has 1%, 3%, and 5% of PLRs. In Fig. 8, line and bar graphs represent the mean values of STVI (right y-axis) and MOS (left y-axis), respectively. Note that STVI value is bounded by 0 to 1; close to zero indicates not only that the temporal quality is good but also that the left and right are well synchronized.



(a) Newspaper

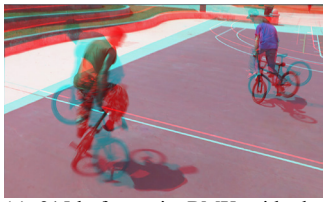
(b) BMX

Fig. 8. The line and bar graphs for mean values of STVI (right y-axis) and MOS (left y-axis) on videos, respectively

All methods based on DCF result in low quality because of a large number of damaged and lost frames. Compared to DCF, all methods based on QMM are not only close to zero in STVI, but also close to 5 in MOS as excellent because frame loss in QMM is significantly reduced. Although the videos on DCF are compensated by IC and IR, the videos have lower quality than the videos using the conventional method and ID in terms of STVI and MOS. The reason is that the videos have freeze-frame artifacts because of lots of copied and replicated frames. All videos with ID outperform other videos because ID deletes not only damaged frames but also asynchronous frames. In the case of QMM, frame damage and loss are significantly reduced, as shown in Fig. 8(a) even in PLRs are 10% and 20%.

As shown in Fig. 8, there are different results between low- and high-resolution videos. Especially, STVI and MOS on IC and IR have better quality than the conventional methods as seen in Fig. 8(b). Since the PLRs of high-resolution videos are lower than those of low-resolution videos, the number of lost frames is reduced. When the network condition is below 10% of PLR, IC and IR in CM are good to compensate 3D videos since the videos do not frequently appear freezing frames.

STVI has a similar pattern to MOS as shown in Fig. 8. The correlation coefficient value of STVI and MOS for Newspaper and BMX are 0.83 and 0.87, respectively. Since a correlation



(a) 215th frame in BMX with the conventional method for DCF

(b) 215th frame in BMX with IC for QMM

Fig. 9. Example scenes of 3D videos

coefficient greater than 0.8 is generally described as strong [28], there exist a strong correlation between STVI and MOS.

As shown in Fig. 9(a), temporal asynchrony occurs in the conventional method running on DCF for both video sources. The application of TSS with IC for QMM restores Fig. 9(a) to Fig. 9(b). As a result, TSS improves temporal asynchrony in live 3D video streaming over wireless networks compared to the conventional method using DCF.

As a consequence, the QMM outperforms the DCF for all the cases. The mean values of STVI and MOS for ID are significantly better than those for the other methods. Note that temporal asynchrony is not completely resolved even with the proposed TSS scheme. Even though IC minimizes temporal asynchrony in 3D video, it still has some freeze-frame artifacts in congested networks due to copied frames. Using ID, the temporal asynchrony and freeze-frame artifacts in 3D video are removed, but, sudden scene-to-scene transitions occur due to deleted frames. Lastly, IR reduces the number of freeze-frame artifacts and sudden scene-to-scene transitions compared to IC and ID, however, it also reduces the depth perception.

VI. Conclusion

This paper proposed the Temporal Synchronization Scheme (TSS) to reduce and compensate temporal asynchrony in live 3D video streaming over wireless networks. TSS consists of a QoS Mapping Module (QMM) to minimize frame damage and loss over IEEE 802.11 and a Compensation Module (CM) to compensate for temporal asynchrony due to frame damage and loss. In addition, the Stereoscopic Temporal Variation Index (STVI) was developed to measure the degree of temporal asynchrony in 3D video.

TSS is the first work to address the temporal asynchrony problem in live 3D video streaming over wireless networks. Our experimental results show that TSS significantly improves the visual quality of 3D videos even when frame damage and loss occur. In addition, STVI shows strong correlation to subjective MOS rating.

Our proposed TSS can be improved in a number of ways, better restoration of lost frames, recovery of damaged frames, and error resilience and concealment for 3D encoding and decoding. As a future work, we plan to develop a compensation scheme for the other factors which lead to visual fatigue for users. In addition, we will perform and evaluate our scheme using a test-bed.

VII. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant founded by the Korea government (MEST) (No. 2012R1A2A2A01008244).

References

[1] J. Li, C. Blake, D.S.J.D. Couto, H.I. Lee, and R. Morris, “Capacity of Ad Hoc Wireless Networks,” in Proc. ACM MOBICOM, 2001.

[2] N. Ozbek, B. Gorkemli, A.M. Tekalp, and T. Tunali, “Adaptive Streaming of Scalable Stereoscopic Video Over DCCP,” in Proc. IEEE ICIP, 2007.

[3] M.O. Bici, D. Bugdayci, G.B. Akar, and A. Gotchev, “Mobile 3D Video Broadcast,” in Proc. IEEE ICIP, 2010.

[4] C.G. Gurler, K.T. Bağci, and A.M. Tekalp, “Adaptive Stereoscopic 3D Video Streaming,” in Proc. IEEE ICIP, 2010.

[5] I.A. Howard and B. Rogers, “Binocular Vision and Stereopsis,” Oxford University Press, Oxford, 1995.

[6] L. Goldmann, J.S. Lee, and T. Ebrahimi, “Temporal Synchronization in Stereoscopic Video: Influence on Quality of Experience and Automatic Asynchrony Detection”, in Proc. IEEE ICIP, 2010.

[7] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE 802.11 WG, Jun. 2007.

[8] *PSNR*, http://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio.

[9] Subjective Video Quality Assessment Methods for Multimedia Applications, ITU-T P.910, 1999.

[10] P. Merkle, K. Muller, and T. Wiegand, “3D Video: Acquisition, Coding, and Display,” IEEE Transactions on Consumer Electronics, vol. 56, no. 2, pp. 946–950, Jan. 2010.

[11] G.M. Su, Y.C. Lai, A. Kwasinski, and H. Wang, “3D Video Communications: Challenges and Opportunities,” International Journal of Communication Systems, vol. 24, no. 10, pp. 1261–1281, Nov. 2011.

[12] I.E. Richardson, “The H. 264 Advanced Video Compression Standard,” Wiley, 2011.

[13] Q. Huynh-Thu, P.L. Callet, and M. Barkowsky, “Video Quality Assessment: from 2D to 3D Challenges and Future Trends”, in Proc. IEEE ICIP, 2010.

[14] Y. Nojiri, H. Yamanoue, A. Hanazato, M. Emoto, and F. Okano, “Visual Comfort/Discomfort and Visual Fatigue Caused by Stereoscopic HDTV Viewing,” in Proc. SPIE Electronic Imaging, 2004.

[15] R.G. Kaptein, A. Kuijsters, M.T.M. Lambooij, W.A. IJsselsteijn, and I. Heynderickx, “Performance Evaluation of 3D-TV Systems,” in Proc. SPIE Image Quality and System Performance V, 2008.

[16] ″The Guidebook for Stereoscopic 3D Contents Production”, 3DTV Broadcast Promotion Center, Korea Communications Commision, 2012.

[17] A. Chan, A. Pande, E. Baik and P. Mohapatra, “Temporal Quality Assessment for Mobile Video,” in Proc. ACM Mobicom, 2012.

[18] U. Choudhry and J.W. Kim, “Performance Evaluation of H. 264 Mapping Strategies over IEEE 802.11e WLAN for Robust Video Streaming,” Advances in Multimedia Information Processing-PCM, 2005.

[19] A. Ksentini, M. Naimi, and A. Gueroui, “Toward an Improvement of H.264 Video Transmission over IEEE 802.11e through a Cross-layer Architecture,” IEEE Communications Magazine, vol. 44, no. 1, pp. 107-114, Jan. 2006.

[20] C. Hewage, S. Nasir, S. Worrall, and M. Martini, ″Prioritized 3D video distribution over IEEE 802.11 e,” IEEE Future Network and Mobile Summit, 2010.

[21] C. Lee, M. Kim, B. Lee, S.J. Hyun, K. Lee, and S. Lee, "OEFMON: An Open Evaluation Framework for Multimedia Over Networks," IEEE Communications Magazine, vol. 49, no. 9, pp. 153-161, Sep. 2011.

[22] *Variance*, http://en.wikipedia.org/wiki/Variance.

[23] *BMX*, http://www.merl.com/pub/avetro/mvc-testseq/stereo-interlaced/.

[24] *3D video database*, http://sp.cs.tut.fi/mobile3dtv/stereo-video/.

[25] *JM Software*, http://iphome.hhi.de/suehring/tml/.

[26] *Stereoscopic Player*, http://www.3dtv.at/.

[27] Methodology for the subjective assessment of the quality of television pictures, ITU-R BT.500-1, 2002.

[28] *Correlation and dependence*, http://en.wikipedia.org/wiki/Correlation_and_dependence..