Visual-Attention-Aware Progressive RoI Trick Mode Streaming in Interactive Panoramic Video Service

Joo Myoung Seok and Yonghun Lee

In the near future, traditional narrow and fixed viewpoint video services will be replaced by high-quality panorama video services. This paper proposes a visual-attention-aware progressive region of interest (RoI) trick mode streaming service (VA-PRTS) that prioritizes video data to transmit according to the visual attention and transmits prioritized video data progressively. VA-PRTS enables the receiver to speed up the time to display without degrading the perceptual quality. For the proposed VA-PRTS, this paper defines a cutoff visual attention metric algorithm to determine the quality of the encoded video slice based on the capability of visual attention and the progressive streaming method based on the priority of RoI video data. Compared to conventional methods, VA-PRTS increases the bitrate saving by over 57% and decreases the interactive delay by over 66%, while maintaining a level of perceptual video quality. The experiment results show that the proposed VA-PRTS improves the quality of the viewer experience for interactive panoramic video streaming services. The development results show that the VA-PRTS has highly practical real-field feasibility.

Keywords: Human visual system, Scalable Video Coding, panoramic video, RoI, trick mode, visual attention, CSF.

I. Introduction

Video services are becoming smarter and more realistic and are evolving from a traditional narrow and fixed viewing environment to immersive multimedia services. In particular, immersive video technologies are already being commercialized around 3D and UHD technologies and are spreading widely, even to general users. Furthermore, interest in a new immersive video technology called high-quality panoramic video is gradually increasing. High-quality panoramic video can make viewers feel an immersive sensation, as if they are actually at the site, through a wide field of view (FOV) that is larger than the human visual angle [1]. Panoramic video provides spatiotemporal views by stitching together images from multiple video cameras pointed in different directions and allows viewers to tour all angles of an open space or structure. Panoramic video services are already being provided through panoramic photo applications in smartphones and digital cameras, as well as through panoramic images in Google Maps. Exhibition halls and theaters are using a graphics-based panorama. Meanwhile, real panoramic video technology is being partly used in the surveillance domain through low-quality panoramic video, while high-quality real panoramic video technology is being researched and developed by HHI [1], the Immersive Media Corporation [2], and other organizations [3]-[5]. This paper also aims to present high-quality real panoramic video technology for streaming services.

Panoramic video gives the feeling of immersion owing to its wide visual angle, but it can be used only in a special environment because of its wide resolution. In other words, there are few projection devices that are able to display panoramic views on a screen simultaneously under a general viewing environment, requiring a huge bandwidth for

Manuscript received May 29, 2013; revised Nov. 10, 2013; accepted Nov. 19, 2013.

This work was supported by government financing of Electronics and Telecommunications Research Institute (ETRI), Daejeon, Rep. of Korea, entitled "Human-Centric Panorama Technology".

Joo Myoung Seok (phone: +82 42 860 1216, jmseok@etri.re.kr) is with the Broadcasting &Telecommunications Media Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Yonghun Lee (lhk6975@gmail.com) is with the Agency for Defense Development (ADD), Daejeon, Rep. of Korea.

streaming service.

For this reason, it requires special solutions suitable for a panoramic video streaming service. Another problem is that there is a significant waste of bandwidth when an entire panoramic video is sent using a conventional image-based panorama service. Therefore, to view panoramic video in a general viewing environment, we need an interactive spatial trick mode service such as region of interest (RoI) trick mode [6]-[8]. However, interactive services have a low-quality service experience (QoSE) owing to interactive delay, which is also called a zapping delay [9]. An interactive delay appears in temporal trick modes (for example, channel switching and random access) as well as in RoI spatial trick modes. Furthermore, owing to the limited screen size, viewers need to make frequent navigations to view panoramic videos with a wide FOV, and RoI trick mode is more sensitive to interactive delay. Several proposals have been made to solve the interactive delay problem on a multi-channel streaming service such as IPTV.

In [10], channels that were not currently viewed were transmitted along with the currently viewed channel to reduce interactive delay. In [11], bandwidth savings was obtained by transmitting only the preferred channels based on the viewing history of the viewers. Because the methods used in [10] and [11] require additional bandwidth, [12] presented a method that transmits low-quality video data with a lower amount of bandwidth for fast channel switching. Reducing the bandwidth using lower video quality decreases the viewer dissatisfaction in terms of interactive delay but increases the viewer dissatisfaction regarding the video quality.

Therefore, this paper proposes a priority-based progressive streaming method using the characteristics of the human visual system (HVS) to minimize a waste of resources and reduce the degradation of QoSE resulting from an interactive delay.

Muntean and others worked on the characteristics of an HVS-based RoI streaming service in [13] and [14]. This method introduced the RoI-based adaptive scheme, which adjusts the regions differently within each frame. In more detail, each server is ready to send the encoded video through a visual perception based on eight quality states of RoI. The clients then receive RoI video data from one of the eight state servers depending on the suitable RoI state feedback report regarding user interest as a result of eye tracking and network status.

The proposed visual-attention-aware progressive ROI trick mode streaming service (VA-PRTS) is similar to previous works [13], [14] in its approach of considering the HVS characteristics and applying them to RoI streaming. However, the RoI composition method of previous works has a concentric circle shape similar to a target and can be applied only if the screen size is identical to the video resolution. Thus, it cannot provide a flexible composition of RoI for dynamic RoI selection, which is required for RoI trick mode for the panoramic video of this study. Furthermore, resource utilization is low if a video is pre-encoded in eight states of RoI and distributed to each server. In addition, a client must frequently access the server with suitable RoI states according to the eye tracking and changes in network status while receiving streaming video. Moreover, the complicated process in which a client selects one stage among various types of stages, such as the eight RoI stages, is not efficient in terms of its implementation.

To overcome the limitation of previous methods, this paper exploits the characteristic of visual attention in which the perceptible spatial frequency decreases dramatically away from the point of gaze. This paper proposes VA-PRTS, which transmits and receives video data progressively by corresponding to changes in visual attention based on the viewing environment. VA-PRTS decreases the interactive delay without degrading the perceptual video quality.

The remainder of this paper is organized as follows. Section II introduces the background of related technologies. Section III describes an interactive panoramic video service environment, in which RoI spatial trick modes are used, and presents the proposed VA-PRTS in further detail. Section IV describes an experiment for evaluating the effectiveness of the proposed VA-PRTS in increasing the QoE. Finally, section V provides some concluding remarks regarding this proposal.

II. Background

The characteristics of the HVS should be taken into account when developing accurate video quality metrics. A fundamental characteristic is usually expressed using a contrast sensitivity function (CSF) derived by Daly [15]. A widely used CSF usually peaks at a certain frequency level and decreases drastically following a frequency increase. This phenomenon also applies to a temporal domain with a temporal velocity. On the other hand, previous researchers have found that the contrast sensitivity also has the highest resolution around the gaze point of the eyes and extremely decreases its sensitivity to other objects away from this point [16], [17]. Our eyes can easily recognize a simple icon from the center of view to 60 degrees in the visual field. However, letters are not recognizable within this same range. In addition, our eyes cannot recognize text outside of an area within 20 degrees of the gaze point [18].

This phenomenon is caused by visual attention and visual acuity owing to the non-uniform distribution of the photo receptors (cones and rods) on the retina. Experimentally, several researchers previously improved the original CSF, as

shown in (1), by following the visible contrast threshold function for visual attention [16], [17]:

$$CT\left(f_S, e\right) = CT_0 \exp\left(\alpha f_S \frac{e+e_2}{e_2}\right), \quad (1)$$

where f_S denotes the spatial frequency based on the cycle per degree (cpd), CT_0 is the minimal contrast threshold, α is the spatial frequency decay constant, and e and e_2 are the eccentricity and half-resolution eccentricity, respectively. Both e and e_2 are represented in terms of degree.

As previously mentioned in the introduction, a panorama is any wide-angle view or representation of a physical space whether in painting, photography, or video. Multiple images and videos with overlapping fields of view are combined to produce a segmented panorama or high-resolution image. A complex sequence of steps is needed to make a panoramic video. The first stage involves setting up the camera and configuring it to capture all videos. The second stage involves a stitching process, that is, the shifting, rotating, and distortion of each of the images such that both the average distance between all sets of control points is minimized, and the chosen perspective is still maintained. The next stage is then cropping the panorama so that it adheres to a given rectangular image dimension [1], [19].

III. Proposed VA-PRTS

The VA-PRTS proposed in this paper is a progressive streaming method that sets the priority of the RoI video data using the capability of visual attention based on the viewer's level of visual attention, which has a lower quality recognition sensibility as the distance from the focus point increases. In addition, only the perceptible video data based on the RoI quality is sent initially for a fast start up. VA-PRTS can provide a practical RoI trick mode for an interactive panoramic video streaming service because it can improve the QoSE with no degradation in the quality of the video experience (QoVE).

VA-PRTS's service scenario for interactive panoramic video is as follows. VA-PRTS provides low-quality panoramic navigation to allow viewers to select the point of view that they want to see and plays the video on the RoI main screen in line with the RoI screen size when they select a point of interest (PoI).

1. Viewing Environment of VA-PRTS for Panoramic Video

To easily understand the principle of VA-PRTS, above all, the characteristic of the viewing environment must be understood, as shown in Fig. 1. The most important factors of the viewing environment are the screen size, video resolution, and viewing distance.

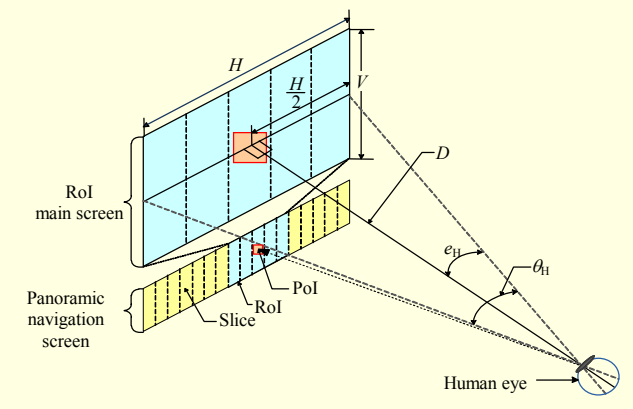


Fig. 1. Viewing environment of VA-PRTS for panoramic video.

As shown in Fig. 1, H denotes the horizontal screen size and V denotes the vertical screen size. D is the viewing distance. The available visual angle, θ_H, is computed using (2). D is calculated by multiplying V by d, which is a multiple of (D=d×V) [14]. Visual angle θ_H of the screen shown in Fig. 1 is

$$\theta_H = \left[\theta_{H,\text{radians}} = 2\arctan\left(\frac{H}{2d \times V}\right) = 2\arctan\left(\frac{H}{2D}\right)\right] \times \frac{180}{\pi}. \quad (2)$$

As described above, when viewers watch the RoI video of a panoramic video, as they are farther from the center of the RoI main screen, their visual attention gradually decreases, and the degree of this change is denoted as e(deg). Visual perception is more sensitive to horizontal changes, and a type of cylindrical projection is therefore of interest in this work. Hence, only horizontal changes in eccentricity are considered, that is, e=e_H= θ_H/2, as they are more suitable for the characteristics of a panoramic video [17], [18].

When the point of gaze is directed to the PoI, visual attention decreases away from the PoI as a function of eccentricity. Therefore, to determine a viewing environment in line with the visual perception, we can find the optimized viewing distance in tune with the screen size and video resolution through (2) and determine the appropriate video resolution in tune with the viewing distance and visual acuity.

It has been said that in the HVS, it is difficult to visually perceive spatial frequencies of 60 cpd or higher [17]. According to the Nyquist sampling theorem, the frequency is one cycle, which is sampled as two pixels. Based on the results of existing studies, for full HDTV, the proper horizontal visual angle at a viewing distance (d=3) equal to three times the height is 32 degrees (30 degrees to 33 degrees). In other words, the proper video resolution for full HDTV at the viewing distance of d=3 is 32 degrees × 30 cpd × 2 pixels or 1,920 pixels for the horizontal size, which is identical to

1,920 × 1,080 pixels, which is the current standard resolution of full HDTV with a 16:9 ratio. If 60 cpd is used, which is the maximum spatial frequency perceivable by humans at the same screen size, the horizontal resolution becomes about 4,000 pixels, which is the standard for UDTV and digital cinema.

In the final analysis, if the viewing distance is determined by the screen size using (1) and (2), the video quality in tune with human visual acuity, as well as the available viewing distance appropriate for the video quality, can be determined. Moreover, to configure the RoI with flexibility and fine granularity, VA-PRTS needs a special video encoding method rather than a conventional simple video coding method for interactive panoramic video services. Accordingly, this study uses the slice encoding method, which encodes a video by dividing it in the vertical direction, as shown in Fig. 1, for a wide panoramic video and flexible RoI composition. In other words, multiple slices in tune with the RoI main screen size with the PoI at the center are rendered as a one-slice group.

2. Visual-Attention-Based Priority Position

As mentioned above, visual attention is closely associated with the viewing environment, and the determination of e_H as shown in Fig. 1 is an important variance. As e_H increases, the visual attention decreases and the perceptible spatial frequency decreases. The spatial frequency that a person with 20/20 vision perceives according to changes in e_H is called the cutoff spatial frequency and is denoted as $f_{S,\text{cutoff}}$, which is presented in (3).

$$f_{S,\text{cutoff}} = f_S(e) = \frac{e_2 \ln(1/CT_0)}{(e_H + e_2)\alpha}. \quad (3)$$

The constant values of α, e_2, and CT_0 are 0.106, 2.3, and 1/64, respectively. At the same screen size, the smaller the available visual angle becomes, the smaller e_H becomes at a far viewing distance [20].

Figure 2 shows the basic concept of VA-PRTS for a determination of the priority position of RoI video data with no degradation of the QoVE. As the visual attention becomes less capable of discerning the quality as the visual point becomes farther from the fovea point, it can be expressed as a $f_{S,\text{cutoff}}$ curve, as shown in Fig. 2. More specifically, assuming that the RoI main screen is a full HDTV screen as in the current viewing environment, the viewing environment is d=3, and e_H = 16 degrees. By (3), $f_{S,\text{cutoff}}$ = 40 cpd at the fovea point, which is the center of visual attention. Figure 2(a) shows a case in which the viewing distance is d=3, and Fig. 2(b) shows the far viewing distance of d=5. $f_{S,\text{cutoff}}$ is normalized to the same resolution to compare the variations of $f_{S,\text{cutoff}}$ based on the viewing distance.

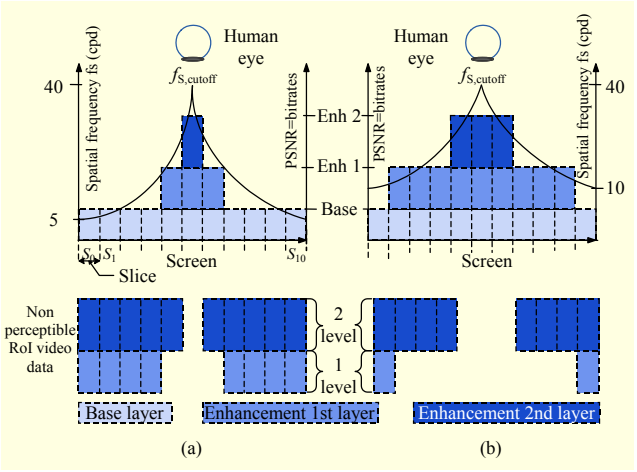


Fig. 2. Proposed VA-PRTM based on visual attention: (a) d=3 and (b) d=5.

The shape of the $f_{S,\text{cutoff}}$ curve is gradual when the viewing distance is far, as shown in Fig. 2(b), because the entire video is shown at a single glance, and the video quality of the entire video must be similar for viewers to feel that the quality is good at their available viewing distance. In contrast, when the viewing distance becomes shorter than in Fig. 2(a), the $f_{S,\text{cutoff}}$ curve falls sharply. In conclusion, VA-PRTS can determine the priority of RoI video data required in this study when the video quality of each position is determined according to the choice of suitable video quality and the visual attention at the available viewing distance according to the visual acuity.

As mentioned above and shown in Fig. 2, each video frame is divided into a number of vertical slices (S_n, that is : S_0 to S_{10}) to flexibly configure the RoI along the horizontal axis and determine different qualities according to $f_{S,\text{cutoff}}$. Each slice supports various quality levels (for example, basic, intermediate, and high) using a layered encoding [21] scheme such as Scalable Video Coding (SVC) [22]. According to [23], SVC has a high adaptability to a streaming network. Therefore, this paper also works based on the SVC scheme.

As shown in Fig. 2, the quality layer determines three spatial layers: a base layer, first enhancement layer, and second enhancement layer. For the proposed VA-PRTS, this paper defines the CVAM algorithm to determine the quality of the encoded video slice by $f_{S,\text{cutoff}}$ and the progressive streaming method based on the priority of the RoI video data. In this case, if there are too many video quality layers, only the encoding and service complexity may increase despite there being no perceptible visual sensitivity.

A. CVAM

An RoI is composed of h slices with PoI $P(x, y)$ as the center.

http://dx.doi.org/10.4218/etrij.14.2113.0012

The number of slices in an RoI corresponds to that of the horizontal and vertical sizes of the RoI main screen (that is, H and V) in Fig. 1.

Once an RoI is determined, the average spatial cutoff frequency for each slice position $f_{\text{avg},i}$ is computed using $f_{\text{S, cutoff}}$ in (4) and slice size S_{H} (S_{H}=H/h) to determine the quality level for each slice in the RoI. The smaller the value of S_{H} is, the closer the value to the shape of the $f_{\text{S, cutoff}}$ curve that can be obtained, although the encoding bitrates increase.

$$f_{\text{avg},i} = \frac{1}{S_{\text{H}}} \sum_{x=1}^{S_{\text{H}}} f_{\text{S}}(e), \qquad e = x\left(i \cdot S_{\text{H}}\right) \frac{e_{\text{H}}}{\left(W_{\text{opt}} / 2\right)}, \qquad (4)$$

$$i = 0,1,\ldots,(h-1),$$

where W_{opt} denotes the width of the optimal resolution at the given viewing environment. To adapt the image quality according to the visual attention, maximum spatial frequency F_i for the l-th quality level is computed. When optimal width W_{opt} of optimal spatial frequency F_{opt} is set, the spatial frequency for the l-th quality level F_i is computed by multiplying F_{opt} by the ratio of the width of the image at the l-th quality level, W_i to W_{opt}. L is the maximum number of l-th layers.

$$F_i = F_{\text{opt}} \cdot \frac{W_i}{W_{\text{opt}}}, \text{ subject to } l = 0,1,\ldots,(L-1). \qquad (5)$$

In Full HD videos, $e_{\text{H}} \approx 16(\text{deg})$ when $W_{\text{opt}}/2 = 960$ pixels and d=3 based on F_{opt} of 30 cpd. In CVAM, the quality level for each slice corresponding to its visual attention is determined by minimizing the difference between $f_{\text{avg},i}$ and F_i. As a result, CVAM-based RoI video data, L^*, is determined. This is expressed in (6).

$$L^* = \arg \min_{l \in (L-1)} \left| F_i - f_{\text{avg},i} \right|, \; i = 0,1,\ldots,(h-1). \qquad (6)$$

Meanwhile, RoI can be determined continuously through eye tracking and the motion-based RoI selection as suggested by Ciubotaru and others [14] and by Azad and others [24], respectively. However, in the viewing environment considered in this study, the eye tracking procedure should not only support high-performance camera-based long distance eye tracking to ensure high-accuracy RoI selection but also a very difficult camera calibration between the line of vision and the tracking camera. The motion-based RoI selection requires previous determination of the RoI of a specific area of the content that is expected to receive high interest by motions. The method of encoding only the pre-defined RoI area in desirable quality has the problem of lower RoI selection freedom because viewers can select only pre-defined RoIs. Therefore, the pre-defined RoI method is not suitable to apply to the viewing environment considered in this study because it is difficult to determine all RoIs in advance for panoramic videos that have many meaningful contexts.

On the other hand, as shown in Fig. 1, the VA-PRTS offers the advantage of selecting the RoI freely because when you choose a PoI through the panoramic navigation map in which a panorama frame is divided into many vertical slices, the PoI slice is placed at the center and the neighbor slices are placed for the screen size. Furthermore, because the PoI slice chosen by the viewer is the highest interest point, we can assume that the PoI slice is always identical to the center of the screen according to the common human behavior without the complex process.

3. Progressive Streaming Based on CVAM

Figure 3 shows the basic concept of the progressive streaming method, which minimizes the interactive delay through the progressive streaming of the CVAM-based priority of RoI video data. As shown in Fig. 3, two seconds of RoI video data is buffered, and the propagation delay time, T, is assumed to be the sum of the packetizing time, transmission time, and intra frame period time [9]. When a viewer requests RoI trick mode, the conventional method (as shown in Fig. 3(a)) renders RoI video data for two seconds at the startup

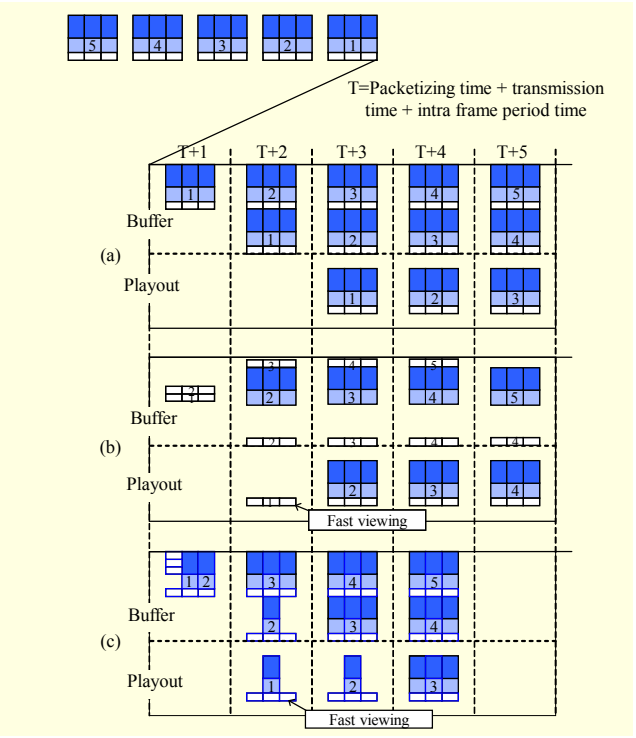


Fig. 3. Transmission of RoI video data (d=3): (a) conventional streaming, (b) low-quality base layer progressive streaming, and (c) CVAM-based progressive streaming.

time (T+3) after the buffering time (up to T+2). In this case, even though the video quality is not the original video quality, as in Figs. 3(b) and 3(c), because the two-second RoI video data is made up of small amounts of video data, the propagation delay time is reduced to T–Δ and the startup time is faster. Furthermore, if VA-PRTS transmits RoI video data at the same throughput as in Fig. 3(a), the receipt of video data for buffering is completed at T+1, and play startup is possible at T+2. As a result, the interactive delay time decreases compared to that shown in Fig. 3(a), and the QoSE is improved. However, Fig. 3(b) uses only the base layer of RoI video data, and the QoSE is improved, but the QoVE is decreased. However, CVAM-based progressive streaming, as shown in Fig. 3(c), improves the QoVE for a fast startup while the change in QoVE is not recognized.

Furthermore, the time when the original video quality is completed, in terms of QoSE, is called the convergence time. To prevent any problems from occurring due to a change in focus point after the convergence time, the video data (frames 3, 4, and 5, as shown in Fig. 3) excluding the initial buffered RoI video data is progressively transmitted at the original quality to achieve the practicality of the proposed RoI trick mode streaming.

IV. Simulation Results and Discussion

In this study, two experiments are conducted to verify the proposed method. First, the QoVE and QoSE are tested for VA-PRTS. To analyze the effects of the proposed method in detail, it is assumed that the viewer previously requested the RoI video data through a panoramic navigation. The reason for this is that the measurement results can be objectified because the QoE results after the generation of a user request must be differentially measured, and the proposed method can be generalized if it can be tested based on HD content with various genres rather than limited panoramic videos. Second, we develop a VA-PRTS player to check the practical effectiveness of the proposed method in the real world and discuss the implantation details through our panoramic video content.

1. QoVE and QoSE Assessment

In the experimental environment, a 60-inch display with a resolution of 1,920 × 1,080 is used, and D = 2.37 m in (1). As mentioned above, the SVC scheme is used to provide layered encoding that adapts the image quality to changes in visual attention. There are three SVC scalability schemes: temporal, quality, and spatial. The temporal scalability controls the frame rate for layered encoding. The quality scalability performs layered encoding according to the quantization parameter (QP) stepsize, but it is difficult to regularize the difference in the QP stepsize because the QoVE varies from video to video. In this study, the spatial scalability that has high correlation with visual attention is used for layer encoding. Three-layer spatial scalability (L =3) is used. The spatial resolution of each layer is as follows: l =2 is 1,920 × 1,080 (1,080 p), l =1 is 1,280 × 720 (720 p), and l =0 is 640 × 480 (480 p). QP for the layers is set to 26, which means a high quality uniform encoding. To flexibly configure an RoI along the horizontal axis, wipe slices (the slice number of each layer, h =10) are used for the RoI main screen. The maximum spatial frequency for each layer (l) is computed using (4). l =2 is 30 cpd (W_{L-1}=W_{opt}), l=1 is 20 cpd, and l=0 is 10 cpd. Four test sequences in [25] (Old Town Cross, Sunflower, Touchdown Pass, and Tractor) are used in the experiment, as shown in Table 1. All of the test sequences have an average bitrate of 8 Mbps (±5%) and higher quality (42 dB) than that of general services (that is, 35 dB), which clearly verifies the effectiveness of VA-PRTS.

Two metrics are used to compare the performance of the proposed method with that of conventional methods. First, a foveal video quality assessment is used for measuring the video quality that the HVS can perceive. The interactive delay is then used for measuring the time interval from a trick mode request to the rendering of the requested video.

A. Foveal Video Quality Assessment

The peak signal-to-noise ratio (PSNR) is commonly used to measure the video quality. This paper employs the foveal-mean square error (FMSE) in [26] and [27] to measure the perceptual video quality. To quantify the video quality based on the visual attention curve, we divide each frame into a given number of slices and apply distortion sensitivity $\rho_s(m, n)$ according to the position of each slice.

$$FMSE = \sum_{S=0}^{H/S_H-1} \frac{1}{S_H V} \sum_{m=0}^{S_H-1} \sum_{n=0}^{V-1} \left[\rho_S(m,n) \{ I(m,n) - R(m,n) \} \right]^2,$$

$$\rho_S(m,n) = \frac{f_S(e)}{f_{S,MAX}}, \quad 0 < \rho_S(m,n) \le 1, \tag{7}$$

where $f_{S,MAX}$ denotes the CSF at PoI, and, thus, $f_{S,MAX}$= $f_S(0)$ in (3). $f_S(e)$ is the CSF of a point (m, n) distant from the PoI. $I(m, n)$ and $R(m, n)$ denote the original and reconstructed images, respectively. FMSE computed using (7) is a measure for the human perception of the reconstructed video quality, and it therefore cannot be higher than the highest quality of the original encoded video, denoted as MSE_{MAX}(MSE_{MAX}= MSE_{L-1}). Therefore, FMSE is modified to $\overline{FMSE}$ through the

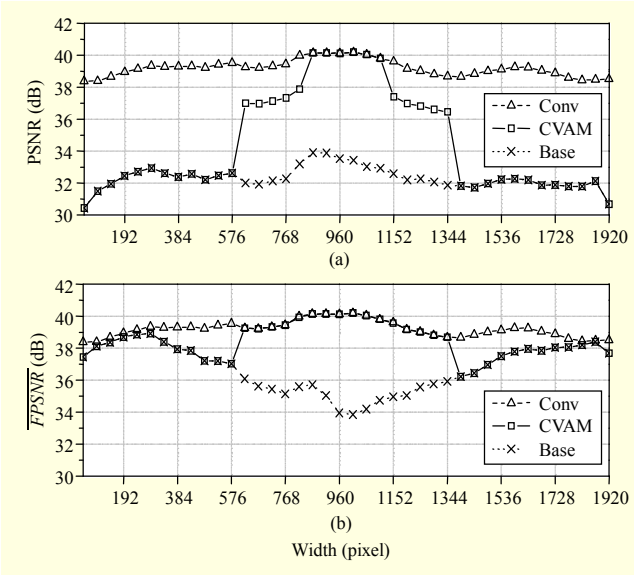


Fig. 4. PSNR and $\overline{FPSNR}$ (10-th frame of Old Town Cross).

following condition.

$$\overline{FMSE} = \begin{cases} MSE_{\text{MAX}}, & FMSE \le MSE_{\text{MAX}}, \\ FMSE, & FMSE > MSE_{\text{MAX}}, \end{cases} \quad (8)$$

$\overline{FMSE}$ can be converted into the foveal PSNR ($\overline{FPSNR}$) using the same conversion as MSE into the PSNR.

In Fig. 4, three different methods are compared. In Conv, SVC-encoded video signals that are necessary to reconstruct video content with a given level of quality are received and decoded. In Base, the reconstruction with a minimum quality level is made. CVAM is the proposed method using (6) to adapt the quality of the coded video data to the visual attention. Figure 4 shows the PSNR and $\overline{FPSNR}$ of these three methods.

Figure 4(a) compares the three methods in terms of the PSNR. This is an objective quality measure that determines the video quality solely based on the scalability layer. Compared to the quality (PSNR) of Base and CVAM in Fig. 4(a), the perceptual quality ($\overline{FPSNR}$) of Base and CVAM in Fig. 4(b) is higher in the large eccentricity areas where the distortion sensitivity is low. However, the quality of Base and CVAM in Fig. 4(b) is not significantly different from that in Fig. 4(a) in the areas near the PoI because the distortion sensitivity is increased in these areas.

Figure 5 shows the captured images. Figure 5(a) shows a conventional method that receives all layers for decoding and rendering. The image frame in Fig. 5(b) has a decreased video quality by the proposed CVAM, but the difference in the video quality is difficult to recognize unless it is enlarged and has a 69% bitrate saving. Compared to the encoding bitrates of non-layer sliced-based H.264, the encoding bitrates of the three-spatial-layer sliced-based SVC increases by 112%. Consequently, our proposed method offers a 57% bitrate saving.

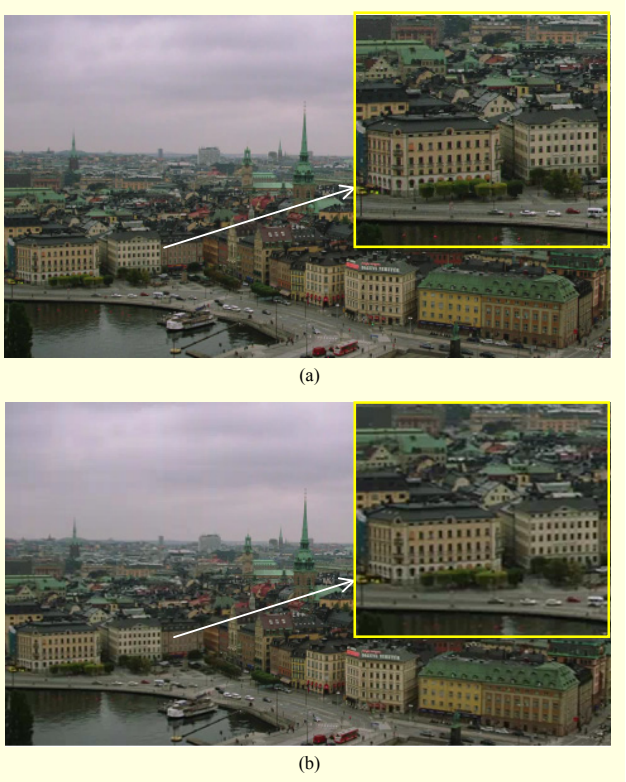


Fig. 5. Captured image (10-th frame of Old Town Cross): (a) conventional method and (b) proposed method.

B. Interactive Delay Assessment

When spatiotemporal trick mode is applied to interactive video streaming services, the QoE of the three methods is examined by measuring the interactive delay and perceptual quality ($\overline{FPSNR}$). The network throughput is assumed to be the average bitrate of the encoded bitstream. The transmission delay is computed over the average bitrate. For jitter relaxation and error recovery, the receiver receives transmitted video data for two seconds before performing decoding and rendering.

In the experiment, startup quality Q_S, startup time T_S, convergence quality Q_C, and convergence time T_C are measured when a request for a trick mode occurs (T=0).

Compared to Conv, which transmits all the layers, Base and CVAM have small amounts of video data to transmit, and the delay from a trick mode request to the startup time is thus shorter in Base and CVAM. Base has the smallest amount of video data to transmit, and its startup time ($T_{S,\text{Base}}$) is thus the earliest. However, its perceptual quality is much lower than that of Conv and CVAM. In Base, the viewer is likely to experience a quality degradation owing to a large gap between $Q_{S,\text{Base}}$ and

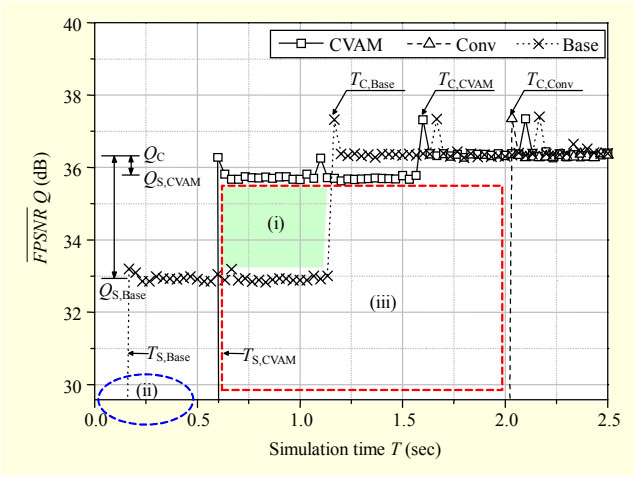
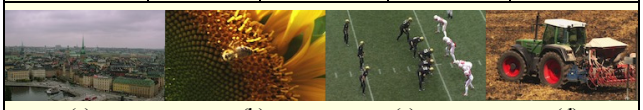


Fig. 6. Trick mode delay and quality (Old Town Cross sequence).

Table 1. Transition time and transition quality gain.

Test sequence	CVAM		Base	
	T_s	Q_s (%)	T_s	Q_s (%)
(a) Old Town	0.56 s	35.74 (93.77)	0.12 s	32.94 (50.36)
(b) Sunflower	0.88 s	42.17 (98.86)	0.30 s	41.16 (70.25)
(c) Touchdown	0.47 s	36.42 (84.85)	0.12 s	34.34 (54.26)
(d) Tractor	0.56 s	38.82 (92.91)	0.25 s	36.89 (63.17)



(a) (b) (c) (d)

Q_C. The startup time ($T_{S,CVAM}$) of CVAM is slower than that of Base, but a quality degradation is less likely to be experienced, that is, the difference between $Q_{S,CVAM}$ and Q_C is not large. This means the quality degradation is not perceptible by humans. The shaded area labeled "(i)" in Fig. 6 represents the extent of the QoE (both QoVE and QoSE) improvement in CVAM over Base. The blue-colored circle labeled "(ii)" represents the 0.5-second interactive delay, evaluated as "fair" or "good," according to the mean opinion score (MOS) [28]. The red-colored rectangle labeled "(iii)" shows that CVAM has a better QoE than that of Conv.

Table 1 presents the performance of CVAM and Base for the four test sequences used in the experiment. The parenthesized percentage next to startup quality Q_S is the proportion of Q_S in convergence quality Q_C (100%). The test sequences, each of which has different characteristics, have varying performance results. On average, CVAM achieves 93% of the perceptual quality of Conv and speeds up the startup time by 66% ($\approx$0.61 s, $T_{S,Conv} = T_{C,Conv} = 1.81$ s).

Video quality has traditionally been measured either subjectively (based on human experience), such as according to the MOS and the structural similarity, or objectively (based on computerized algorithms), such as according to encoding bitrates and PSNR. As subjective video quality is relative to a viewer's perception, it reflects his or her opinion on a particular video sequence. The important factors considered in this study to improve QoSE for panoramic video streaming service are QoVE and interactive delay. Therefore, we have to perform the subjective test of the proposed method on the connection between QoVE and interactive delay. However, subjective video quality tests are quite expensive in terms of time (preparation and running) and human resources, which limits the possibility of performing broad subjective tests in this study. Therefore, this study has a simple subjective measurement of QoVE when QoSE is improved, owing to low interactive delay because low interactive delay indicates desirable quality regarding the QoSE aspect.

The subjective video quality measurement test is conducted with 15 participants who are engaged in related jobs. For the test environment, in the viewing environment as shown in Fig. 1, the participants stare at the center of a 60-inch TV, and two-second test videos with 1) original video quality and 2) CVAM-based video quality are shown alternately. Then, participants are asked to identify the video suffering from video quality degradation by selecting "No. 1," "No. 2," or "I don't know."

To remove prejudices about the comparison, this test is repeated three times by changing the order of 1) and 2). The reason for limiting the repetition to three is to minimize the measurement error that results from the tester losing attention when repetitive tests are performed with the same test sequences. The test results show that the perception of the video quality degradation is insignificant: Old Town (4%), Sunflower (2%), Touchdown (6%), and Tractor (11%) for 2). In addition, the test results show that the content type has an influence on the capability of visual attention. For Tractor, as the object is so large that it occupies the entire screen, the testers detect the difference much more easily owing to high correlation of the object. However, when the object moves and when playing times of the degraded video data are shorter, the difference is difficult to recognize. As a result, the QoSE is improved by VA-PRTS with minimized interactive delay and without degrading the perceptual video quality. Furthermore, for panoramic video, VA-PRTS is more important because the frequency of an RoI interaction is high owing to a limited viewing environment and a wide resolution of the content.

2. Development Results of VA-PRTS

As shown in Fig. 7, a meaningful area of 6,912 × 1,088 (as

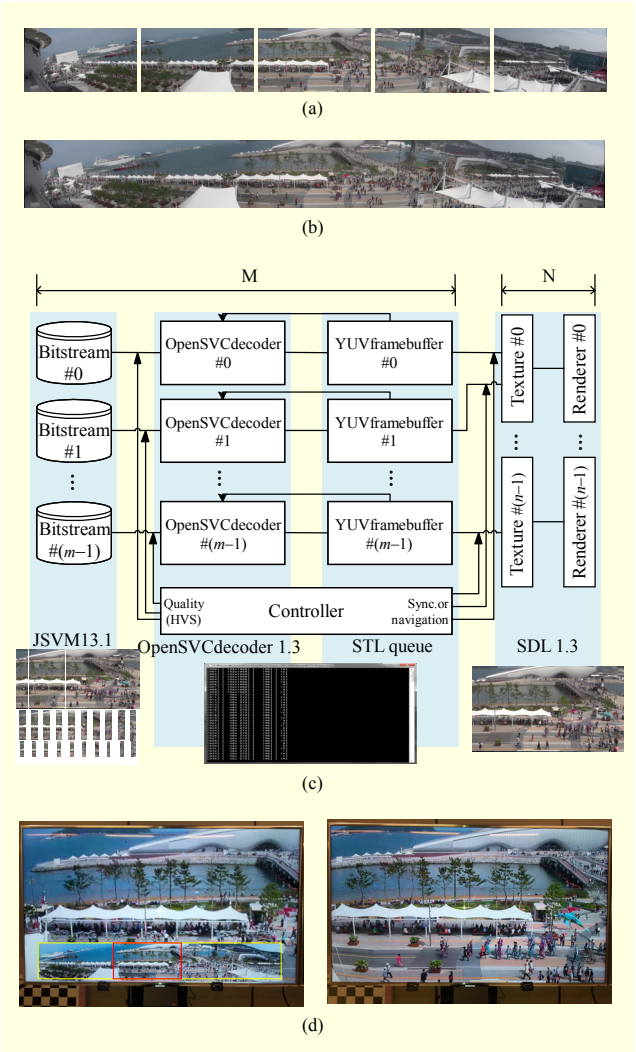


Fig. 7. Development results of VA-PRTS: (a) snapshot of five input HD videos; (b) snapshot of stitched and cropped panoramic video; (c) design of VA-PRTS player; and (d) VA-PRTS player on 60-inch HDTV.

shown in Fig. 7(b)) in size is cropped from 8,581 × 1,102 pixels, obtained through a cylindrical stitching of an Expo parade captured with five HD cameras at 180 degrees (as shown in Fig. 7(a)). The configuration of a slice is S_H =192 and h=36, and a slice size is thus defined as 192 × 1,088. Three layers for spatial scalability (L=3) are used. The spatial resolution of each slice layer is as follows: l =2 is 192 × 1,088, l =1 is 96 × 1,088, and l =0 is 48 × 1,088. According to the IPTV service configuration, each intra period is a 15-frame interval, and the types of encoding are only I and P types based on JSVM 13.1 [22]. The running time of the panoramic video is 120 seconds at 30 frames per second.

The off-line version of the VA-PRTS player is completed by C/C++, as shown in Fig. 7(d), and the on-line version is under development. Figure 7(c) shows the design of the VA-PRTS player. In summary, the M section transfers the video data to the N section after decoding the inputted bitstreams and converting the color format. The N section assigns computer resources for rendering. Notably, considering the future upgrade of the M section, we implement a consistent input interface for the off-line version (file-based) and on-line version (streaming-based). Moreover, the decoder and frame buffer are assigned to each input member of the bitstream to improve the player's performance, and the VA-PRTS player is developed on the basis of the SDL1.2.15 [29] framework. Actually, the same number of textures as the allocated number of slices is created in advance, and the multiple pieces of video data inputted from the M section are then connected to each texture for rendering so that there will be no problem in the playback speed and synchronization. The controller block, which is the brain of VA-PRTS, sets the configuration of the viewer's viewing environment for optimal CVAM operation and controls resources. The results of the implementation are reflected in Fig. 7(d): the left image shows a navigation to select the RoI through a panoramic navigation based only on the base layer, and the right image is a rendering of VA-PRTS based on the RoI.

V. Conclusion

This paper proposed VA-PRTS, an effective RoI trick mode for an interactive panoramic video service. The proposed VA-PRTS exploits the characteristics of visual attention in that the perceptual quality decreases away from the point of gaze. It prioritizes RoI video data to transmit according to changes in visual attention and transmits the prioritized data progressively so that a picture in visual attention-sensitive areas is reconstructed first.

The proposed VA-PRTS decreases the interactive delay by over 66% without degrading the perceptual video quality and increases the bandwidth utilization by over 57%. The CVAM algorithm, which utilizes layered encoding to adapt the video quality to the visual attention, was suggested. CVAM-based progressive streaming for RoI video data is able to reduce the interactive delay. We also developed an RoI trick mode panoramic video player based on CVAM to confirm its feasibility in the real world and modified FMSE according to a perceptual quality measure by applying eccentricity to the MSE measure.

The proposed VA-PRTS can be utilized for temporal trick modes (for example, channel switching and random access) in general interactive video service environments, such as IPTV and mobile video services, as well as in panoramic video service environments.

References

[1] O. Scheer et al., "Ultrahigh-Resolution Panoramic Imaging for Format-Agnostic Video Production," *Proc. IEEE*, vol. 101, no. 1, Jan. 2013, pp. 99-114.

[2] *London Calling Demo Contents*, Immersive Media Company, Kelowna, BC, Canada, 1994. Accessed Jan. 15, 2013. http://immersivemedia.com/demos

[3] H.S. Quershi et al., "Quantitative Quality Assessment of Stitched Panoramic Images," *IET Image Proc.*, vol. 6, no. 9, Dec. 2012, pp. 1348-1358.

[4] A. Ahmed et al., "Geometric Correction for Uneven Quadric Projection Surfaces Using Recursive Subdivision of Bézier Patches," *ETRI J.*, vol. 35, no. 6, Dec. 2013, pp. 1115-1125.

[5] J. Yoo et al., "Regional Linear Warping for Image Stitching with Dominant Edge Extraction," *KSII Trans. Internet Inf. Syst.*, vol. 7, no. 10, Oct. 2013, pp. 2464-2478.

[6] H. Kimata, K. Fulazawa, and N. Matsuura, "Partial Delivery Method with Multi-bitrates and Resolutions for Interactive Panoramic Video Streaming System," *ICCE IEEE Int. Conf.*, Las Vegas, NV, USA, vol. 4, no. 1, Jan. 9-12, 2011, pp. 891-892.

[7] M. Makar et al., "Real-Time Video Streaming with Interactive Region of Interest," *17th IEEE Int. Conf. Image Process.*, Hong Kong, China, Sept. 2010, pp. 4437-4440.

[8] J. Seok et al., "A Visual Perception Based View Navigation Trick Mode in the Panoramic Video Streaming Service," *IEICE Trans. Commun.*, vol. E94-B, no. 12, Dec. 2011, pp. 3631-3634.

[9] P. Siebert, T.N.M. Van Caenegem, and M. Wagner, "Analysis and Improvements of Zapping Times in IPTV Systems," *IEEE Trans. Broadcast.*, vol. 55, no. 2, June 2009, pp. 407-418.

[10] I. Kopilovic and M. Wagner, "A Benchmark for Fast Channel Change in IPTV," *IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, Mar. 31 - Apr. 2, 2008, pp. 1-7.

[11] C.Y. Lee, C.K. Hong, and K.Y. Lee, "Reducing Channel Zapping Time in IPTV Based on User's Channel Selection Behaviors," *IEEE Trans. Broadcast.*, vol. 56, no. 3, Sept. 2010, pp. 321-330.

[12] E. Kurutepe, M.R. Civanlar, and A.M. Tekalp, "Client-Driven Selective Streaming of Multi-view Video for Interactive 3DTV," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, Nov. 2007, pp. 1558-1565.

[13] G. Muntean, G. Ghinea, and T.N. Sheehan, "Region of Interest-Based Adaptive Multimedia Streaming Scheme," *IEEE Trans. Broadcast.*, vol. 54, no. 2, June 2008, pp. 296-303.

[14] B. Ciubotaru, G. Muntean, and G. Ghinea, "Objective Assessment of Region of Interest-Aware Adaptive Multimedia Streaming Quality," *IEEE Trans. Broadcast.*, vol. 55, no. 2, June 2009, pp. 202-212.

[15] S. Daly, "Engineering Observations from Spatiovelocity and Spatiotemporal Visual Models," *Proc. SPIE Human Vis. Electron. Imag.*, vol. 3299, Jan. 17, 1998, pp. 180-191.

[16] J. You et al., "Visual Contrast Sensitivity Guided Video Quality Assessment," *IEEE Int. Conf. ICME*, Melbourne, VIC, Australia, July 9-13, 2012, pp. 824-829.

[17] S. Winkler, "Issue in Vision Modeling for Perceptual Video Quality Assessment," *Signal Process.*, vol. 78, no. 2, Oct. 1999, pp. 231-252.

[18] Y. Ishiguro and J. Rekimoto, "Peripheral Vision Annotation: Noninterference Information Presentation Method for Mobile Augmented Reality," *Proc. 2nd Augmented Human Int. Conf.*, Tokyo, Japan, Mar. 12, 2011, article no. 8.

[19] R. Szeliski, "Image Alignment and Stitching: A Tutorial," Microsoft Research, Technical Report MSR-TR-2004-92, Oct. 2004. http://research.microsoft.com

[20] Z. Wang et al., "Foveated Wavelet Image Quality Index," *Proc. SPIE Appl. Digital Image Process. XXIV*, vol. 4472, Aug. 2001, pp. 42-52.

[21] T.M. Bae et al., "Multiple Region-of-Interest Support in Scalable Video Coding," *ETRI J.*, vol. 28, no. 2, Apr. 2006, pp. 239-242.

[22] Text of ISO/IEC 14496-10:2005/FDAM 3 Scalable Video Coding, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG, Lausanne, N9197, Sept. 2007.

[23] H. Kim et al., "Reducing Channel Capacity for Scalable Video Coding in a Distributed Network," *ETRI J.*, vol. 32, no. 6, Dec. 2010, pp. 863-870.

[24] S. Azad, W. Song, and D. Tjondronegoro, "Measuring Bitrate and Quality Trade-off in a Fast Region-of-Interest Based Video Coding," *Springer Adv. Multimedia Modeling*, vol. 6524, 2011, pp. 442-453.

[25] C. Montgomery et al., *Xiph.org Video Test Media (derf's collection)*, the xiph open source community, 1994. Accessed Aug. 16, 2013. http://media.xiph.org/video/derf

[26] S. Rimac-Drlje, M. Vranješ, and D. Žagar, "Foveated Mean Squared Error — A Novel Video Quality Metric," *Multimedia Tools Appl.*, vol. 49, no. 3, Sept. 2010, pp. 425-445.

[27] S. Lee, M.S. Pattichis, and A.C. Bovic, "Foveated Video Quality Assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, Mar. 2002, pp. 129-132.

[28] R. Kooij, K. Ahmed, and K. Brunnström, "Perceived Quality of Channel Zapping," *Proc. 5th IASTED Int. Conf. Commun. Syst. Netw.*, Palma de Mallorca, Spain, Aug. 28-30, 2006, pp. 155-158.

[29] S. Lantinga et al., *SDL version 1.2.15 open source*, Simple DirectMedia Layer forum, 2012. Accessed Jan. 15, 2013. http://www.libsdl.org/download-1.2.php

Joo Myoung Seok received his MS and PhD degrees in electronics from Kyung Hee University (KHU), Sunwon, Rep. of Korea, in 1999 and 2011, respectively. He has worked as a senior member of the research staff with the Realistic Broadcasting Media Research Department, ETRI, Daejeon, Rep. of Korea, since 1999. He served as a detached staff member for the Task Force Team of the Korea Communication Commission (KCC) in 2009. He served for promoting policies for broadcasting and telecommunication convergence contents and received the Achievement Award from the KCC. He was involved in developing the data broadcasting systems and personalized digital mobile broadcasting systems. Now, he is involved in developing the high quality panoramic video system. His research interests are in the areas of multimedia streaming, interactive media, and panoramic video cameras.

Yonghun Lee received his MS, BS, and PhD degrees in electrical engineering from Kyung Hee University, Suwon, Rep. of Korea, in 2006, 2008, and 2012, respectively. He has worked as a senior member of the research staff with the Agency for Defense Development (ADD), Daejeon, Rep. of Korea, since 2012. His research activities include image recognition, multimedia streaming, and wireless communications.