IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 13, 2018, accepted April 18, 2018, date of publication April 23, 2018, date of current version August 15, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2829194

Internet of Things (IoT) for Seamless Virtual Reality Space: Challenges and Perspectives

DONGHO YOU, (Student Member, IEEE), BONG-SEOK SEO, EUNYOUNG JEONG, AND DONG HO KIM, (Senior Member, IEEE)

Graduate School of Nano IT Design Fusion, Seoul National University of Science and Technology, Seoul 01811, South Korea

Corresponding author: Dong Ho Kim (dongho.kim@seoultech.ac.kr)

This work was supported by an Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea Government (MSIT) (No. 2016-0-00144, Moving Free-viewpoint 360VR Immersive Media System Design and Component Technologies).

ABSTRACT This paper addresses a novel virtual reality (VR) system that is based on the real world in which we live. The ultimate goal is to implement it as though a VR user freely exists in a place. To this end, it is most important to reconstruct a VR space that provides six degree-of-freedom (DOF), namely, yaw, pitch, roll, surge, sway, and heave. However, most currently released VR services that are based on the real world limit users' movements to three DOF. Even if the services support six DOF, most are highly complex and based on computer graphics. To overcome this problem, we first assume that there is a full Internet of things (IoT) infrastructure for collecting important data for VR space reconstruction. This assumption is realistic because many researchers expect that in the near future, IoT technology will lead to a world that connects not only people to people but also things to things. In this paper, we propose an end-to-end system architecture for the VR space that is based on the real world along with the element technologies that constitute the proposed system. This paper also includes a detailed survey of both conventional and emerging studies by other researchers.

INDEX TERMS Internet of Things, Internet of Multimedia-Things, virtual reality, virtual space, social media, immersive media, free-viewpoint video, 360° video.

I. INTRODUCTION

Approximately two decades ago, a film was released that brought a considerable shock to many people around the world. This film is ''The Matrix'', in which there are two worlds: a real world (i.e., Zion) and a virtual reality (VR) space (i.e., the Matrix). This movie has motivated the technical interest of many researchers in the representation and implementation of VR space [1]–[6]. The first VR machine was Sensorama Simulator, which was devised in the late 1950s by M. L. Heilig, who is called the ''Father of VR'' [7], [8]. This simulator was an amusement machine that simulated the nervous system using three-dimensional images, stereo sound, and smell. In the middle of the 1960s, a paper titled ''The ultimate display'' was released by I. E. Sutherland [9]. This paper predicted the direction of research and development of VR, and in the ultimate display space, computers can control the existence of things. Sutherland also released another paper, in which a head-mounted display (HMD) was designed with two small cathode-ray tubes (CRTs) surrounding the user's eyes [10]. This display was the first visual display device for the perception of VR. In later years, auditory display devices (e.g., Headphones) were widely used in combination with HMD. This technology led to the development of current HMDs, such as Oculus Rift, which has popularized modern VR technology.

According to [11], the size of the market for VR hardware is predicted to reach approximately 25 billion US dollars by 2021. Therefore global companies such as Facebook, Google, Samsung, LG, and HTC are competing to dominate this wired/wireless VR hardware market. Furthermore, in terms of VR service, various companies are introducing their own VR applications. For example, in May 2016, IKEA introduced ''IKEA VR Experience'', in which users can explore and interact with IKEA kitchens in VR [12]. Virtual Xperience Inc. also introduced a VR platform, which helps people explore and experience real estate by providing immersive VR visualizations [13]. These VR service platforms provide 6 full degrees-of-freedom (DOF) (rolling, pitching, yawing, surging, swaying, and heaving), but are based on computer graphics (CG), not the real world. In contrast, Google

introduced ''Google Expeditions'', which enables users to lead or join immersive virtual trips all over the real world [14]. NextVR also provides live broadcasting in virtual reality, in which users can experience sporting events, concerts, cinematic productions and so on [15]. These services have the advantage of being based on the world in which we live, but only 3-DOF (rolling, pitching, and yawing) are provided for users. In other words, translation, such as moving forward and backward, moving left and right, and moving up and down, cannot be experienced.

In recent years, various algorithms and methods have been studied for the 6-DOF VR space that is based on the real world, not CG. According to [16], it is difficult for common users to capture 6-DOF VR videos because it requires very complicated camera setups such as camera arrays or light-field cameras. Hence, structure-from-motion (SFM) [17]–[19] is used to reconstruct motion and 3D scenes, and a novel spherical panorama warping method for minimizing distortions when the scene is played has been proposed. However, it has a limitation that the algorithm assumes the scene to be fixed. Thus, a dynamic subject is perceived on a fixed geometry. Lytro, Inc. also introduced ''Lytro Immerge'', in which 6 DOF are provided, even if the subject moves [20], [21]. It is characterized by the use of a light-field camera and acquired data to seamlessly blend live action (i.e., real world) and CG. However, the light-field camera has the chronic disadvantage of low spatial resolution. This is because that the light-field camera shares imaging sensors for capturing spatial and angular information, which results in a trade-off between spatial and angular resolution. Therefore, for high-spatial-resolution performance, a very expensive light-field camera is required, which is not easy to popularize.

In this paper, we propose an end-to-end system architecture and its element technologies for providing a 6-DOF VR space that is based on the real world. The proposed system is realizable with the help of IoT infrastructure. We posit that the IoT environment is the main factor in overcoming the huddle of 6-DOF VR space. The proposed architecture especially considers commercialized smartphone cameras and video-sharing websites and social network services (SNSs) such as Youtube and Facebook for acquiring and storing the captured real-world images/videos.

In the acquisition and storage process, metadata are assumed to be exchanged between the camera and a variety of subjects. Such an assumption is feasible in the IoT infrastructure in which every subject has its own communication module. Based on the large number of images/videos that are stored on the image/video-sharing server, various virtual positions and viewpoints are reconstructed and free-viewpoint 360° VR real-time images/videos can be provided for users regardless of the time and place. The element technologies that constitute the proposed system must be developed further. We hereby describe related technologies for constructing the proposed system. In this regard, we make the following contributions:

- **System Architecture:** In contrast to previous work on VR space service that is based on the real world, we design a fully distributed architecture for acquisition, classification, virtual image/video reconstruction, transmission, and consumer processing.
- **Movement:** The proposed system can provide VR space with the movement of the watching user, in other words, 6 DOF.
- **Reasonableness:** The proposed VR space service can be realized with commercialized equipments and networks such as smartphone, 360° camera, wired/wireless HMD and SNS, and advanced equipment such as multiple arrayed or light-field cameras.

The remainder of this paper is organized as follows: In Section II, we present a detailed description of the proposed VR space service architecture, which is based on the real world. In Section III, various proposed methods are evaluated to determine their feasibilities. Finally, we present conclusions in Section IV.

II. PROPOSED SYSTEM ARCHITECTURE

A. OVERVIEW AND FINAL OBJECTIVE

Figure 1 shows the basic concept of the VR service architecture that is proposed in this paper, in which 360° VR images can be reconstructed from images/videos that are captured and stored on an image/video sharing server, and provided to VR users. The proposed end-to-end architecture can be divided into five parts: acquisition, classification, virtual image/video reconstruction, transmission and consumer processing. These are described in the following subsections. Furthermore, to incorporate convincing sound, signal processing methods for VR audio are described in the audio subsection.

Figure 2 shows the final goal of this paper, in which free-viewpoint 360° virtual views are provided for users, who can move all directions (i.e., 6-DOF). To achieve this objective, novel 360° VR view synthesis methods are required. We discuss such methods in the following subsections.

B. ACQUISITION AND CLASSIFICATION

As mentioned before, our objective is to reconstruct a VR world from the real world. This is different from virtual 3D modeling for games that are based on CG and requires a very large number of images or videos. We assume that a variety of images and videos, which are taken by anonymous users, are stored in image/video servers. In the reconstruction of the VR world, efficient classification is necessary. Therefore, metadata such as location, direction, time and weather are also necessary for efficient classification. For example, suppose that millions of images of the city of Paris are captured and stored in a media server. When we try to reconstruct a VR space in the city of Paris with those images, we must first classify the images that were taken near the Eiffel Tower to reconstruct the VR space near the Eiffel Tower. If there are images of another place (e.g., Louvre Museum), it is difficult to reconstruct a proper VR space.

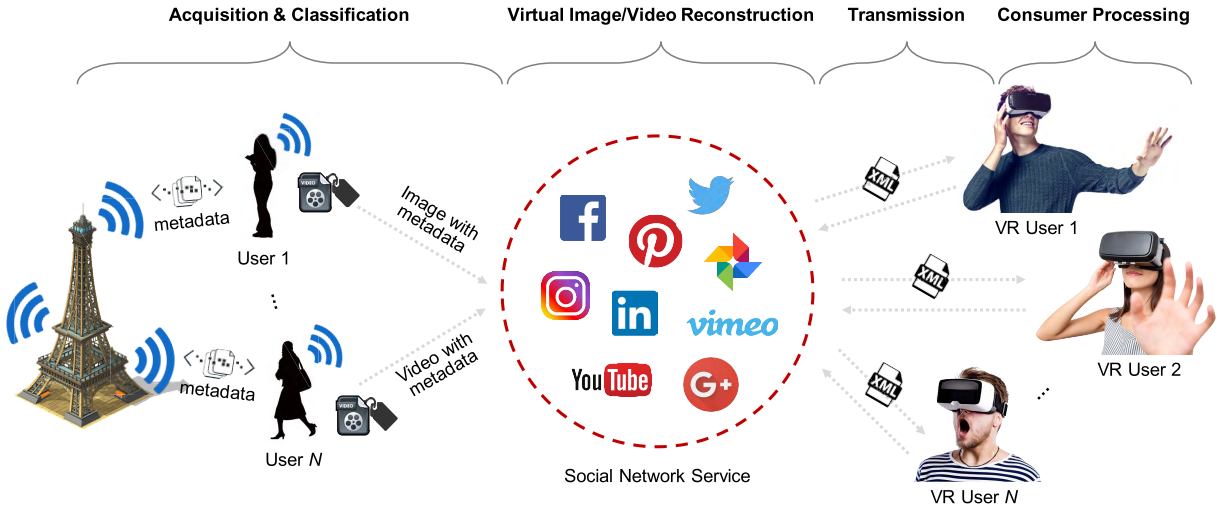


FIGURE 1. Basic concept of the proposed system architecture.

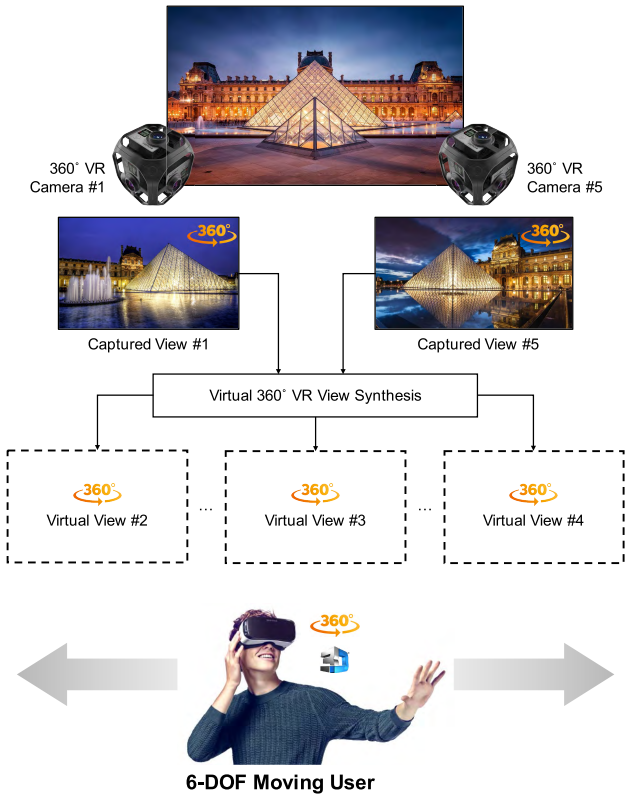


FIGURE 2. Main purpose of the proposed VR service.

1) ACQUISITION WITH METADATA

Figure 3 shows an overview of the acquisition stage, in which a user wants to take (or record) a 360° picture (or 360° video) of a specific subject. At this time, users can store basic information, such as their location with the global positioning system (GPS), image/video type, resolution, duration, focal length, and camera vendor (hereafter referred to as ''internal metadata''), without communication with the subject,

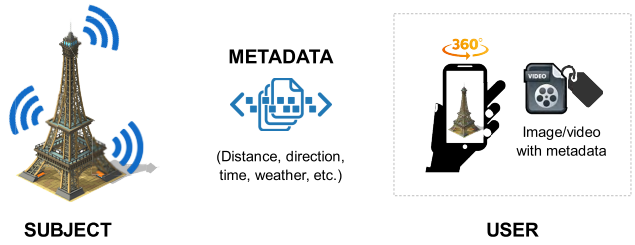


FIGURE 3. Example of the acquisition stage.

because this information can be obtained through the camera that is used. According to the studies of free-view video (FVV) and free-viewpoint TV (FTV) [22]–[24], the quality of virtually synthesized views depends on the baseline distance and angle. Thus, knowing the accurate parameters is very helpful for efficient 360° VR view synthesis. Hence, the users also obtain special information such as the relative position of the subject (e.g., baseline distance and angle) and a variety of attributes of the subject (e.g., name, year, producer), in addition to weather and time, etc. (hereafter referred to as ''external metadata'') from the subject with the help of various networks such as Wi-Fi, LTE and IoT networks. For this, the subjects of the pictures are assumed to be connected to various networks. This assumption will be satisfied in the future when the IoT infrastructure has been built. It is also assumed that 360° cameras are equipped with network function. The considered information on time and weather is also very important for high-quality virtual view synthesis because this information determines the color and brightness. For example, if an image was taken at night, it would be generally darker than an image that was taken during the day. According to [25]–[27], for seamless virtual view synthesis, the color and brightness should be properly adjusted. Therefore, this paper considers both types of information (i.e., internal + external) that were obtained in the acquisition stage.

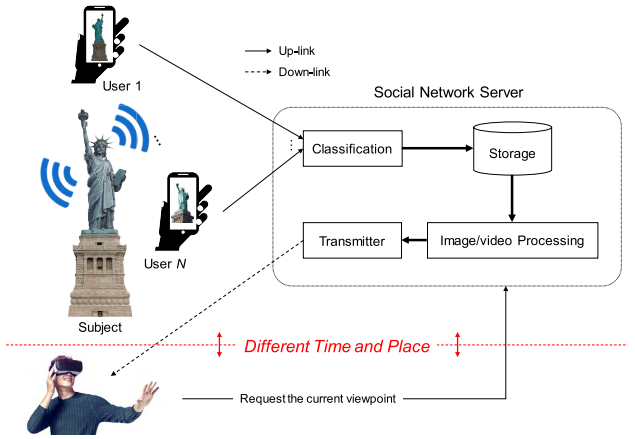


FIGURE 4. Example of the proposed end-to-end process.

2) CLASSIFICATION

The captured images/videos with the internal and external metadata are uploaded to the image/video sharing server and the server classifies the uploaded data in accordance with various methods. As shown in Figure 4, this classification process is handled before image/video processing (e.g., virtual view synthesis) in the server. However, since it is very closely related to the acquisition, we cover this process here.

As mentioned earlier, the quality of virtually synthesized views depends on the baseline distance and angle. In other words, knowing this information (i.e., the distance and angle between the subject and the camera) is very helpful for efficient 360° VR view synthesis. Unfortunately, this information can not be obtained by internal information only such as the user's GPS information. Several methods for finding the relative distance and angle between two parties have been proposed. Most of the methods can produce better results if they exchange information with one another. It is well known that Vincenty's formula is a method for calculating the distance and direction between two points (i.e., inverse problem) [28]. If the captured images/videos include both internal and external GPS information, the distance and angle between the subject and the user can be calculated by the Vincenty method, and images/videos that are captured nearby can be approximately classified in the image/video sharing server. In addition to the Vincenty method, there are many other ways to measure the distance and angle, and among them, the time-of-arrival (TOA), time-difference-of-arrival (TDOA) [29]–[31], received-signal-strength (RSS) [32]–[34], and angle-of-arrival (AOA) [35]–[37] methods are traditional and popular. In recent years, positioning methods that are based on radio cellular (e.g., LTE) such as observed time difference of arrival (OTDOA), enhanced cell ID (E-CID), and assisted global navigation satellite system (A-GNSS), have also been considered. Therefore, if the external metadata contain the required parameters for the above positioning methods and the classification is performed based on the metadata, the quality of the virtually synthesized views is expected to improve remarkably.

For seamless virtual view synthesis, it is also desirable to classify the color and brightness information, which can be easily obtained from the internal information. As mentioned earlier, time can be the biggest factor in determining the color and brightness because the sun rises in the daytime, but goes down at night. In addition, the weather information at the time of capture is very important because images/videos are bright on a clear day, but are generally dark and hazy on a cloudy day. According to [38]–[40], focal length is considered an important parameter for stitching image frames and performance improvements have been already demonstrated. However, if the colors and brightness of the images/videos that are stored in the image/video that are sharing server are random, as we considered in this paper, it is preferable to perform primary classification based on time and weather in advance. In Figure 5, we present several examples of classification sequences with metadata in which the quality of the synthesized image/video is determined. No detailed algorithms are currently proposed, but it is expected that there is an optimal classification combination in terms of quality and complexity.

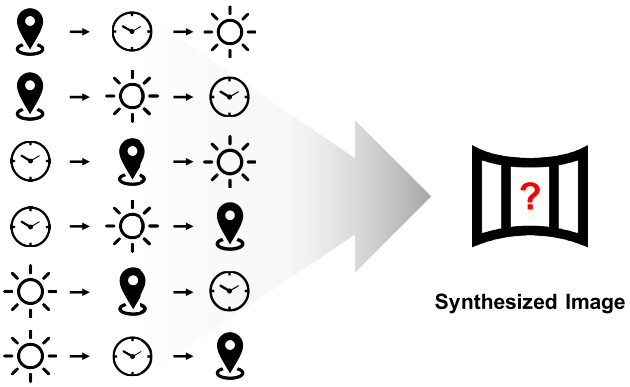


FIGURE 5. Example of a synthesized image in accordance with a classification sequence.

C. RECONSTRUCTION OF VIRTUAL IMAGE/VIDEO

As shown in Figure 4, the classified images/videos are stored in the media server. Next, these files are processed to reconstruct a novel 360° VR view. There are various methods for reconstructing the virtual view. Among them, the image stitching method is the most simple and popular. Figure 6 shows an example of the reconstruction of a virtual viewpoint, in which five 2D images are stitched by feature-based image stitching [38] and users can freely change their locations and viewpoints within the stitched image. Although it allows limited viewpoint changes, it is the easiest way to provide virtual viewpoints for VR users. Moreover, if the stored images are numerous and various, it can provide a wider range of viewpoints. It is well known that image feature detection is the first step in image stitching. The traditional feature detection methods can be categorized into four types: edge

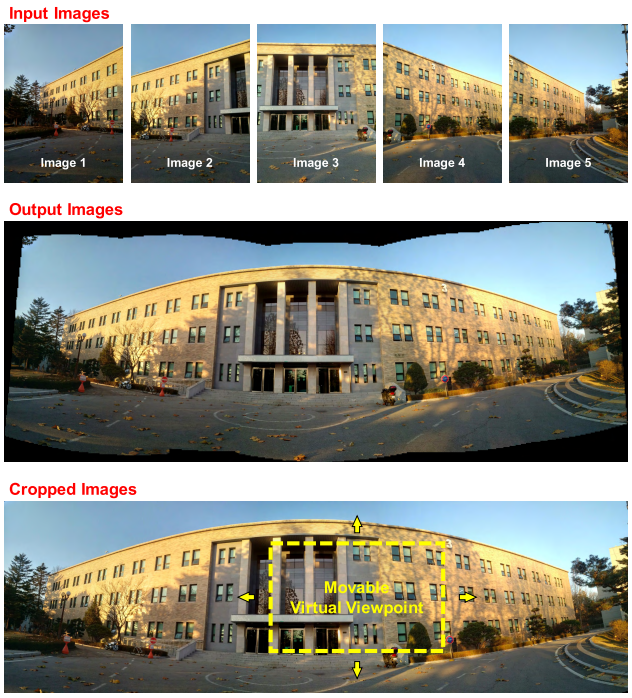


FIGURE 6. Example of feature-based image stitching for movable virtual viewpoints.

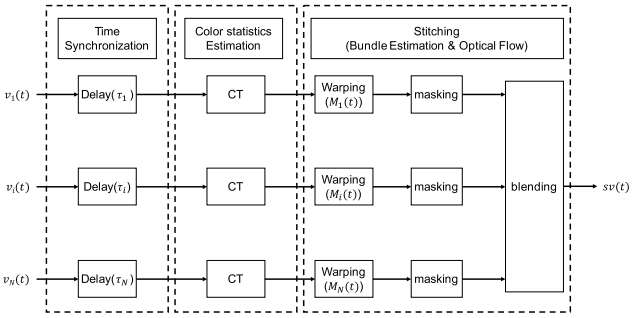


FIGURE 7. Framework for video stitching.

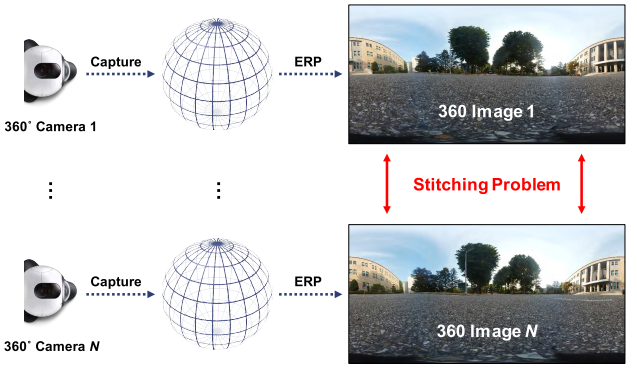


FIGURE 8. Example of the stitching problem of equirectangularly projected 360° images.

detection [41]–[43], corner detection [44]–[46], interest point detection [47]–[49], and interest region detection [50]–[52]. According to [53], the edge detection methods are related to object boundaries and are required for image interpretation tasks such as object recognition, image segmentation, visual tracking, and action analysis. In contrast, corner, interest point and region detection methods are favored in wide baseline matching and stereo analysis because the corners and blobs (i.e, interest points and regions) are unique in local image regions. Therefore, detection methods that are suitable for the services and applications to be provided may be used in the media server.

For virtual video reconstruction, we consider the video stitching method. For simplicity, we assume that a large number of video images that were captured at similar times and locations are stored in the media server. This assumption is reasonable for a sporting event or music concert. In Figure 7, the video stitching framework, which consists of time synchronization, color statistics estimation and video stitching techniques, is presented. To reconstruct a virtual video using many video images that are stored in the server, it is important to synchronize the videos with respect to time first. Among a variety of methods, we consider a coarse synchronization with metadata and fine synchronization with audio data [54]. Each video image in the media server may have been captured with different type of camera and the color statistics may be different. We consider color statistics estimation and color matching [55]. In the video stitching stage, we consider warping with homography transformation and camera stabilization, and image blending techniques.

In recent years, 360° cameras have become widespread around the world and are expected to become as popular as smartphones. The captured 360° images/videos are generally represented by simple equirectangular projection (ERP), which was invented by Marinus of Tyre in approximately AD 100 and is still used in map creation today. However, as shown in Figure 8, stitching equirectangularly projected 360° images is still challenging. This is because 360° images contain all directions, unlike traditional 2D images. Therefore, it is difficult to detect the feature points of the 360° images in all directions and match them. Furthermore, since ERP contains redundant pixels at both ends, the required data rate is increased compared to the original. Hence, in this paper, we consider cubemap projection (CMP) [56], which is considered as a solution to the data rate problem, and present another solution to the stitching problem for equirectangularly projected 360° images by using the CMP. As illustrated in Figure 9, we assume that the captured and classified 360° images are projected onto a cubemap which is a combination of the six faces of the cube (i.e., top, base, left, right, front, and back). Next, each face is separately stitched to reconstruct six stitched images, as in the conventional method, which is shown in Figure 6. The stitched six faces are remapped to the cube. However, this time, it is important to match six common points to seamlessly represent virtual 360° viewpoints. It is expected that the presented solution can sufficiently provide virtual 360° viewpoints for VR users by stitching many 360° images.

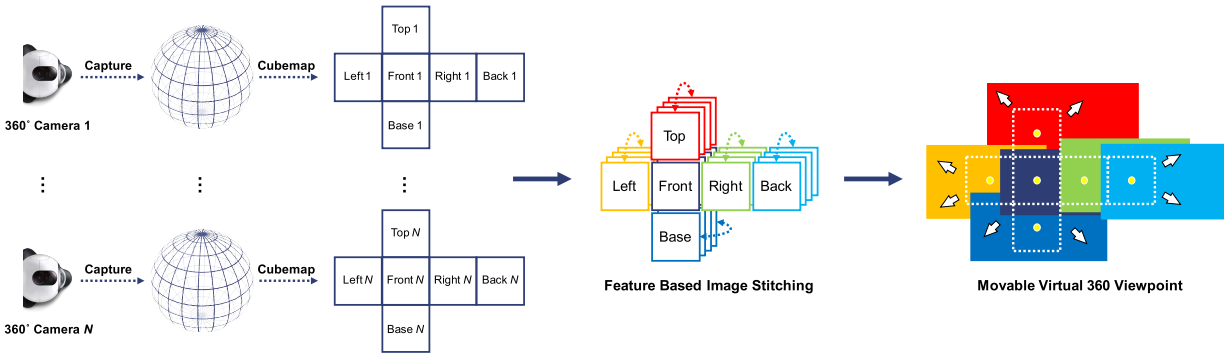


FIGURE 9. 360° image stitching using the cubemap projection and feature-based image stitching.

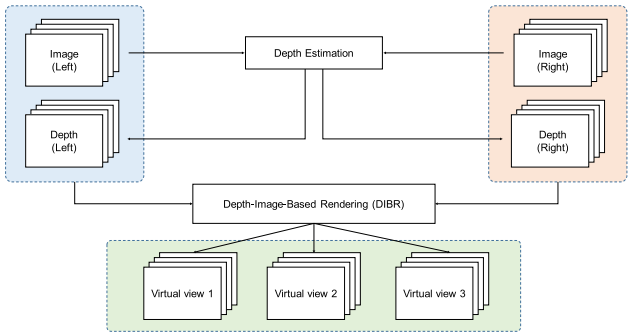


FIGURE 10. Example of depth-image-based rendering.

In addition, the media server has other roles in reconstructing a virtual space and viewpoint, such as SFM [17]–[19] and depth-image-based rendering (DIBR) [57]–[59]. Through the SFM method, 3D structures can be estimated from 2D image sequence by using the projected 2D motion field of a moving object. Since this approach also depends on the features between the captured images that are used to estimate camera poses (i.e., camera location and angle), if our proposed classification methods (i.e., methods based on internal and external metadata) are used, the quality can be remarkably improved. In the case of DIBR, as shown in Figure 10, virtual views are generated by using 2D images/videos and their associated depth information. According to [59], DIBR also uses the convergence distance and angle for the zero-parallax setting (ZPS) [60], which is an important parameter for the pre-processing of depth estimation. Therefore, our proposed classification methods can also be used to improve the DIBR performance.

D. TRANSMISSION AND CONSUMER PROCESSING

1) TRANSMISSION

It is well-known that the transmission of high-quality 360° VR image/video requires high resolution (12K, 11520×6480) and frame rate (up to 100fps). As an example, an HEVC-encoded 8K video with 60 fps requires approximately 100 Mbps [61]. Furthermore, low latency is also a very important factor for a seamless user experience, and at least a 10ms is required to achieve the seamless quality of experience (QoE) [62].

Hence, to reduce the required bandwidth, research on adaptive transmission for 360° VR images/videos has been conducted in recent years [63]–[65]. In [63], [64], Hosseini and Swaminathan proposed a novel tiling method for 360° VR video, in which six 3D meshes are mapped into a hexaface sphere 3D geometry to decrease the bandwidth requirements of 360° VR videos. Furthermore, since the tiling information (e.g., spatial relationship) can be described by the spatial representation description (SRD) [66], it is easy to apply to the conventional MPEG-DASH standard. X. Corbillon *et al.* also proposed a novel algorithm for viewport-adaptive 360° video delivery, in which the concepts of quality emphasized region (QER) and quality emphasis center (QEC) is first defined. The QER is a region of the 360° video with better quality than the original, and the QEC is the center of the QER. In this system, the users first select and request the QER in accordance with their viewports. Then, they select and request the DASH representation in accordance with their network environments. It is similar to the conventional DASH concept, but additionally provides adaptive QER-based streaming by considering the viewports of the users.

According to [67], head movement prediction is also a solution for reducing the required bandwidth because if the server knows where the user moves, a proper virtual video sequence (or tiles) can be reconstructed and provided. This study uses three prediction approaches: average, linear regression (LR) and weighted linear regression (WLR). The three approaches show satisfactory performance for short-term prediction (i.e., 0.5 sec), but only LR and WLR show improved performance for long-term prediction (i.e., from 1 sec to 2 sec), regardless of the tested 360° video sequences. Furthermore, this study shows that the proposed scenario using the prediction approaches can reduce bandwidth consumption by up to 80%. Similarly, in this paper, we propose a head movement prediction method that uses sound information, which is called sound localization information description (SLID). We assume that the SLID is described in media presentation description (MPD) with

Essential Parameter	Description
sound_R	This provides information on the right phase of the sound source in 360VR video.
sound_L	This provides information on the left phase of the sound source in 360VR video.
sound_ spatial_hori	This provides horizontal angle information to the user for sensing the sound source in 360VR video.
sound_ spatial_verti	This provides vertical angle information to the user for sensing the sound source in 360VR video.

FIGURE 11. Sound localization information description (SLID) for viewport prediction.

SRD, and consists of four parameters, as shown in Figure 11: sound_R, sound_L, sound_spatial_hori, and sound_spatial_verti. sound_R and sound_L denote the right and left output phases of the sound source in the 360° VR video, respectively, and sound_spatial_hori and sound_spatial_verti denote the horizontal and vertical angles of the 360° VR video, respectively. Based on this information (i.e., SRD and SLID), users can request both current viewport segments and the next predicted segments, as shown in Figure 12. At the next time point, if the user changes his or her viewport in accordance with the sound (e.g., $t+1$ timeslot), the prediction is successful and seamless QoE is guaranteed. However, if the user does not change his or her viewport in accordance with the sound (e.g., $t+2$ timeslot), the prediction is erroneous and seamless QoE is not guaranteed because the user should request the changed viewport segments from the server. The evaluation is discussed in Section III.

Finally, to provide a more immersive VR experience, transmission of stereoscopic 360° VR video should be considered, rather than that of mono 360° VR video. Therefore, in this paper, we also propose hybrid 360° VR transmission using scalable HEVC (SHVC). Figure 13 shows the proposed blocks for hybrid 360° VR transmission, in which high-quality stereoscopic 360° VR videos are considered. The down-sampled right and the original left 360° VR videos are used for the base and enhancement layers, respectively, and are encoded by scalable HEVC (SHVC). The base layer is provided for users in bad network environments for seamless QoE, whereas the enhancement layer is provided for users in good network environments for high-quality QoE. Moreover, the base and enhancement layers can be combined to generate stereoscopic 360° VR (i.e., 3D perception). This method enables three services through one SHVC encoder, and its detailed MPD signaling is also described in Section III.

2) CONSUMER PROCESSING

Unlike conventional video streaming, 360° VR video streaming requires the interaction between the server and the user

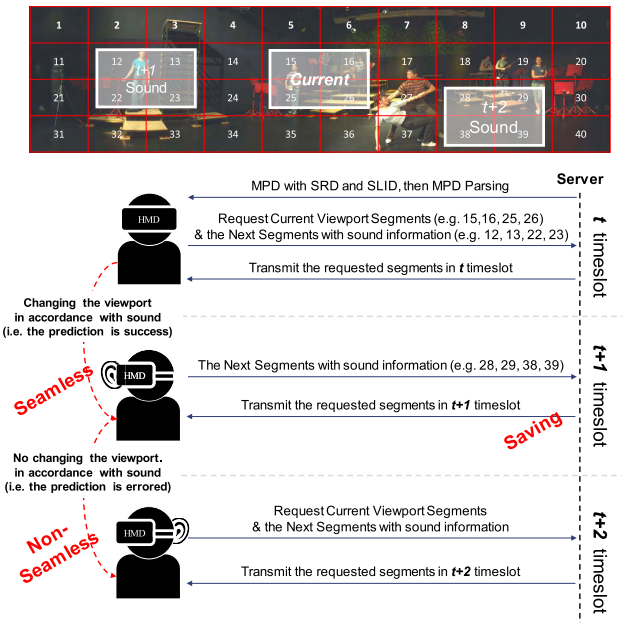


FIGURE 12. Workflow of the proposed viewport prediction method.

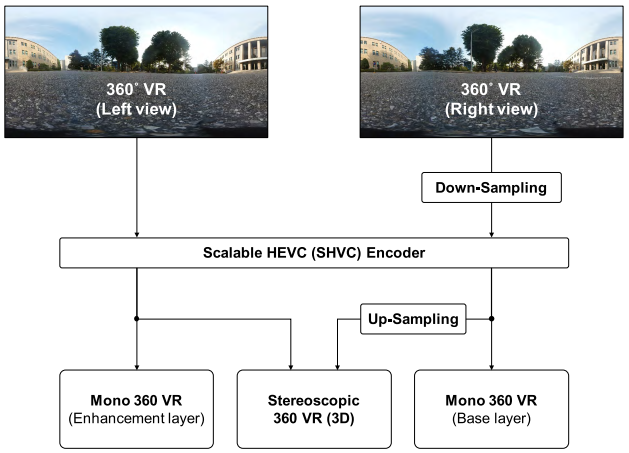


FIGURE 13. Proposed hybrid 360° VR transmission.

whenever the user moves. Then, the server transmits the requested virtual viewport segments to the user. Upon receiving a compressed video stream, it is decoded and stored in buffer(s) which is an area of memory that temporarily holds the segment data while transferring data to the player. In the recent years, using the buffer, transmission methods have been studied, and buffer-based streaming has been widely conducted. In buffer-based streaming methods, the buffer size can be used to provide seamless 360° VR video to users. In the example that is shown in Figure 14, if the buffer does not reach a specified threshold, the user requests a lower-quality video to fill the buffer faster, whereas if the buffer has exceeded the threshold, the user requests a higher-quality video. Similar concepts and algorithms are proposed in [68]–[70]. Additionally, Feuvre and Concolato of [70] proposed tile priority methods for avoiding oscillations in

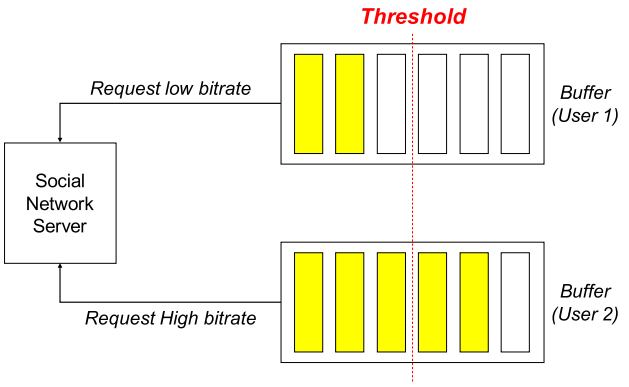


FIGURE 14. A concept of buffer-based adaptive streaming.

video quality, in which the priority is allocated per tile by the user when the tiled sets are received and identified.

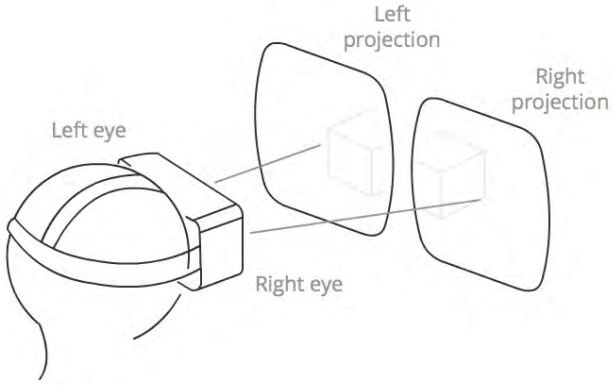


FIGURE 15. Example of projection for stereoscopic 360° VR video [71].

In addition, the receiver side is responsible for generating the proper images/videos to be displayed. According to [23], processes such as depth estimation and view synthesis can be conducted in the receiver side, but we assume that the process is only conducted in the media server to reduce the burden on the receiver. In general, monoscopic 360° VR videos are served by content providers, and displayed on VR devices of users, but cannot provide 3D perception to the users. Hence, as mentioned in Section II-D.1, providing and displaying stereoscopic 360° VR video is very important for making the VR experience more immersive. An example of projection for stereoscopic 360° VR video is illustrated in Figure 15 [71], in which left and right videos are separately projected and displayed. According to [72], since poorly implemented stereoscopic 360° VR video causes major discomfort such as headaches, eye strain and nausea, it is necessary to reduce this problem even if there are many variables and moving parts, such as in extreme sports, in terms of consumer processing.

III. CONSIDERATIONS AND EVALUATION

In this section, we consider and evaluate two issues that were described in the previous section. The first issue that we consider is MPEG-DASH MPD because the most recent research in the field of multimedia communication is related

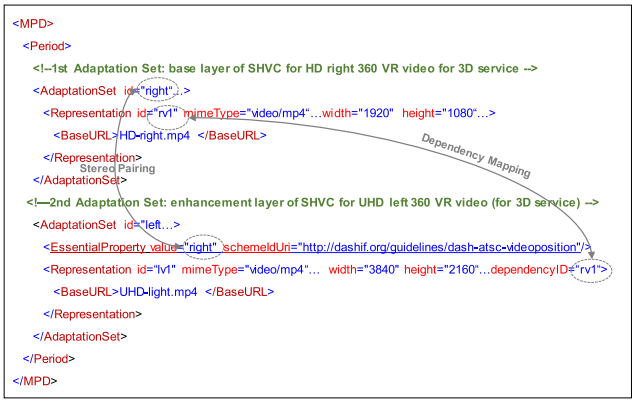


FIGURE 16. Example of MPD signaling for the proposed hybrid 360° VR transmission.

to the MPD. According to [73], the MPD contains a description of the provided content, its alternatives (e.g., the same content but of different quality), its URL addresses, and its segments, which are the transmitted bitstreams. Users first receive the MPD and parse it to receive the content by requesting their desired alternative and segment from to the media server. This series of processes is called MPD signaling. Therefore, it is very important that the characteristics of the provided content be well documented within the MPD. Figure 16 shows an example of MPD for the proposed hybrid 360° VR transmission that is shown in Figure 13. To realize this transmission, `EssentialProperty` in the second adaptation set for the left video should be first considered for the stereoscopic pairing between the left and right 360° videos. In addition, the relationship between the base and enhancement layers should be considered by `dependencyID`, as described in the second adaptation set for left video. The most important requirement here is that the value of `dependencyID` be the same as the ID of the first adaptation set for the right video. By considering SRD in conjunction with this MPD, we can improve the performance of the proposed hybrid 360° VR transmission. However we do not investigate this approach in this paper.

The next aspect to consider is the proposed viewport prediction method, which uses SLID, and is shown in Figures 11 and 12. Since the latest MPEG-DASH standard [74] extends to SRD only, it is not easy to implement the proposed prediction method using SLID with SRD within MPD. Therefore, in this paper, we consider and evaluate the feasibility of the proposed method based on the consistency value (c), which is described as follows:

$$c^k = \frac{\left|u_1^k - u_2^k\right| + \left|u_1^k - u_3^k\right| + \ldots + \left|u_{N-1}^k - u_N^k\right|}{C_{(N,2)}} \quad (1)$$

where, k denotes the DOF (yaw, pitch, roll) and c^k denotes the consistency value of DOF k and indicates the nearness of each angle among users. In other words, if the value is close to 0, the users' viewports are almost the same. u_N^k and $C_{(N,2)}$ denote the k angle value of N-th user and $N!/2!(N-2)!$, respectively. For the evaluation that is shown in Figure 17,

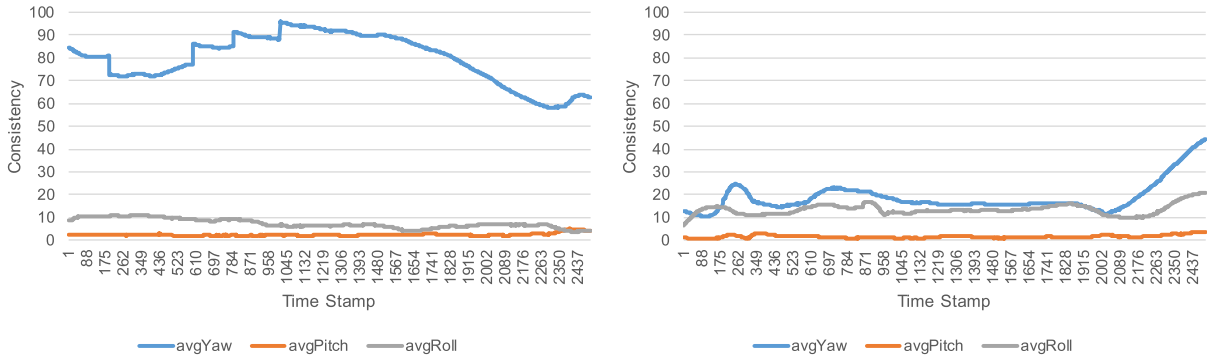


FIGURE 17. The consistence value versus time stamp, in which 1 time stamp is defined as 250 HZ.

the Mars 360° VR sequence by NASA Jet Propulsion Laboratory [75] is used. OpenTrack 2.3.9 [76] is also used to track five users' head movements. Figure 17(a) shows the consistency value as a function of the time stamp (1 time stamp := 250Hz), when the sound information is not used. The consistency value of the yaw angle is higher than the others because the users see points of view from different yaws. In contrast, if the sound information is used, the consistency value is generally low, as shown in Figure 17(b), because most users move their viewports toward where the sound is heard. Consequently, since the low consistency value means that the sound information has a substantial influence on the users' viewports, if SLID is used with SRD within MPD, it is expected to improve the performance of head movement prediction.

IV. CONCLUSIONS

In this paper, an end-to-end system architecture that underlies IoT networks is proposed for the reconstruction of seamless VR space and the proposed architecture is divided into five stages: acquisition, classification, virtual image/video reconstruction, transmission, and consumer processing. In each stage, conventional and novel element technologies are discussed and proposed, respectively, that constitute the proposed system architecture. We assume that the IoT infrastructure is already in place, and it is used to collect important data for seamless VR space reconstruction. Furthermore, we present the directions of future research that are related each stage. Among these, two proposed methods are evaluated to determine their feasibilities. In addition to the evaluated methods, it is challenging to specify and implement the methods that are proposed in this paper (e.g., Figure 5, 7 and 9). Finally, we expect that the continued integration of the VR space service into the IoT domain will cause significant transformation of social media services, thereby bringing about new business models.

REFERENCES

[1] Y. Kohsaka, K. Hashimoto, Y. Shibata, and M. Katsumoto, "Flexible multimedia lecture supporting system based on extended virtual reality space," in *Proc. Int. Workshops Parallel Process.*, Aizu-Wakamatsu, Japan, Sep. 1999, pp. 614–619.

[2] M. Slater, J. Howell, A. Steed, D.-P. Pertaub, and M. Garau, "Acting in virtual reality," in *Proc. Int. Conf. Collaborat. Virtual Environ.*, San Francisco, CA, USA, Sep. 2000, pp. 103–110.

[3] M. Nakazato and T. S. Huang, "3D MARS: Immersive virtual reality for content-based image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo*, Tokyo, Japan, Aug. 2001, pp. 44–47.

[4] B. Spear, "Virtual reality: Patent review," *World Patent Information*, vol. 24, no. 2, pp. 103–109, Jun. 2002.

[5] S. Prince *et al.*, "3D live: Real time captured content for mixed reality," in *Proc. Int. Symp. Mixed Augmented Reality*, Darmstadt, Germany, Oct. 2002, pp. 317-1–317-7.

[6] M. J. Tarr and W. H. Warren, "Virtual reality in behavioral neuroscience and beyond," *Nature Neurosci.*, vol. 5, pp. 1089–1092, Nov. 2002.

[7] M. L. Heilig, "Stereoscopic-television apparatus for individual use," U.S. Patent 2 955 156 A, Oct. 4, 1960.

[8] M. L. Heilig, "Sensorama simulator," U.S. Patent 3 050 870 A, Jan. 10, 1962.

[9] I. E. Sutherland, "The ultimate display," in *Proc. IFIPS Congr.*, vol. 62. New York, NY, USA, 1965, no. 2, pp. 506–508.

[10] I. E. Sutherland, "A head-mounted three dimensional display," in *Proc. AFIPS*, San Francisco, CA, USA, Dec. 1968, pp. 757–764.

[11] Digi-Capital.Com. (2017). *After Mixed Year, Mobile AR to Drive $108 Billion VR/AR Market by 2021*. [Online]. Available: https://www.digi-capital.com/news/2017/01/after-mixed-year-mobile-ar-to-drive-108-billion-vrar-market-by-2021/#.WkIBw99l-Um

[12] Ikea.Com. (2016). *IKEA VR Experience*. [Online]. Available: http://www.ikea.com/ms/en_US/this-is-ikea/ikea-highlights/Virtual-reality/index.html

[13] Virtual-Xperience.Com. (2015). *Virtual Xperience*. [Online]. Available: https://www.virtual-xperience.com/

[14] Edu.Google.Com. (2016). *Google Expeditions*. [Online]. Available: https://edu.google.com/expeditions/#header

[15] NextVR.Com. (2016). *NextVR*. [Online]. Available: https://www.nextvr.com/

[16] J. Huang, Z. Chen, D. Ceylan, and H. Jin, "6-DOF VR videos with a single 360-camera," in *Proc. IEEE Virtual Reality*, Los Angles, USA, Mar. 2017, pp. 37–44.

[17] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge Univ. Press, 2003.

[18] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 835–846, 2006.

[19] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. Int. Conf. 3D Vis.*, Seattle, WA, USA, Jun./Jul. 2013, pp. 127–134.

[20] Lytro.Com. (2016). *Lytro Immerge*. [Online]. Available: https://www.lytro.com/immerge

[21] T. Milliron, C. Szczupak, and O. Green, "Hallelujah: The world's first lytro VR experience," in *Proc. ACM SIGGRAPH VR Village*, Los Angeles, CA, USA, Jul./Aug. 2017, p. 7.

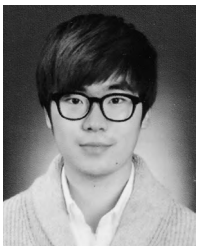
[22] K. Oh, S. Yea, and Y.-S. Ho, "Hole filling method using depth based inpainting for view synthesis in free viewpoint television and 3-D video," in *Proc. Picture Coding Symp.*, Chicago, IL, USA, May 2009, pp. 1–4.

[23] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 67–76, Jan. 2011.

[24] T. Kanade, P. Rander, and P. J. Narayanan, "Virtualized reality: Constructing virtual worlds from real scenes," *IEEE MultimediaMag.*, vol. 4, no. 1, pp. 34–47, Jan. 1997.
[25] X. Yang, J. Liu, J. Sun, X. Li, W. Liu, and Y. Gao, "DIBR based view synthesis for free-viewpoint television," in *Proc. 3DTV Conf., True Vis. Capture, Transmission Display 3D Video*, Antalya, Turkey, May 2011, pp. 1–4.
[26] J. Gautier, O. L. Meur, and C. Guillemot, "Depth-based image completion for view synthesis," in *Proc. 3DTV Conf., True Vis. Capture, Transm. Display 3D Video*, Antalya, Turkey, May 2011, pp. 1–4.
[27] S. Shimizu, H. Kimata, and Y. Ohtani, "Adaptive appearance compensated view synthesis prediction for multiview video coding," in *Proc. IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 2009, pp. 2949–2952.
[28] T. Vincenty, "Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations," *Survey Rev.*, vol. 22, no. 176, pp. 88–93, Apr. 1975.
[29] K. W. Cheung, H. C. So, W. K. Ma, and Y. T. Chan, "Least squares algorithms for time-of-arrival-based mobile location," *IEEE Trans. Signal Process.*, vol. 52, no. 4, pp. 1121–1130, Apr. 2004.
[30] Y.-T. Chan, W.-Y. Tsui, H.-C. So, and P.-C. Ching, "Time-of-arrival based localization under NLOS conditions," *IEEE Trans. Veh. Technol.*, vol. 55, no. 1, pp. 17–24, Jan. 2006.
[31] R. Kaune, "Accuracy studies for TDOA and TOA localization," in *Proc. Int. Conf. Inf. Fusion*, Singapore, Jul. 2012, pp. 408–415.
[32] K. Kaemarungsi and P. Krishnamurthy, "Properties of indoor received signal strength for WLAN location fingerprinting," in *Proc. Int. Conf. Mobile Ubiquitous Syst., Netw. Services*, Boston, MA, USA, Aug. 2004, pp. 14–23.
[33] G. Wang, H. Chen, Y. Li, and M. Jin, "On received-signal-strength based localization with unknown transmit power and path loss exponent," *IEEE Wireless Commun. Lett.*, vol. 1, no. 5, pp. 536–539, Oct. 2012.
[34] X. Li, J. Liu, Q. Yao, and J. Ma, "Efficient and consistent key extraction based on received signal strength for vehicular ad hoc networks," *IEEE Access*, vol. 5, pp. 5281–5291, Mar. 2017.
[35] Q. H. Spencer, B. D. Jeffs, M. A. Jensen, and A. L. Swindlehurst, "Modeling the statistical time and angle of arrival characteristics of an indoor multipath channel," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 347–360, Mar. 2000.
[36] P. Rong and M. L. Sichitiu, "Angle of arrival localization for wireless sensor networks," in *Proc. IEEE Commun. Soc. Sensor Ad-Hoc Commun. Netw.*, Reston, VA, USA, Sep. 2006, pp. 374–382.
[37] I. Kazemi, M. R. Moniri, and R. S. Kandovan, "Optimization of angle-of-arrival estimation via real-valued sparse representation with circular array radar," *IEEE Access*, vol. 1, pp. 404–407, Jun. 2013.
[38] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, Aug. 2007.
[39] R. Szeliski, "Image alignment and stitching: A tutoral," *Found. Trends Comput. Graph. Vis.*, vol. 2, no. 1, pp. 1–104, Jan. 2006.
[40] A. Levin, A. Zomet, S. Peleg, and Y. Weiss, "Seamless image stitching in the gradient domain," in *Proc. Eur. Conf. Pattern Recognit.*, 2004, pp. 377–389.
[41] P. Perona and J. Malik, "Detecting and localizing edges composed of steps, peaks and roofs," in *Proc. Int. Conf. Comput. Vis.*, Osaka, Japan, Dec. 1990, pp. 52–57.
[42] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.
[43] W. McIlhagga, "The Canny edge detector revisited," *Int. J. Comput. Vis.*, vol. 91, no. 3, pp. 251–261, 2011.
[44] C. Harris and M. Stephens, "A Combined corner and edge detector," in *Proc. Alvey Vis. Conf.*, 1988, pp. 147–151.
[45] S. M. Smith and J. M. Brady, "SUSAN—A new approach to low level image processing," *Int. J. Comput. Vis.*, vol. 23, no. 1, pp. 45–78, May 1997.
[46] X. Zhang, H. Wang, A. W. B. Smith, X. Ling, B. C. Lovell, and D. Yang, "Corner detection based on gradient correlation matrices of planar curves," *Pattern Recoginition*, vol. 43, no. 4, pp. 1207–1223, Apr. 2010.
[47] D, Lowe, "Distinctive image features from scala-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[48] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *Comput. Vis. Image Understading*, vol. 110, no. 3, pp. 346–359, Jun. 2008.
[49] Z. Miao and X. Jiang, "Interest point detection using rank order LoG filter," *Pattern Recognit.*, vol. 46, no. 11, pp. 2890–2901, Nov. 2013.
[50] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust widebaseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, Sep. 2004.
[51] T. Kadir, A. Zisserman, and M. Brady, "An affine invariant salient region detector," in *Proc. Eur. Conf. Pattern Recognit.*, 2004, pp. 228–241.
[52] H. Deng, W. Zhang, E. Mortensen, T. Dietterich, and L. Shapiro, "Principal curvature-based region detector for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
[53] Y. Li, S. Wang, Q. Tian, and X. Ding, "A survey of recent advances in visual feature detection," *Neurocomputing*, vol. 149, no. 3, pp. 736–751, Feb. 2015.
[54] S. Bano and A. Cavallaro, "ViComp: Composition of user-generated videos," *Multimedia Tools Appl.*, vol. 75, no. 12, pp. 7187–7210, Jun. 2016.
[55] Y. Qian, D. Liao, and J. Zhou, "Manifold alignment based color transfer for multiview image stitching," in *Proc. IEEE Int. Conf. Image Process.*, Melbourne, VIC, Australia, Sep. 2013, pp. 1341–1345.
[56] E. Kuzyakov and D. Pio. (Oct. 2015). *Under the Hood: Building 360 Video*. [Online]. Available: https://code.facebook.com/posts/1638767863078802/under-the-hood-building-360-video/
[57] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," *Proc. SPIE*, vol. 5291, pp. 93–104, May 2004.
[58] C. Fehn, "A 3D-TV approach using depth-image-based rendering (DIBR)," in *Proc. IASTED Conf. Vis., Imag. Image Process.*, Benalmadena, Spain, Sep. 2003, pp. 482–487.
[59] L. Zhang and W. J. Tam, "Stereoscopic image generation based on depth images for 3D TV," *IEEE Trans. Broadcast.*, vol. 51, no. 2, pp. 191–199, Jun. 2005.
[60] A. J. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," *Proc. SPIE*, vol. 1915, pp. 36–48, Sep. 1993.
[61] U, Pal and H. King, "Effect of UHD high frame rates (HFR) on DVB-S2 bit error rate (BER)," in *Proc. SMPTE Aust. Conf. Exhib.*, Sydney, NSW, Australia, Jul. 2015, pp. 1–11.
[62] H. Sanneck, "5G Network slicing management for challenged network scenarios," in *Proc. ACM Workshop Challenged Netw.*, Snowbird, UT, USA, Oct. 2017, pp. 1–28.
[63] M. Hosseini and V. Swaminathan, "Adaptive 360 VR video streaming: Divide and Conquer," in *Proc. IEEE Int. Symp. Multimedia*, San Jose, CA, USA, Dec. 2016, pp. 107–110.
[64] M. Hosseini and V. Swaminathan, "Adaptive 360 VR video streaming based on MPEG-DASH SRD," in *Proc. IEEE Int. Symp. Multimedia*, San Jose, CA, USA, Dec. 2016, pp. 407–408.
[65] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017, pp. 1–7.
[66] O. A. Niamut, E. Thomas, L. D'Acunto, C. Concolato, F. Denoual, and S. Y. Lim, "MPEG DASH SRD: Spatial relationship description," in *Proc. Int. Conf. Multimedia Syst.*, Klagenfurt, Austria, May 2016, p. 5.
[67] F. Qian, B. Han, L, Ji, and V. Gopalakrishnan, "Optimizing 360 video delivery over cellular networks," in *Proc. Workshop Things cellular, Oper. Applications Challenges*, New York, NY, USA, Oct. 2016, pp. 1–6.
[68] R. Skupin, Y. Sanchez, D. Podborski, C. Hellge, and T. Schierl, "HEVC tile based streaming to head mounted displays," in *Proc. IEEE Annu. Consum. Commun. Netw. Conf.*, Las Vegas, NV, USA, Jan. 2017, pp. 613–615.
[69] S. Gudumasu, E. Asbun, Y. He, and Y. Ye, "Segment scheduling method for reducing 360° video streaming latency," *Proc. SPIE Appl. Digit. Image Process. XL*, vol. 10396, p. 103960X, Sep. 2017.
[70] J. L. Feuvre and C. Concolato, "Tiled-based adaptive streaming using MPEG-DASH," in *Proc. 7th Int. Conf. Multimedia Syst.*, Klagenfurt, Austria, May 2016, p. 41.
[71] A. J. Lehrer. (Nov. 2016). *What We Need to Know About AR/VR*. [Online]. Available: https://higheredrevolution.com/what-we-need-to-know-about-ar-vr-db6b9eb9f440
[72] M. Rowell. (Sep. 2015). *Stereo vs Mono 360 Video for VR*. [Online]. Available: http://360labs.net/blog/stereo-vs-mono-360-video-vr
[73] I. Sodagar, "The MPEG-DASH standard for multimedia streaming over the Internet," *IEEE Multimedia*, vol. 18, no. 4, pp. 62–67, Apr. 2011.
[74] MPEG-DASH. *Dynamic Adaptive Streaming over HTTP (ISO/IEC 23009)*. [Online]. Available: https://mpeg.chiariglione.org/standards/mpeg-dash

[75] NASA Jet Propulsion Laboratory. (2016). *NASA's Curiosity Mars Rover at Namib Dune (360 View)*. [Online]. Available: https://www.youtube.com/watch?v=ME_T4B1rxCg
[76] S. Halik. (2017). *OpenTrack 2.3.9*. [Online]. Available: https://github.com/opentrack/opentrack/releases

DONGHO YOU (S'14) received the B.S. and M.S. degrees from the Seoul National University of Science and Technology, South Korea, in 2012 and 2014, respectively, all in media IT engineering, where he is currently pursuing the Ph.D. degree with the Department of Broadcasting and Communication Program, Graduate School of Nano IT Design Fusion. He was a Researcher with the IoT Convergence Research Center, Korea Electronics Technology Institute, in 2014. His current research interests include distributed algorithms for robust and reliable multimedia communications.



BONG-SEOK SEO received the B.S. degree in electronics and media IT engineering from the Seoul National University of Science and Technology, South Korea, in 2017, where he is currently pursuing the M.S. degree with the Department of Broadcasting and Communication Program, Graduate School of Nano IT Design Fusion. His current research interests include communication for virtual reality and augmented reality.

EUNYOUNG JEONG received the B.S. degree in electronics and media IT engineering from the Seoul National University of Science and Technology, South Korea, in 2017, where she is currently pursuing the M.S. degree with the Department of Electronic and IT Media Engineering. Her current research interests include improving accuracy of view-port prediction for guarantee of quality of experience.

DONG HO KIM (M'04–SM'13) received the B.S. degree from Yonsei University, South Korea, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, South Korea, in 1999 and 2004, respectively. He was with the 4G Wireless Technology Laboratory, Samsung Advanced Institute of Technology from 2004 to 2006, and the Mobile Communications Research Institute, Samsung Electronics, from 2006 to 2007. Since 2007, he has been with the Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, South Korea. His current research interests include convergence of broadcasting and mobile communication: joint source-channel coding, multimedia signal processing, and all aspects of OFDM-MIMO systems.