IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 30, 2021, accepted June 24, 2021, date of publication July 7, 2021, date of current version July 15, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3095284*

# Analysis of Object-Centric Visual Preference in 360-Degree Videos

**MIN-SEOK LEE[1], SEOK HO BAEK[2], YOO-JEONG SHIM[2], AND MYEONG-JIN LEE[2,3], (Member, IEEE)**

[1]EO/IR System Research and Development, LIG Nex1, Giheung-gu Yongin-si, Gyeonggi-do 16911, Republic of Korea
[2]School of Electronics and Information Engineering, Korea Aerospace University, Goyang-si, Gyeonggi-do 10540, Republic of Korea
[3]Department of Smart Drone Convergence, Korea Aerospace University, Goyang-si, Gyeonggi-do 10540, Republic of Korea

Corresponding author: Myeong-jin Lee (artistic@kau.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant through the Korean Government [Ministry of Science and ICT (MSIT)] under Grant NRF-2018R1D1A1B07050603, and in part by the Gyeonggi-do Regional Research Center (GRRC) Program of Gyeonggi Province under Grant GRRC-KAU2021-B01.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Public Institutional Review Board designated by the Ministry of Health and Welfare of the Korean Government under Approval No. P01-201908-13-002.

**ABSTRACT** To provide information for 360-degree visual space exploration, we design experiments to measure and analyze object-centric visual preference. After defining the static and dynamic properties of the objects of interest, we collect real-shot 360-degree videos and synthesize computer-generated 360-degree videos so that the objects have different combinations of static and dynamic properties. From head movement trajectories of subjects wearing head-mounted displays and watching 360-degree videos, we compare visual preference between objects with different static and dynamic properties. The experimental results indicate that subjects have visual preference for certain static and dynamic properties of objects over others; with this knowledge we can construct visually salient viewports by detecting and comparing static and dynamic properties of objects in a 360-degree video.

**INDEX TERMS** 360-degree video, visual attention, visual preference, static property, dynamic property.

## I. INTRODUCTION

Recently, realistic media services such as virtual reality, augmented reality, and 360-degree videos are drawing attention as state-of-the-art services in the 5G era. Realistic media refers to media that maximizes realism and immersion. It is being used in various industries such as entertainment, education, training, commerce, and health care to create new types of services. The core of the realistic media, 360-degree videos, is the stitching of videos taken simultaneously in all directions using multiple cameras with wide angles. Such 360-degree videos are rapidly being developed in various fields, including content production, video rendering platforms, and video imaging equipment.

Since 360-degree videos can be viewed from all directions, viewing methods are also changing to Head-Mounted Displays (HMDs) or mobile handsets. Users watching conventional videos with normal field-of-view (FoV) are free from the burden of choosing viewports because these videos deploy stories to the intent of video directors or producers on fixed screens. However, 360-degree videos feature realism and immersion and tell stories by user-driven video space exploration. Because 360-degree videos can have time-varying contents, the value of the contents can be increased if viewers can explore the 360-degree visual space that they are interested in at a specific time.

However, some papers have presented negative findings on HMD-based 360-degree video viewing and have called for improvement [1], [2]. If a network cannot meet the high bandwidth demand during 360-degree video playback, the viewport response speed in response to head movement may slow down, resulting in simulation sickness (SS). In addition, the following hypothesis has been verified: When the use of an HMD completely obscures the surrounding field of vision, motion sickness is more common than when the surrounding area is partially visible [1]. Although an HMD provides an excellent sense of immersion for studying people's 360-degree video viewing patterns, it can cause simulation sickness, physical discomfort, and high cognitive burden [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaogang Jin.

While various playback environments of 360-degree videos are likely to coexist, many users are believed to prefer viewing over flat screens, except for short-length video applications in particular fields, due to the aforementioned limitations in the HMD playback environment. Besides, 360-degree video users have to choose which direction to watch, but if the camera moves or fast storytelling takes place, they may miss critical or exciting events, leading to user confusion. Therefore, a service study for 360-degree videos is needed to automatically select or recommend viewports on mobile or large flat displays rather than an HMD, depending on the visual preference of the general public or an individual user.

Some studies produced visual saliency maps in a way that statistically measures or predicts visual saliency in classical images or videos [3]–[10], [13]. Research societies have provided head or eye movement datasets and reference visual saliency maps for visual saliency study of 360-degree image/video [3]–[7]. Several research teams proposed prediction methods of visual saliency for 360-degree images and videos [8]–[10]. A visual saliency map represents which pixels or areas are more visually attractive, and can be used to optimize a source coding algorithm, reduce bandwidth for viewport rendering, protect important areas during transmission, and enhance quality of experience to viewers for 360-degree video services. A visual saliency map is the statistical figure for visual attention not the explicit fixation or preference for specific objects or regions. Therefore, it is limited to use for constructing a viewport among multiple visually salient regions far from each other in 360-degree videos. It does not reflect the tendency of a group or an individual user to prefer a specific color, object type, or event existing in 360-degree videos.

Foreground objects or events with these objects can be the visual targets for watching, i.e. the viewport construction. No visual saliency study has considered high-level abstraction of scenes, especially the object-level visual preference by users. Although the static and dynamic properties of objects, such as type, color, and movement, may impact visual saliency, a visual saliency map cannot provide information on specific properties of objects in reverse.

The field-of-view of the 360-degree video is beyond the human eye, which has increased the burden of viewers' exploration in 360-degree visual space. In the case of 360-degree images with sufficient viewing time, the viewer's exploration in 360-degree image space may not be burdensome. However, the viewer's burden for exploring the 360-degree videos increases proportionally with the speed of change in contents, such as the number of events, the pace of events, or environmental changes over time in the 360-degree visual space. Therefore, if high-level information or visual preference for objects could be provided, it would help users to efficiently explore 360-degree videos or service providers to generate viewports for specific groups or users.

In this study, we measured and analyzed the visual preference for objects and events, according to the properties of the objects and events, to provide 360-degree video exploration information. We defined an object's static and dynamic properties by analyzing the objects and events in existing 360-degree videos. 360-degree videos were collected from the Internet or created by computer graphics tools to measure the visual saliency between different properties or different property values within the same property of objects. When subjects watch these 360-degree videos, their head trajectories for viewport change are measured and used to determine the visual preference. Finally, we analyzed the visual preference for objects of static and dynamic properties through intra-static, inter-static, and dynamic property comparison.

This paper is organized as follows. Section II explains the existing studies related to a viewer's behavior measurement and analysis of 360-degree image/video and the studies related to visual saliency prediction. Section III defines the static and dynamic properties of objects in the 360-degree visual space and proposes an experimental method and a list of experimental contents designed for visual preference measurement. Section IV presents the comparison results of visual preference for intra-static, inter-static, and dynamic properties of objects. Finally, section V concludes this paper.

## II. RELATED WORKS

Visual saliency prediction aims to estimate the areas of an image that attract the attention of people. It is useful for applications such as content-aware image/video compression and transmission, object and motion detection, and image/video search. The positions from which human observers watch images/videos are often used as the ground truth of image/video saliency. The computational model that estimates saliency value at each pixel of the image/video is called the saliency model.

### A. VIEWER BEHAVIOR DATASETS FOR 360-DEGREE IMAGES/VIDEOS

Several studies have helped the 360-degree research community to study users' exploration behavior as they watch omnidirectional images or videos [3]–[7]. These studies presented datasets consisting of 360-degree images or videos, their associated head and eye trajectories, scan paths, and visual saliency maps.

Rai *et al.* presented a dataset of sixty 360-degree images with the associated head-eye tracking data recorded from 63 viewers [3]. The authors examined the variation of exploration strategies with time and the viewer's expertise and the effect of eye-movement within the viewport. They found several differences in the final head-eye saliency map compared to the head-movement only saliency map.

Lo *et al.* presented datasets consisting of 360-degree videos, the associated sensor data, the saliency maps, and the motion maps for possible use in the design of 360-degree video systems and algorithms [4]. The authors collected ten 360-degree videos from the Internet consisting of computer-generated (CG) and natural video sequences. They measured head position and orientation from 50 viewers

wearing HMDs and generated saliency maps and motion maps from the 360-degree videos using the Convolutional Neural Network (CNN)-based method in [14] and the optical flow based method in [15].

David *et al.* presented a dataset of head and eye tracking data from 57 observers for research on exploring behaviors with 360-degree videos [5]. The authors classified nineteen 360-degree videos gathered from YouTube into several scene categories, such as indoor/outdoor, urban/rural, people faces, and water to indicate some high-level properties. They compared the head-only and head-and-eye saliency maps by normalizing the saliency maps and argued that the differences are from the information lost in head-only saliency maps, but that they can be simplistically modeled by a larger Gaussian kernel during the creation of the saliency maps. They argued that the saliency map difference caused by different longitudinal starting positions disappears after 5 seconds of exploration.

Corbillon *et al.* presented a head movement dataset sampled from 59 users watching five 70-sec long 360-degree videos on an HMD [6]. The authors extracted statistics from the dataset to provide information on the users' behavior and the video characteristics for a viewport adaptive streaming scenario.

Fremerey *et al.* recorded a dataset for twenty 360-degree videos with 48 participants wearing VR headsets and analyzed head rotation and exploration over time [7]. More than half of the subjects explored more than 330 degrees of the horizontal area of the videos. Almost 90% of the subjects were comfortable within a pitching angle of up to 100 degrees. However, participants spent 90% and 40% of time on areas from -30 to 30 degrees of pitch and yaw, respectively. Hence, almost half of the time, people keep watching the 360-degree video in the initial position and do not explore the video further.

### B. VISUAL SALIENCY PREDICTION FOR 360-DEGREE IMAGES/VIDEOS

Several studies have recently been performed to predict visual saliency maps for 360-degree images and videos [8]–[12]. These studies mostly apply or modify conventional visual saliency prediction algorithms such as boolean map based saliency (BMS), operational block description length (OBDL), or saliency in context (SALICON) models for 360-degree videos [16]–[18].

Monroy *et al.* presented an architecture extended from the CNN to refine traditional 2D saliency prediction for 360-degree images [8]. The authors subdivided the 360-degree images into undistorted patches and predicted these patches' saliency by providing the spherical coordinates of each pixel in the patches to the CNN. Assens *et al.* proposed a deep neural network, SaltiNet, to predict the scan path on 360-degree images. The network is based on a temporal-aware representation of saliency information named saliency volume [9]. Lebreton *et al.* addressed the problem of extending 2D saliency prediction algorithms to support 360-degree images [10]. Djemai *et al.* proposed a framework to apply any 2D saliency prediction method to 360-degree images [11]. 2D saliency models are used for generating saliency maps for faces in a cube map projection (CMP) 360-degree image format, and the CNN merges them to obtain one saliency map. Based on the observation that the moving objects draw more visual attention, Jiang *et al.* designed an object-to-motion CNN and a saliency structured convolutional Long Short Term Memory (LSTM) network to generate saliency maps for 360-degree videos [12].

Especially for videos, conventional studies on saliency prediction only encoded each video pixel's probability of capturing the user's visual attention. These studies did not provide information on the order in which these pixels are scanned and the duration of the fixation. The generated or predicted visual saliency map may have information on salient objects or events at the pixel level and can be used for efficient video compression or streaming of 360-degree videos. However, in other 360-degree applications such as automatic viewport generation or multiple story-telling, these pixel-level saliency maps are limited because they do not have the high-level cues that draw visual attention, such as foreground objects and events that occur in 360-degree videos.

With the era of virtual reality with 360-degree video, recent studies have called attention to the need for order in which pixels may be scanned or the fixation duration. Predicting the navigation pattern that people follow in 360-degree visual space can be more accurate if high-level cues can be used in addition to the pixel or region based saliency.

### C. AUTOMATIC VIEWPORT GENERATION

Studies are currently underway to automatically select and play the viewports of 360-degree videos. This frees viewers of the burden of exploring the 360-degree visual space and provides one or multiple videos with a regular field-of-view for rendering on normal displays.

Su *et al.* defined the Pano2Vid problem to automatically control the pose and motion of a virtual narrow-field-of-view (NFOV) camera within a 360-degree video [19]. In [20], the authors generalized the task of Pano2Vid to allow changes in the FOV in addition to the spatial selections within the 360-degree video. They also presented a coarse-to-fine search approach that iteratively refines the camera control while reducing the search space in each iteration. Hu *et al.* proposed a deep learning-based agent that navigates a 360-degree sports video and smoothly captures interesting moments [21]. The authors used an object detector to identify candidate objects of interest, recurrent neural networks (RNNs) to select the main object among the candidates and predict the viewing angle. Xu *et al.* proposed a deep reinforcement learning (DRL) based head movement prediction (DHP) approach with offline and online versions [22]. In the offline DHP, an HM heatmap is generated using the HM positions determined by multiple DRL workflows at each 360-degree video frame. In the online DHP, the next HM position of a subject is estimated given the current HM position. Wang *et al.* proposed an attention-based deep

reinforcement learning approach which incorporates saliency detection to select NFOVs in a 360-degree video [23]. The authors proposed a new reward function for the DRL framework which considers the saliency values, ground truth and smooth transition for NFOV selection.

These existing studies took different approaches for automatic viewport generation; most of them learned viewports from NFOV videos [19], [20], main objects [21], and saliency [22], [23]. It is still challenging to predict viewports online for a generic panoramic video, which may include multiple competing salient objects. None of them studied which object or regions of objects should be selected for viewports in the event of competing multiple salient objects.

## III. VISUAL PREFERENCE EXPLORATION FOR STATIC AND DYNAMIC PROPERTIES OF OBJECTS

### A. MOTIVATION

Fremerey *et al.*'s experiment [7] found that almost half of the time people keep watching the 360-degree videos around the initial position and do not explore the video further. On the other hand, the viewers' behavior when watching limited areas within a 360-degree visual space can be interpreted as viewers continuing to look at the area they first saw because they do not know where else to look. For example, when some people go to a festival or event site full of fun things to see, they may sometimes get information from accompanying friends looking at something other than what they are looking at. Sometimes their friends' recommendations may be more interesting than what they have seen, so they may change their visual target to these.

In other words, the field-of-view of 360-degree videos is too wide for viewers to judge where to look. It would be possible to efficiently explore the 360-degree visual space if more high-level abstraction of viewers' visual preferences were provided.

### B. STATIC AND DYNAMIC PROPERTIES OF OBJECTS

In this study, we propose to provide high-level abstraction of viewers' visual preference in the forms of events to help select viewports based on events in 360-degree visual spaces. For event-oriented viewport selection in 360-degree videos, it is necessary to analyze which event has a high visual preference in various in-video events that coincide. Events in 360-degree videos are defined as acts caused by a single object or multiple objects, including standstills. Each object in events has static and dynamic properties. An object's static properties do not change over time, such as the type, the sex, and the color of the object. An object's dynamic properties represent a type of the object's movement over time, such as appearing, disappearing, approaching, receding, etc.

Existing studies [3]–[6] classified 360-degree videos with high abstraction levels, such as indoor, outdoor, people-containing scenes, graphic contents, and real-shot contents, which cannot measure the amount of visual attention on a particular object appearing in the 360-degree videos. To investigate which properties of objects draw more visual attention by analyzing the available real-shot or CG 360-degree videos from the Internet including foreground objects, we defined the static and dynamic properties of objects in the video frames as given in Table 1. When describing scenes in analyzed videos, nouns, adjectives, and verbs are usually used for describing foreground objects in the scenes. We classified these parts of speech into static and dynamic property groups. Associations and independence among these parts of speech within each property group are analyzed to establish final properties and property values. Methods of describing an object's behavior can define myriad dynamic properties depending on what level of abstraction it should be. Thus, we defined the objects' movement methods in videos with respect to a camera or a specific object as dynamic properties. A sequence of defined dynamic properties can represent an object's high-level behavior. For 360-degree videos without foreground objects, few changes were made in the video, or only changes were present due to the camera's movement. We did not consider these videos because it was unlikely that the viewer will miss the time-varying event in the video.

**TABLE 1. Static and dynamic properties of objects.**

(a) static properties

| property | property values |
|---|---|
| crowd | single, group (three) |
| sex | male, female |
| color | R, G, B |
| type | person, vehicle, animal |

(b) dynamic properties

| property | property values |
|---|---|
| activity | standstill, loiter |
| proximity | approach, recede |
| existence | appear, disappear |
| density | gather, scatter |

In existing 360-degree videos, various objects, such as people, cars, airplanes, and animals, appear and generate various events, such as conversation, dance, music performance, exercise, driving, and flying. Each static property of objects has property values: the size of a crowd has 1 or 3, type of colors has R, G, or B, type of objects has people, vehicle, or animal, and sex has male or female for human objects. We classified the dynamic properties of objects into four categories: activity, proximity, existence, and density. Each dynamic property had the property values standstill and loiter, approach and recede, appear and disappear, gather and scatter.

### C. VISUAL PREFERENCE MEASUREMENT FOR DIFFERENT OBJECT PROPERTIES

For the analysis of viewers' viewing patterns for the properties of objects appearing in 360-degree videos, we designed experiments of visual preference measurement consisting of intra-static, inter-static, dynamic property comparison. Intra-static property comparison measures the visual preference for property value changes within an object's static property. Inter-static property comparison measures the visual preference across the object's static properties. Dynamic property comparison measures the visual preference between two dynamic property values. For these experiments, we collected real-shot 360-degree videos and synthesized CG 360-degree videos. Some examples are

**FIGURE 1.** Real-shot and computer-generated 360-degree video sequences. Real-shot: the 1st and the 2nd columns, CG: the 3rd column.

presented in Fig. 1, including multiple objects with static and dynamic properties as defined in Table 1.

Real-shot 360-degree videos with different background complexities were collected with the following criteria: First, 360-degree videos with little background change were collected to focus only on comparing objects' properties in the videos. Second, videos of two or more objects were collected to compare which objects' properties were more visually attractive. Third, video clips with no scene change were collected to measure successive changes in visual saliency. Fourth, high stitching quality videos were collected for viewers' immersion. Fifth, a crowd of people at a distance was considered part of the background. The final real-shot 360-degree videos collected totaled 30, as shown in Table 2, with a frame rate of 30 Hz, resolutions of 2880 $\times$ 1440, 3840 $\times$ 1920, 3840 $\times$ 2160, and lengths of 17-45 seconds.

Computer-generated 360-degree videos were created to measure which objects' properties draw more visual attention than others. Eight 360-degree test videos for intra-static property comparison, twenty-four inter-static property comparison, and twenty-eight for dynamic property comparison were produced. Object 1 and object 2 with different static and dynamic properties are located on the left and right sides of the center of the initial viewport, respectively, to receive visual selection by the subject's head movement.

To determine whether the background complexity affects the visual attention on the object's static properties, we graphically synthesized test videos with simple and complicated backgrounds for the same foreground objects in intra-/inter-static property experiments. We created simple background 360-degree videos with objects with different static properties in the default empty background. We also created complicated background 360-degree videos by combining these simple background videos with a complex background. To create CG 360-degree videos, we used a 3D graphic authoring tool that can configure 3D visual space and set properties of 3D object models.

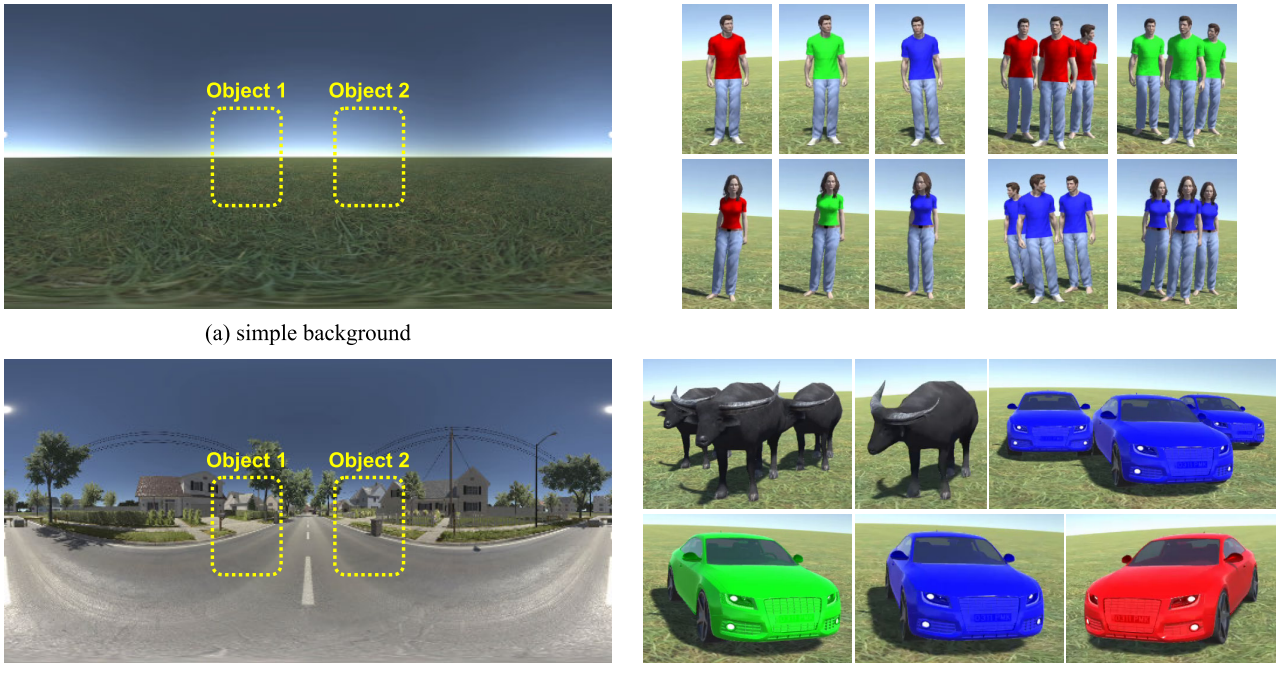
Viewers' visual fixation in 360-degree videos occurs more at the equator and has a uniform distribution in the longitudinal direction [3]. Thus, in this study, the objects were initially placed at 0 degree of latitude and $\pm$ 25 degrees of longitude, as shown in Figs. 2(a) and (b), so that viewers would pay attention only to object properties, excluding equatorial centric bias during their visual preference measurement.

Ninety-two CG 360-degree videos were created, as shown in Table 3. Their frame rate, resolution, and lengths were 30 Hz, 4096 $\times$ 2048, and 5 or 10 seconds, respectively, depending on the object's dynamic properties.

CG01b-CG08b sequences were created for comparison of intra-static properties. In this case, static properties other than

**TABLE 2.** The list of real-shot 360-degree video sequences.

| seq. | length (sec) | resolution | static property | | | | dynamic property | source |
|---|---|---|---|---|---|---|---|---|
| | | | crowd | sex | color | type | | |
| RS01 | 23 | 3840x2160 | x | x | rgb | person, animal | approach, recede, loiter | [22] |
| RS02 | 31 | 3840x1920 | x | x | rgb | person | gather, loiter | [22] |
| RS03 | 25 | 2880x1440 | x | x | rgb | vehicle | approach, recede, loiter | [22] |
| RS04 | 26 | 3840x1920 | o | o | rgb | person | loiter, gather, scatter, approach, recede, standstill | [22] |
| RS05 | 20 | 3840x1920 | x | o | rgb | person | standstill | [22] |
| RS06 | 20 | 3840x2160 | x | o | rgb | person, animal | standstill, approach | [22] |
| RS07 | 20 | 3840x1920 | o | o | rgb | person, vehicle | approach, recede, standstill, loiter | [22] |
| RS08 | 30 | 3840x1920 | o | o | rgb | person | appear, disappear, approach, recede, standstill | [22] |
| RS09 | 45 | 3840x1920 | x | x | rgb | person | loiter | [22] |
| RS10 | 29 | 2880x1440 | x | x | rgb | animal | approach, recede, gather, scatter | [22] |
| RS11 | 30 | 3840x1920 | x | x | rgb | person, vehicle | loiter, approach, recede | [22] |
| RS12 | 20 | 3840x1920 | x | x | rgb | person | approach, recede | [22] |
| RS13 | 17 | 3840x1920 | o | o | rgb | person | standstill, gather, scatter | [22] |
| RS14 | 32 | 3840x1920 | o | x | rgb | person | gather | [22] |
| RS15 | 10 | 3840x1920 | o | x | rgb | person | standstill, loiter | [22] |
| RS16 | 20 | 3840x1920 | x | o | rgb | person | standstill | [22] |
| RS17 | 26 | 3840x1920 | o | o | rgb | person | standstill, gather, loiter | [22] |
| RS18 | 22 | 3840x1920 | o | o | rgb | person | standstill, loiter | [22] |
| RS19 | 30 | 3840x1920 | o | o | rgb | person | standstill, approach, recede, appear, disappear | [22] |
| RS20 | 20 | 3840x1920 | x | o | rgb | person | standstill | [6] |
| RS21 | 20 | 3840x1920 | x | x | rgb | person, animal | standstill, loiter | [6] |
| RS22 | 20 | 3840x1920 | x | o | rgb | person | standstill | [6] |
| RS23 | 20 | 3840x1920 | x | o | rgb | person | standstill | [6] |
| RS24 | 20 | 3840x1920 | x | o | rgb | person | standstill, loiter, approach, recede | [6] |
| RS25 | 20 | 3840x1920 | x | o | rgb | person | standstill, loiter | [6] |
| RS26 | 20 | 3840x1920 | o | o | rgb | person | standstill | [6] |
| RS27 | 20 | 3840x1920 | o | x | rgb | animal | loiter, approach, recede | [22] |
| RS28 | 20 | 3840x1920 | x | o | rgb | person, animal | standstill, approach, loiter | [6] |
| RS29 | 30 | 3840x1920 | o | x | rgb | animal | standstill, approach | [6] |
| RS30 | 30 | 3840x2048 | x | x | rgb | animal | standstill | [6] |




(a) simple background

(b) complicated background

(c) foreground objects

**FIGURE 2.** Background frames and foreground objects for CG 360-degree video.

the one being compared were set to have the same property values. CG01b is a sequence for measuring the visual preference between three people and a single person in the crowd property. CG02b is for measuring visual preference between male and female in the sex property. CG03b-CG05b are for measuring visual preference among R, G, and B colors in the color property. Each static property has at least two property values. Increasing the number of property values increases the number of property value combinations for measurement of visual attention. Therefore, only a few property values clearly distinguished in each static property were included for effective experiments.

**TABLE 3. Computer generated 360-degree video sequences.**

(a) intra-/inter-static property comparison

| seq. | properties | object 1 | object 2 |
|---|---|---|---|
| CG01b | crowd | 1 | 3 |
| CG02b | sex | male | female |
| CG03b | color | R | G |
| CG04b | color | R | B |
| CG05b | color | G | B |
| CG06b | type | person | vehicle |
| CG07b | type | person | animal |
| CG08b | type | vehicle | animal |
| CG09b | sex, crowd | male, 1 | female, 3 |
| CG10b | sex, crowd | male, 3 | female, 1 |
| CG11b | color, crowd | R, 1 | G, 3 |
| CG12b | color, crowd | R, 1 | B, 3 |
| CG13b | color, crowd | G, 1 | R, 3 |
| CG14b | color, crowd | G, 1 | B, 3 |
| CG15b | color, crowd | B, 1 | R, 3 |
| CG16b | color, crowd | B, 1 | G, 3 |
| CG17b | crowd, type | 1, person | 3, vehicle |
| CG18b | crowd, type | 1, person | 3, animal |
| CG19b | crowd, type | 1, vehicle | 3, person |
| CG20b | crowd, type | 1, vehicle | 3, animal |
| CG21b | crowd, type | 1, animal | 3, person |
| CG22b | crowd, type | 1, animal | 3, vehicle |
| CG23b | sex, color | R, male | G, female |
| CG24b | sex, color | R, male | B, female |
| CG25b | sex, color | G, male | R, female |
| CG26b | sex, color | G, male | B, female |
| CG27b | sex, color | B, male | R, female |
| CG28b | sex, color | B, male | G, female |
| CG29b | color, type | R, person | G, vehicle |
| CG30b | color, type | R, person | B, vehicle |
| CG31b | color, type | R, vehicle | B, person |
| CG32b | color, type | G, vehicle | B, person |

(b) dynamic property comparison

| seq. | object 1 | object 2 |
|---|---|---|
| CG33 | loiter | standstill |
| CG34 | approach | standstill |
| CG35 | recede | standstill |
| CG36 | appear | standstill |
| CG37 | disappear | standstill |
| CG38 | gather | standstill |
| CG39 | scatter | standstill |
| CG40 | approach | loiter |
| CG41 | recede | loiter |
| CG42 | appear | loiter |
| CG43 | disappear | loiter |
| CG44 | gather | loiter |
| CG45 | scatter | loiter |
| CG46 | recede | approach |
| CG47 | appear | approach |
| CG48 | disappear | approach |
| CG49 | gather | approach |
| CG50 | scatter | approach |
| CG51 | appear | recede |
| CG52 | disappear | recede |
| CG53 | gather | recede |
| CG54 | scatter | recede |
| CG55 | appear | recede |
| CG56 | gather | appear |
| CG57 | scatter | appear |
| CG58 | gather | disappear |
| CG59 | scatter | disappear |
| CG60 | scatter | gather |

For intra-/inter-static property comparison, each sequence's foreground object composition was combined with either a simple or a complicated background to produce two test 360-degree videos, respectively,Simple background (b=s), complicated background (b=c). For dynamic property comparison, only a simple background is used to produce test 360-degree videos.

CG09b-CG32b sequences were created for comparison of inter-static properties to measure the visual attention according to each property value change for two selected static properties. Whether one static property draws more visual attention than another can be measured by analyzing the measurement results on two compared objects having every combination of the compared static property values.

For each dynamic property, we defined two opposite property values. Although an object can have multiple static properties and values, it cannot simultaneously have multiple dynamic properties. Unlike static properties, it is possible to compare property values across dynamic properties. Thus, with two property value combinations from the dynamic properties, CG33-CG60 sequences were created for visual preference measurement.

## IV. EXPERIMENTAL RESULTS

### A. EXPERIMENTAL ENVIRONMENTS

To calculate the visual saliency map and visual preference for real-shot and CG 360-degree videos, we sampled the head movement trajectory of subjects wearing an HMD using the visual fixation collection system in Fig. 3. The HMD had a resolution of 2100 × 2100, a field of view of 110-degree, and a frame rate of 90 Hz. The system was designed with a high-performance multi-core processor and a GPU for smooth playback of 360-degree videos.

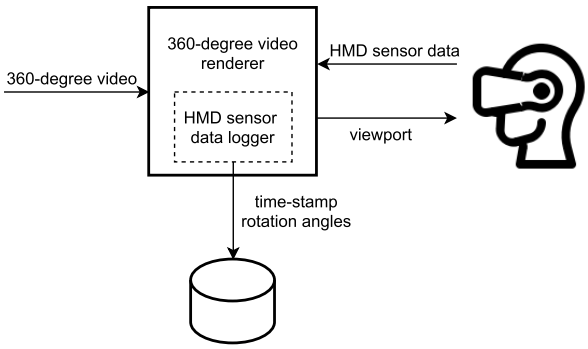



**FIGURE 3. Visual fixation collection system.**

The subjects consisted of 25 men and 25 women. The average age of this group was 24.04 years old, with an average of 1.02 VR experience. Before starting the experiment, we explained the purpose of this study and the experimental procedure to the subjects.[1]

Experiments were conducted on subjects, each designed differently for real-shot and CG 360-degree videos. The subjects were allowed to watch real-shot 360-degree videos freely. On the other hand, for CG 360-degree videos, subjects

[1]This experimental study for human subjects was approved by Public Institutional Review Board Designated by Ministry of Health and Welfare of the Korea government (P01-201908-13-002).

were guided to momentarily determine objects of greater visual preference in the videos being played and move their heads or bodies so that they could look directly at them. For a free viewing experiment, a 360-degree rotating chair was provided. A free rest was given to the subjects after watching each 360-degree video. The 360-degree videos were played without sound so that the measured data was only affected by the visual preference.

The experiment consisted of four parts: a free viewing experiment on real-shot 360-degree videos, two experiments comparing objects' intra- and inter-static properties for CG 360-degree videos with the simple and complicated backgrounds, and an experiment comparing objects' dynamic properties for CG 360-degree videos with the simple background. Each subject participated in the experiment for about 50 minutes, including a break.

### B. VISUAL FIXATION COLLECTION SYSTEM

A visual fixation collection system was constructed to compare objects' static and dynamic properties in real-shot and CG 360-degree videos. Because the HMD used for this study has no eye-tracking sensor, we regarded the sampled head pose vectors as both the centers of the viewports and the eye fixations. Therefore, in this study, we defined visual fixation as the state of a head in which the sampled head pose vectors remain at a particular visual target for more than a specified period.

Subjects' HMD sensor data are sampled while the 360-degree videos in Table 2 and 3 are played. HMD sensor data include time stamps and subjects' head pose vectors consisting of the roll, pitch, and yaw angles.

After the speed between two consecutive sampled head pose vectors in the spherical coordinate system is calculated, the visual fixation can be detected if the speed is below a threshold; the speed of the head movement is calculated by dividing the spherical distance between two samples with the sampling time interval [5] as follows.

$$v = \arccos(\mathbf{n_1} \cdot \mathbf{n_2}) \cdot f_s, \tag{1}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are normals to the unit sphere at the two sample positions 1 and 2, and $f_s$ is the sampling frequency. Each sampled position is considered the fixation if its arriving speed is less than $80°/s$ [5].

The visual fixation collection system was written in software using a 3D graphic authoring tool and an HMD's SDK. Time stamp and HMD head pose vectors were recorded in text file format at 75 Hz.

### C. VISUAL SALIENCY MAP GENERATION

Visual fixation detected from the sensor data is used to analyze the visual saliency change due to the static and dynamic properties of objects and to calculate the visual preference between two compared properties.

To visually analyze the visual saliency change, we created a visual saliency map for each 360-degree ERP frame, which is commonly used in existing studies. A visual saliency map is generated by the accumulation of per-subject fixation maps and its convolution operation with a Gaussian kernel [13].

The visual fixation map of subject $i$ for an ERP video frame is given as follows.

$$f_i(\mathbf{p}) = \sum_{k=1}^{M} \delta(\mathbf{p} - \mathbf{p}_f(k)), \tag{2}$$

where $\mathbf{p}_f(k)$ and $M$ represent the location of the $k^{th}$ fixation and the number of fixations in an ERP frame, respectively. $\delta(\mathbf{p})$ represents the Kronecker delta function.

The visual fixation map for $N$ subjects for the ERP video frame is given as follows.

$$f(\mathbf{p}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{p}). \tag{3}$$

The visual saliency map is calculated by the convolution of a two-dimensional Gaussian kernel with the visual fixation map [2].

$$S(\mathbf{p}) = f(\mathbf{p}) * G_\sigma(\mathbf{p}), \tag{4}$$

where $G_\sigma(\cdot)$ is the 2D Gaussian kernel with the standard deviation of $\sigma$. The standard deviation was set to 1 degree in our experiment.

### D. QUALITATIVE ANALYSIS ON VISUAL PREFERENCE FOR REAL-SHOT 360-DEGREE VIDEOS

To analyze the time-varying visual attention in 360-degree videos according to the static and dynamic properties of objects, we generated a visual saliency map for each ERP video frame using the method in subsection IV-C. Fig. 4 shows the visual saliency map generated for the real-shot 360-degree video sequences RS05, RS06, RS10, RS19, and RS29.

In the RS05 sequence in Fig. 4, five person objects with different static properties perform at fixed positions. All objects have the dynamic property of standstill. From the visual saliency map of the RS05 sequence, subjects initially have great visual attention to human objects located at the center of the frame. However, over time they find their preferred human objects and fixate visual targets.

In the RS06 sequence, a human object feeds two lion animal objects, with dynamic properties close to standstill. The visual saliency map of the RS06 sequence tends to focus on preferred objects, the lions, as time goes by.

In the RS10 sequence, two shark animal objects and thousands of small fish animal objects move fast. The RS10 sequence has static properties such as crowd, type, etc., and dynamic properties, such as approach, recede, gather, etc. Its visual saliency exists mainly around the movements of large sharks and the small fish group.

In the RS19 sequence, six human objects with different static properties perform dancing. The objects' actions are to appear at the door, dance, move from one another, and disappear behind the door. The objects' dynamic properties are combinations of standstill, loiter, approach, recede, appear,

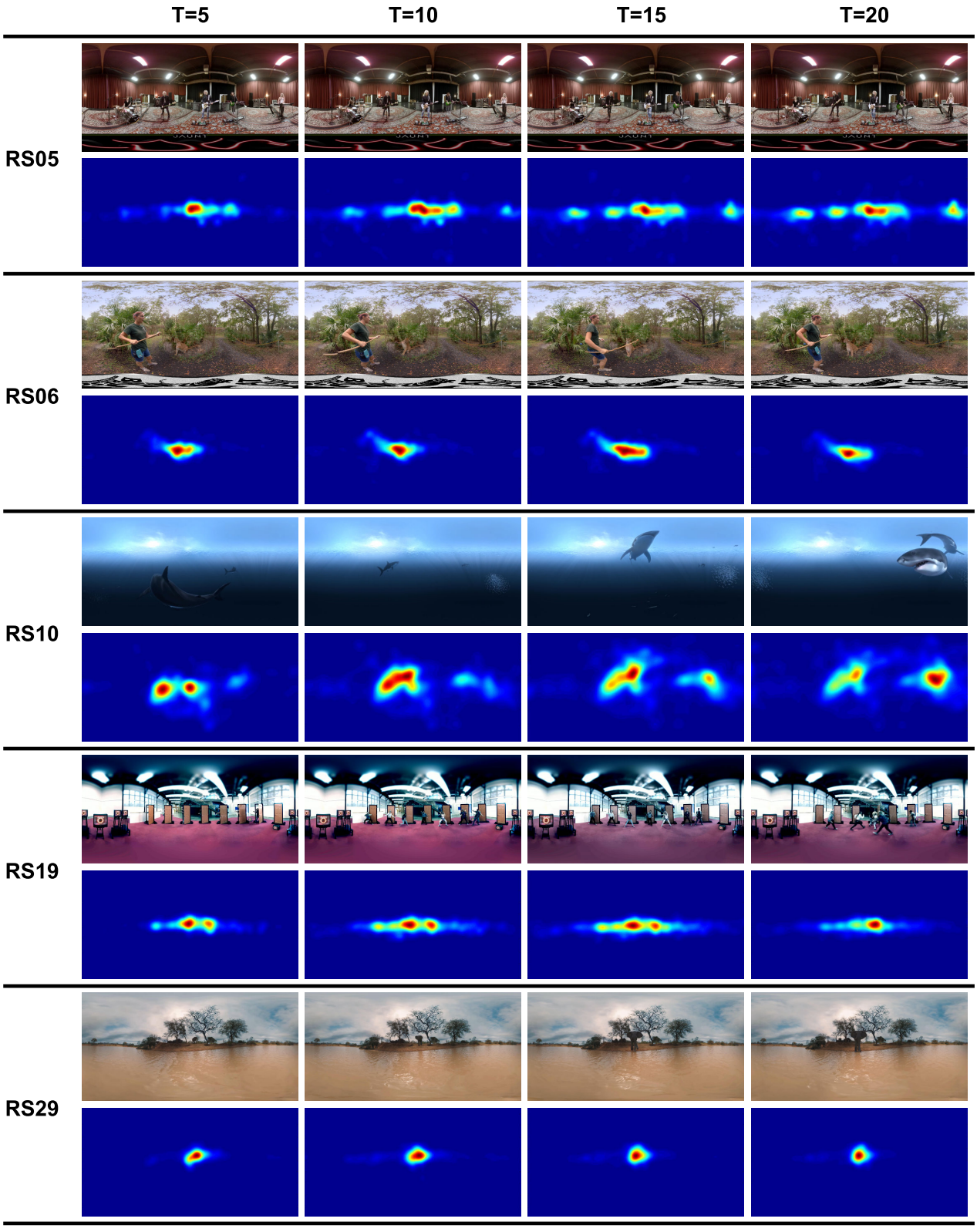



**FIGURE 4.** Saliency map for real-shot 360-degree videos.

and disappear. The approach dynamic property in which human objects move to the camera at the center of the ERP frame has relatively higher visual preference than the other dynamic properties. Thus, unlike the visual saliency map of the RS05 sequence, the visual saliency map of RS19 tends to spread less left and right and concentrates more in the

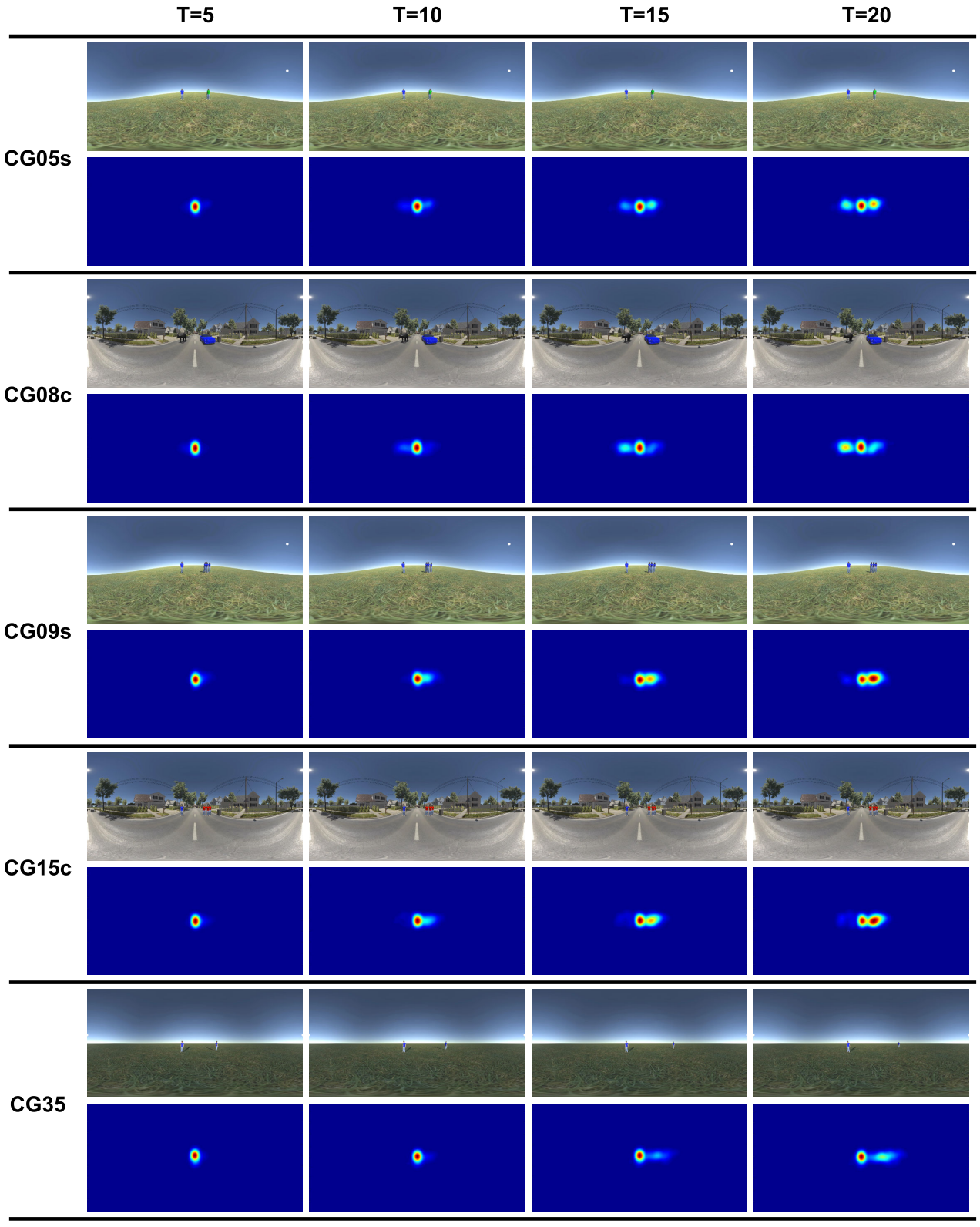



**FIGURE 5.** Saliency map for computer generated 360-degree videos.

middle. This tendency is because objects' movements mostly happened at the center gate, and some objects approached the camera for dancing.

In the RS29 sequence, dozens of elephant animal objects of the same static properties exist. While one elephant object approaches the camera, others move from left to right. The

**TABLE 4.** Visual preference comparison for intra-static property.

| seq. | object 1 vs. object 2 | selection ratio(%) |
|---|---|---|
| CG01b | 1 <3 | 21:79 |
| CG02b | male <female | 32:68 |
| CG03b | R >G | 65:35 |
| CG04b | R ≒ B | 49:51 |
| CG05b | G ≒ B | 50:50 |
| CG06b | human <vehicle | 18:81 |
| CG07b | human <animal | 29:71 |
| CG08b | vehicle <animal | 47:53 |

elephant objects far from the camera may have either single or group static properties. The visual saliency map of the RS29 sequence shows great visual attention to the elephant object approaching the camera.

In real-shot 360-degree video sequences, foreground objects are found to draw greater visual attention than background because most of the strong saliency is formed on the foreground objects. However, it is not easy to measure the visual preference between two properties because more than two objects of different static property combinations may coexist and their dynamic properties vary over time. Moreover, the objects appear far from each other, which makes it difficult for viewers to determine their visual targets. When objects do not move, visual preference is determined by their static properties. As objects move, the visual preference is formed on an object or group of objects approaching the camera in the sequences. Quantitative analysis was not possible because it was hard to find real-shot 360-degree videos containing such objects with the compared properties. This qualitative analysis is aimed to identify which parts of the ERP frame get more visual preference. This analysis can be the starting point for comparing visual attention between two properties in CG 360-degree videos.

### *E. QUANTITATIVE ANALYSIS OF VISUAL PREFERENCE FOR CG 360-DEGREE VIDEOS*

In CG 360-degree video viewing experiments, it was possible for us to directly compare which object's property affects visual attention more because subjects visually select more stimulating objects by moving their heads or bodies. This section analyzes the subjects' viewing behavior for CG 360-degree videos containing two objects with different properties for visual preference comparison.

Tables 4 to 6 show the experimental results after we compared objects' static and dynamic properties for CG 360-degree videos that summarized how object 1 with a particular combination of static properties and a dynamic property value draws more visual attention than object 2 with another combination. The comparison results are expressed using inequality or approximately equivalent symbols. We determined that two objects draw similar visual attention if the subjects' selection ratio for a particular object is within 5% of that for another object.

Each CG sequence in Table 3 is synthesized for an experiment selecting a more visually attractive object or group from the scene. So, the statistically significant level of the comparison results in Tables 4 to 6 is related to the size of the set of the subjects. Although each experiment requires a different size of the set of the subjects to guarantee the same significant level of the comparison result, we used the same size of the set for all the experiments because of the limited research resources. The comparison results of most experiments in Tables 4 to 6 are statistically significant with the current size of fifty subjects. However, in the experiments with a selection ratio of about 50%, a larger set of the subjects should be constructed for the statistical significance of the comparison results.

**TABLE 5.** Visual preference comparison for inter-static property.

| seq. | object 1 vs. object 2 | selection ratio(%) |
|---|---|---|
| CG09b | male, 1 <female, 3 | 17:83 |
| CG10b | male, 3 <female, 1 | 44:56 |
| CG11b | R, 1 <G, 3 | 47:53 |
| CG12b | R, 1 ≒ B, 3 | 51:49 |
| CG13b | G, 1 <R, 3 | 22:78 |
| CG14b | G, 1 <B, 3 | 36:64 |
| CG15b | B, 1 <R, 3 | 16:84 |
| CG16b | B, 1 <G, 3 | 29:71 |
| CG17b | 1, human <3, vehicle | 18:82 |
| CG18b | 1, human <3, animal | 15:85 |
| CG19b | 1, vehicle >3, human | 53:47 |
| CG20b | 1, vehicle <3, animal | 38:62 |
| CG21b | 1, animal >3, human | 53:47 |
| CG22b | 1, animal <3, vehicle | 37:63 |
| CG23b | R, male <G, female | 41:59 |
| CG24b | R, male ≒ B, female | 49:51 |
| CG25b | G, male <R, female | 23:77 |
| CG26b | G, male <B, female | 40:60 |
| CG27b | B, male <R, female | 16:84 |
| CG28b | B, male <G, female | 26:74 |
| CG29b | R, human <G, vehicle | 31:69 |
| CG30b | R, human ≒ B, vehicle | 52:48 |
| CG31b | R, vehicle >B, human | 91:09 |
| CG32b | G, vehicle >B, human | 79:21 |

**TABLE 6.** Visual preference comparison for dynamic property.

| seq. | object 1 vs. object 2 | selection ratio(%) |
|---|---|---|
| CG33 | move >standstill | 96:04 |
| CG34 | approach >standstill | 96:04 |
| CG35 | recede >standstill | 92:08 |
| CG36 | appear >standstill | 82:18 |
| CG37 | disappear >standstill | 88:12 |
| CG38 | gather >standstill | 92:08 |
| CG39 | disperse >standstill | 86:14 |
| CG40 | approach >move | 70:30 |
| CG41 | recede <move | 30:70 |
| CG42 | appear >move | 70:30 |
| CG43 | disappear ≒ move | 50:50 |
| CG44 | gather >move | 80:20 |
| CG45 | disperse >move | 78:22 |
| CG46 | recede <approach | 28:72 |
| CG47 | appear <approach | 32:68 |
| CG48 | disappear <approach | 44:56 |
| CG49 | gather >approach | 68:32 |
| CG50 | disperse <approach | 32:68 |
| CG51 | appear >recede | 86:14 |
| CG52 | disappear >recede | 60:40 |
| CG53 | gather >recede | 66:34 |
| CG54 | disperse >recede | 58:42 |
| CG55 | appear >disappear | 56:44 |
| CG56 | gather >appear | 82:18 |
| CG57 | disperse ≒ appear | 50:50 |
| CG58 | gather >disappear | 82:18 |
| CG59 | disperse >disappear | 56:44 |
| CG60 | disperse <gather | 26:74 |

#### 1) INTRA-STATIC PROPERTY COMPARISON

The results of measuring the visual preference for property value changes within an object's static property are shown in Table 4. A group object has higher visual preference than a single object for the crowd static property. The visual preference for a female object was higher than that for a male object in the sex static property. In the color static property, blue objects have a similar visual preference with other color objects, while red objects have nearly twice as high visual preference as green objects. In the type static property, the visual preference was measured higher in order of animal, vehicle, and human objects.

#### 2) INTER-STATIC PROPERTY COMPARISON

The results of comparing the visual preference across the object's static properties are shown in Table 5. The comparison across static properties composes object 1 and object 2 to have property values from two different static properties respectively, and we asked subjects to determine which static property affects the visual preference more. For example, from the CG09b sequence experiments, three females had higher visual preference than one male. In this case, it is not known which of the sex and crowd static properties affects the visual preference more. However, from the CG10b sequence experiments, we can infer that the sex static property affects the visual preference more than the crowd property because one female's visual preference is higher than that of three males.

The experimental results of CG09b-CG32b sequences indicated that sex over color, crowd over type, sex over color, and type over color had a higher visual preference. In summary, the visual preference was higher in order of sex, crowd, type, and color.

#### 3) DYNAMIC PROPERTY COMPARISON

It is impossible to express an object's behavior with the combination of several dynamic properties as in static properties. Thus, for the dynamic property comparison, we conducted experiments for two different property values from any dynamic properties.

The experimental results of CG33-CG39 sequences, as shown in Table 6, clearly show that objects with dynamic properties, such as loiter, approach, recede, appear, disappear, gather, and scatter, have a significantly higher visual preference than standstill objects. Moreover, the experimental results of the CG33-CG60 sequences revealed that the visual preference was higher in order of gather, approach, appear, scatter, loiter, and disappear.

### F. DISCUSSION

Real-shot 360-degree videos should only be used for experiments as given, and cannot be used for arbitrary property comparison as in CG 360-degree videos. However, we commonly obtained the following results from both real-shot and CG 360-degree video experiments:

1) First, foreground objects draw greater visual attention than background.
2) Second, objects with dynamic properties draw greater visual attention than stationary objects.
3) Third, for objects with the same static properties, each object's visual attention is determined by its dynamic property.
4) Fourth, the visual preference for foreground objects is not affected by the complexity of the background.
5) Fifth, there exist visual preference for certain static or dynamic properties over others.

It was common in both real-shot and CG 360-degree videos that viewers had greater visual preference for dynamic property values, such as gather and approach, and static property values, such as female and animal. Furthermore, the visual preference was higher in order of gather, approach, appear and scatter, loiter and disappear, and recede.

### G. PROPOSAL TO USE THE VISUAL PREFERENCE FOR 360-DEGREE VIDEO SERVICES

Existing visual saliency studies have focused on the probability that each pixel will receive visual attention and they cannot provide explicit information on which objects or regions to watch in the video. Suppose objects' static and dynamic properties can be obtained from 360-degree videos using object detection, tracking, and behavior analysis techniques. In that case, this study's results can be used to recommend viewers multiple object-centric viewports in the order drawing more visual attention for 360-degree video exploration or to generate object-centric viewports automatically for streaming services. If specific preference by groups or users on object properties, such as color, object type, or event, can be given, group- or user-specific automatic viewport generation is possible, and it will generate different viewports for groups or users from the same 360-degree videos.

## V. CONCLUSION

We measured and analyzed the visual preference for objects of different static and dynamic properties to provide 360-degree video exploration information. We defined an object's static and dynamic properties by analyzing the objects and events that appear in existing 360-degree videos. Our experiments were designed for intra-static, inter-static, and dynamic property comparison, which compared the visual preference between two objects with different static and dynamic properties in real-shot and CG 360-degree videos.

Our experimental results confirmed that foreground objects draw greater visual attention than background, and objects with dynamic properties draw more visual attention than stationary objects. In addition, by analyzing our experimental data, we determined the order of static and dynamic properties that draw more visual attention. Our results can be used to recommend viewers objects that attract more visual attention or generate object-centric viewports of normal

field-of-view if the static and dynamic properties of objects can be obtained from 360-degree videos using object detection and behavior analysis techniques.

## REFERENCES

[1] J. D. Moss and E. R. Muth, "Characteristics of head-mounted displays and their effects on simulator sickness," *Hum. Factors J. Hum. Factors Ergonom. Soc.*, vol. 53, no. 3, pp. 308–319, Jun. 2011, doi: 10.1177/0018720811405196.

[2] M. V. D. Broeck, F. Kawsar, and J. Schöning, "It's all around you: Exploring 360° video viewing experiences on mobile devices," in *Proc. 25th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA: Association for Computing Machinery, 2017, pp. 762–768, doi: 10.1145/3123266.3123347.

[3] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proc. 8th ACM Multimedia Syst. Conf. (MMSys)*, New York, NY, USA: Association for Computing Machinery, 2017, pp. 205–210, doi: 10.1145/3083187.3083218.

[4] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "360° video viewing dataset in head-mounted virtual reality," in *Proc. 8th ACM Multimedia Syst. Conf. (MMSys)*, New York, NY, USA: Association for Computing Machinery, 2017, pp. 211–216, doi: 10.1145/3083187.3083219.

[5] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, "A dataset of head and eye movements for 360° videos," in *Proc. 9th ACM Multimedia Syst. Conf. (MMSys)*, New York, NY, USA: Association for Computing Machinery, 2018, pp. 432–437, doi: 10.1145/3204949.3208139.

[6] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proc. 8th ACM Multimedia Syst. Conf. (MMSys)*, New York, NY, USA: Association for Computing Machinery, 2017, pp. 199–204, doi: 10.1145/3083187.3083215.

[7] S. Fremerey, A. Singla, K. Meseberg, and A. Raake, "Avtrack360: An open dataset and software recording people's head rotations watching 360° videos on an HMD," in *Proc. 9th ACM Multimedia Syst. Conf. (MMSys)*, New York, NY, USA: Association for Computing Machinery, 2018, pp. 403–408, doi: 10.1145/3204949.3208134

[8] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "SalNet360: Saliency maps for omni-directional images with CNN," *Signal Process., Image Commun.*, vol. 69, pp. 26–34, Nov. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596518304685

[9] M. Assens, X. Giro-i-Nieto, K. McGuinness, and N. E. O'Connor, "SaltiNet: Scan-path prediction on 360 degree images using saliency volumes," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2331–2338.

[10] P. Lebreton and A. Raake, "GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images," *Signal Process., Image Commun.*, vol. 69, pp. 69–78, Nov. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596518302406

[11] I. Djemai, S. Fezza, W. Hamidouche, and O. Deforges, "Extending 2D saliency models for head movement prediction in 360-degree images using CNN-based fusion," 2020, *arXiv:2002.09196*. [Online]. Available: http://arxiv.org/abs/2002.09196

[12] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: A deep learning based video saliency prediction approach," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 602–617.

[13] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: Strengths and weaknesses," *Behav. Res. Methods*, vol. 45, no. 1, pp. 251–266, Mar. 2013.

[14] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3488–3493.

[15] G. Bradski, "The OpenCV library," Dr. Dobb's Journal of Software Tools, 2000.

[16] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: A Boolean map approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 889–902, May 2016.

[17] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5501–5510.

[18] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 262–270.

[19] Y.-C. Su, D. Jayaraman, and K. Grauman, "Pano2Vid: Automatic cinematography for watching 360° videos," in *Computer Vision—ACCV 2016* (Lecture Notes in Computer Science), vol. 10114, S. H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham, Switzerland: Springer, 2016, pp. 154–171, doi: 10.1007/978-3-319-54190-7_10.

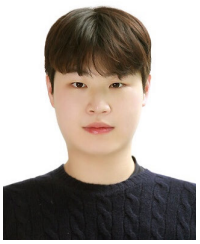
[20] Y.-C. Su and K. Grauman, "Making 360° video watchable in 2D: Learning videography for click free viewing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1368–1376.

[21] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun, "Deep 360 pilot: Learning a deep agent for piloting through 360° sports videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1396–1405.

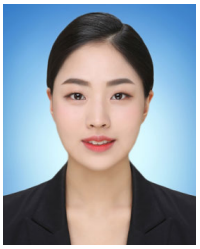
[22] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2693–2708, Nov. 2019.

[23] J. Wang, M. Xu, L. Jiang, and Y. Song, "Attention-based deep reinforcement learning for virtual cinematography of 360° videos," *IEEE Trans. Multimedia*, early access, Sep. 4, 2020, doi: 10.1109/TMM.2020.3021984.

**MIN-SEOK LEE** received the B.S. degree in software and computer engineering and the M.S. degree in electronics and information engineering from Korea Aerospace University, Gyeonggi, South Korea, in 2018 and 2020, respectively. He is currently a Research Engineer with the EO/IR System Research and Development Team, LIG Nex1, Gyeonggi. His current research interests include video signal processing and visual communications.



**SEOK HO BAEK** is currently pursuing the bachelor's degree with the School of Electronics and Information Engineering, Korea Aerospace University, Gyeonggi, South Korea. His current research interests include video signal processing and multimedia systems.



**YOO-JEONG SHIM** received the B.S. degree in electronics and information engineering from Korea Aerospace University, Gyeonggi, South Korea, in 2019, where she is currently pursuing the master's degree with the Department of Electronics and Information Engineering. Her current research interests include video surveillance and 360-degree video processing.

**MYEONG-JIN LEE** (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1994, 1996, and 2001, respectively. From 2001 to 2004, he was a Senior Engineer with System LSI Biz., Samsung Electronics, Gyeonggi, South Korea. From 2004 to 2007, he was an Assistant Professor with the Department of Electrical Engineering, Kyungsung University, Busan, South Korea. In 2007, he joined the School of Electronics and Information Engineering, Korea Aerospace University, Gyeonggi, where he is currently a Professor. His current research interests include visual communication, computer vision, and video surveillance. He is a member of the Korean Information Science Society.