Outside-In: Visualizing Out-of-Sight Regions-of-Interest in a 360 Video Using Spatial Picture-in-Picture Previews

Yung-Ta Lin* **Yi-Chi Liao*** **Shan-Yuan Teng*** **Yi-Ju Chung*** **Liwei Chan†** **Bing-Yu Chen‡**
*‡National Taiwan University †National Chiao Tung University
*{lynda, yichi, tanyuan, violetachung}@cmlab.csie.ntu.edu.tw
†liweichan@cs.nctu.edu.tw ‡robin@ntu.edu.tw

ABSTRACT
360-degree video contains a full field of environmental content. However, browsing these videos, either on screens or through head-mounted displays (HMDs), users consume only a subset of the full field of view per a natural viewing experience. This causes a search problem when a region-of-interest (ROI) in a video is outside of the current field of view (FOV) on the screen, or users may search for non-existing ROIs.

We propose Outside-In, a visualization technique which re-introduces off-screen regions-of-interest (ROIs) into the main screen as spatial picture-in-picture (PIP) previews. The geometry of the preview windows further encodes a ROI's relative location vis-à-vis the main screen view, allowing for effective navigation. In an 18-participant study, we compare Outside-In with traditional arrow-based guidance within three types of 360-degree video. Results show that Outside-In outperforms in regard to understanding spatial relationship, the storyline of the content and overall preference. Two applications are demonstrated for use with Outside-In in 360-degree video navigation with touchscreens, and live telepresence.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
360 Video; Telepresence; Picture-in-picture; Off-screen Targets;

INTRODUCTION
Recently, 360-degree videos (hereafter referred to as "360 video(s)") have increased in popularity as a new video standard, especially on mobile and in virtual reality (VR) space, providing users with a more immersive visual experience [9]. While 360 video can display full-field content, the limitations of viewing generally restrict views to only a subset of the full field of view thus giving users a natural visual experience. As such, users will navigate the viewport in a 360 video by orientating the screen or, in the case of using HMDs, their head. While it is intuitive, the fact that users need to search for potential events in grand 360 videos degrades the user experience, leading the users to missing some important events while they are still searching or exploring the view.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

UIST 2017, October 22-25, 2017, Quebec City, QC, Canada
Copyright © 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4981-9/17/10$15.00
https://doi.org/10.1145/3126594.3126656

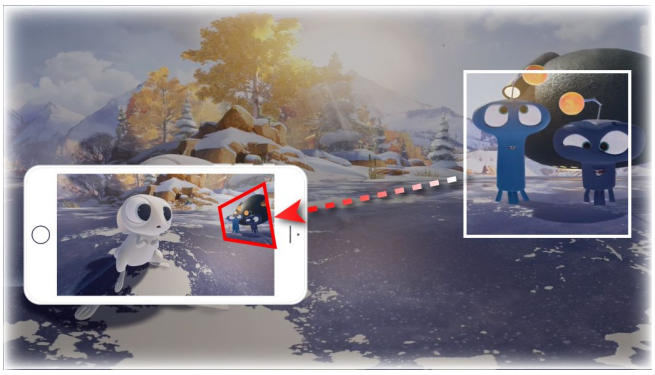

Figure 1. Outside-In is a visualization technique that re-introduces off-screen ROIs onto the main screen as spatial PIP previews, allowing users to make sense of where off-screen ROIs are.

Such circumstances searching for ROIs is more problematic for those videos that contain multiple ROIs. The viewers have to switch between ROIs to catch the whole story, and it is possible for them to be unaware of the positions of all the ROIs or getting lost in the process of searching.

We propose Outside-In, a visualization technique that re-introduces off-screen ROIs as spatial picture-in-picture (PIP) previews onto the main screen. The design of the PIP preview serves two functions. First, it allows users to pre-examine the content in the preview windows, enabling them to develop a better strategy when navigating multiple previews. Second, the spatial representation of a preview window encodes the corresponding ROI's relative location vis-à-vis the current viewport in 360 degree space, providing visual guidance for users to follow.

Our implementation minimizes the occlusion on the main screen due to PIPs by placing the previews in the peripheral regions of the screen. The spatial guidance design is mainly derived from *perspective projection*, in which we transform the PIP plane into 3D space based on the corresponding location of ROIs. To sustain readability, we constrain the level of transformation and mitigate mutual occlusion through manipulating the depth of the PIPs.

Herein, we evaluate Outside-In by comparing it to a traditional arrow-based guidance where three types of multiple-ROIs videos are presented to 18 participants. Based on sub-

jective rating and qualitative analysis, Outside-In provides a better experience to viewers in perceiving the spatial relationship and understanding the storyline, which further leads to a higher preference rating, whereas the arrow-based guidance results in difficulties in use and severe distraction because it lacks detailed information on the indicated destination.

We also implemented two applications: one applying Outside-In on a touchscreen with multi-finger manipulation, and the other demonstrating a 360 degree tele-meeting assisted by our interface.

RELATED WORK

Visualizing Offscreen Targets

There is considerable research on visualizing off-screen ROIs for conditions where the visual context is much wider than a user's viewport. On mobile devices, Halo [2] uses arcs to provide spatial cues, including directions and distance, allowing for effective navigation. EdgeRadar [6] and Wedge [5] were later introduced to better indicate off-screen moving targets and mitigate problems of overlap. These works simply provide directional and distance guidance in 2D scenarios without further information; thus, they are not compatible to highly diverse visual content.

Providing both positions and context of the off-screen targets has been applied to the field of multi-camera surveillance. Girgensohn *et al.* [3] proposes a spatial video player which embeds nearby video feeds around the main video to aid understanding of spatial relationships between the cameras. Their study shows that both static and rotating maps are effective in tracking activities between the cameras. Contextualized Videos [19] identify and characterize an overall class of visualization techniques that combine the video with 3D spatial context. The 3D rotation of videos in this work inspired Outside-In in the context of 360 video navigation.

Navigating Panorama with Limited FOV

360 video allows the receipt of information from a 360 degree surrounding view. In VR/AR, it is particularly beneficial to enable immersive experiences [11, 9] and augment human vision [1]. To allow more effective search in such a wide FOV, overview-plus-detail visualization [17], commonly used in map applications, helps users gain wider understanding of the spatial context while exploring a detailed sub-view.

Some works have expanded a user's FOV and led to more efficient navigation. FlyVIZ [1] replaces the user's normal FOV in VR with omnidirectional graphics, and Xiao *et al.* [20] augment HMDs with LED sparse peripheral displays.

Other works also have attempted to provide more explicit hints to interested targets in the VR environment. Lin *et al.* [10] compare two focus assistance techniques, *i.e.*, auto-pilot and visual guidance. The former one actively brings the user to the target, and the latter one uses an arrow to indicate the direction of the target. Pavel *et al.* [13] introduce re-orientation techniques preventing users from getting lost when shots change in 360 videos. SwiVRchair [4] actively rotates the users' heading (*i.e.*, their viewport) toward off-screen interests in the panoramic view with an actuated swivel chair.

Different from previous works, Outside-In delivers off-screen information into the user's viewport using PIP previews, so that users can browse and navigate freely while keeping track of the out-of-sight targets. Moreover, PIPs can be displayed on a range of videos and devices; hence, Outside-In is also easy to be applied to a wide range of usages.

DESIGN CONSIDERATIONS

Before going into detailed description of the implementations, we here list the following design principles:

1. Minimizing the impact to the main content

Adding visual cues to the main screen would degrade the immersive experience. Also, the assistance should not "steal the show" from the main screen. Hence, the guidance should be placed in the peripheral regions of the screen.

2. Effectively guiding users to reach off-screen targets

In addition to giving information about what is out-of-screen, viewers should know how to find that. To guide the viewers, spatial information should be displayed, which includes both directional and distance information. This guidance should be natural and intuitive.

3. Providing detailed information of off-screen content

In the circumstance of watching 360 videos, the ROIs differ. The meanings of ROIs are also subjective and diverse among viewers. Therefore, the guidance should provide previews of the context. With the help of previews, viewers can make decisions based on the content and develop a better watching strategy.

OUTSIDE-IN

Outside-In is a visualization technique that brings ROIs outside of the viewport back into the viewport and displays them as PIPs on the border of the screen. The concept behind Outside-In is mainly inspired by *perspective projection* that transforms shapes of objects according to their distance to the viewer. Figure 2 shows our design that the PIP planes are floating in the 3D space right between the spherical video and the screen. Thus users can infer the positions of these ROIs in 360 degree space in a natural way according to the facing of the PIP planes.

Spatial guidance consists of both the directional and distance information. For the former, the position and the rotation of the PIP are used to indicate the direction of ROIs. As for the latter, the tilt and the depth of the PIP are adopted to encode information on distance. The details are explained further hereafter in the following sections.

1. PIP as Directional Guidance

The directional information is commonly expressed with an arrow. Instead of adding an extra arrow to the PIP, we use the PIP itself as an arrow.

The Position of PIP

Firstly, we encode the directional information of the ROI in the PIPs position on the main screen. To do this, we calculate

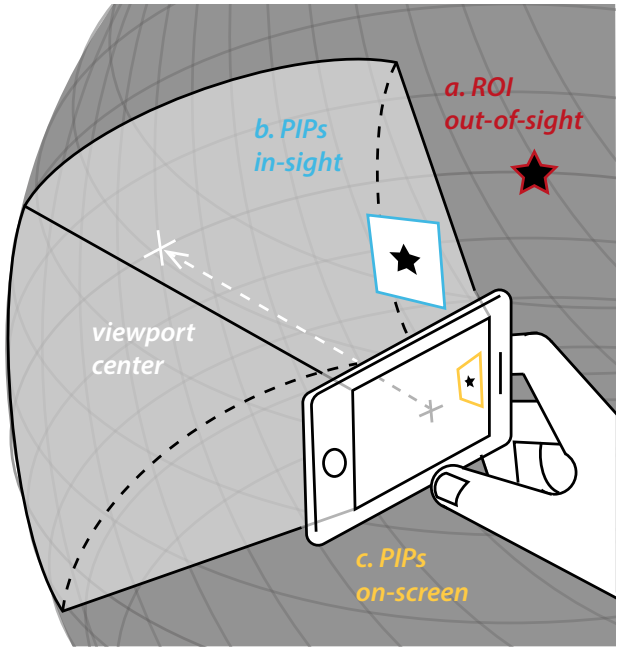

Figure 2. Outside-In uses the effect of perspective projection to bring off-screen ROIs onto the screen as PIPs. (a) The black star represents an ROI out-of-sight on the spherical surface. (b) The corresponding PIP is generated in the field-of-view. (c) Through perspective projection, the final appearance of the PIP is on the screen.

the direction based on the equirectangular projection of the video. In this projection, meridians map to vertical straight lines and circles of latitude map to horizontal straight lines with constant spacing.

Assuming the screen center is at a point $U(longitude of U, latitude of U)$ (e.g., the user viewpoint at U), and the out-of-sight ROI is located at $R(longitude of R, latitude of R)$ (Figure 3). Then, the PIP is placed on the connection from U to R (Figure 4a). This allocation of PIPs is intuitive in guiding users to locate corresponding ROIs, because the equirectangle map better matches the freedom in the vertical and horizontal orientation of the human neck and body respectively.

The Rotation of PIP

To further enhance the directional hints, we also rotate the PIPs around its center. As shown in Figure 4b, this rotation leads the *inner edge* to always face the center of the viewpoint while the *outer edge* is always facing the ROIs. Moreover, the derived *outer edge* is further serving as the tilting shaft for distant guidance which will be explained in "PIP as a Distance Guidance" shortly hereafter.

Additionally, we want to keep the horizon horizontal in the PIP previews to retain readability. Therefore, while the PIP rotates around the center of the viewpoint, we also rotate the virtual camera in charge of the content of the PIP synchronously, to keep the orientation of the content in the PIP unchanged.

Overall, we provide clear directional information with position and rotation while preserving readability, which meets principles 2 and 3 aforementioned.

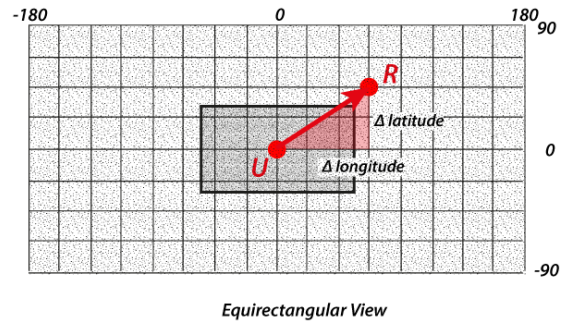

Figure 3. A red arrow shows the connection from view center to an out-of-sight ROI in the equirectangular projection. A PIP is placed on this connection to indicate the direction of the ROI.

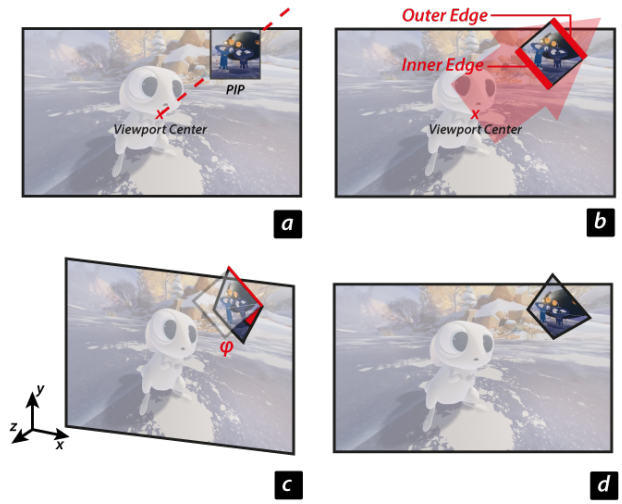

Figure 4. (a) PIP is put on the connection from the viewport center to the out-of-sight ROI. (b) PIP rotates around its own center. The outer edge is defined as the edge facing the ROI and the inner edge is facing the center of the viewport. PIP is like a pointer toward the ROI out-of-sight. (c) The PIP tilts around its outer edge. (d) The final appearance of PIP after applying tilt.

2. PIP as a Distance Guidance

Our design in spatial guidance also encodes the distance of ROIs to the viewport, *i.e.*, how far should users rotate to find corresponding ROIs. To reveal the distance, we make use of the effect of *perspective projection*.

The Tilt of PIP

When users are approaching an off-screen ROI, the gradual change in the tilt of the corresponding PIP gives users substantial evidence of getting close to the ROI. We tilt the PIPs along their *outer edges*, *i.e.*, the red axis in Figure 4c, according to the normalized Euclidean distance of corresponding ROIs to the viewpoint in equirectangular projection.

The normalized Euclidean distance ($dist$) is calculated by the following equation:

$$dist = \sqrt{\left(\frac{\Delta latitude}{90}\right)^2 + \left(\frac{\Delta longitude}{180}\right)^2} \quad (1)$$

Figure 5. (a) Without the tilt, flat 2D PIPs degrade the immersive experience. (b) With the tilt, 3D PIPs can blend into a 3D environment.

The further a ROI is from the viewport, the more inclined the corresponding PIP is to be tilted.

In addition to visual cues, the perspective projection with the scaled tilt allows for more realistic and easy-to-catch 3D positions of the ROIs onto the PIPs (Figure 5). Without scaled tilt, the PIPs would look like photos scattered on a 2D screen which greatly degrades the immersive experience.

Tilting PIPs, however, will distort the content; thus a maximal tilting angle must be determined in order to preserve readability. We set the thresholds for the maximal and minimal values of the tilt angle as 105 and 0 degrees. The tilting angle therefore is linearly interpolated between $maxTilt$ (105) and $minTilt$ (0):

$$\phi = \max Tilt + (0 - \max Tilt) \cdot (maxDist - dist) \tag{2}$$

The Depth of PIP

Since PIPs are stored in the border field of the viewport, mutual occlusion becomes inevitable when the number of PIPs grows. When there are overlapping PIPs, the tilting effect may not be able to clearly differentiate the distances among the PIPs, so we introduced manipulation in depth to further distinguish and resolve the issue of mutual occlusion.

Occlusion is particularly severe when there are ROIs coming from the same direction. For instance, two ROIs apart far off in the off-viewport space however appear to be co-located at the boundary as shown in Figure 6a.

To separate overlapping PIPs and enhance the discernibility of the distance guidance between them, we introduce a new parameter, *the depth of the PIP*. The depths are added to each PIP according to their distance from the viewport center. The depth of PIP is the outcome of the linear interpolation between the maximal depth ($maxDepth$) and the minimal depth ($minDepth$) by following equation:

$$\begin{aligned} depth = \\ \min Depth + \\ (\max Depth - \min Depth) \cdot (maxDist - dist) \end{aligned} \tag{3}$$

The depth range ([$minDepth$, $maxDepth$]) is set to the peripheral region of the entire frame. Closer ROIs will generate PIPs with greater depth values, and following perspective projection, PIPs at greater depths would appear closer to the viewport center (Figure 6b). Thus, the mutual occlusion of

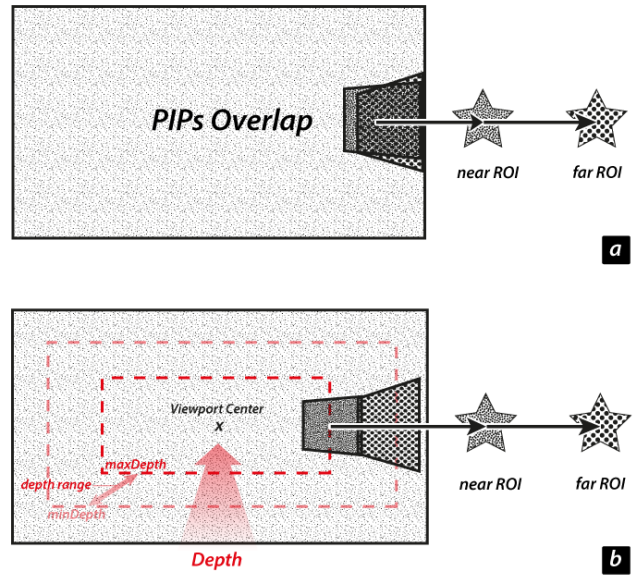

Figure 6. (a) Two PIPs are overlapping since their connection from the center of the viewport to ROI is close. (b) To separate these two PIPs, depths are added to the PIPs. The side-by-side PIPs reveal their relationship as out-of-sight ROIs.

PIPs is resolved, and from the order of PIPs, we can better understand the spatial relationship of multiple ROIs.

However, adding depths to PIPs makes them appear closer to the screen center, which would impact the main content. This becomes a tradeoff, in our design, between minimizing the intrusion to the main screen and minimizing the mutual occlusion of the PIPs. In the end, we suggest to keep the depth range, and the size of the PIP adjustable since the extent of impact differs from person to person and even from video to video.

Distance information is expressed through manipulating the tilt angles and the depths of the PIPs. Plus, readability is preserved through setting the maximal tilt angle and mitigating the mutual occlusions. Last but not least, the depth range dropping at the peripheral region does minimize the intrusion to the main content. To summarize, we can see this design follows the design principles 1, 2, and 3 aforementioned.

USER STUDY

This study aims at investigating the navigational behavior of viewers of 360-degree video using Outside-in. Moreover, we compare user preferences of Outside-in system to a typical arrow-based guidance. To achieve this participants were given three videos to view on a mobile phone, and then they were asked to describe what they were looking at, and how the interfaces worked for navigating the videos. After that, a post-study survey and a semi-structural interview were conducted.

Study Considerations

360 Video Contents and Selection

Based on the number of ROIs and their movement, we categorized 360-degree panoramic videos into five types and listed

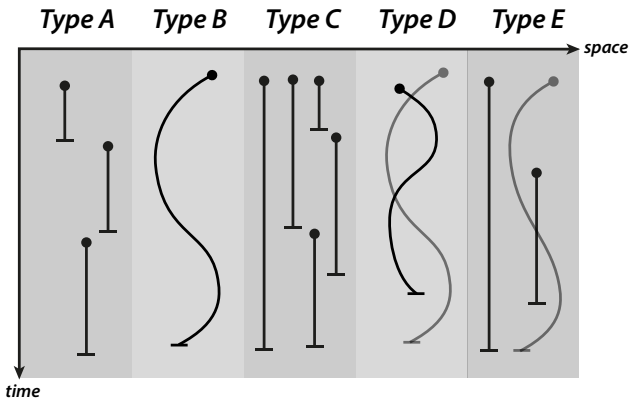

Figure 7. This figure shows the 5 types of 360 videos with the movement of their ROIs in time and space. The x-axis represents that the space is only a longitude angle. (a) Concurrent single ROI with static position (b) Concurrent single ROI with dynamic position (c) Concurrent multiple ROIs with static position (d) Concurrent multiple ROIs with dynamic position (e) Concurrent multiple ROIs with dynamic and static position

examples as per the descriptions found in Figure 7. Also, for the following study, we especially edited videos into short clips to use as Types C, D and E.

For the single-ROI cases, *Type A* is commonly seen in tour-guided videos, where a guide tells the audience about the concurrent target. *Type B* can be found in extreme sports videos, for example, one athlete moves fast in the scene. Lin *et al.* [10] have designed focus assistance for the aforementioned two types of videos. As for the multiple-ROIs cases, *Type C* is commonly seen in videos of entertainment shows, such as a group of performers playing music or dancing around the camera [18]. *Type D* includes films in which characters move and chase each other [16, 15]. *Type E* is a mixture of Types C and D, in that the characters sometimes moves around and sometimes stay still [12].

Because multi-ROIs video are more likely to have issues of out-of-sight ROIs, this study has disregarded single-ROI videos at first. There are also a few conditions for selecting these videos for study. First, the selected videos must be long enough to be segmented into three 1-minute clips, and the split clips should contain similar characters, and story. Second, the selected videos are popular on YouTube which is testimony of their easy comprehension. We labeled the center of ROIs manually within an equirectangular view of the video where every clip has two to three ROIs. We also manually marked each ROI with a semitransparent star for the viewers to more easily identify all the targeted ROIs. Sample screenshots of the three selected videos are displayed in Figure 8, and described in detail hereafter.

Concurrent Multiple ROIs with Static Position

A Pokémon [14] 360 video was selected, and the viewers were asked to look for Pokémons, which barely moved throughout the video. However, the viewer is teleported to other remote scenes for a period of time. Above all, this video is a typical explorative video with scene switches. In the 3-minute length of film, the scene switches 5 times pushing the user to re-explore the environment.

Concurrent Multiple ROIs with Dynamic Position

Made by Google Spotlight, *Help* [15], was selected for this category. It is a short action movie with an intense storyline. The video is basically about a monster attacking people in a subway station where the monster and several human characters were marked as ROIs. As the monster chases those characters, the ROIs position shifts quickly which also leads to the camera moving fast to capture the whole scene.

Concurrent Multiple ROIs with Static and Dynamic Position

In the video *School of Rock* [12], a musical band of fourteen students and a teacher are rehearsing in a classroom. The camera is placed in the center of the scene and the characters are placed around the camera. Although the position of the camera is fixed, this video is actually more complicated and intriguing because a part of the ROIs are static while some are dynamic. We labeled only the characters dialoging or performing as ROIs. In addition, lyrics displayed on the ceiling of the classroom was also marked, which easily went unnoticed without visual guidance.

Interfaces

As shown in Figure 9, two interfaces are compared: arrow-based guidance, and Outside-in guidance. The arrow-based guidance is the other common approach to indicate out-of-sight objects, which is our primary interface for comparison. We applied the identical directional equation of Outside-In to the arrow-based guidance and placed arrows on the rectangular orbit attached to the border of the screen similar to that done for Outside-in guidance. The arrow-based guidance provided only the directional information but nothing regarding distance.

Participants and Apparatus

18 paid participants (8 males) age from 22 to 29 were recruited from our university. This study was conducted in the well-controlled environment. We implemented our system in the Unity 3D game engine, and all the videos of our study were displayed on an iPhone 6s with 4.7 inch (104 mm x 58 mm) sized screen.

Experiment Design

A repeated measures within-subject 2x3 factorial design was conducted with the independent variables being two interfaces, *Outside-In*, traditional *arrow-based guidance* and three videos, *Concurrent Multiple ROIs with static position*, *Concurrent Multiple ROIs with dynamic position*, and *Concurrent Multiple ROIs with static and dynamic position*. The study was divided into three sections based on the three different videos, where the order is fixed for all the participants. However, to prevent the users from predicting the storyline of the next clip in advance, we segmented all the videos into three 1-min clips and presented them randomly in a discontinuous sequence, *e.g.*, a video originally composed of 1-2-3 clips might be presented with one of the following sequences: 3-2-1, 2-1-3, or 1-3-2. The sequence of interfaces always began with the baseline, *i.e.*, no guidance, and then the other two interfaces were assigned with a counter-balanced order. Overall, we collected data on *3 videos x 3 clips x 18 participants = 162 data of clips.*

Figure 8. This figure shows the equirectangular frame of each selected video. Each circle represents the ROI in that frame.

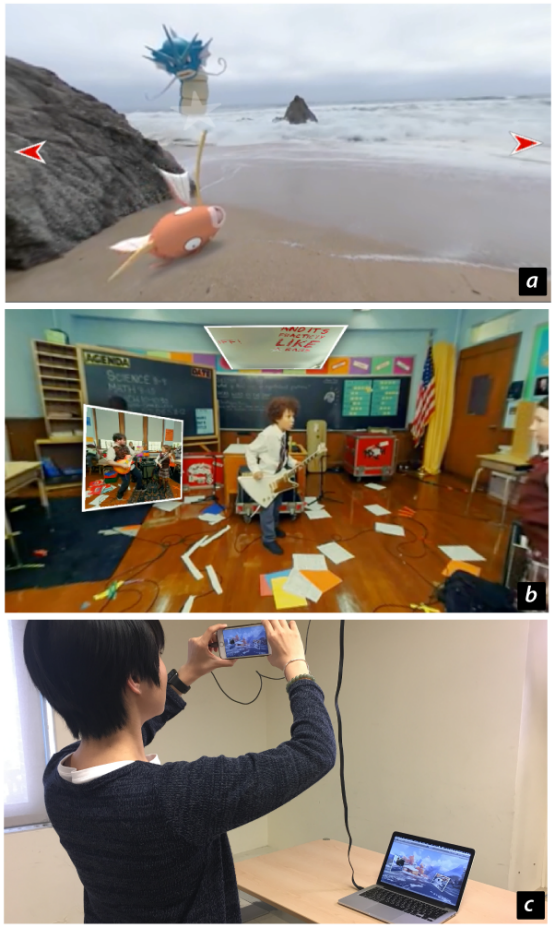

Figure 9. (a) The *Pokemon* video with arrow-based guidance. In this screenshot, two red arrows are placed on the right and left borders indicating that there are two other ROIs in these directions. (b) The *School of Rock* video with Outside-In. In this screenshot, two PIPs are placed, on the left and upward, showing that there is another character playing guitar on the left-hand side and that there are lyrics on the ceiling. (c) This photo shows the setting for the study. Participants watched the video via a smartphone while standing. That smartphone is connected to the laptop via a cable suspended above the participant to allow greatest freedom of movement and least interference during the process.

PROCEDURE

There was a practice session before the formal study was conducted. In this practice session, we gave detailed explanations of how the interfaces would work while participants are viewing the video [16]. In this session, the same video was played repeatedly using different interfaces, and each lasted for 30s. The video could be replayed if the participants needed it. This video was not included in the formal study.

The formal study was then conducted after a short break. As the users were asked to think aloud while watching the clips for us to better understand their behavior, we recorded the whole process in term of video and audio. There was a semi-structural interview for rating their experience of the interfaces to assess factors such as distraction level and helpfulness of understanding the storyline. Also, through the interviewing, we attempted to understand the mental models behind their behavior during tasks. The previous records were useful for the participants to recall the process in this survey. The whole process including practice and interview lasted for around one hour.

SUBJECTIVE RATING ANALYSIS AND RESULTS

During the 4-question post-study questionnaire, the participants were asked to give their subjective rating, as a score between 1 (the least) and 7 (the most), *on every video in the two visual-guiding interfaces*. Thus, the subjective rating analysis design is: *18 participants x 2 interfaces x 3 videos x 4 questions = 432 data points.* The results have been further analyzed with a two-way repeated measures ANOVA.

The four questions are as follow:
Q1: The interference level of the interface.
Q2: The understanding level of the spatial relationship of ROIs
Q3: The awareness level of the storyline
Q4: The overall preference level of the interface

Results

The overall ratings of arrow-based guidance are 2.89 (s = 1.58), 4.31 (s = 1.23), 4.89 (s = 1.49), and 3.96 (s = 1.20) for Q1, Q2, Q3, and Q4 respectively, and the ratings of the Outside-In guidance are 3.63 (s = 1.59), 5.20 (s = 1.48), 5.63 (s = 1.16), 4.93 (s = 1.57) respectively.

Interference Level: The results show no interaction between the VIDEOS and INTERFACES ($F_{2,34}$=0.114, p >0.05) on the interferences, which is indicated by Q1 in Figure 10. Main effects analysis also reveal no significant effects between interfaces ($F_{1,34}$ = 3.825, p >0.05) and no effects between videos ($F_{2,34}$ = 1.393, p >0.05).

Perceiving Spatial Information: The results show no interaction between the factors on *perceiving spatial information* ($F_{2,34}$ = 0.114, p >0.05), which is indicated by Q2 in Figure 10. Significant differences between interfaces was found

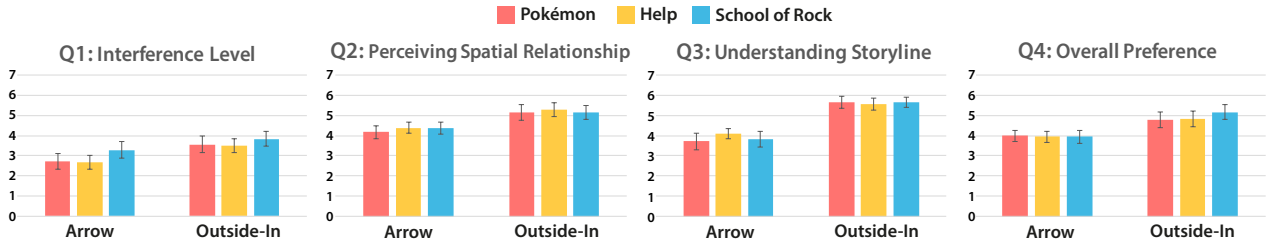

Figure 10. Estimated Marginal Means for four questions.

($F_{1,34}$ = 7.556, p <0.05) (out:5.20 vs. arrow:4.31) and there are no differences between the perception of the videos ($F_{2,34}$ = 0.135, p >0.05).

Understanding Storyline: There is no interaction between the factors on *understanding storyline* ($F_{2,34}$ = 0.693, p >0.05), which is indicated by Q3 in Figure 10. There are significant differences between interfaces ($F_{1,34}$ = 28.666, p <0.01) and no effects between videos ($F_{2,34}$ = 0.103, p >0.05).

Preference Level: Again, no interaction is found between the two factors ($F_{2,34}$ = 0.434, p >0.05), which is indicated by Q4 in Figure 10. Outside-In has significant higher preferences ($F_{1,34}$ = 4.876, p <0.05), and there are no differences between the videos ($F_{2,34}$ = 0.378, p >0.05).

Discussion

Overall, Outside-In outperforms the arrow-based guidance under most conditions for all the videos. It provides clear spatial relationships with ROIs to the viewers, which helps them attain a better comprehension of the storylines, and finally leads to a better navigational experience. Although the PIPs occupied more space than the arrows which potentially might lead to occlusion issues, the results of interference level reveal no significant differences between the two guidance methods.

QUALITATIVE ANALYSIS AND RESULTS

The qualitative data from interviews is transcribed and the sentences are divided with distinctive meanings into 606 quotes. Afterward, the transcriptions were coded via an iterative process and themes were identified.

Watching 360 Videos Using Arrow

At first, participants found it *easier to search for interested targets* by following the arrows. Participant 15 (P15) said, "*Arrows indicated to me the direction that might contain more worthwhile watching.* " P13 also noted, "*The arrows did help me catch on to the interesting things and made it easier to follow.* " P11 added, "*The arrows helped me in viewing the content of the 360-videos, yet without making it more interesting.* " Also, in the case of less ROIs or fixed-character scenarios, (*e.g.*, the video *Help* has only 3 characters) the viewers can keep track of the context as pointed out by each arrow. As P3 mentioned, "*The characters keep popping in-and-out in the Pokémon video. However, it's easier to memorize what arrows points to which characters in the Help video because there are only three characters and three arrows.* "

However, in most cases, arrows were used with difficulty as most participants report *a serious problem not knowing what the arrows point to*. This problem forces them to constantly search the ROIs, causing them difficulty in concentrating on the main screen. As P6 reported, "*I felt I was cheated by the arrows. Even I chased the arrows all the time, still feeling that I was missing the points of the video.* " P10 also agreed, "*It (the arrow) provides the direction, but sometimes it's something different to my expectation.* " Such problem become even worse in a *multi-ROIs or moving-ROIs* scenario, as P12 noted, "*I cannot decide which arrow I should follow when there are two arrows shown on the screen at once.* "

In addition, even though the arrows occupied only a little space near the border they were still deemed a *distraction* by participants. Many participants stated that they just could not ignore the arrows. P1 explained, "*I would noticed the cues (arrows) once they pop out. It's really distracting.* " After watching *Help*, P11 also reflected that, "*After I found the monster, I wanted to concentrate on it. Yet, I couldn't help but look where those arrows pointed to.* " Multiple arrows also pose a bigger distraction. P12 suggested, "*There should not be two apparent cues leading to different directions simultaneously. That really confused me as to where to look.* "

At some point, the participants wholly *gave up on following the guidance*. P18 said, "*No matter what angles I turned to, there's always another arrow. Eventually, I felt annoyed and decided to ignore all the arrows and decide what to watch myself.* " Some decided to follow only under certain circumstances. P5 said, "*In the beginning, I would follow the arrows to find possible characters, but once I found the main character, I would ignore all the other arrows.* " In addition, P11 said, "*After finding the monster (in Help), I ignored other stuff. The first time I noticed there's an arrow was when I felt bored with the monster; then, I realized that there was a police officer behind me the whole time.* "

Discussion

At first, arrows seem to be an effective means of guiding viewers to interesting targets, but there are advantages only in the simpler scenarios. For example, videos with less ROIs, or videos with ROIs in fixed positions. For more complicated videos, the arrows fail to provide effective assistance due to lack of information toward the destination, and become a severe distraction. Which sometimes makes the users *exhaustively search for unknown targets*, or *abandon the guidance for remaining more consistent watching experiences*. Above observation and finding explain why arrows have poorer rating in the previous subjective rating analysis.

Watching 360 Videos Using Outside-In

In general, the participants considered Outside-In an *effective assistance for understanding the content and storyline* of the videos due to the multi-screen design. As P13 stated, "*It's convenient because I wouldn't miss anything important.* " P9 also agreed, "*Because I knew what's there (in the PIPs), it's easier to catch all the points of the scene.* " P3 said, "*I felt more comfortable with the assistance of picture-in-pictures because I could focus on one point and also perceive the other events from the picture-in-pictures. I wasn't afraid to miss anything interesting.* " P3 further added, "*I could comprehend the video content faster because I could watch all the characters at the same time.* " Some of the participants developed a time-division method to acquire the information across screens. P16 said, "*Most of the time, I stayed focused on the main screen, while a fraction of time I looked for the picture-in-pictures. Then, I scanned the content of them every 3 to 4 seconds.* " P10 also said, "*I watched the main screen and picture-in-pictures alternately.* "

Many participants also claimed that they could *concentrate on the main screen more and waste less time looking around*. P5 pointed out, "*With the help of the preview, I didn't have to turn around all the time.* " Also, participants tended to view less-important-ROIs on the PIPs after locating the primary character on the main screen. After watching *Help*, P3 said, "*I've already identified the main characters, the monster and two men, so I put one of them on the main screen and watched others on the picture-in-pictures.* " P7 also said, "*I could concentrate more on what I wanted to focus on. As long as the character in the picture-in-picture showed no special actions, I didn't have to turn to any particular direction.* " For instance, in the third video, *School of Rock*, the lyrics on the ceiling were less important than the teacher and students. P17 said, "*I could see the lyrics through the picture-in-picture, so I didn't have to raise my head.* "

Outside-in is also a great help for *making decisions about switching ROIs*. After watching *School of Rock*, P1 said, "*(When I needed to decide which ROI to turn to,) it feels like this interface gives me more information in advance.* " Based on the content in the PIPs, participants are able to determine the importance level of each ROI. P12 said, "*To me, picture-in-picture shows how important the character is, for example, when the explosion and the man appeared simultaneously, I chose the explosion because it's more important.* "

Outside-In also helped the participants to *understand the spatial relationship* between characters. P10 noted "*I could approximately guess the relative positions of the characters by the position of the main screen and picture-in-pictures.* " After watching *Help*, P2 said, "*When the camera moved downstairs, I knew the police were chasing behind from the preview.* "

Finally, some participants mention the *distraction issues* raised by Outside-In. P4 said, "*The main screen was broken into pieces because of the occlusion of some previews. I couldn't get the full view of the main screen, and I had to dodge the picture-in-picture.* " The level of distraction also altered according to the complexity of the scene. For example about the video of *Pokémon*, P9 explained, "*Because the spacious environment of this video, and the few and scattered things to focus on, picture-in-picture didn't occlude the Pokémons. Therefore the interference is acceptable.* " In contrast, in the video of *Rock*, P11 found, "*It's disturbing. Because there were lots of supporting roles around the scene, the picture-in-picture blocked them.* " P15 commented, "*There were too many picture-in-pictures plus the main screen to watch, I didn't know what I needed to concentrate on.* "

Discussion

Overall, Outside-in enables participants to quickly understand the storyline and the spatial relationship of the ROIs. Such benefits allow them to pay more attention to the main screen and easily perceive out-of-sight information from the PIPs. Even when the participants wanted to switch to other targets, they can accomplish their searching task faster due to perceiving the approximate location in advance. Although there's some distraction and occlusion issues raised by the system, the interface allows a better watching experience under most conditions.

APPLICATIONS

We demonstrate Outside-In with two applications for a 360 video player on touchscreens and a telepresence interface through a laptop with a webcam.

360 Video Player on Touchscreens

There are an increasing number of 360 videos to be shared and consumed on mobile touchscreens. Here, the 360 video player integrated touch interactions with the Outside-In technique for a more effective navigation experience. (Figure 11a)

Auto-Piloting

When users click on a PIP preview, the player will auto-pilot the viewing window, bringing in the corresponding ROI at the screen's center. An easing function, which accelerates at the beginning and decelerates at the end, is applied to provide smooth piloting.

Show/Hide One and All

On-screen PIP previews impact the main screen view, especially when users prefer an immersive view for current on-screen ROI. Pinching-out allows users to dismiss previews, keeping full screen for improved immersion. Pinching-in again brings back the previews. Two interactions are further proposed to help users engage in ROIs of focus. Long-press-to-Focus allows users to dismiss all previews except the one for long-press select. Swipe-to-dismiss allows users to remove uninterested ROIs by swiping them out.

Tele-Meet: Telepresence Interface

Tele-Meet integrates the Outside-In technique for telepresence interface. Here, we target the scenario of attending a remote roundtable discussion. However, the interface can be applied to other teleportation applications with similar needs.

As displayed in Figure 11b, the Tele-Meet device which includes a 360 degree camera and a turnable screen is set at

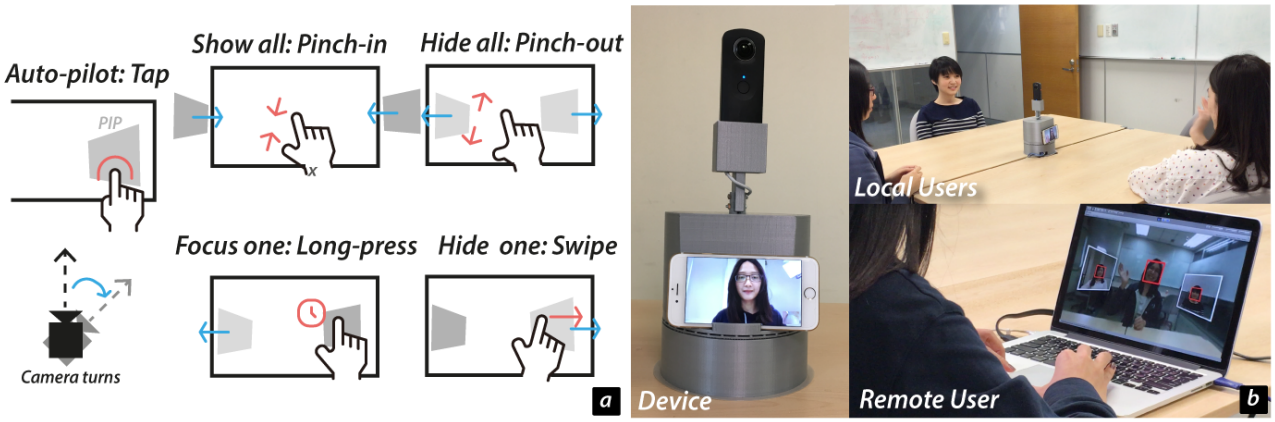

Figure 11. (a) Touchscreen interaction: Five multi-finger gestures are introduced. Auto-pilot triggered by tap on the PIP makes the main camera auto-navigate to the selected target. Pinch-in and pinch-out shows and hides all the PIPs respectively. Long press leaves one PIP on the screen while swipe dismisses the selected PIP. (b) Tele-Meet: Device: a 360 degree camera for capturing omnidirectional live stream video and a step motor carrying a mobile phone representing the remote user. The direction of the mobile device was controlled by the remote attendee during the Tele-Meet so she could face any other local attendee. Remote User: The Outside-In interface of the remote end. A remote attendee rotated his viewport by swipe gestures while the out-of-sight attendee was revealed via the PIPs.

the center of a table in a remote meeting place. The 360 degree camera live streams a full field of images from the surrounding, allowing the virtual attendee to see physical attendees face-to-face from the table center. The screen set next to the 360 degree camera displays the virtual attendees face, and syncs its physical orientation with the virtual attendee's viewing direction in 360 degree space.

We detect human faces in 360 videos as ROIs and create corresponding PIPs in real time. PIPs dynamically appear or disappear when people join or leave the table. The Outside-In technique enables the virtual attendee to maintain spatial awareness of off-screen people during the roundtable discussion. The face detections are done in the panoramic format of the 360 video and implemented using a third-party OpenCV library for Unity.

DISCUSSION AND FUTURE WORKS

Automatic ROI Labeling

In this work, all ROIs in the tested videos were marked manually to ensure correct annotations. In addition, we also presented OpenCV face detection in the Tele-Meet application to show the possibilities of automatic annotation. Such annotations could be replaced by video salience detection methods [21, 8]; furthermore, adopting deep learning can effectively enhance the accuracy of object detection in 360 videos [7].

Investigating Proper Video Contents

While the three professional 360 videos are used in our study and showed clear need for the assistance for videos containing multi-ROIs or moving-ROIs, however, this might not be desirable for some types of videos. For example, "highly immersive VR scenes" 360 videos, *e.g.*, horror-VR and treasure-hunting, often require the viewers to explore the scene on their own with no need for visual guidance. Here, we encourage future research to explore and identify the proper usage of Outside-In.

Adaptive PIPs for Dynamic ROI Size

The size and FOV of PIPs are fixed in the current Outside-In system. The design might not be effective for a video containing ROIs with extremely varied sizes. For instance, a small ROI might be too small to be carefully watched in a PIP, and a huge ROI might exceed the border of the PIP. This issue can be alleviated by implementing auto-adjusting FOV for PIP, and auto-adjusting the distance of the camera according to the actual size of ROIs to maintain an approximately similar size inside the PIP view. Future PIP design should consider merging nearby ROIs automatically to avoid the occlusion problem.

Displaying the Importance Levels of ROIs

Understanding the importance levels inherent to the ROIs is crucial for watching a video with storylines, *e.g.*, recognizing the main character, or identifying an urgent event. Such an importance level can be visualized via Outside-In by adding different colors on the borders of the PIPs or changing opacities. In addition to importance levels, other information can also be visualized through proper design; such as the status of characters, *e.g.*, talking, moving or leaving.

LIMITATIONS

From the feedback of the user study, we find that users form their own navigational strategies to use the PIPs, *e.g.*, some users used PIPs to view the whole story without pursuing them, and some will do just the opposite. These strategies also vary according to different types of videos. Due to the complexity of the possible scenarios, we did not quantitatively examine the influence of factors such as length of searching time, familiarity of storyline and the correctness of spatial relationship of the targets. Instead, we identified these factors based on participants' qualitative feedback and on our observation of them during the study.

The distraction raised by the PIPs is another issue. However, the results of our questionnaire showed no differences in the distraction level between the PIPs and the arrows. Also, the potential for distraction and occlusion are inevitable for all

visual guiding systems. To minimize this drawback, as described in Implementation, our algorithm allows the flexibility to adjust the factors such as the range of depth and the size of the PIPs, and our first application demonstrated the "Show/Hide One and All" feature toward the PIPs. Future research should explore other solutions and improvements.

CONCLUSION

In this paper we have proposed Outside-In, a visualization technique that re-introduces the out-of-sight ROIs back onto the main screen when watching 360 degree videos. We use PIPs to provide previews to the ROIs and indicate the spatial relationship of ROIs using the PIPs' geometry. Our user study conducted reveals that Outside-In outperforms arrow-based visualization in both comprehension to the overall storyline and perception of the spatial relationship of the content. Two applications were presented: A touchscreen 360 video display combined with several advanced gestural manipulations, and a 360 degree telepresense with auto face recognition.

We encourage future researchers to consider implementing automatic ROI-labeling, adaptive PIPs and encoding importance levels of ROIs. We also look forward to the exploration of wider usages of Outside-In, such as integration with HMDs for 360 degree awareness of surroundings in the VR social network.

ACKNOWLEDGEMENTS

We thank Yu-Cheng Chen and Yu-Hsiang Tseng for their kind supports. This work was partly supported by Ministry of Science and Technology and MediaTek Inc. under Grants MOST105-2622-8-002-002, 103-2218-E-002-024-MY3, 106-3114-E-002-010, 106-2923-E-002-013-MY3, and 106-2221-E-002-211-MY2.

REFERENCES

1. Jérôme Ardouin, Anatole Lécuyer, Maud Marchal, Clément Riant, and Eric Marchand. 2012. FlyVIZ: A Novel Display Device to Provide Humans with 360° Vision by Coupling Catadioptric Camera with Hmd. In *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology (VRST '12)*. ACM, New York, NY, USA, 41–44. DOI: **http://dx.doi.org/10.1145/2407336.2407344**
2. Patrick Baudisch and Ruth Rosenholtz. 2003. Halo: A Technique for Visualizing Off-screen Objects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 481–488. DOI: **http://dx.doi.org/10.1145/642611.642695**
3. Andreas Girgensohn, Frank Shipman, Thea Turner, and Lynn Wilcox. 2007. Effects of Presenting Geographic Context on Tracking Activity Between Cameras. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1167–1176. DOI: **http://dx.doi.org/10.1145/1240624.1240801**
4. Jan Gugenheimer, Dennis Wolf, Gabriel Haas, Sebastian Krebs, and Enrico Rukzio. 2016. SwiVRChair: A Motorized Swivel Chair to Nudge Users' Orientation for 360 Degree Storytelling in Virtual Reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1996–2000. DOI: **http://dx.doi.org/10.1145/2858036.2858040**
5. Sean Gustafson, Patrick Baudisch, Carl Gutwin, and Pourang Irani. 2008. Wedge: Clutter-free Visualization of Off-screen Locations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 787–796. DOI: **http://dx.doi.org/10.1145/1357054.1357179**
6. Sean G. Gustafson and Pourang P. Irani. 2007. Comparing Visualizations for Tracking Off-screen Moving Targets. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems (CHI EA '07)*. ACM, New York, NY, USA, 2399–2404. DOI: **http://dx.doi.org/10.1145/1240866.1241014**
7. Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. 2017. Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Video. *CoRR* abs/1705.01759 (2017). **http://arxiv.org/abs/1705.01759**
8. Laurent Itti and Pierre F. Baldi. 2006. Bayesian Surprise Attracts Human Attention. In *Advances in Neural Information Processing Systems 18*, Y. Weiss, P. B. Schölkopf, and J. C. Platt (Eds.). MIT Press, 547–554. **http://papers.nips.cc/paper/2822-bayesian-surprise-attracts-human-attention.pdf**
9. Shunichi Kasahara and Jun Rekimoto. 2014. JackIn: Integrating First-person View with Out-of-body Vision Generation for Human-human Augmentation. In *Proceedings of the 5th Augmented Human International Conference (AH '14)*. ACM, New York, NY, USA, Article 46, 8 pages. DOI: **http://dx.doi.org/10.1145/2582051.2582097**
10. Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell Me Where to Look: Investigating Ways for Assisting Focus in 360° Video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 2535–2545. DOI: **http://dx.doi.org/10.1145/3025453.3025757**
11. Hajime Nagahara, Yasushi Yagi, and Masahiko Yachida. 2003. Wide Field of View Head Mounted Display for Tele-presence with An Omnidirectional Image Sensor. In *2003 Conference on Computer Vision and Pattern Recognition Workshop*, Vol. 7. 86–86. DOI: **http://dx.doi.org/10.1109/CVPRW.2003.10070**
12. School of Rock the Musical. 2015. SCHOOL OF ROCK: The Musical - "You're in the Band" (360 Video). **https://www.youtube.com/watch?v=GFRPXRhBYOI&feature=youtu.be**. (2015).

13. Amy Pavel, Maneesh Agrawala, and Bjorn Hartmann. 2017. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *In proceedings of the 30th annual ACM symposium on User Interface Software and Technology (UIST '17)*. New York, NY, USA. DOI:http://dx.doi.org/https://doi.org/10.1145/3126594.3126636
14. Freakin Rad. 2016. Pokémon 360 - CATCH 'EM ALL in VR! https://www.youtube.com/watch?v=pHUVS_GrIeM&feature=youtu.be. (2016).
15. Google Spotlight Stories. 2016. 360 Google Spotlight Story: HELP. https://www.youtube.com/watch?v=G-XZhKqQAHU&feature=youtu.be. (2016).
16. Baobab Studios. 2016. INVASION! 360 VR Sneak Peek. https://www.youtube.com/watch?v=gPUDZPWhiiE&t=12s. (2016).
17. Laura A. Teodosio and Michael Mills. 1993. Panoramic Overviews for Navigating Real-world Scenes. In *Proceedings of the First ACM International Conference on Multimedia (MULTIMEDIA '93)*. ACM, New York, NY, USA, 359–364. DOI:http://dx.doi.org/10.1145/166266.168422
18. Verest 360 VR. 2015. Bambino Down Mode. https://www.youtube.com/watch?v=ujYyE0lfSUk. (2015).
19. Yi Wang, David M. Krum, Enylton M. Coelho, and Doug A. Bowman. 2007. Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1568–1575. DOI:http://dx.doi.org/10.1109/TVCG.2007.70544
20. Robert Xiao and Hrvoje Benko. 2016. Augmenting the Field-of-View of Head-Mounted Displays with Sparse Peripheral Displays. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1221–1232. DOI:http://dx.doi.org/10.1145/2858036.2858212
21. Yun Zhai and Mubarak Shah. 2006. Visual Attention Detection in Video Sequences Using Spatiotemporal Cues. In *Proceedings of the 14th ACM International Conference on Multimedia (MM '06)*. ACM, New York, NY, USA, 815–824. DOI:http://dx.doi.org/10.1145/1180639.1180824