applied sciences

MDPI

Article

MonoMR: Synthesizing Pseudo-2.5D Mixed Reality Content from Monocular Videos

Dong-Hyun Hwang and Hideki Koike *

Department of Computer Science, School of Computing, Tokyo Institute of Technology, Tokyo 152-8550, Japan; hwang.d.ab@m.titech.ac.jp
* Correspondence: koike@c.titech.ac.jp

Abstract: MonoMR is a system that synthesizes pseudo-2.5D content from monocular videos for mixed reality (MR) head-mounted displays (HMDs). Unlike conventional systems that require multiple cameras, the MonoMR system can be used by casual end-users to generate MR content from a single camera only. In order to synthesize the content, the system detects people in the video sequence via a deep neural network, and then the detected person's pseudo-3D position is estimated by our proposed novel algorithm through a homography matrix. Finally, the person's texture is extracted using a background subtraction algorithm and is placed on an estimated 3D position. The synthesized content can be played in MR HMD, and users can freely change their viewpoint and the content's position. In order to evaluate the efficiency and interactive potential of MonoMR, we conducted performance evaluations and a user study with 12 participants. Moreover, we demonstrated the feasibility and usability of the MonoMR system to generate pseudo-2.5D content using three example application scenarios.

Keywords: augmented reality; computer vision; human-computer interaction

Citation: Hwang, D.-H.; Koike, H. MonoMR: Synthesizing Pseudo-2.5D Mixed Reality Content from Monocular Videos. *Appl. Sci.* **2021**, *11*, 7946. https://doi.org/10.3390/app11177946

Academic Editor: Chang-Hun Kim

Received: 1 August 2021
Accepted: 25 August 2021
Published: 27 August 2021

1. Introduction

Mixed reality (MR) head-mounted devices (HMDs) are display devices that can overlay virtual content in the real world, and a user can watch it on a free viewpoint. Specifically, compared with 2D media, such as photos and videos, 3D content synthesized from real-world objects has higher immersiveness. Hence, many methods for synthesizing MR content from real-world objects have been proposed [1,2], and several systems are already commercialized [3,4]. A common method in creating MR content is to place multiple monocular RGB or depth cameras around an object, synchronously capture the images, and then reconstruct 3D shapes of the object from the captured images. The content synthesized using this method is accurate and impressive and is utilized in various industry fields, such as sports broadcasting and entertainment. However, since most previous systems based on this method require multiple cameras and synchronization devices, configuring the system is complex, and the operating environment is restricted [2,5,6]. Thus, these factors make it difficult for end-users to use these systems. In addition, estimating the intrinsic and extrinsic parameters of the cameras from prerecorded videos is a challenging task. Some methods try to estimate the parameters using landmarks in images [7]. However, this limits the types of video that the method can process.

Nowadays, we can record monocular videos through smartphones and digital cameras, and a large number of various monocular videos, such as sports, entertainment, and daily life, have been uploaded to the Internet. If there is an easy way for end-users to create MR content from these videos, various MR content can be provided without any special equipment. Moreover, the created content can be shared freely, similar to the present video-sharing websites on the Internet. To explore the usability and feasibility of MR content made from monocular content, we propose MonoMR, a system for synthesizing MR content from monocular videos. MonoMR is an end-to-end system (Figure 1) that

uses simple yet effective methods, different from the previous complex systems. In order to create MR content, first, the system detects people from a monocular video through a deep neural network (DNN), then calculates the homography matrix between real-world and image distances using an interactive user interface, and estimates pseudo-3D positions of the detected people. Next, person textures are extracted and placed at the estimated positions. Finally, the MR content is synthesized on these elements. The content synthesized by the proposed system is played on Microsoft HoloLens; the user can freely place the content in the actual world and view it from a free viewpoint. The paper's outline is summarized as follows:

- We propose MonoMR, a system to synthesize MR content from single or multiple monocular videos.
- We evaluate the quantitative performance of our system.
- We assess the impact of the synthesized content through a user study.
- We develop suitable sample applications using the proposed system.

Figure 1. MonoMR enables users to easily synthesize pseudo-2.5D mixed reality content from monocular videos uploaded on the Internet or taken with common imaging equipment such as smartphones and cameras. With the MonoMR system, the user can create and experience immersive mixed reality content from various monocular videos, such as (**a**) sports broadcasting videos and (**b**) entertainment videos. (**c**) The synthesized content can be displayed in the real world through a mixed reality head-mounted display.

2. Related Work

Our proposed system is related to monocular video-based content synthesis, free-viewpoint video systems, and human analysis. In this section, we briefly discuss these related works.

2.1. Monocular Video Based Content Synthesis

Algorithms for enhancing 3D information from monocular images and methods that utilize monocular images as Augmented Reality (AR) or MR content have been explored. The algorithm proposed by Ballan et al. [8] creates an optimal viewpoint path among two monocular videos and an interpolated video between the videos based on the optimal path; therefore, users can recognize the spatial information with a change in the viewpoint of the video. Algorithms [9,10] have also been proposed to recover 3D information from monocular videos, construct meshes, reconstruct videos based on recovered information, and freely convert viewpoints.

Langlotz et al. proposed a smartphone-based AR system [11]. In this system, the user manually specifies the moving person of the video, and the system synthesizes the AR content by extracting the designated person portion from the video and synthesizing it on

other videos. The learning support system proposed by Mohr et al. [12] extracts features from a monocular video, and the features are transformed based on the previously defined three-dimensional (3D) model and projected to the real world with the AR content.

2.2. Free-Viewpoint Video System

Since the method of capturing a real-world object using multiple cameras and converting the object into a 3D object produces a high-quality result, systems utilizing this method are mainstream in free-viewpoint video systems. Initially, Kanade et al. proposed a virtualized reality system [13], in which 51 monocular cameras are arranged in a dome-shaped structure. In this system, a target object can be converted into free-viewpoint content, and since the system has scalability, this system was used for early free-viewpoint sports broadcasting [14]. This multi-view method is applied to various free-viewpoint systems [5,15–18], and since these systems consist of RGB cameras, they are less restrictive to the environment. In addition, free-viewpoint sports broadcasting systems, such as Intel True View [4] and Canon's free-viewpoint video system [3], based on this method are actively commercialized by various companies. However, the systems based on this method require many cameras, synchronization devices, and a large amount of computation. Consequently, utilizing these systems by individuals or small groups is still challenging.

Studies on generating highly detailed free-viewpoint video using depth cameras have been conducted. The introduction of commercial depth cameras, such as RealSense (https://software.intel.com/en-us/realsense, accessed on 24 August 2021) and Kinect (https://developer.microsoft.com/en-us/windows/kinect, accessed on 24 August 2021), facilitates these studies. Accordingly, the systems [6,19–21] that generate high-quality free-viewpoint video through depth cameras have been proposed. Collet et al. proposed a system [1] that generates free-view video with 106 RGB and depth cameras and compresses the video for real-time free-viewpoint video streaming. Based on this research, Orts-Escolano et al. presented the Holoportation [2] system that enables real-time telepresence on MR HMDs. However, operating this method outdoors is difficult because most depth cameras are not suited for natural light; thus, the capturing environment is restricted to indoor environments. Furthermore, since the systems still require many cameras and synchronization devices, the end user's accessibility is still limited.

Various methods for generating a free-viewpoint video from a monocular video have been proposed. However, it's still challenging because estimating the depth from a single camera is not easy, and the visual information captured by the monocular camera is limited. One of the early proposed systems, Tour into the picture (a system proposed by Horry et al. [22]), transforms artwork into a 3D scene using a perspective transformation based on user interaction. Recently, DNNs [23,24] have been proposed to estimate a monocular image's depth information, which is an essential basis in generating stereoscopic content from monocular videos. As one of the state-of-the-art technologies, Rematis et al. proposed a free-viewpoint soccer video system [25] by restoring the player's position and 3D mesh from a single soccer broadcasting video using multiple DNNs.

Research on how to view the generated free-viewpoint videos has been conducted. Before the development of the HMD, the user controls the viewpoint of the generated free-viewpoint videos through the primary input interface, such as a keyboard, mouse, and joystick. In order to improve the usability of these non-intuitive methods, interactive systems that control the viewpoint using markers [26] and multi-touch gestures [27] have been proposed. With the advancement of HMD technology, Inamoto proposed an early-type interactive MR system [28] that displays a free-viewpoint video in the real-world using a video see-through HMD, and the system proposed by Rematas et al. [25] can intuitively change the viewpoint in a free-viewpoint video using a MR HMD.

2.3. Human Analysis

Because deep learning technology has been rapidly developed recently, human analysis problems, which are difficult to solve using conventional image processing algorithms, have been addressed. Person detection [29,30] and pose estimation [31,32] provide an important basis in recording free-viewpoint videos from monocular videos. Semantic segmentation that utilizes deep learning, such as mask R-CNN [33], estimates not only the person's pose but also the segmentation mask at the pixel level. However, it requires a very time-consuming calculation.

Another method is to detect the pose of a person in the image, which is applied to the previously generated 3D human body model. The initial human body model [34] requires a separate network in recognizing the person's pose in the image, and an improved method [35] can fit the 3D body model to the person in the image. Pavlakos et al. [7] proposed a system that can determine the silhouette information of the human body in the image and generates a mesh model based on this acquired information.

In this work, we propose the MonoMR, a simple yet effective system to generate MR content. Compared to existing complex systems, the proposed system can generate MR content from various videos with minimal user interaction; hence end-users without expert knowledge can easily use it. Furthermore, since our approach has a high generalization capability, various types of video (e.g., sports, daily life, surveillance, and more) can be converted to immersive MR content.

3. MonoMR System

As shown in Figure 2, the MonoMR system consists of a personal computer-based scene synthesizer to synthesize content and a HoloLens-based client player to play the content. This section describes how the scene synthesizer generates MR content from monocular videos, and the client player displays the generated scene.

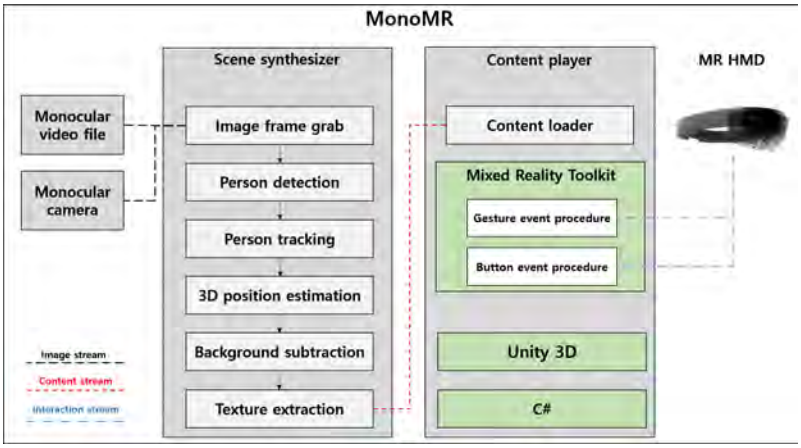

Figure 2. Configuration diagram of the MonoMR system.

3.1. Person Detection and Tracking

As the first step of the scene synthesizer, people in video frames are detected. This procedure was a very challenging problem in computer vision until a few years ago. However, recent dramatic advances in DNNs allow accurate person detection and their body keypoints in a monocular image. We use OpenPose [32], one of the state-of-the-art person detectors to detect persons and body keypoints from video frames.

Our system provides normal mode (656 × 368) and precision mode (1312 × 736) according to the input resolution of the network. In the normal mode, the network detects normal-sized human bodies. However, small people in the video cannot be detected because of the low input resolution. In precision mode, the network accurately detects small-sized human bodies with slow inference speed. Then, the bounding boxes are defined based on the detected keypoints.

Even if the system uses a state-of-the-art person detector, the results may contain false positives and false negatives depending on the quality of the video. In order to detect persons robustly, each detected person should be continuously tracked in subsequent frames. In tracking an object, the system should solve the $X \in \mathbb{R}^{2\times k}$ assignment problem. Therefore, the following equation is minimized using the Kuhn-Munkres algorithm [36].

$$E = \sum_{m}^{M}\sum_{n}^{N}\|X_t^m - X_{t-1}^n\|\cdot C_{mn} \tag{1}$$

$$C_{mn} = \begin{cases} 1 & \text{Person } m \text{ assigned to person } n \\ 0 & \text{Otherwise.} \end{cases}$$

Here, X is a set of k-joints and M and N are the number of detected people in time t and $t-1$. After the assignment procedure, the moving average filter is applied to the coordinates of each object's bounding box to remove the jitter of each tracking object's trajectory. Moreover, the filter can estimate the undetected person's position based on previously observed values.

3.2. Pseudo-3D Position Estimation

In order to capture the depth information of humans in the real world, multi-view stereo vision, depth cameras, and DNNs have been used in existing systems. However, these systems require a complicated configuration or special equipment and are difficult to use. In this study, we propose a simple depth estimation method using the homography matrix [37], calculated based on the detected person's ankle position and minimal user interaction, as illustrated in Figure 3. The system receives the four vertices and approximate distance of the real world between the vertices in the first frame of a video from users. Then, a homography matrix H for mapping the image coordinate system i, j into the real-world x, z coordinate system is calculated. Then, the pseudo-3D position on the real world $X_t(x, z)$ is calculated using the following equation.

$$X_t(x, z) = H \cdot \mathcal{A}(X_t) \tag{2}$$

where $\mathcal{A}$ is the average position of ankles i, j in the X_t set. The moving average filter is applied to the calculated pseudo-3D position to remove the noise.

3.3. Extracting Person Texture Using Background Subtraction

A segmentation procedure is performed to extract the texture of the detected person in the video. Graph cuts [38] and mask R-CNN [33] are the standard algorithms for the segmentation, however these algorithms have high computational complexity. We propose a simple method to extract the person's texture using a background subtraction algorithm for efficient texture extraction.

Given that most videos constantly change foreground and background, the foreground objects extracted with the vanilla background subtraction algorithm based on image difference do not have sufficient quality. We use a k-nearest neighbor (KNN)-based background subtraction method [39], one of the Gaussian mixture model (GMM)-based methods. GMM-based algorithms are robust to repeated, slow motion, and constantly changing lighting conditions. Thus, these algorithms can be used in most videos that have constantly changing foregrounds and backgrounds. Notably, the KNN-based background subtraction algorithm automatically updates the parameters in real-time and selects only the components required by each pixel. Therefore, the processing time of the KNN-based algorithm is reduced compared with the existing GMM algorithms, even with better quality. In addition, the background image without foreground objects can be acquired with this algorithm, and this image can be utilized in the generated content.

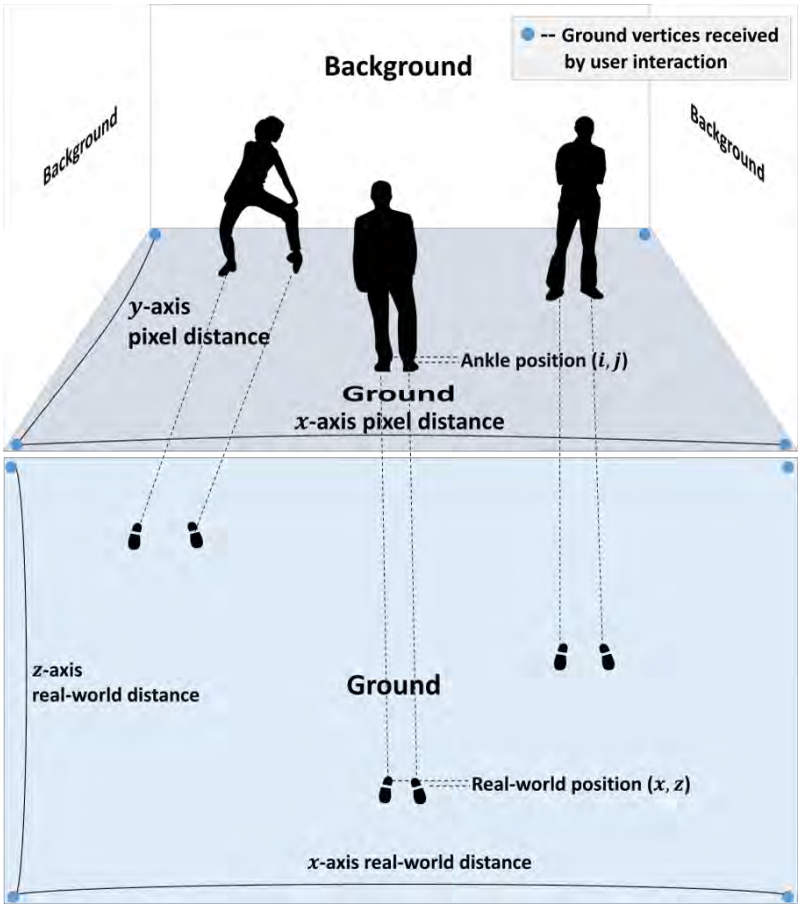

Figure 3. Proposed method for estimating the pseudo-3D position of a person in the image. The ankle position detected in the image coordinate system (i, j) is mapped using a homography matrix to estimate the real-world coordinate system (x, z).

The foreground image extracted by the KNN method may contain noises and holes. In order to remove them, morphological operations are applied to the foreground image. After noise removal, only the moving objects' textures are extracted by masking with the bounding boxes. Figure 4 shows the results of the background subtraction procedure.

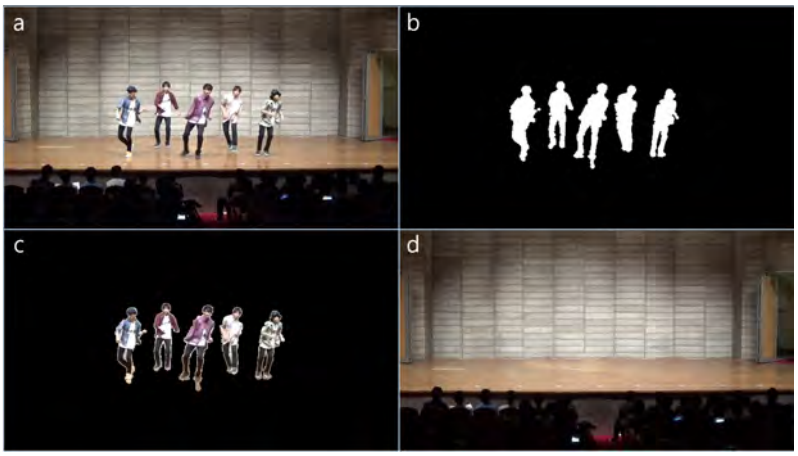

Figure 4. Results of the texture extraction procedure. (**a**) Input image, (**b**) foreground mask, (**c**) foreground segments, and (**d**) background image.

3.4. Texture Size Correction Using Weak-Perspective Projection and Content Synthesis

If the extracted texture is directly applied to the scene, then the size of the same person is different according to the position of the perspective. To minimize this perspective distortion, we use a weak-perspective projection-based correction method. First, the pixel per meter in the real world at the corresponding position of the person is calculated from the homography matrix H and the position of the image coordinate system $X_t(i, j)$. Then, the texture size is recalculated, and the distortion is corrected as shown in Figure 5.

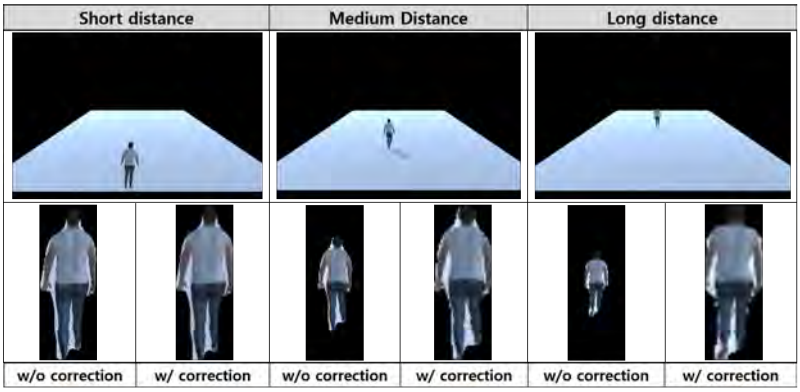

Figure 5. Result of the texture size correction.

Finally, the corrected textures are placed in the 3D world based on the pseudo-3D position of each texture, the extracted background image or a custom image is set to the ground texture, and then the MR content is synthesized.

3.5. Billboard Rendering

The MR content is synthesized by extracting the textures from image frames of a monocular video. Therefore, the camera's viewpoint is fixed, and we cannot obtain the information not captured in the original video (e.g., an information loss on the part not facing the camera), and the user notices the unnaturalness when the viewpoints of the camera and user are different (See Figure 6a).

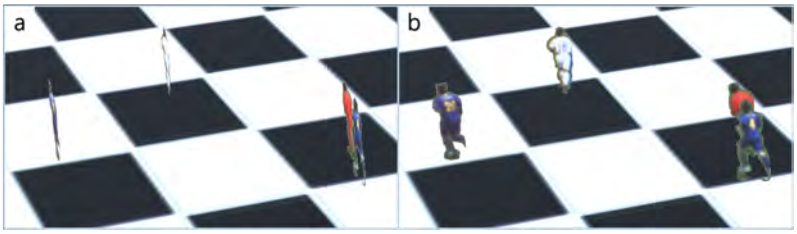

Figure 6. Result of billboard rendering. (**a**) Billboard rendering disabled and (**b**) billboard rendering enabled.

In order to address this problem, the system has a function that applies billboard rendering [40] to every texture, as shown in Figure 6b. Billboard rendering is a simple technique in which the textures are rotated toward the user's viewpoint, thereby reducing the unnaturalness of placing 2D textures in a 3D space. With this function, even if the user changes their viewpoint from the original camera's viewpoint, they notice less unnaturalness.

3.6. Playing Synthesized Content on MR HMDs

The generated content is played using a client player application. The client player is based on Unity and Mixed Reality Toolkit (https://github.com/Microsoft/MixedRealityToolkit-Unity, accessed on 24 August 2021) and runs on the HoloLens MR HMD. The content is displayed in the real world, and users can control the playback, pause, and billboard rendering functions and activation/deactivation through buttons.

Given that the content consists of a minimal number of polygons, rendering and producing multiple textures and polygons in 30 fps even using a standalone MR HMD, which has limited processing power, is possible. The user can enjoy the MR content while changing their position and viewpoints freely, and the content can be placed or resized in the real world through user gestures, as shown in Figure 7.

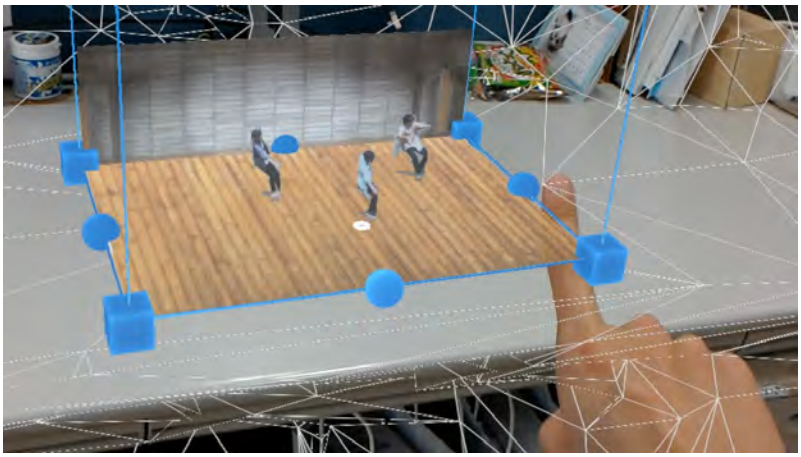

Figure 7. Display of the generated content in the real world.

4. Performance Evaluation

In this section, we performed performance assessments on the depth estimation accuracy, quality of textures, and processing speed, factors that directly affect generated content's quality.

4.1. Accuracy of Depth Estimation

First, we evaluated the depth accuracy of the proposed method. We perform performance evaluations based on two capturing scenarios (small space and large space).

In the case of a small space, the subject freely walked in a square, with a space of approximately 1.7 m in width and approximately 3 m in height, and a Kinect was used to obtain the ground truth data. Five subjects (two female) were the participants in the preparation of the ground truth set. We obtained 600 frames of full HD images and depth information for each subject. The mean absolute error between ground truth and estimated results is 24.57 cm, and the results for each subject are shown in Figure 8.

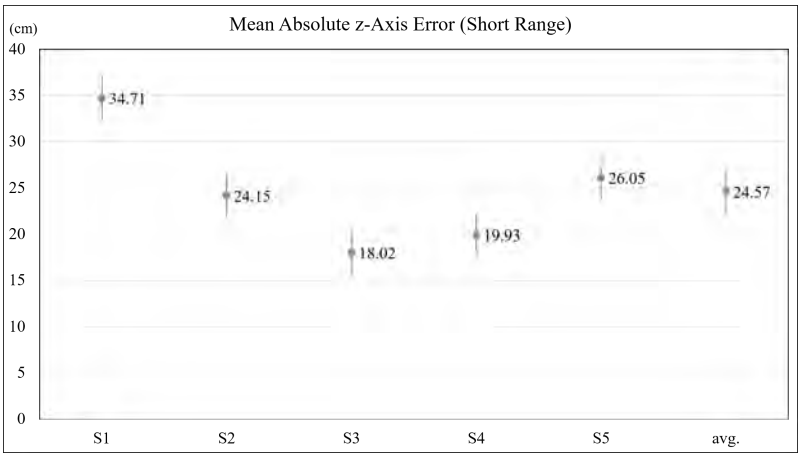

Figure 8. Mean absolute error of depth estimation for each subject in a short-range space.

In the case of a large space, it's not easy to acquire the accurate ground truth from a real-world scene. Therefore, we rendered some synthetic ground truth composed of 3000 frames of images and the depth information using computer graphics. The mean

absolute error between ground truth and estimated results is 76.04 cm, and the results for each distance section are shown in Figure 9.

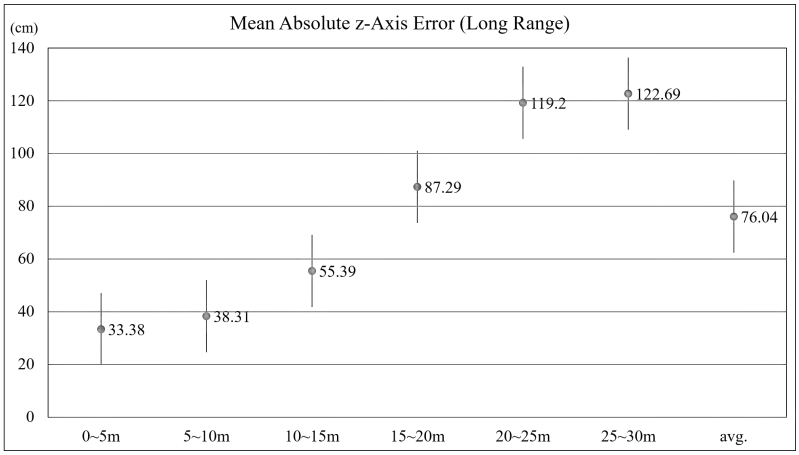

Figure 9. Mean absolute error of depth estimation for each distance section in a long-range space.

At a distance of less than 20 m, depth errors were less than 1 m. As shown in Figure 9, the error increased as the model moved away from the camera because as the distance of the ground truth image increases, the number of pixels that represent the same distance decreases. Thus, the error increases as the quantization error increases.

4.2. Accuracy of Texture Extraction

We evaluated the accuracy of the person texture extraction method applied to MonoMR. We created ground truth data for 500 images through the mask R-CNN [33], one of the state-of-the-art algorithms of segmentation, and measured the mask intersection over union (IoU). Consequently, the obtained average mask IoU is 0.72 (sd = 0.05). Figure 10 is the visualization result of extracting texture through the mask R-CNN and the proposed method.

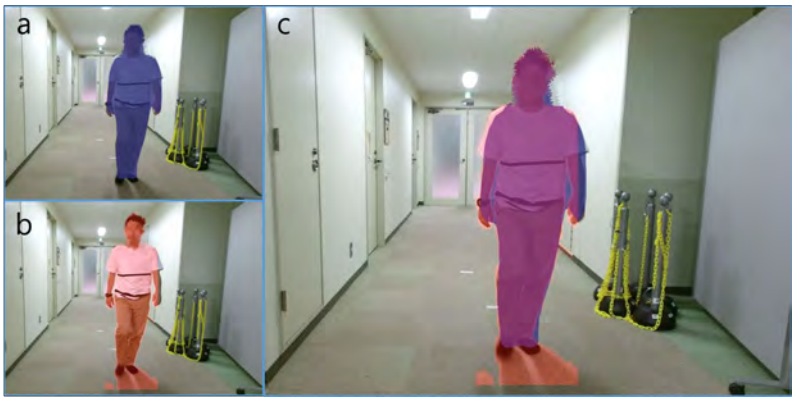

Figure 10. Visualization results of texture extraction methods. (**a**) Mask R-CNN (blue region), (**b**) ours (red region), and (**c**) overlapped two methods (purple region is the intersection area).

The mask image obtained through our method generally does not have significant artifacts. However, compared with the mask obtained through the mask R-CNN, the results of our method may contain noise such as shadow and missing body parts. Nevertheless, given that the proposed method has more advantages than the mask R-CNN regarding processing speed, we applied the background extraction-based texture extraction method in this study to enhance the possibility of real-time processing.

4.3. Processing Speed

We measured the processing time of MonoMR to synthesize the content. All experiments were conducted on a desktop with a Core i7 processor (4-core, 8-thread), 16 GB RAM, and GTX 1080 with 6 GB memory. We used the video sequence (full HD resolution) of ISSIA-CNR [41] (soccer video dataset) as the synthesizing target and measured the average processing time to generate the content from 500 frames of the video. The results are shown in Table 1.

The system consumed approximately 179 ms per frame to generate the content in normal mode; hence, we can confirm that the proposed system generates images at a processing speed of approximately 5 fps (the precise mode for detecting small people has a performance of two frames per second). In addition, the processing speed of the mask R-CNN and proposed texture extraction method was measured, and the results are listed in Table 2. The proposed method extracts textures at a very high speed compared with mask R-CNN, and it can increase the possibility of real-time processing. Based on these results, we confirmed that MonoMR could generate MR content at a relatively fast speed because all the procedures are not mainly composed of DNNs but use a combination of common image processing algorithms.

Table 1. Processing time for each procedure.

Procedure	Time (ms)
Person detection with normal/precision modes	137.83/406.94
Person tracking	0.07
Pseudo-3D position estimation	0.03
Texture extraction with background subtraction	41.43
Total processing time with normal / precision modes	**179.36/448.47**

Table 2. Processing time of two texture extraction methods.

Procedure	Time (ms)
Texture extraction with background subtraction (Ours)	**41.43**
Texture extraction with mask R-CNN	2545

5. Small-Scale User Study

In order to measure the effectiveness of the synthesized content, we conducted a user study with a two-by-three design with the two independent factors content types and display types. Twelve participants (three females; mean age = 26, SD = 9.23) volunteered in our experiment, and they had no experience using MR and Virtual Reality (VR) devices.

5.1. Experiment Design

We attempted to confirm the effectiveness of the content by qualitatively evaluating the following items:

- Depth perception: How much of a stereoscopic degree the user feels in the content.
- Immersiveness: How immersed is the user in the content.
- Attractiveness: How interested is the user in the content.

We used two types of content, sports broadcasting (soccer) and entertainment (dancing), for the experiment. The comparison conditions are as follows:

- C1: Monocular videos displayed on a flat-panel display.
- C2: Monocular videos displayed on a MR HMD.
- C3: Synthesized content displayed on a MR HMD.

Each subject experienced each comparison condition in a random order, and the evaluation was performed using a 5-point Likert-based questionnaire sheet (where 1 = Strongly Disagree to 5 = Strongly Agree). The detailed questions are listed in Figure 11.

Category	Question
Depth perception	Q1. It was easy to recognize a visual-depth of the sports content.
	Q2. It was easy to recognize a visual-depth of the entertainment content.
Immersiveness	Q3. It was easy to immerse in the sports content.
	Q4. It was easy to immerse in the entertainment content.
Attractiveness	Q5. It was attractive to watch the content with this system.

Figure 11. Questions used in the user study.

5.2. Results

5.2.1. Depth Perception

First, we performed the Friedman test [42] for multiple comparisons to assess the differences between the group means in the experimental results. As presented in Table 3, the test result shows a significant difference in the subject's assessment depending on the method. We conducted the Wilcoxon signed-rank test [43] as post hoc analysis to identify which factors have a significant difference, and the result is illustrated in Figure 12. In order to verify whether the randomization controls order effects, the two-way ANOVA test [44] was used to check whether there were significant differences in assessment results between the ordering groups. The ANOVA test shows p-value = 0.163 for sports broadcasting and p-value = 0.589 for entertainment content, which means that the randomization works well because no significant differences were detected across the ordering conditions.

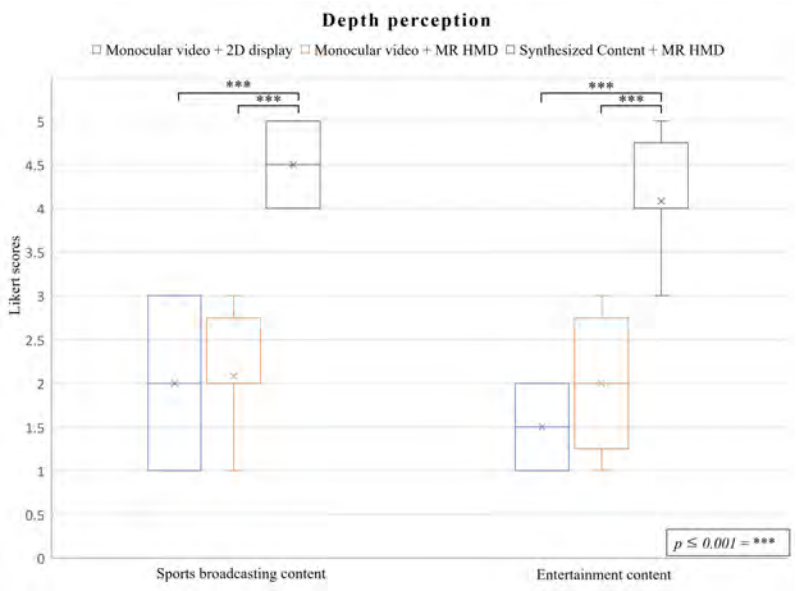

Figure 12. Evaluation of the depth perception for each condition.

Table 3. Friedman test table of subjective score with the depth perception.

Source	Sum Sq.	d.f.	Mean Sq.	Chi Sq.
Method (*)	315.65	2	157.82	46.87
Content	3.56	1	3.56	2.95

*: $p \leq 0.05$.

Based on the post hoc test result, the user can recognize the spatial information through the proposed method C3 more easily than C1 and C2 ($p \leq 0.001$). Hence, we assumed that the content generated by the proposed system allows the subjects to perceive the depth information of the content based on positive feedback. This result showed that the user could feel an improved stereoscopic effect on the content created by the proposed system compared with other experimental conditions.

5.2.2. Immersiveness

The results of the Friedman test and Wilcoxon post hoc analyses are shown in Table 4 and Figure 13, respectively. We can observe a significant difference in the results of the method. The ANOVA test shows p-value = 0.605 and 0.389 for sports broadcasting and entertainment content, respectively, which means that the randomization is valid.

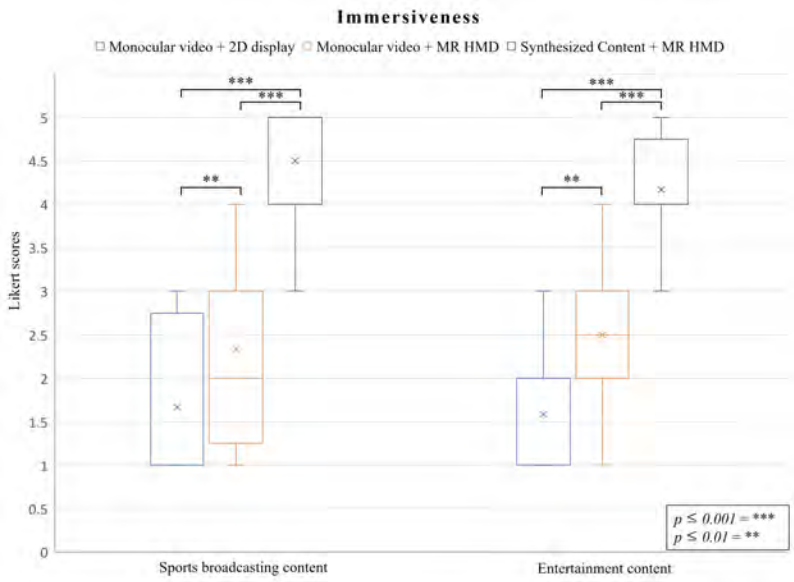

Figure 13. Evaluation of the immersiveness for each condition.

Table 4. Friedman test table of the subjective score with the immersiveness.

Source	Sum Sq.	d.f.	Mean Sq.	Chi Sq.
Method (*)	166.54	2	83.27	24.17
Content	0.13	1	0.13	0.1

*: $p \leq 0.05$.

Specifically, a meaningful difference also exists between C3 and other conditions and between C2 and C1. The users responded that C2 was more immersive than C1($p \leq 0.01$) because the size of the virtual screen of C2 was more significant than the physical screen of C1. The subjects evaluated the proposed method (C3) as the most immersive method ($p \leq 0.001$). We assumed that the content generated by MonoMR could be watched at a free-viewpoint, and this feature affects the immersiveness of users. Based on these results, we confirmed that the proposed system could increase the immersiveness of monocular content.

5.2.3. Attractiveness

Attractiveness was comprehensively evaluated regardless of the content type. The Friedman test and Wilcoxon post hoc analysis results are shown in Table 5 and Figure 14, respectively. The ANOVA test shows p-value = 0.319, and the ordering groups do not affect assessment results.

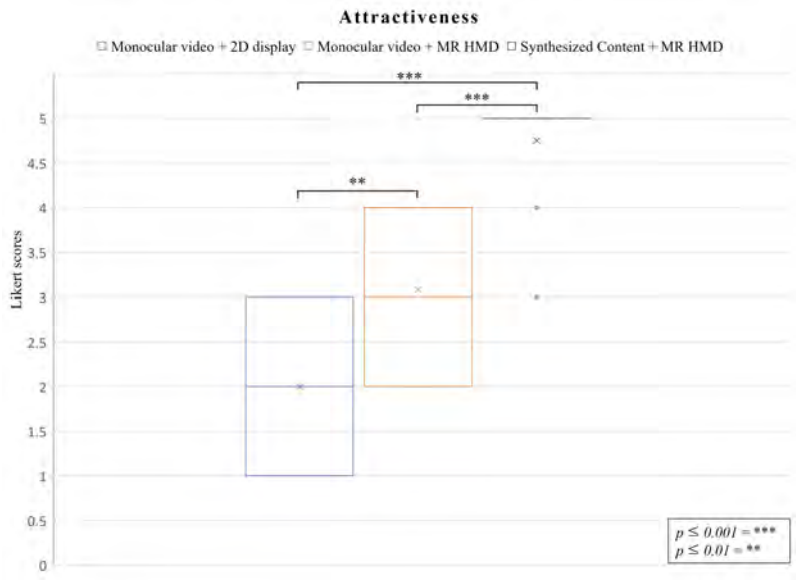

Figure 14. Evaluation of the attractiveness for each condition.

Table 5. Friedman test table of the subjective score with the attractiveness.

Source	Sum Sq.	d.f.	Mean Sq.	Chi Sq.
Method (*)	166.54	2	83.27	24.17

*: $p \leq 0.05$.

Subjects responded that C2 was more interesting than C1 ($p \leq 0.01$). Based on the users' verbal feedback, we confirmed that the reason is that the video player's size in MR HMD can be freely adjusted according to the context of the content.

The subjects reported that C3 provided the most exciting experience among the methods ($p \leq 0.001$). Apart from the questionnaire, we asked the subjects which factor most increased the attractiveness of the synthesized content. Seven subjects answered the improved depth perception, and five subjects answered the advantage of free viewpoint.

6. Applications

In order to demonstrate the applications of MonoMR, we synthesized prototype content from various monocular videos. Because the proposed system can generate content from monocular videos with small environmental constraints, it can be applied to various fields, as shown in Figure 15.

6.1. Immersive Sports Broadcasting

Sports broadcasting is one of the areas where free-viewpoint video systems are most actively applied. Users who watch sports broadcasts want to watch the game from a different viewpoint; hence, systems such as Eye Vision (http://diva.library.cmu.edu/Kanade/kanadeeye.html, accessed on 24 August 2021) have been applied to sports broadcasting to meet these demands. However, most of the existing methods are difficult to set up because these methods require many cameras to be placed around the target and are complicated to use for users who are not experts.

Because the MonoMR system can easily synthesize the MR content from a single monocular camera, the user can create sports content from 2D videos, view the content with a free-viewpoint, and recognize the 3D positions of players intuitively (Figure 15a). In addition, given that the system can generate a single scene from multiple monocular videos, capturing a large stadium, which is difficult to capture using a single camera, is possible by dividing the capture area into multiple cameras and merging the videos into a single content.

Figure 15. Application examples of the proposed system. (**a**) Sports broadcasting and (**b**) entertainment content provide improved stereoscopic effect and an immersive feeling to users than original monocular videos. (**c**) Surveillance systems based on MonoMR allow users to easily recognize situations and spatial information of multiple cameras.

6.2. Dynamic Entertainment Content

Entertainment content, such as performance and theater, is also one of the areas where the proposed system can be utilized. In the case of performance and theater content in DVD and Blu-ray media, the videos have been recorded from various viewpoints. Hence, it allows the user to select and enjoy scenes from a specific viewpoint. As described in the experiment section, the entertainment content created by our system is more attractive than a monocular video displayed on a flat-panel display. Therefore, with our system, the user can enjoy the content while freely changing their viewpoint (Figure 15b).

6.3. Effective Surveillance System

The existing surveillance systems display videos through a single display with divided windows or multiple displays. However, these methods are complex for a user to observe the plurality of screens simultaneously. Mainly, recognizing the place displayed on the monitor is not intuitive. In order to address this problem, a system [45] has been proposed where multiple camera images are synthesized into a single 360-degree image and displayed to a VR HMD. However, the entire scene cannot be recognized within a limited field of view of the HMD, and it's hard to recognize the depth of the target object.

Constructing an efficient surveillance system with the MonoMR is possible because the proposed system can synthesize a single scene from several monocular cameras. We synthesized the four surveillance videos of CMUSRD [46] into a single scene using MonoMR as the demo application of the surveillance system (Figure 15c). The user can intuitively perceive a place where a specific target is located. In addition, even if the target moves from the camera viewpoint to another camera's viewpoint, the user can track the moving target intuitively. Since the system can synthesize content from various kinds of videos, there are many other potential applications besides the ones proposed.

7. Discussion

We present the MonoMR system that generates the pseudo-2.5D MR content from monocular camera videos. The system can render MR content from a large number of previously captured videos of various types. Our system does not require special imaging equipment, such as multiple monocular cameras or depth cameras, similar to most conventional systems and complicated settings, such as the synchronization among the cameras. In addition, because the proposed system can generate MR content using a single monocular camera with minimal user interaction, this has higher usability than any previously proposed systems. Therefore, users without expert knowledge can easily create MR content.

Our system consists of not only a DNN but also uses typical image processing algorithms. Based on the experimental results, we confirmed that our proposed method has reasonable performance for depth information estimation and texture extraction required for producing MR content from monocular videos. In addition, the proposed method is processed relatively high speed, except for the DNN process. Therefore, if a high-performance GPU is used and parallel optimization for image processing is applied, the proposed system can reach real-time performance.

We conducted the small user study, and the results show the feasibility of converting existing monocular videos into more exciting and immersive content with the proposed system. Although the MonoMR system has good performance and usability, we describe some technical challenges and limitations of the system based on the conducted experiment and system implementation.

Limited camera posture. MonoMR does not use a global motion compensation algorithm and the camera posture estimation method using specific landmarks [7,47] because of high computation and low generalization capacity. Therefore, the input video's viewpoint, which is converted into MR content, should be fixed.

Pseudo-3D position. The system estimates the person's x and z positions based on the correlation between the ankle position of the human and the ground. However, if the subject moves on the y-axis, such as jumping and tumbling, it's difficult to estimate the correct 3D position. To address this problem, we will attempt to apply the global depth estimation DNN [23,24] to our system.

Texture quality. MonoMR extracts textures using a background subtraction algorithm to increase the entire processing speed. As can be observed from the previous experiments, MonoMR extracts person textures with acceptable quality. However, if the texture quality reduces due to a detection failure of the human detector or drastic illumination changes in the capturing environment (Figure 16), then some artifacts could be observed.

Figure 16. Artifacts of the extracted textures and content caused by abrupt illumination change, non-detection of body parts, and overlapping people.

Mask R-CNN, the current state-of-the-art algorithm, shows excellent quality; however, as mentioned in Section 4.2, it has drastically reduced the system's entire processing speed. Therefore, this method has not been applied in this study. If a segmentation network with good accuracy and performance is proposed, then we will consider applying the network to our system.

2D texture. The system displays 2D textures instead of 3D mesh models. Hence, it's difficult for the human vision to recognize that the generated model is planar if the user views the content at a certain distance [48]. However, if the user views the content at a short-range, then they feel the unnaturalness. Therefore, we are considering applying methods to recover the information not captured by the camera, such as generative adversarial nets [49] and 3D mesh recovery network [50], as future work.

Large content size. The content consists of the human textures and location data of each frame. The size of the content is large because it does not use any compression method for the real-time operation in low I/O performance of the standalone HMD (e.g., the size of the 30 s of soccer content is approximately 200 MB). We expect that if the HMD's network

bandwidth and I/O performance are increased and the texture compression method is applied, then real-time streaming could be applied without difficulty.

Small-scale user study. We conducted the user study with a small subject group. Because of the small number of subjects in the experiment, the statistical power is insufficient to prove significance. A post hoc power analysis revealed that the effect size and statistical power observed in the user study are 0.74 and 0.75, respectively. Therefore, more than 14 participants would be required to obtain statistical power at the recommended 0.80 level [51]. We plan to conduct a user study with a large number of participants and perform accurate statistical analysis to prove the usability of the system.

Despite the limitations, MonoMR is a potentially powerful system in which anyone can easily convert monocular videos into immersive and exciting MR content. To the best of our knowledge, no system that can convert videos of various genres captured using a single monocular camera into MR content has been proposed yet. In addition, it will be possible for a DNN to be developed and applied to the system to produce better quality content when the hardware limitations have been addressed.

8. Conclusions

This paper presents MonoMR, the system synthesizing the pseudo-2.5D MR content from monocular videos for MR HMD. Our approach can generate MR content from only a single monocular camera or many different videos uploaded on the Internet. In addition, the system requires only minimal user interaction during content creation, and end-users without expertise can easily use this system. Users can enjoy the synthesized content at a free-viewpoint through MR HMD and freely arrange and adjust contents via hand gestures.

We confirmed that the generated content is more immersive and attractive than the original monocular video through user studies. Based on these evaluations, we believe that the proposed system converts a lot of existing monocular content into MR HMD optimized content. We hope that the proposed system will contribute to the distribution of MR content regarding the increase in content demands owing to the commercialization and expansion of MR HMD.

Author Contributions: All authors contributed to the study conception and design. Methodology: D.-H.H.; Software: D.-H.H.; Validation: D.-H.H., H.K., Formal analysis: D.-H.H.; Writing—Original Draft: D.-H.H.; Writing—Review and Editing: D.-H.H., H.K.; Visualization: D.-H.H.; Funding acquisition: H.K.; Supervision: H.K. Both authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JST CREST Grant Number JPMJCR17A3, Japan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study, and written informed consent has been obtained from the subjects to publish this paper.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest, and the funders had no role in this study.

References

1. Collet, A.; Chuang, M.; Sweeney, P.; Gillett, D.; Evseev, D.; Calabrese, D.; Hoppe, H.; Kirk, A.; Sullivan, S. High-quality Streamable Free-viewpoint Video. *ACM Trans. Graph.* **2015**, *34*, 69:1–69:13. [CrossRef]
2. Orts-Escolano, S.; Rhemann, C.; Fanello, S.; Chang, W.; Kowdle, A.; Degtyarev, Y.; Kim, D.; Davidson, P.L.; Khamis, S.; Dou, M.; et al. Holoportation: Virtual 3D Teleportation in Real-time. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16), Tokyo, Japan, 16–19 October 2016; ACM: New York, NY, USA, 2016; pp. 741–754. [CrossRef]
3. Canon Announces Development of the Free Viewpoint Video System Virtual Camera System That Creates an Immersive Viewing Experience. Available online: https://global.canon/en/news/2017/20170921.html (accessed on 16 August 2021).
4. Intel® True View—See More Game Than Ever. Available online: https://www.intel.com/content/www/us/en/sports/technology/true-view.html (accessed on 24 July 2021).

5. Goorts, P.; Maesen, S.; Dumont, M.; Rogmans, S.; Bekaert, P. Free viewpoint video for soccer using histogram-based validity maps in plane sweeping. In Proceedings of the 2014 International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, Portugal, 5–8 January 2014; Volume 3, pp. 378–386.
6. Ye, G.; Liu, Y.; Deng, Y.; Hasler, N.; Ji, X.; Dai, Q.; Theobalt, C. Free-Viewpoint Video of Human Actors Using Multiple Handheld Kinects. *IEEE Trans. Cybern.* **2013**, *43*, 1370–1382. [CrossRef] [PubMed]
7. Pavlakos, G.; Zhu, L.; Zhou, X.; Daniilidis, K. Learning to Estimate 3D Human Pose and Shape From a Single Color Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
8. Ballan, L.; Brostow, G.J.; Puwein, J.; Pollefeys, M. Unstructured Video-based Rendering: Interactive Exploration of Casually Captured Videos. *ACM Trans. Graph.* **2010**, *29*, 87:1–87:11. [CrossRef]
9. Chen, J.; Paris, S.; Wang, J.; Matusik, W.; Cohen, M.; Durand, F. The video mesh: A data structure for image-based three-dimensional video editing. In Proceedings of the 2011 IEEE International Conference on Computational Photography (ICCP), Pittsburgh, PA, USA, 8–10 April 2011; pp. 1–8. [CrossRef]
10. Russell, C.; Yu, R.; Agapito, L. Video Pop-up: Monocular 3D Reconstruction of Dynamic Scenes. In *Computer Vision–ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 583–598.
11. Langlotz, T.; Zingerle, M.; Grasset, R.; Kaufmann, H.; Reitmayr, G. AR Record & Replay: Situated Compositing of Video Content in Mobile Augmented Reality. In Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12), Melbourne, Australia, 26–30 November 2012; ACM: New York, NY, USA, 2012; pp. 318–326. [CrossRef]
12. Mohr, P.; Mandl, D.; Tatzgern, M.; Veas, E.; Schmalstieg, D.; Kalkofen, D. Retargeting Video Tutorials Showing Tools With Surface Contact to Augmented Reality. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17), Denver, CO, USA, 6–11 May 2017; ACM: New York, NY, USA, 2017; pp. 6547–6558. [CrossRef]
13. Kanade, T.; Rander, P.; Narayanan, P.J. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMed.* **1997**, *4*, 34–47. [CrossRef]
14. Kitahara, I.; Ohta, Y.; Saito, H.; Akimichi, S.; Ono, T.; Kanade, T. Recording of multiple videos in a large-scale space for large-scale virtualized reality. *Kyokai Joho Imeji Zasshi J. Inst. Image Inf. Telev. Eng.* **2002**, *56*, 1328–1333.
15. Koyama, T.; Kitahara, I.; Ohta, Y. Live mixed-reality 3D video in soccer stadium. In Proceedings of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality, Tokyo, Japan, 10 October 2003; pp. 178–186. [CrossRef]
16. Kameda, Y.; Koyama, T.; Mukaigawa, Y.; Yoshikawa, F.; Ohta, Y. Free viewpoint browsing of live soccer games. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), Taipei, Taiwan, 27–30 June 2004; Volume 1, pp. 747–750. [CrossRef]
17. Grau, O.; Hilton, A.; Kilner, J.; Miller, G.; Sargeant, T.; Starck, J. A Free-Viewpoint Video System for Visualization of Sport Scenes. *SMPTE Motion Imaging J.* **2007**, *116*, 213–219. [CrossRef]
18. Grau, O.; Thomas, G.A.; Hilton, A.; Kilner, J.; Starck, J. A Robust Free-Viewpoint Video System for Sport Scenes. In Proceedings of the 2007 3DTV Conference, Kos, Greece, 7–9 May 2007; pp. 1–4. [CrossRef]
19. Bogomjakov, A.; Gotsman, C.; Magnor, M. Free-viewpoint video from depth cameras. In Proceedings of the International Workshop on Vision, Modeling and Visualization (VMV), Aachen , Germany, 22–24 November 2006; pp. 89–96.
20. Kuster, C.; Popa, T.; Zach, C.; Gotsman, C.; Gross, M.H. FreeCam: A Hybrid Camera System for Interactive Free-Viewpoint Video. In Proceedings of the International Workshop on Vision, Modeling and Visualization (VMV), Berlin , Germany, 4–6 October 2011; pp. 89–96.
21. Matsumoto, K.; Song, C.; de Sorbier, F.; Saito, H. Free viewpoint video synthesis using multi-view depth and color cameras. In Proceedings of the IVMSP 2013, Seoul, Korea, 10–12 June 2013; pp. 1–4. [CrossRef]
22. Horry, Y.; Anjyo, K.I.; Arai, K. Tour into the Picture: Using a Spidery Mesh Interface to Make Animation from a Single Image. In Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97), Los Angeles, CA, USA, 3–8 August 1997; ACM Press/Addison-Wesley Publishing Co.: New York, NY, USA, 1997; pp. 225–232. [CrossRef]
23. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Montreal, QC, Canada, 8–13 December 2014; pp. 2366–2374.
24. Godard, C.; Aodha, O.M.; Brostow, G.J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6602–6611. [CrossRef]
25. Rematas, K.; Kemelmacher-Shlizerman, I.; Curless, B.; Seitz, S. Soccer on Your Tabletop. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
26. Watanabe, T.; Kitahara, I.; Kameda, Y.; Ohta, Y. 3D Free-viewpoint video capturing interface by using bimanual operation. In Proceedings of the 2010 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video, Tampere, Finland, 7–9 June 2010; pp. 1–4. [CrossRef]
27. Kashiwakuma, J.; Kitahara, I.; Kameda, Y.; Ohta, Y. A Virtual Camera Controlling Method Using Multi-touch Gestures for Capturing Free-viewpoint Video. In Proceedings of the 11th European Conference on Interactive TV and Video (EuroITV '13), Como, Italy, 24–26 June 2013; ACM: New York, NY, USA, 2013; pp. 67–74. [CrossRef]

28. Inamoto, N.; Saito, H. Free Viewpoint Video Synthesis and Presentation of Sporting Events for Mixed Reality Entertainment. In Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE '04), Singapore, 3–5 June 2004; ACM: New York, NY, USA, 2004; pp. 42–50. [CrossRef]
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
30. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
31. Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.; Schiele, B. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4929–4937. [CrossRef]
32. Cao, Z.; Simon, T.; Wei, S.; Sheikh, Y. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310. [CrossRef]
33. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
34. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graph. Proc. SIGGRAPH Asia* **2015**, *34*, 248:1–248:16. [CrossRef]
35. Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; Black, M.J. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Computer Vision–ECCV 2016*; Lecture Notes in Computer Science; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 561–578.
36. Kuhn, H.W. The Hungarian Method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
37. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2003.
38. Boykov, Y.; Funka-Lea, G. Graph Cuts and Efficient N-D Image Segmentation. *Int. J. Comput. Vis.* **2006**, *70*, 109–131. [CrossRef]
39. Zivkovic, Z.; van der Heijden, F. Efficient Adaptive Density Estimation Per Image Pixel for the Task of Background Subtraction. *Pattern Recogn. Lett.* **2006**, *27*, 773–780. [CrossRef]
40. Akenine-Moller, T.; Haines, E.; Hoffman, N. *Real-Time Rendering*, 3rd ed.; A. K. Peters, Ltd.: Natick, MA, USA, 2008.
41. D'Orazio, T.; Leo, M.; Mosca, N.; Spagnolo, P.; Mazzeo, P.L. A Semi-automatic System for Ground Truth Generation of Soccer Video Sequences. In Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09), Genova, Italy, 2–4 September 2009; IEEE Computer Society: Washington, DC, USA, 2009; pp. 559–564. [CrossRef]
42. Daniel, W. *Applied Nonparametric Statistics*; Duxbury Advanced Series in Statistics and Decision Sciences; PWS-KENT: Boston, MA, USA, 1990.
43. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
44. Fisher, R.A., Statistical Methods for Research Workers. In *Breakthroughs in Statistics: Methodology and Distribution*; Kotz, S., Johnson, N.L., Eds.; Springer: New York, NY, USA, 1992; pp. 66–70._6. [CrossRef]
45. Du, R.; Bista, S.; Varshney, A. Video Fields: Fusing Multiple Surveillance Videos into a Dynamic Virtual Environment. In Proceedings of the 21st International Conference on Web3D Technology (Web3D '16), Anaheim, CA, USA, 22–24 July 2016; ACM: New York, NY, USA, 2016; pp. 165–172. [CrossRef]
46. Hattori, K.; Hattori, H.; Ono, Y.; Nishino, K.; Itoh, M.; Boddeti, V.; Kanade, T. *Image Dataset for Researches about Surveillance Camera-CMUSRD (Surveillance Research Dataset)*; Technical Report; Carnegie Mellon University: Pittsburgh, PA, USA, 2014.
47. Yao, Q.; Sankoh, H.; Nonaka, K.; Naito, S. Automatic camera self-calibration for immersive navigation of free viewpoint sports video. In Proceedings of the 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP), Montreal, QC, Canada, 21–23 September 2016; pp. 1–6.
48. Shade, J.; Gortler, S.; He, L.W.; Szeliski, R. Layered Depth Images. In Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98), Orlando, FL, USA, 19–24 July 1998; ACM: New York, NY, USA, 1998; pp. 231–242. [CrossRef]
49. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
50. Choi, H.; Moon, G.; Lee, K.M. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
51. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1988.