Technical Disclosure Commons

Defensive Publications Series

May 16, 2017

Enabling Social Virtual Reality Experiences Using Pass-Through Video

Christian Frueh

Recommended Citation

Frueh, Christian, "Enabling Social Virtual Reality Experiences Using Pass-Through Video", Technical Disclosure Commons, (May 16, 2017)
http://www.tdcommons.org/dpubs_series/522

Enabling Social Virtual Reality Experiences Using Pass-Through Video

Abstract:

Appropriately segmented portions from a video stream are captured by an outward-facing camera mounted on a virtual reality (VR) device and inserted into a VR environment. The outward-facing camera can be the onboard camera on the VR device or it can be a separate camera mounted on the VR device. Because the camera view from the point of view of the VR user is similar to a full three-dimensional (3D) model of the user's environment, rendered from the user's location, the video stream can be inserted directly into the VR environment by segmenting out the relevant pixels and placing them into the VR environment as a 3D object. In this way, a high-quality VR experience, including desired aspects of the user's physical and social environment, can be provided in most settings without expensive 3D modeling or avatar generation.

Keywords: virtual reality, VR, social VR, social setting, group setting, head mounted display, HMD, goggles, headset, camera, head-mounted camera, player perspective, pass-through video, 2D billboard, 2.5D billboard, merging video

Background:

Providing a high-quality virtual reality (VR) experience in a social setting is a challenge because the VR participants cannot easily interact with each other. When playing a game or exploring a VR environment together, the players or explorers cannot see each other in the physical environment because they are wearing VR goggles or a head-mounted display (HMD). Further, when a single player is playing with observers in a group setting, the player cannot see the observers' reactions to the game play.

Current solutions to this problem rely on using an avatar to represent each player or require the interaction to take place in a customized physical environment. Avatar-based solutions lack personalization because it is difficult for the avatar to accurately represent the physicality of the user it represents, especially facial expressions, hand movements, and body posture. To the extent that an avatar attempts to provide this information, detection and display errors may lead to false presentations of even this limited physical and social presence. Further, the avatar-based approach works only when the other players and observers are also wearing a VR head-set or are otherwise connected to the VR environment in a way that enables an avatar to be generated.

Environment-based solutions require a space equipped with multiple cameras and/or three-dimensional (3D) sensors that construct a dynamic 3D model of the physical environment in real time. Because a real-time adaptive 3D display consumes significant computational resources, and often must be professionally installed and calibrated, it is prohibitively expensive for the typical VR user. Even then, these systems typically have problems with video artifacts, poor reconstruction quality, or occlusion problems.

Description:

To address this problem, appropriately segmented portions from a video stream, captured by an outward-facing camera mounted on a virtual reality (VR) device, are inserted into a VR environment. The outward-facing camera can be the onboard camera on the VR device or it can be a separate camera mounted on the VR device. Because the camera view from the point of view of the VR user is similar to a full three-dimensional (3D) model of the user's environment, rendered from the user's location, the video stream can be inserted directly into the VR environment by segmenting out the relevant pixels and placing them into the VR environment as a 3D object. In this way, a high-quality VR experience, including desired aspects of the user's physical and social environment, can be provided in most settings without expensive 3D modeling or avatar generation.

Figure 1 is a diagram that illustrates an example process for enabling social VR experiences using pass-through video. Additional details of the steps are described after the figure.

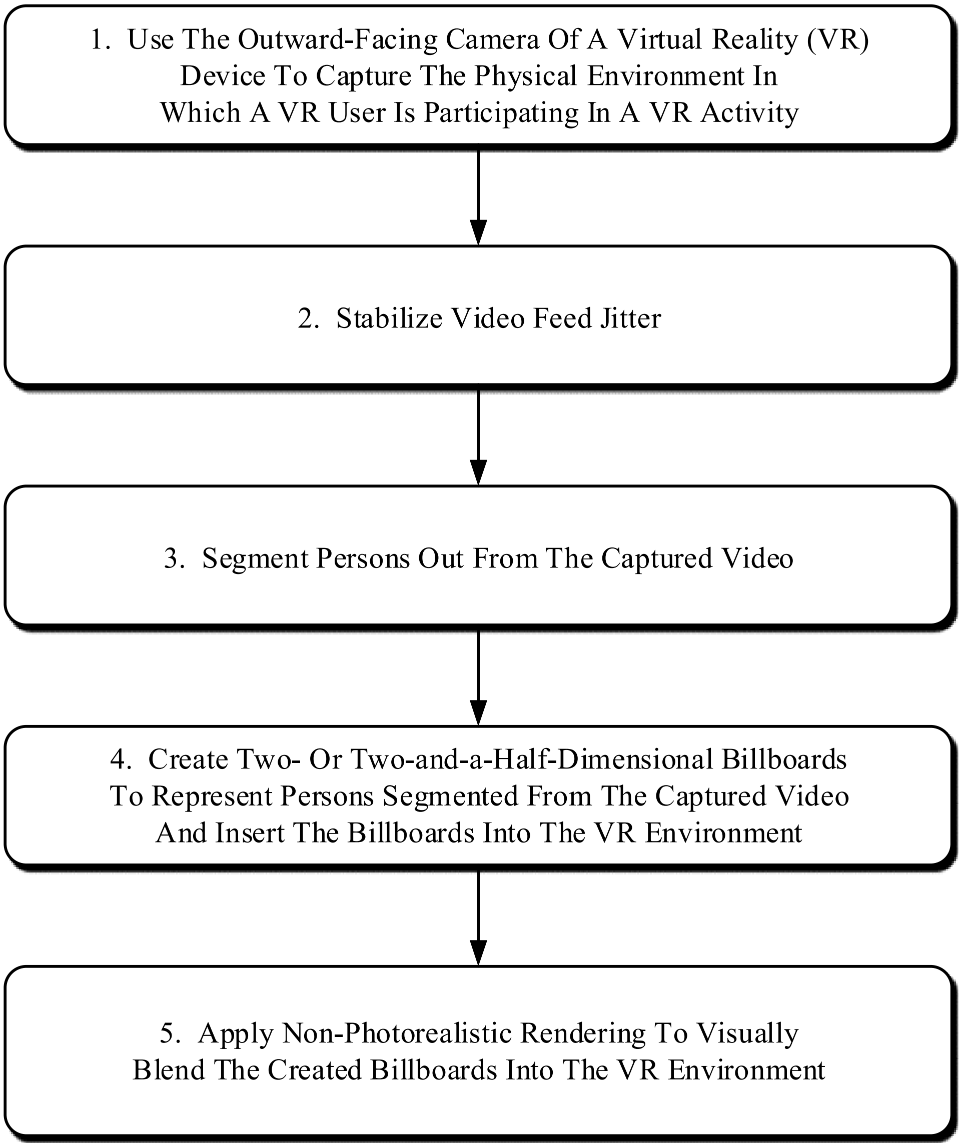

Figure 1

The example process illustrated in Figure 1 comprises five steps. The first step is to capture video of the physical environment in which the VR user is participating in a VR activity. As noted, a camera view from a participant's point of view is similar to a full 3D model of the view rendered from that location. Consequently, an outward-facing camera mounted on the VR head-set can be used to capture the video. The HMD used with many VR systems includes an integrated outward-facing camera, and for systems without an HMD-mounted camera, a small video camera (*e.g.*, GoPro™ or Blackmagic™) can be mounted on the HMD or near it – so long as the viewing perspective from the camera is similar to the user's physical perspective. The captured video stream is appropriately time-stamped to accurately track the HMD pose for each frame. While not necessary, multiple cameras can aid in developing better stereoscopic representations of the physical environment.

The second step is to optionally stabilize video feed jitter (or packet delay variation). Video feed jitter can introduce flickering and other video artifacts, which reduce the quality of a VR experience. In some cases, video feed jitter can be stabilized using the HMD pose information. The HMD pose information describes the three-dimensional position and orientation of the HMD. When the video time stamp and HMD pose information are accurate, the camera feed jitter can be stabilized using a straight-forward pose lookup process. In other cases, small errors and delays introduce jitter in the pose information. This "residual motion" jitter can be addressed using feature-tracking points in the video feed or optical flow techniques similar to those used for video stabilization.

At the third step, co-players, observers, or other objects that the VR user wants to see in the VR environment are segmented out of the video feed. Various segmentation or object extraction methods can be used. One technology involves using a green screen background, as used in mixed reality (MR) or augmented reality (AR) applications. In a typical scenario, in which the VR application is used in a home or office, this approach is impractical. In that case, segmentation can be performed using, for example, foreground-background subtraction or region-growing segmentation using the known HMD pose as a seed. Additionally, if depth information from 3D sensors or stereo cameras is available, it can also be incorporated in the segmentation algorithm. Figure 2 shows an example video frame that includes a person.

Figure 2

Figure 3 shows the person segmented out from the example video frame using the techniques described above.

Figure 3

The fourth step is to create a two-dimensional (2D) or two-and-a-half-dimensional (2.5D) billboard element to represent the person or object segmented out of the video feed. A billboard is a 2D element that is placed in a 3D VR environment at a particular location and automatically oriented so that it always faces the VR user. If only a single outward-facing camera is available, the 2D billboard can be created and located at the depth estimated from the observer's or co-player's HMD pose. In either case, the billboard is then inserted into the VR environment.

If a stereo camera pair or a 3D sensor is available, a 2.5D profile can be sensed and generated. A 2.5D model provides left-eye-right-eye parallax to simulate 3D, and is therefore computationally inexpensive to produce. Further, if the outward-facing camera is close to the VR user's HMD, and reasonably approximates the user's viewpoint, the 2.5D model will be displayed from the same viewpoint from which it is captured and is therefore sufficient to provide a good visual experience. This is in contrast to approaches that need a true 3D model. Because the subject of a 3D model is typically viewed from an angle different from the angle from which the model was captured, true 3D is computationally expensive and can be resolution-limited and artifact-prone. A 2.5D billboard can also be created with a single camera using depth-synthesis techniques based on deep learning (a type of machine learning algorithm that features extraction using multiple layers of nonlinear processors). These techniques, however, typically require additional computing resources.

Figure 4 shows the billboard model of the person segmented out from the example video frame and inserted into the VR environment using the techniques described above. In Figure 4, the insertion has been blended into the VR environment so that the person's legs are behind the building.

Figure 4

The fifth step is to optionally apply one or more techniques for non-photorealistic rendering. Depending on the VR application, non-photorealistic rendering may be used to visually blend the created billboard into the VR environment. For example, a color filter may help the billboard either blend in with or stand out from the VR environment, depending on the application. Feathering techniques can help blend or hide seams and segmentation errors around the edges of the billboard. Non-photorealistic rendering can be used to simulate environmental "special effects" in the VR environment for the billboard. For example, a tinted red color, along with alpha-blending or feathering around the seams can be used to create a hologram-like appearance.

Figure 5 shows the billboard model of the person inserted into the VR environment with an example non-photorealistic technique applied to give the billboard a hologram effect.

Figure 5

Depending on the VR application and the available computing resources, other techniques in the segmentation process could be used to apply other effects or, if desired, to remove the HMD from the model.

Compared with options that are available in typical VR systems with true 3D modeling, the described techniques for capturing and displaying the physical environment of VR users provide similar, though not identical, output. True 3D systems, however, require more sensors, use far more computing resources and streaming bandwidth, and may introduce additional video artifacts because of the complex modeling.