January 2022

PSEUDO-3D VIDEOCONFERENCING USING BACKGROUND PARALLAX

Jonathan Mounty

PSEUDO-3D VIDEOCONFERENCING USING BACKGROUND PARALLAX

AUTHORS:
Jonathan Mounty

ABSTRACT

Current video conferencing technology relies on sending a video stream to a flat, single layer screen, with no depth perception possible, as the image is fixed regardless of viewing angle. This reality results a compromised experience compared to what should be possible with modern day technology. Presented herein are novel techniques through which existing technologies can be applied in order to separate a participant (in the foreground) from a known state of an unchanging background, and then introduce a parallax effect using face tracking of the observer to control the angle of parallax, thus, adding a three-dimensional (3D) effect to an otherwise two-dimensional (2D) video, thereby providing a more engaging and lifelike user experience.

DETAILED DESCRIPTION

Over the past 18 months, videoconferencing solutions have become very familiar to most people, but the ubiquity of videoconferencing has laid clear some of its disadvantages. Often described as providing a 'window' into other peoples' lives, anyone who has used a video product will quickly realize that it is less of a 'window', and more of a flat, featureless picture. While full 3D/virtual reality style video solutions are generating a large amount of attention at the moment, such solutions are still in their infancy, and, such, their adoption, as of yet, is niche and is likely only experienced by a very small proportion of users that have appropriate equipment to operate such solutions.

This proposal provides novel techniques through which some pseudo-3D aspects of video may be realized by users who may not have the luxury of expensive equipment by utilizing a fusion of modern and established techniques such as machine learning (ML) and parallax scrolling. The techniques of this proposal do not seek to replicate a traditional system, nor a full virtual reality alternative, but may help at least provide one important feature missing from contemporary videoconferencing solutions—depth perception.

Although the techniques of this proposal will work in a 1-to-many environment (in terms of participants), provided such a background subtraction algorithm is available and the user experiencing background parallax is on the end with only one participant, examples provided herein are discussed with reference to the 1-1 use case.

The first step of applying the techniques of this proposal involves traditional background subtraction, which provides for the ability to identify the foreground (e.g., the participant(s) features), and separate it from the background. The foreground and the background can both be stored and transmitted separately, with transparency encoded, or alternatively can be sent as a normal video stream, but alongside metadata describing the line between the foreground and the background. Alternatively, the background subtraction algorithm could be performed solely on the recipient end, though this end will have less background information available. For the second step, facial tracking can be used on the recipient's end in order to track the viewing users' facial position with low latency relative to the screen, which provides for the ability to identify how much distortion (or parallax) to apply to the background.

The novel features of this proposal involve separating the foreground and background of the video stream, slightly distorting the background (to allow for a certain portion of the background to be 'stretched' behind the participant), and then moving the background in relation to the foreground by a proportional amount reference to the viewer's head position, as further illustrated in Figure 1, below

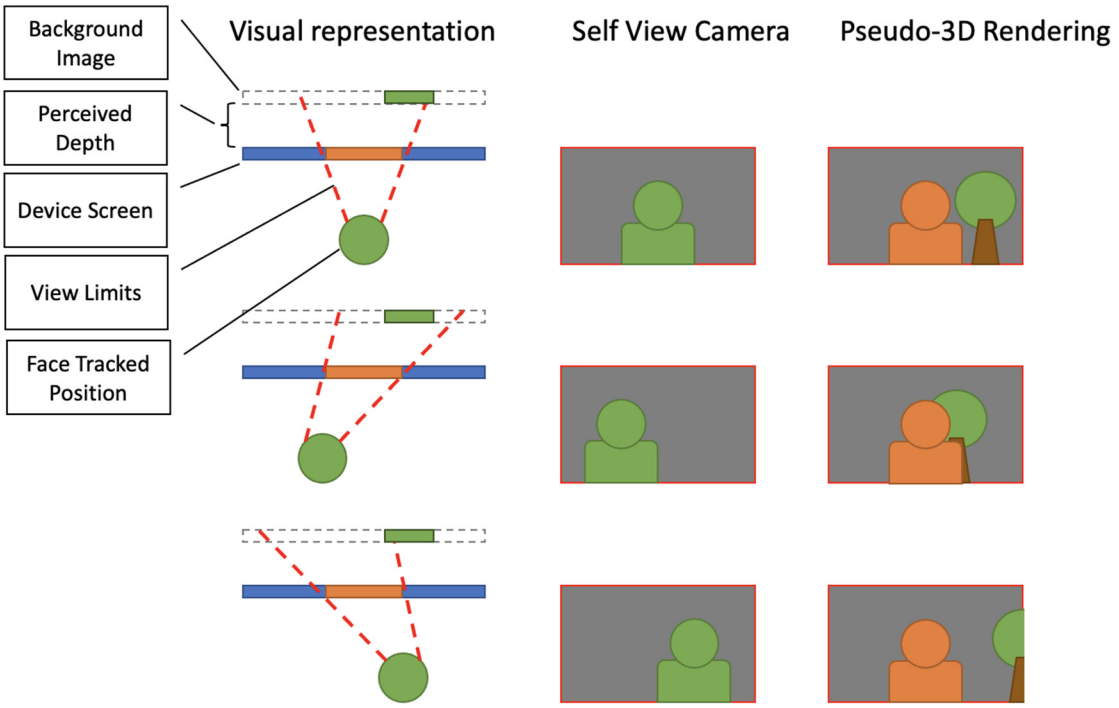

Figure 1: Pseudo-3D Rendering

Such techniques differ from current solutions in that the changes are provided 'full frame'; that is, there is no blank area around the background or the foreground. This is made possible through the background manipulation of 'stretching' the background image larger than the size in the frame, and the parallax effect being introduced by background scrolling rather than rendering the panes in a 3D effect (which also reduces the 'window' effect due to the 3D distortion).

Further experimentation may be involved to determine an effective and believable way of performing this 'stretching' of the background; however, it is possible that advanced machine learning techniques could be used to synthetically extend a background beyond the view bounds of the camera. The use of background scrolling, as opposed to manipulating the image in 3D space could also save resources and eliminate some latency. Furthermore, the substitution of the real background for a virtual background in this scenario may allow an even higher quality background scrolling effect.

One way to implement the techniques proposed herein could involve providing a simple 'multiplier', whereby for every unit a viewer's face view position tracks in one direction in the self-view, the background is moved proportionally the opposite direction, by some multiplier. A larger multiplier will result in the background being perceived to be

further away, and a smaller multiplier will result in a background that is perceived to be closer. It is also possible to achieve the same effect up and down on the other axis.

Of course, despite this visual trickery, the view of the participant and the background will remain two-dimensional and displayed on a normal screen, so there may be a 'billboard' effect visible with no change in angle of the foreground face, however the '2.5D' experience is likely to be novel for the user, and likely to improve the 'window' effect by at least enabling some form of depth in the video.

With the addition of more hardware, it would theoretically be possible to include multiple layers of background, all being moved by different amounts, for a more comprehensive three-dimensional effect (for example a tree 10 meters away would be on a layer much 'closer' than a windmill on a hill 1 kilometer away). In an office setting, this could be achieved through the use of infrared laser measuring devices, such as Light Detection and Ranging (LIDAR), or in a virtual background case, it could be pre-baked into the virtual 'image'. The value, in terms of perceived quality by the user, of the background scrolling effect beyond 2 layers may be reduced.

Another possible hardware change could involve including a dedicated 'fisheye' camera lens with a particularly large field-of-view, which could be used to capture a wider proportion of a participant's background. This would reduce the need to manipulate the user's video as much to 'stretch' the parallax effect.

In contrast to current solutions that provide a '3D style' of rendering (which can be rather off-putting), techniques herein provide a full-frame image and full-frame background, which can utilize virtual backgrounds fetched separately from the server, and may provide a reduction of the main video bandwidth.

A variation on this system may include utilizing a pre-captured representation of the background, which may be practical, for example, on room video conference systems, or static endpoints. Before the conference begins, it should be possible to establish a 'neutral background' containing no subjects of the video. In this case, it would not be necessary to distort the background to recreate the area 'behind' the participant.

This effect would also be completely possible to apply to virtual backgrounds, in which case, there is no need to send a full frame background image, as the client endpoint can apply the effect entirely by itself based on the participant in the foreground. For

example, when using multiple cameras in a holographic or other Augmented Reality/Virtual Reality (AR/VR) system, each of the extreme (left-right) cameras could be used to accurately determine the background behind a subject in order to piece together an accurate background to parallax through, and further, to provide additional information to isolate the subject from the background in order to further improve edge detection around the subject itself. Thus, in such implementations, the techniques proposed herein could also be used to solely send the content of a video subject, with no background information, in which the content could then be used to generate a 3D effect of the subject on a known virtual background. Such features may also have applications beyond multi-camera use cases; as such enhancements could also be used to omit unwanted attendees from being recorded in a meeting without being visible from the far end, as full background information would be available for a portion of the video.

However, it is noted that such holographic or AR/VR type solutions that involve the use of two cameras may not be useful in the context of single-camera applications, such as those involving smartphones and/or webcams, as depth information may be unavailable in such single-camera applications. Thus, such single-camera applications would rely solely on the use of edge-detection of a user's face to separate them from the background, and then artificially move that background 'back' by some arbitrary distance due to the lack of depth information.

The techniques proposed herein could be extended or otherwise enhanced to encompass different features.

It is likely that the rendering operations discussed herein would be best performed at a client end—that is, the same end of the call at which the facial tracking is running, due potential latency considerations. This may also reduce the load on video teleconference servers, and should not be a major performance consideration, as even the most basic mobile phones should have the processing capabilities to perform the techniques discussed herein. In some instances, it may be possible to provide such features with multiple 'performance levels' of facial tracking, which may be varied depending the processing power of a client device, or to offer a server rendering option, accepting the increased latency, on underpowered devices.

In terms of the information sent from the far end, it should be sufficient to supply a normal 2D video stream from a client's device, which, when received by a server, can be separated into the subject and the background. A decision could then be made as to whether to send a new background keyframe (e.g., if the user has moved and/or the background has changed), or to extend the lifetime of the existing background (in which case metadata describing that there has been no update would be sent). In such cases, the subject in the foreground would be sent, which may also significantly reduce bandwidth requirements. However, such an implementation represents only one potential implementation of the techniques discussed herein. Other implementations can be envisioned. For example, it would also be possible, though potentially possibly less desirable to perform this on the far end, and send the background and foreground to the server separately.

In summary, novel techniques are proposed herein through which existing technologies can be applied in order to separate a participant (in the foreground) from a known state of an unchanging background, and then introduce a parallax effect using face tracking of the observer to control the angle of parallax, thus, adding a 3D effect to an otherwise 2D video, thereby providing a more engaging and lifelike user experience.