Microsoft takes you through the steps in HoloLens video creation

July 29 2015, by Nancy Owano

How does Microsoft record holographic video content for the HoloLens? A team has come forward with a video that steps you through the system they use for creating high quality free-viewpoint video that can be compressed for bandwidth fit for consumer applications. They use the example of a two-man traditional Maori performance in their video.

The Microsoft researchers, reported Executive Editor Ben Lang in *Road to VR*, have also published a paper detailing the technique used to record live action 'holographic' video for the HoloLens headset.

The paper is titled "High-Quality Streamable Free-Viewpoint Video" by Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk and Steve Sullivan.

It appears in the ACM Transactions on Graphics journal, which carries proceedings of the ACM SIGGRAPH event in Los Angeles.

In the video they said, "We start by capturing performances with 106 synchronized RGB and infrared cameras on a calibrated green-screen stage. We subtract the background to compute silhouettes, then schedule the data for processing."

The first step generates a 3D point cloud starting with stereo depth maps from RGB and IR pairs. Points from the depth maps are merged and refined locally. Then the cloud is refined globally using a multi-view stereo algorithm.

The next step creates a mesh per frame. They modify surface reconstruction to produce meshes constrained by the silhouettes. Typically, these meshes have topological artifacts and spurious components.

They apply topological de-noising as part of their approach for cleanup. They reach a stage where they have 1 million triangles per frame.

They next see which areas contain perceptually important details such as hands or faces. They preserve the geometry of texture in these areas; the example in the video is reduced to 20,000 triangles per frame.

They said they establish temporal coherence by choosing mesh key frames and working on them to fit ranges of the performance. .

The team mentioned a split into four subsequences. The final step is to enwrap the meshes, generate a texture atlas, then compress and encode the data in a single streamable file at our target bit rate.

If the video announcer does not convince one that this process is no easy stroll in the park, then Ben Lang's comments strengthen the understanding that this feat is impressive. Lang said, "Inserting a computer-generated object (say, a cube) in AR is simple enough as new views of the object can be computed on the fly as the user moves about. Capturing and playing back a live-action scene in AR is however quite a bit more challenging as the subjects of the scene not only need to be captured from every angle, but also need to be extracted from the background of the capture space such that they can be transplanted into the user's own environment for convincing AR.

Lang described what is special about the Microsoft approach. Motion capture systems have been doing something similar for years, he said, but "most systems only capture the motion of performers which is then used to animate digital models." Essentially, Microsoft's technique captures both the performance and generates a model at the same time.

More information: dl.acm.org/citation.cfm?id=2809654