Breakout 1.1: Characteristics of realistic Digital Avatars – ARPA-E questions

Tuesday, April 26, 11:00-12:30

Objective: Imagine a digital telepresence avatar in the year 2030 that is fungible with in-person communication, eventually offsetting >1 quad of transportation energy. Describe the avatar's key characteristics for capture/control, data back-end, and rendering/display.

Avatar Capture & Control

Effective telepresence avatars will need to be as easy to use as Skype/Hangouts/FaceTime and as low-cost as other laptop peripherals. The state-of-the-art systems for avatar capture are bulky, require special makeup, and have awkward camera and lighting rigs (see image).

How hard will it be to create the target system? Is it an engineering challenge, research building on existing theory, or a theoretical research challenge?

1. What existing theory is foundational?
2. What physics-based hard-limits will constrain capabilities of the target system? E.g. light-fields illuminating and capturing subject, optical resolution, ambient lighting
3. Can estimators or other techniques be used to circumvent these constraints?
4. What does the path forward look like for real-time (<10ms) processing?

Avatar Data Back-End

The current structure of the internet (TCP/IP) struggles to handle even today's real-time communications due to limited bandwidth, long & irregular latency, and small packets.

- Describe **changes to the internet** that will simplify and loosen constraints on telepresence avatars. Who will be implementing these changes?

Describe industry roadmaps for **codecs and real-time routing** and streaming technologies, and possible off-roadmap advancements. Be sure to include key parts of theory and leaders in the field.

Avatar Rendering

Assume that the $10B+ investments in VR/AR displays and processing hardware achieve all imaginable goals. What will be the specifications of these devices and how will those specifications affect real-time rendering and display of Realistic Telepresence Avatars?

Rendering engines: What are the challenges that will face avatar rendering technology?

- Will evolutionary advancements to graphics engines (e.g. Unreal, Unity) be sufficient?
- Will physics engines be a necessary part of the engine?
- How will bandwidth requirements affect rendering methods for avatars?

Information Modalities: Describe characteristics required for the following aspects of avatars to be 'realistic': Visual (2D vs 3D), Haptic (contact vs force), Auditory, Olfactory.

Computing Hardware: What will CPU/GPU architectures and capabilities look like for future telepresence avatars? Can avatar technologies on one architecture be ported easily to another?

Displays: What characteristics of the **perfect** version of the following displays would affect the design of telepresence avatars?:

- 2D wall mounted,
- Desktop 'holographic' (light-field, Pepper's Ghost, etc.)
- VR head mounted display
- AR head mounted display
- AR wall mounted retinal display

General Questions

The Value Add: What characteristics of Digital Avatars would make them preferable to in-person communication? What would this mean for the future of workplace interaction?

Avatars & Teleoperation: How does everything we talked about in this breakout relate to tomorrow's discussion on telepresence technology to effect the physical world?

Breakout Session 1.1: Feedback

Avatar Capture and Control:

- Realistic behavior is more important than realistic looks (e.g. no need to see a razor cut on somebody's face).
- Can do a one-time, high-resolution scan/model of yourself once, and then in real-time do coarse motion/expression capture to animate the high-resolution avatar.
- Whatever the system is, it needs to enable mobility of the person within a room, which is something that the gaming/cinema industry isn't too worried about (yet). Groups are using optical, acoustic, and accelerometer based suits for motion tracking within the room.
- Don't be in a rush to replicate physical world. No need to impose constraints of physical world on the digital.
- Can procedurally generate an avatar in 5 min now, but requires a 100s of cell phone cameras and raspberry pis (costs around $10,000)

Avatar Rendering:

- Rendering isn't solved, and it is really hard and all of the gaming/cinema people are spending billions already to solve this problem.
- Asymmetric Collaborative Environment: when the space you are actually in is different from the virtual space you are in.
- Trust must be considered, for example what happens if we enable puppeteering? Also trust in that a remote expert is able to properly see and experience the situation that you are actually in.
- Almost impossible to render the inside of a mouth properly
- We aren't that far from being able to render facial emotion and capture on a realistic avatar. The path forward is just a reduction in the time and bandwidth and computing power it takes to render.
- Voxel-based rendering instead of triangular rendering. Maybe there's some convergence—if you're scanning people you are developing data sets that are not triangles.
- Variable resolution imaging—high resolution of facial expressions, lower resolution of less critical areas.

Other Questions:

- Visual and auditory stimuli can drown out or make you believe you've experienced haptic feedback.
- Latency in operator head movement is a big problem, especially in telepresence robotics.
- 250 ms latency is too slow. <100ms should be the goal.
- Microexpressions - we don't quite know how many points we need for this to be effective. Study last week – map 15-25 points on the face, can get a lot of expression data out of it.

    - What is needed is knowledge of the underlying emotions, not the microexpressions per say.
- For creating Paul Walker in The Fast and the Furious, goal was to try to trick audience. Took Petabytes of data.

Avatar Data Backend:

- Latency is of utmost importance, use latest UDP packet over retransmitted TCP packets with high latency.

Other Challenges We Haven't Covered

- Synchronization between audio and visual cues is both tough and extremely important. We need to first focus on determining what information is important to convey for particular use cases.
- AI to distill what data is really important to show. When someone drops something, should ignore and filter out the sound. When someone making subtle cues that they want attention, system should know. How does system know when it should isolate, include, etc. Depends on context.
- Spatialized audio - capturing someone moving through an environment. Audio degradation is more distracting then video degradation. Studies have shown this. If there's any latency in audio, people won't tolerate it.
- Smart avatars that can serve as a proxy, only signaling you in when they can't engage itself. Have some autonomous behavior and for people to trust that autonomous behavior. May help you get around the latency
- Displays: there are already $3B of investment in displays—what can $3M do? Only group looking at displays is Magic Leap. Whatever solution Magic Leap has could be one of the solutions. They are focused on glasses that fit on eyes. Wall-sized holographic displays. Very desirable. Nobody in the industry is really looking at this.