HUMANIZING VIRTUAL PEOPLE

MATTHIAS WITTMANN

When we talk about virtual humans in real-time space, we often refer to them as Avatars. However an Avatar is only a digital representation of an actual existing person, driven by that person. Autonomous virtual humans on the other hand are virtual beings that are controlled by "themselves" through artificial intelligence (AI).

Currently there is a lot of talk about AI and we constantly hear buzzwords like "Machine Learning", "Neural Networks" or "Deep Learning". But what is AI? One definition is the theory of developing computer systems able to do tasks that normally require human intelligence.

A lot of AI is being used for understanding processes and then using that knowledge on other merely similar (but not the same) processes. Facial recognition is a good example: Once an AI system has been "trained" on enough human faces it will be able to recognize a human face on a picture, even if it hasn't seen that particular face before. This is possible because during training the AI system is learning more abstract information about these images. For example: humans do have two eyes, a nose between but below the eyes, one mouth centered under the nose … and so on. So what happens if a Zyklop is in the picture? The cool thing with AI's is that they never just give you results based on "yes" or "no". They return probabilities. And they do that very well. So when an AI sees a Zyklop it will most likely return something like a 65% probability that it is a human face. This behavior by itself feels very human. Of course you can also get wrong results from an AI. Its skills to interpret unknown information depends on the complexity of their "training".

So why don't we already have perfectly AI driven robots or virtual humans if AI's are so great? That's simply because reading and interpreting is easier than producing. Reading an emotion from a person's body language or face or even from the content of their spoken words is much easier than coming up with an appropriate emotional reaction: The right body or facial pose, motion or gesture. And that is simply because the training of reading is easier than the training of doing. Reading emotions versus creating emotions. For me it is all about creating emotions. This article will illustrate where I was in that process in 2019 and how I got there.

After receiving my diploma from the Filmakademie Baden Württemberg in Ludwigsburg, Germany, I started in 2000 as a character animator in the movie industry. In 2008 I became the animation lead of "The Curious Case of Benjamin Button". Working a lot with the modeling department I became something like an expert for human facial anatomy, facial muscle movement and skin micro motion. Over the next few years I was mostly involved with digital human animation and development like "Tron Legacy", "Virtual Tupac" (the Coachella hologram) and "Maleficent".

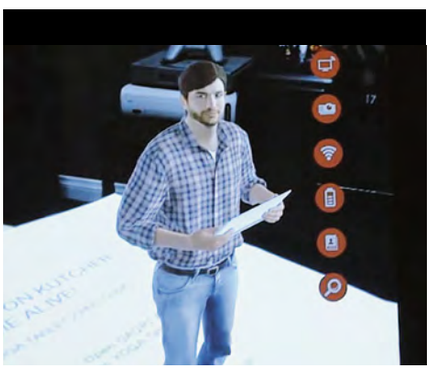
While working in animation I always spent some time on coding tools. Not only because it was helpful, but also because it was a lot of fun! I am not a schooled programmer, so writing code for me is like playing a game. If it works, I won. And as opposed to

animating, coding gives you a clear result. Either it works, or not. Sure, you can optimize code to work more efficiently (which is fun too), but the results stay the same. Animation on the other hand is always open for interpretation.

In 2010 I started using Unity at home to create little games for my iPhone. And while I mentioned before that coding means fun to me, coding for real time engines is really exciting. Four years later in 2014 a friend of mine asked me if I would be interested in joining a company to develop interactive virtual humans for Augmented Reality (AR). That is when I switched from film productions to realtime development.

VIRTUAL HUMANS IN AR

The first project I was involved in was a 30cm virtual Ashton Kutcher presenting a new tablet right in front of you on your desk. You could watch him in AR through your own tablet. I added a "look at us" logic to make him, well, look at us whenever he talked to us. My knowledge from high-end film productions came in handy even in the low-res world of mobile real time. It was important to me that the way he looked and moved would feel natural, which was tricky considering he consisted of only about five polygons.

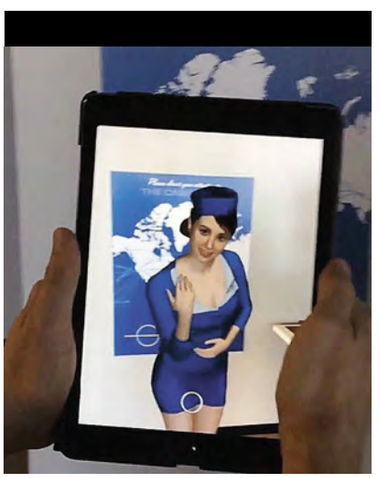


30CM VIRTUAL ASHTON KUTCHER

The next AR project utilizing a tablet as a window to the augmented world was a life size stewardess promoting an airline. She did not have to talk but be friendly and invite us to the airline's booth at a convention. Aside from a 'look at us' mechanism I added reactions to her based on our distance. When we were in reasonable range she would beckon us and if we got closer I let her smile.

Both of these AR projects felt pretty successful from an emotional standpoint. But now I became curious about how far one could go with Virtual Reality (VR). For a start one could use better graphics since serious VR systems were still connected to a PC. And how immersive would it feel sharing the same three dimensional space without just a tablet as a window?

First I tested it on an old film asset: a human head with a few really well modeled shapes for eyelid deformation. It was the first time for me using Unreal Engine and I programmed the eye and head rotation in a way that the head would look in my direction in VR. The results felt mind blowing realistic. Of course, it did not look photorealistic, however the presence of seemingly another being was undeniable.



A LIFE SIZE STEWARDESS PROMOTING AN AIRLINE

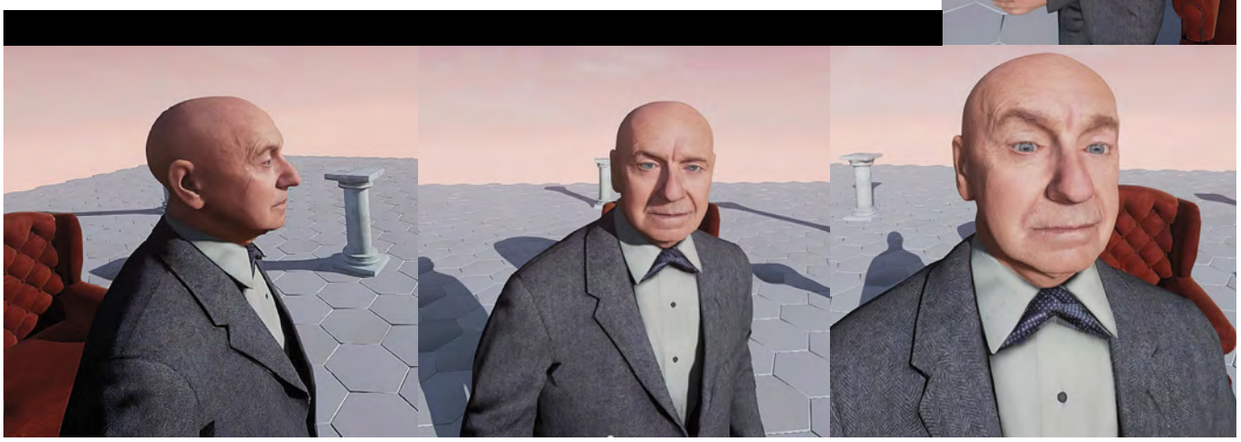
THE SESSION

This confirmed that one could go much further with the given technology and in 2015 for my next project "The Session", I developed a full-size virtual human in VR: An old professor. This time I let the character speak to the player whose reactions and decisions determined how the story proceeded.

To establish communication between player and virtual human I decided to go with gesture recognition. The professor was able to read the player's nodding and head shaking as "Yes" and "No". All AI features were developed based on fuzzy logical decision trees. The whole system was supposed to (technically) fake human behavior.

The professor was able to memorize objects and could also forget them if he would not see them for a while. Just as humans favor certain features in the images they perceive of the world around them (motion, contrasts, vanishing lines) this vision AI was supposed to do the same. Although I did not get to implement the other two elements, the Professor was able to notice motion, especially if it was fast or sudden. While he was not able to notice static objects in his peripheral vision, he did notice them when they moved. He also was aware of the fact if the player was looking at him, but only if he was looking at the player at that moment.

THE SESSION



I did not create a locomotion logic at this point, meaning the Professor would not be able to avoid a player and could potentially walk right through them. But he nearly never did since the story was designed to predict where the player actually was. the AI only walked into spots where it knew the player could not be. For example: The Professor told the player to go to a pillar in the room. If the player did not go there the AI would not change location. However, if the player went over to the pillar the AI would know that the player just freed up space and the Professor could move without the danger of intersecting with the player.

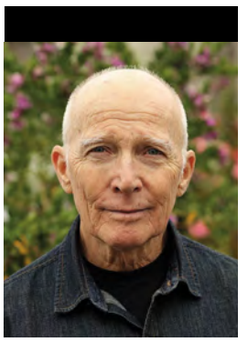
Aside from the actual game part I also created a test-level where the user could just "hang out" with the professor and take a closer look at him. While he would not move or engage in a conversation with the player, he still would react to their presence.

The professor would look at them when they entered his field of view, he would nod or shake his head when the player did and would react confused and even annoyed when the player got too close. These reactions felt very personal giving the player the impression that this being had real feelings.

MAN AT A BUS STOP

In 2018 I got the chance to design and create an even more advanced interactive virtual human in VR. The project was called "Man at a Bus Stop". I had several goals this time. One was having a way more complex environment, to showcase that a complex interactive virtual human in VR could be used in an equally complex environment without overwhelming the computational power of a decent, yet not overly hi-end PC. Other goals were adding hair, voice recognition and developing an emotion engine. Also, I decided to update the user interface giving the player hand representations this time.

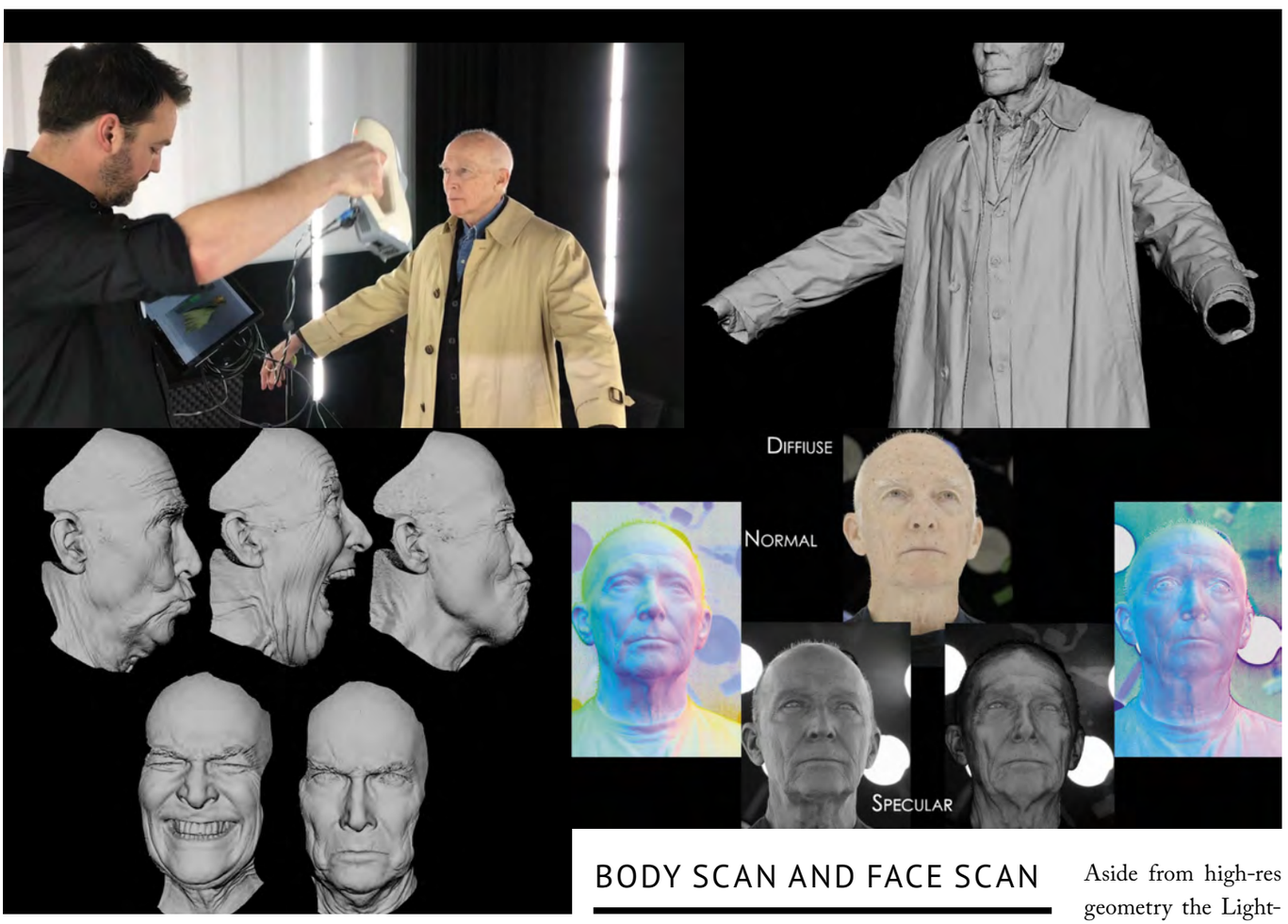
The user experience was based on two principles: firstly the character AI should be as humanlike as possible not only in terms of look and behavior, but also in abilities. This AI should not be all-knowing. The human-AI should only see, hear and feel what a human would. No eyes on the back of the head… Secondly the user should not have to "learn" anything. Anything they would be able to do should come naturally to them.



MAN AT A BUS STOP IS BASED ON A REAL ACTOR, TOM FITZPATRICK.

Like "The Session" before, "Man at a Bus Stop" is based on a real actor, Tom Fitzpatrick. We started as we would for film productions. We did an extensive body scan. This time we experimented with a new method using a mobile hand scanner solution. It was based on photogrammetry but would be moved around the subject. The actor had to stay still for a longer time than in a classical volume, however its flexibility was a huge advantage.

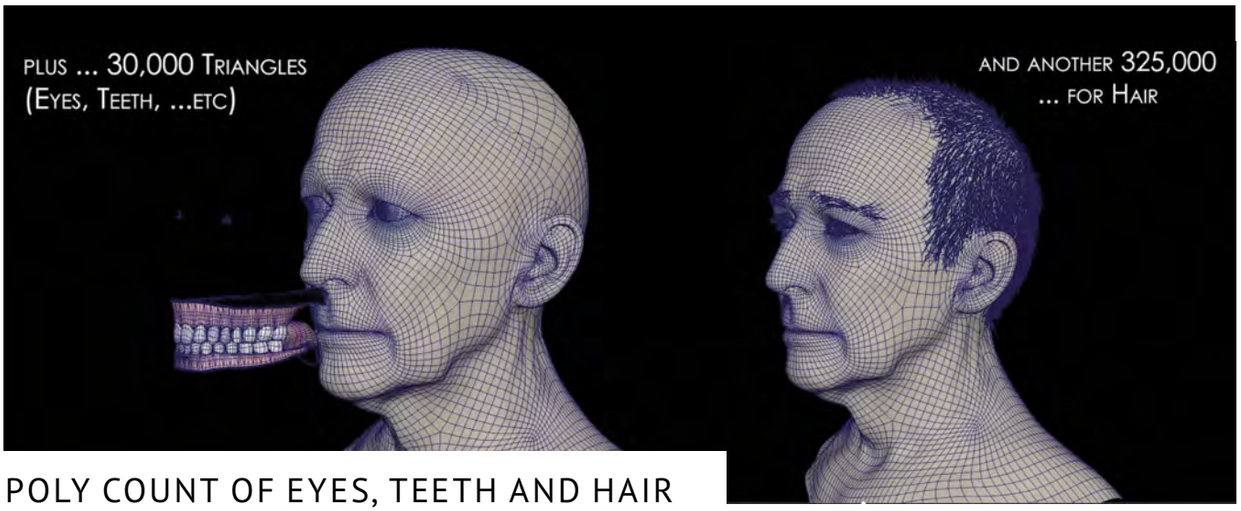
The face was scanned on a light-stage, a system originally invented by Paul Debevec at the Institute of Creative Technologies (ICT) at USC. This process creates geometry that is so precise, it shows every single pore of the actor's skin. The resolution of such scans is very high, 7.5 million polygons in this case. While one would not use the high-resolution geometry for real time, one still gets the necessary information for creating highly detailed displacement maps. Displacement maps can give a low-res geometry the look of a much higher res one, while needing less resources. For that we remodeled the head with a much lower poly-count of 30.000. Then we used a modeling tool that compared the two heads. This produced a black and white image that represented the delta between high res and low-res geometry: the displacement map.



BODY SCAN AND FACE SCAN

Aside from high-res geometry the Light-stage scan produces also other very important maps that are needed for shader creation like defuse, specular and normal maps. To create realistic looking skin, we utilized Unreal Engines ability for sub surface scattering (SSS) shader models. Skin reacts strangely to light. It absorbs a lot of it and reflects it in a more diffuse way. Without an SSS shader it is very difficult to create this look. That is why digital humans looked like "plastic" for a long time. In real time graphics they often still do. The main reason is that even though this kind of shaders do exist now, at least in some game engines, they need more rendering power, which usually results in lower frame rates.

Since you cannot just scan hair, it had to be created separately in Maya using XGen. Luckily Tom had very short hair, so we did not have to deal with dynamics. Nevertheless, in the end the poly count of just the hair was ten times higher than that of the rest of the face, including eyes and teeth.



POLY COUNT OF EYES, TEETH AND HAIR

At this point we should talk briefly about the Uncanny Valley, the effect that we subconsciously accept simplified representations of humans more than representations that are only "almost perfect". The question is, how to overcome the uncanny valley? What makes it uncanny? There is no simple answer. Anything can pull you out of the illusion, look, motion and even the voice.

Let us assume we start from a really good 3D model of a person which has already decent shaders. Most of the time the trouble starts with the eyes. They are, as some may say, the gateway to the soul. Poorly designed eyes can be disturbing even when the character is not moving at all. Motion is another factor. The most perfect looking eyes will still break the illusion if they do not move and deform the surrounding skin like we are used to. And then there is the rest of the face. A human face is very complex and to recreate it with computer graphics is only accepted by the audience if it moves as a whole. If you only pose the mouth to the spoken words but do not move the rest of the face accordingly, the illusion falls apart. Everything in the face, and the whole body for that matter, is interconnected. There are specific ways how our head moves while we are talking, and there is a way how our body moves when our head is moving, and vice versa. If you break any of these relationships, you run into danger of falling deep into the uncanny valley.

Once things are moving, deformation and therefore volume preservation becomes a factor. When pulling a rubber band, it becomes thinner. The same basically happens in the face, you pucker your lips, the cheeks are getting stretched and flatten against your teeth. Moving parts of a face and not taking care of volume preservation will be subconsciously noticed very quickly. There are even more possibilities to break the illusion, yet these belong to the most important ones. If not taken care of properly you get reactions like "Something is off, I can't tell you exactly what it is, but it's not real".

To make sure to have proper facial motion and deformation I went with DI4D to do a facial performance capture. A volumetric scan like that returns a deforming facial geometry of the complete performance. Such performance clips are very large and it was clear that I would not have the bandwidth in VR to run it on high frame rates. So we used an algorithm from an old siggraph paper to convert the complete performance into a small number of animated blend shapes. However, blend shapes and animation created through a process like that can't be edited. The data looks too random. We still managed to enhance the rig to allow for adding emotions and moving his eyes interactively.

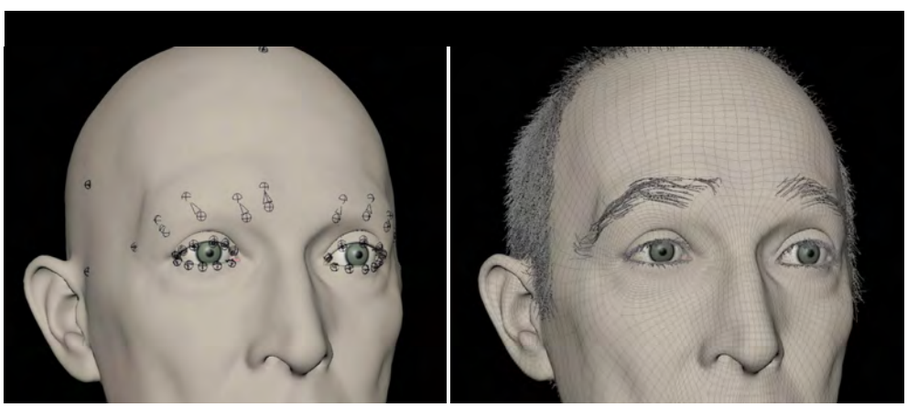
At that time high quality scans like these were not possible to be captured simultaneously with body performance. So we did a separate body motion capture session of the same actions. It is not an easy task for an actor to reproduce the same action twice, once sitting in a chair and once moving on stage. Ideally the body capture should happen before the face capture session and both should happen on the same day. That way it is easier for the actor to understand and remember the context when he has to sit still. And believe me, it is a lot to ask from an actor to perform 'lively' without moving their body.

FACIAL AND BODY PERFORMANCE

Now we had two separate captures. To connect the facial and the body performance, I spent some time in editing. It was important that his facial rhythm would fit to his body motion. One could not just line up the voice recordings of the two captures, since length and rhythm was always slightly different. So the key was to find the most important 'beat' in his performance, and line it up to that.

Once again the hair had to be dealt with separately. A performance scan moves the skin, but the hair geometry does not know about that. So we had to design a special joint setup that would move the hair together with the head. Aside from the scalp that also included the eyebrows and eyelashes.

In order to make a character feel naturalistic within an experience like "Man at a Bus Stop", multiple layers of interactivity come to play. First there is the actual story line. To give the illusion of choice, the story has to branch from time to time. Branching often feels better for the user, but also creates more work. Key to an efficient branching design is that it leads back to the main track as soon as possible. Else the amount of options will increase exponentially.



A SPECIAL JOINT SETUP THAT WOULD MOVE THE HAIR TOGETHER WITH THE HEAD

Aside from this obvious kind of interactivity I added layers of behavior to make Tom feel more like a real human. For once there are actions the character can do simultaneously with the main performance. Turning towards us when we are not right in front of him is one of them, or leaning away from us when we get too close. These motions have to be well balanced. The eyes, head and body turn towards us with individually tuned speed, acceleration and precision. There are also more subtle actions to keep him "alive", that react interactively too. His breathing and eye blinking frequency change based on his current mood. So does his mimic.

Like before with the professor for "The Session" it was a no-brainer to add emotions to the character again. Though this time the emotion engine was more elaborate. Rather than changing the face and pose just based on e.g. distance, Tom would now remember his own emotions and at times even holding a grudge for a while. That indeed was a very powerful update that added a lot to his "humanlike" feel.

ADDING VIRTUAL HANDS

In terms of game play and controlling the action, dealing with the interaction of humans and virtual humans is challenging. Since players in VR are sharing the space with the character it is harder to predict what they may do. Some people new to VR, are often overwhelmed and have a hard time focusing on the story. While others are quickly bored if there is not enough "action". Some players try to break the system, others do not move at all.

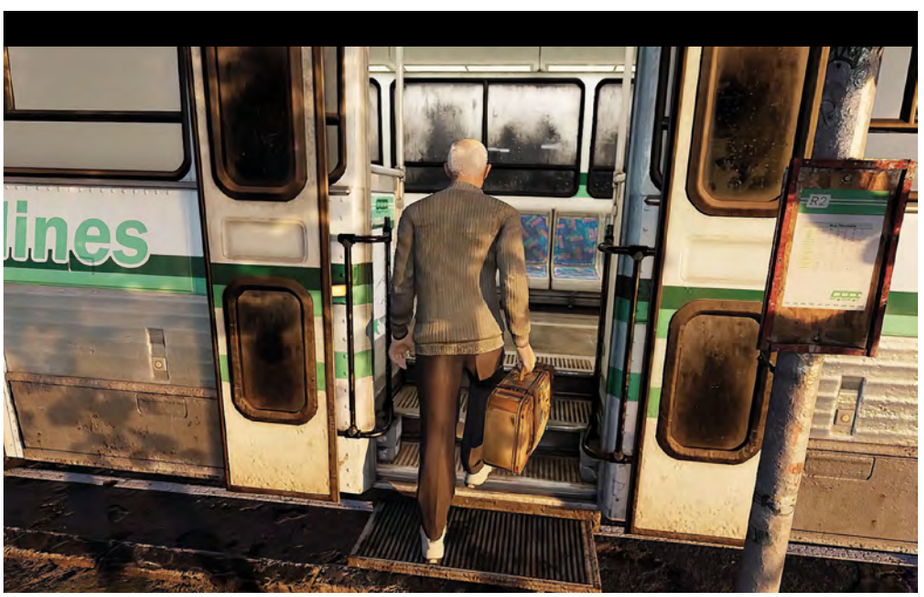
So how to deal with that situation? First of all you need to have a basic understanding of user psychology. What attracts them? What bores them? How can you gain their attention? The easiest way is to follow the concept of reward and punishment: The character asks the user to do something. If the player fulfills the request something entertaining happens. When the user moves too far away from that scene just fade to black. (Don't fade to black, it is mean!)

Adding virtual hands as user representation to "Man at a Bus Stop" was well received. To avoid breaking the immersion the hands were designed with situational awareness, mostly based on the distance to Tom and the area one would reach out to. That way it was not necessary for the user to learn anything. When someone reached out to Tom the hand naturally posed in a way that made sense. An interesting result was that even though the hands clearly changed shape while moving, everybody just accepted this behavior as normal.

Having said this, I mentioned above that players are hard to control in VR. Give them VR hands and see what happens. The reward and punishment principle becomes blurry. The AI reaction to touch is rewarding, but it can also disturb the story, for example when touching his face while he is speaking. Tom reacts clearly unhappy when you poke his nose and while some users are feeling bad about that, others are motivated by it. Most players treated Tom "respectfully", yet some twirled their hand inside his head, only to see if they could break him.

Interesting was that almost nobody would actually try to walk through these virtual humans, neither the "Professor", nor Tom. There seems to be a natural barrier when VR space and characters become too real.

When players used the system as intended, by playing the story and listening to Tom, "Man at a Bus Stop" became very rewarding. The fairly hi-end design and the sophisticated Emotion Engine made it feel very intimate. Tom came already very close to being a very human-like virtual being. Who knows what lies in the future …



THE FAIRLY HI-END DESIGN AND THE SOPHISTICATED EMOTION ENGINE MADE IT FEEL VERY INTIMATE.