VIRTUAL PRODUCTION AND REALTIME FILMMAKING TECHNOLOGIES FOR THE INDEPENDENT FILMMAKERS. AN OVERVIEW.

VOLKER KUCHELMEISTER

Echtzeit-Computergrafik in der Filmproduktion hat sich seit den Machinima-Filmen der 1990er Jahre beträchtlich weiterentwickelt. In Machinima verwendeten Filmemacher Computerspiel-Engines und digitale Inhalte, um narrative Filme zu produzieren, die in Echtzeit generiert wurden. Dies erlaubt es unabhängigen Filmemachern, die Rolle des Regisseurs, Kameramanns, Bühnenbildners und Animators in einer Person zu vereinigen. Das virtuelle Filmemachen wird heute von Film- und VFX-Studios sowie Spieleentwicklern weiträumig eingesetzt. Es erlaubt beispiellose Flexibilität im Produktionsprozess durch schnelle Iterationszyklen und ein Modell des kreativen Experimentierens mit sofortigem Feedback. Der traditionell langwierige Produktionsprozess vom Drehbuch bis zum fertigen Produkt ist damit zum Teil aufgehoben. Virtuelles Filmen bedeutet einen Paradigmenwechsel, wie wir Filme machen und wie wir zusammenarbeiten, um sie zu machen. Dies gilt insbesondere für digitale Produktionen, beeinflusst aber auch die Produktion von Live-Action-Filmen erheblich, indem sich der Pre-Produktionsprozess grundlegend verändert.

Source: University of New South Wales Sydney

Fig. 1: Digital 3D character "Viv" within the virtual 3D set (The Visit. 2019).

Real-time computer graphics in narrative film production has come a long way since Machinima emerged in the 1990s. In Machinima, filmmakers utilised computer game engines and digital assets to create narrative films that are generated on the fly in the computer graphics processor. Empowering independent filmmakers to take on the role of director, cinematographer, editor, set designer and animator, all in one person. Virtual filmmaking and production techniques are now embraced by film and visual effects studios as well as game developers. It promises unprecedented flexibility in the production pipeline and even more significant, it opens up filmmaking to a model of creative experimentation with instant feedback and quick iteration by overcoming the traditional drawn out production process from script to finished product. It signifies a paradigm shift how we make movies and how we collaborate to make them. This is especially the case for computer generated films but also heavily influences how live action films are produced, by fundamentally changing the pre-production process.

Technologies such as performance capture, virtual cinematography, physical based lighting, photogrammetry and 3D scanning become more accessible and emerging technologies such as volumetric capture, real-time Raytracing and AI powered image processing are not far behind. Game engines like Unity 3D or Unreal Engine are at the centre of this revolution. It is where traditional filmmaking ideologies meet real-time. Game engines provide free, easy to use and customisable tools for the creative and technical minded filmmaker.

This article gives an overview of the various concepts, technologies, tools and processes for real-time Filmmaking and Virtual Production. The focus is not on high-end feature film but an introduction to the technologies for independent filmmakers and artists.

To illustrate these concepts, this article takes a recent project by the author (The Visit, 2019. Fig. 1 [1]) as a point of departure. A real-time interactive and non-linear film, for screen and VR, employing many of the technologies outlined in this article. This production emerged from research conducted by artists and psychologists working with a number of women with dementia and is produced by University of

New South Wales Sydney - Felt Experience and Empathy Lab, with creative director Volker Kuchelmeister. It focuses on a single character on screen 'Viv', a woman with dementia with whom visitors interact as she moves about in her kitchen. The character is able to talk, express a range of emotions and make direct eye contact, and by doing so enables a high level of engagement and empathic responses from the viewer. The point of the work is to draw the viewer into the emotional and perceptual world of Viv, ultimately to break down stigma that exists around mental health.

Digital Humans

Digital 3D humanoid characters come in as many shapes and forms as real humans do. From stylised cartoon to a high level of realism. They might be created by a 3D modelling artist, 3D scanned or customised with a character-engine. Table 1 gives an overview of various character related platforms. For the character Viv in *The Visit,* we choose a 3D scanned character from renderpeople.com. These characters come ready to be used in a game-engine, with textures, a skeleton and a basic facial rig. Only minor customisation was required. A processed 3D scanned character can provide a good level of stylised realism within the limitations for real-time rendering.

Working with realistic rendered human characters, it is important to be aware of the concept of the Uncanny Valley, which describes the relationship between the degree of resemblance to a real human being and the emotional response. The concept was identified by the robotics professor Masahiro Mori in 1970. It suggests that humanoid objects which imperfectly resemble actual human beings provoke uncanny feelings of eeriness and revulsion in observers. A stylised character expressing human emotions is more likely to be accepted than an imperfectly rendered realistic digital human. There is a fine line between the character appearing unintentional creepy due to its flaws and the level of realism of the character. In computer games, humanoid non-player characters do often appear lifeless, stiff and blank. While their depiction might be considered realistic, their emotional expression is everything but believable. This is acceptable in many computer games with a focus on game play, but not for a filmmaker with the intention to tell a story with a believable characters at its centre which whom the audience can empathise.

In the author's experience, sometimes less is more when it comes to emotive digital humans. In *The Visit*, the character's depiction is far from photo realistic, partly due to limited production resources but for the most part as a consequence of a conscious decision to work with a stylised look. Her emotional states are limited to a basic repertoire and the voice actress was directed to keep the characterisation contained and not to over-act. This approach worked well for this particular project, but every production has its unique requirements.

To be able to animate a character within a game-engine, it requires an internal skeleton, also called a rig (Fig. 2). The rig is fitted with a weighted and textured skin. This skinned mesh deforms according to the skeleton pose. Game-engines have good support for rigged characters and the mapping of humanoid motion animations. A common file format for characters and animations is FBX.

Motion and Performance Capture

Motion Capture (MoCap) is a process of digitally recording patterns of movement for the purpose of animating a digital 3D character in a film or video game. Unlike a video recording, a MoCap system records spatial data only, no visual representation. There are various MoCap systems on the market, which differ in regard to quality of recording, fidelity and detail, capture volume, robustness against occlusion and of course the price point. High-end optical tracking systems such as Optitrack or Vicon [4,5] utilise reflective tracking markers attached to a MoCap suit worn by the subject(s). An array of infrared cameras pick up the markers and triangulate its position in 3D space. This principle allows for precision tracking at a high frame rate, a large tracking volume and multiple subjects at world coordinates. Mid-range systems (Xsens, Rokoko [6,7]) are based on inertial measurement sensors attached to the joints of a subject, no external sensors are required. These systems can be employed anywhere, no MoCap stage is needed and there is no limit to the capture volume. However, the precision is inferior to an optical system and finger tracking is not supported. Low-end systems are based on a single visual or time-of-flight depth sensor, such as Microsoft Kinect. The software is capable of detecting the human anatomy and mapping a simple skeleton to the body movement in real-time. These systems have a limited tracking volume and are orientation sensitive. A good example is the Brekel Body application for Kinect sensor [8].

The MoCap data is then mapped to the rigged characters' skeleton and the skin deforms according to the movements. To get a curious filmmaker started, Adobe offers a free humanoid animation library through their Mixamo online platform, as well as a limited number of humanoid characters. While the focus of this collection are computer games, one can find a selection of walking, running and idling animations to get started (Table 1).

Performance capture extends on MoCap, by capturing the full spectrum of emotionally rich expressions of an actor. This includes image based performance capture of facial

Table 1: Humanoid character sources.

Adobe Mixamo	Reallusion Character Creator	Autodesk Character Generator	Daz 3D, Genesis	Unity Multipurpose Avatar UMA	3D scanned Characters
Online platform to create, rig and animate characters. No 3D knowledge required. Humanoid animation library.	Create, import and customize stylized or realistic looking character assets.	Create, customize, and download rigged 3D characters from a catalog of body types, outfits, hairstyles, and physical attributes.	The Genesis 8 figure platform is an engine to blend, mix and customise unique characters.	Efficient customizable characters in runtime for the Unity game engine.	Various online platforms offering realistic looking 3D scanned characters. Some ready for use in game engines.
https://www.mixamo.com	https://www.reallusion.com/character-creator/	https://charactergenerator.autodesk.com	https://www.daz3d.com/genesis8	https://assetstore.unity.com	https://renderpeople.com https://3dpeople.com https://humanalloy.com

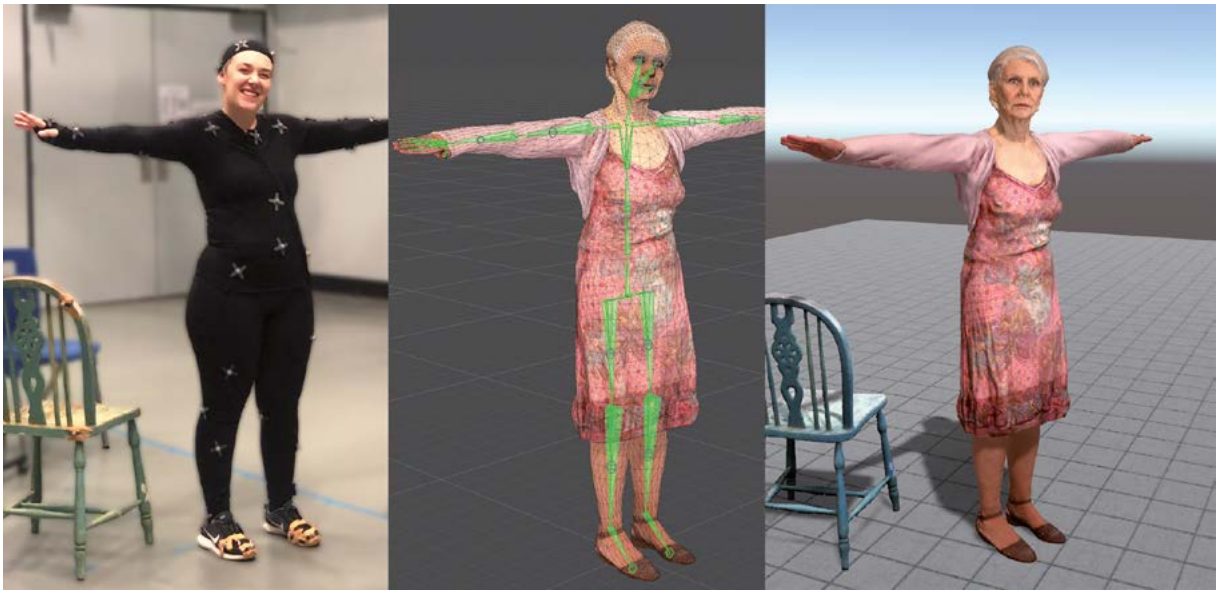


Fig. 2: MoCap pipeline: actress in optical MoCap suit (l), rigged digital character (m), rendered scene (r). (The Visit. 2019).

Source: University of New South Wales Sydney

expressions and allows for true human emotions to be preformed by a non-human character. Unlike in traditional CGI productions, this approach does not require hand animation to bring to life multiple characters scene by scene. James Cameron and Weta digital pioneered this technique for the feature film Avatar in 2009. The humanoid alien characters in this film show an uncanny resemblance with its human performer, particular in facial features and expressions. However, the translation of facial expressions and lip-sync data to a digital model of a non-human character is none trivial. It requires a rich set of customised controls, synthetic muscles, blendshapes and sometimes manual intervention by an animator.

Blendshapes are variations on the position of vertices in a 3D mesh and are created by 3D modelling software. Individual blendshapes can be dynamically adjusted and mixed with other blendshapes to create expressions. The Facial Action Coding System (FACS) [9], an anatomically based taxonomy to break down facial expressions into individual components, is a good starting point to create a set of facial blendshapes for a character. While untested by me, recent advances in machine learning also allow for the automatic generation of a FACS set of facial blendshapes by a service from Polywink [10].

There are now tools available for independent filmmakers to employ performance capture for low budget productions. Apps such as MocapX, Face Mocap or Face Cap [11] for Apple iPhone X or Brekel Face for Kinect [8], utilise a depth sensing camera to record facial expressions and lip movement by varying values of a rich set of FACS based facial blendshapes. Additionally, the apps record the head orientation and voice/video as a reference of the performance. While this is not jet up to the quality level a feature film might require, it is sufficient for stylised and cartoon characters or as a starting point for an animator.

For the emotive expressions of the character in The Visit, I choose to work with a more unconventional method. A physical interface to drive the emotions in realtime. A set of seven basic expressions (neutral, happy, sad, worried, angry, stern, excited) where mapped to a 7DOF Space Mouse controller (Fig. 4). While the digital character performed a scene, adjustments on the interface controls the facial blendshapes

Source: University of New South Wales Sydney

Fig. 3: Emotive facial expressions realised as Blendshapes. (The Visit. 2019).

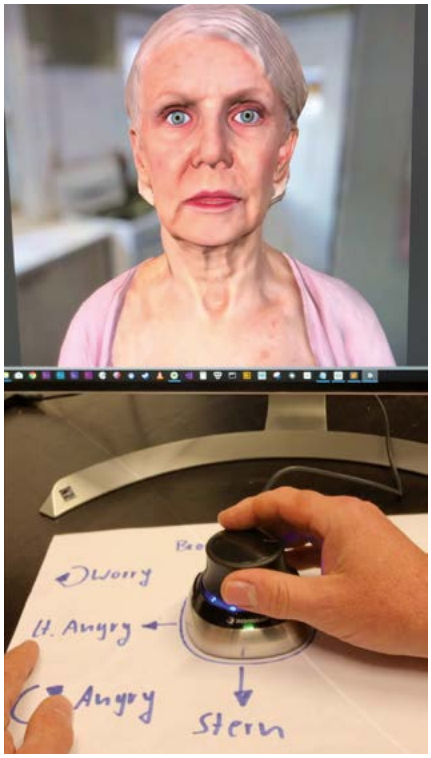
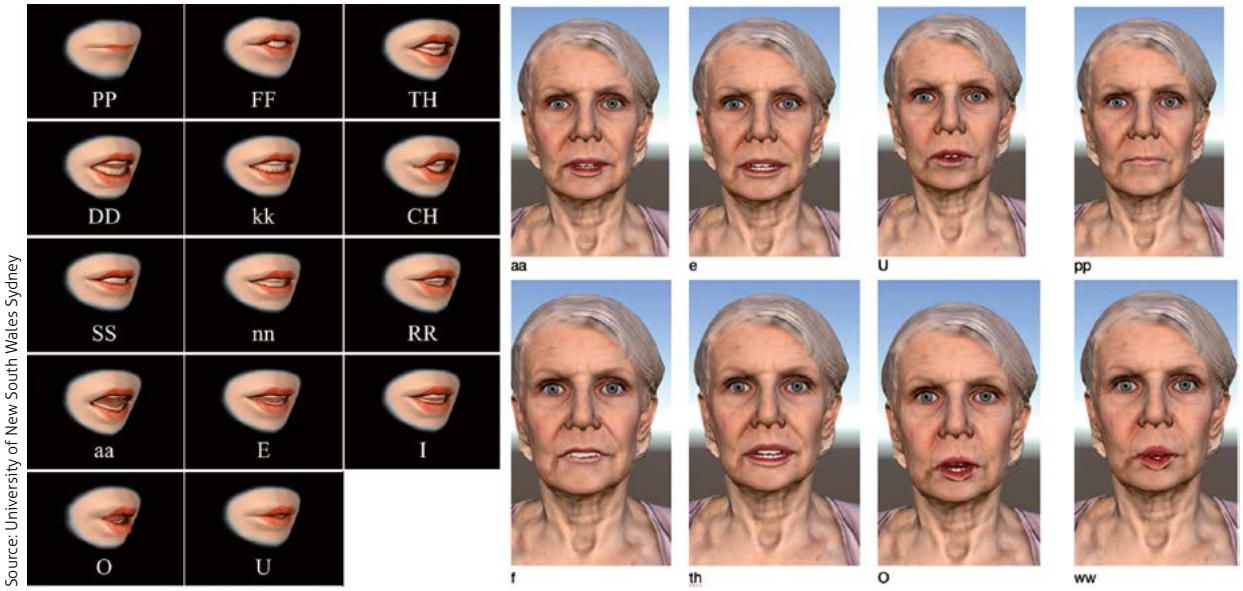


Fig. 4: A physical interface to drive character emotions in real-time (The Visit. 2019).

Source: University of New South Wales Sydney



Source: University of New South Wales Sydney

Fig. 5: Lipsync: Face Animation Parameter (FAP) standard set of 14 visemes, adapted by Oculus (l) and a reduced set for the custom character, realised as facial blendshapes (r). (The Visit. 2019).

dynamically and the changing values are recorded on the fly.

Despite the technological progress in performance capture, working with digital humans is complex. Many factors play into the believability and the quality of engagement with an observer. Rendering, body motion, facial expression, voice acting, lip-syncing and many more form a complex network of interrelated phenomena. For a filmmaker, it is crucial to be aware of limitations and to be prepared to go through numerous iterations and testing stages before their characters fulfil their creative vision.

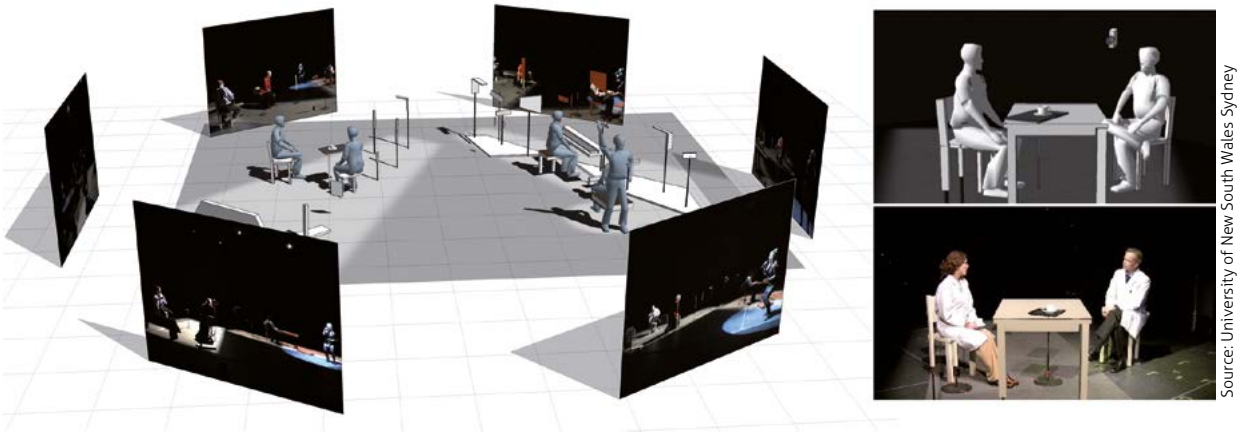
Lipsync

For the character to be able to talk, some form of lipsync is required. Hand-animating lipsync is a drawn-out process and simple jaw movement triggered by the voice recording dynamic level does not produce a satisfactory result. Human utterance is an intricate combination of jaw motion, lip shape and tongue. The Moving Pictures Experts Group (MPEG) published a definition for virtually representing humanoids in a way that adequately achieves visual speech intelligibility, the Face Animation Parameter (FAP) standard. It contains the definition of a set of 14 visemes (Fig. 5), the visual equivalent to phonemes or expressions of the lips and face that correspond to a particular speech sound. For an experienced 3D modeler, creating these visemes in form of blendshapes for an existing character is not too difficult. Important is to test the result frequently with a voice recording. Oculus developed a Lipsync library for their social VR applications. The OVR Oculus Lipsync API. It describes a set of plugins and APIs that can be used to sync avatar lip movements to speech sounds and laughter. Oculus Lipsync analyses the audio input stream from a microphone input or an audio file and predicts a set of viseme values. Utilising this library in the game-engine Unity or Unreal is free and produces great results, provided the visemes are modelled correctly.

Virtual Set and 3D props

A virtual set is a representation of an environment in form of a textured 3D model. This can be utilised at the previs stage (Fig. 6) or in case of a real-time film it represents the final digital set.

Virtual sets might be sourced from a library (such as the Unity asset store), constructed in a 3D modelling application or a 3D scan of a real location (Fig. 7). Whichever technique is used depends on the nature of the environment (interior vs exterior), the extend of the scene, the level of realism required, the constraints for real-time rendering such as polygon count and materials and most important the aesthetics.



Source: University of New South Wales Sydney

Fig. 6: Virtual set of a theatre stage in previs quality (Fragmentation 2011)

Source: University of New South Wales Sydney

Fig. 7: A virtual set with props. 3D scans reconstructed with Photogrammetry (Metashape and RealityCapture). (The Visit. 2019).

In my view, a 3D scan, while relatively complex in its creation, offers a range of advantages over a traditional hand-crafted models. It captures not only the geometry of a scene, but the quality of the light and shadows, natural imperfections and variations of materials, all difficult to achieve with traditional modelling and lighting techniques. One accessible method to scan an environment is Photogrammetry, a technique based on making measurements from photographs by recovering the exact positions of surface points in a scene. With feature detection and correlation amongst multiple photographs, from various points of view, a photogrammetry algorithm estimates the camera intrinsic and extrinsic parameters and, by doing so, generates a textured 3D model of a scene or object. The texture of the model is also derived from the photographs, by re-projecting the images onto the geometry to generate a UV texture. Agisoft Metashape or Reality Capture [13] produce equally good results. But not all subjects are suitable for photogrammetry, reflective surfaces or featureless plain walls are often not reconstructed very well. In this case, it helps to put markers on walls and use dulling spray. Photogrammetry software initially generates very large models with tens of millions of polygons, but also allow to export models with a reduced polygon count. Additionally, a cleanup stage in a 3D modelling application will most likely be required.

Narrative design and sequencing in a game-engine

Video games often use cutscenes to engage the player and tell parts of the story. To make this sequencing of events easier for the developer, game-engines introduced the notion of the timeline to their tool-belt. This is a very familiar concept for filmmakers working with nonlinear editing software. Unity's timeline feature was introduced only in late 2017 and has since grown into a highly functional tool. The Unity timeline contains references to objects in the scene, such as characters, sets, lights, cameras, sounds ... and sequenced events or animations for those objects (Fig. 8). For instance, a camera track might contain numerous virtual cameras all with varying position/orientation and lens settings. In timeline these virtual cameras can blend settings over time and simulate a tracking shot or zoom or whatever else the director might want to work with. The same blending functionality is available for animations, which simplifies the sequencing of multiple MoCap takes. Timeline is able to blend the character pose of a one take into the beginning of the next. It is also possible to blend multiple tracks, for instance for facial expressions and lipsync, both based on blendshapes for the same character mesh. For a filmmaker this is probably the most exciting and enjoyable part in the process. The system encourages creative experimentation with instant feedback and quick iteration.

Virtual Production

Virtual production (VP) is a broad term referring to a spectrum of computer-aided production and visualisation filmmaking methods. VP leverages real-time visualisation of characters and digital sets in combination with live-action capture for cast and crew. VP is where filmmaking and computer game technology or the physical and the digital world meet. It merges traditional practices with current and ongoing advances in real-time technology to enable filmmakers to make better creative choices much earlier in the production process [15]. It goes beyond visual effects; it adds an element of real-time preview to the pre and production stage. A recent feature film example is the production of The Lion King (Disney and Technicolor, 2019). Director Jon Favreau utilised VP to bring live-action actors into a synthetic environment. While directing the action on stage, the director operates a tracked virtual camera (Fig. 9) to frame shots while the preview of the composited image is presented in real-time. This way of working encourages a more iterative, nonlinear and collaborative process.

The term Virtual Cinematography includes aforementioned live-action real-time systems on a MoCap or chromakey stage, but also virtual cameras controlled in a game-engine, without a live-action component. For a filmmaker or cinematographer, this is quite liberating to be able to experiment on the fly. Changes in lighting, selecting a lens, framing shots, using dollies and cranes and even realise shots that simply cannot be filmed for real are all at the filmmaker's fingertip. There are however fundamental differences to

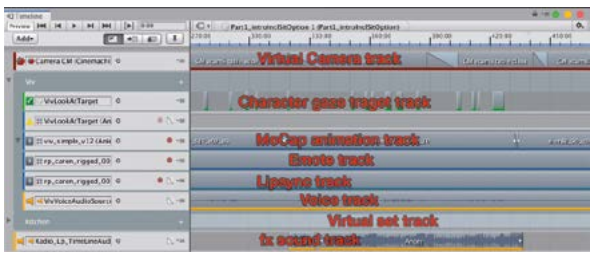


Fig. 8: Unity game-engine Timeline sequencer. Each track holds an object reference and timed events or animations (The Visit, 2019).

Source: University of New South Wales Sydney

Source: University of New South Wales Sydney

Fig. 9: Virtual cinematography: Live-action stereoscopic camera with HTC Vive tracker (l). Virtual camera with optical tracking markers and real-time preview screen (m). Virtual cameras in Unity game-engine (r).

a physical camera-lens system. The concept of shutter speed and aperture does not exist with virtual cameras. Originally a shot will always be in focus throughout the frame, even for fast moving subjects. To give more control over the look of a scene, the game-engines provide a post-processing stack to simulate a physical camera more closely. This processing is also performed in real-time and includes depth-of-field, motion blur, color-grading, vignette and so on.

Conclusion and Outlook

A direct comparison of the production time required to create a live-action vs a real-time computer-generated film is not straight forward. In particular the pre-production stage is more time consuming for live action. Every detail needs to be determined in advance to limit the time required for the principle shoot. For a real-time film, decisions can be made throughout the production process. Character and set design, lighting and cinematography can change at any time. The exemplary project The Visit, with a total duration of 17 minutes, took about 3 months to produce. This is after the initial technical research and with the script finalised. It includes voice recording, MoCap, MoCap cleanup, character development, virtual set creation and integration in the Unity game-engine. It involved me as art-director working as a non-expert in Unity and virtual set creation and a 3D artist for about 6 weeks to work on the MoCap data and create blendshapes for facial expressions. The film was Initially created for screen display, an adaptation for Virtual Reality (Oculus Quest) took less than one week to implement.

Still, producing a real-time computer-generated film with virtual production techniques is a significant undertaking. It requires expertise in various domains and technical resources. But this technology is still in its infancy. Current research indicates that character creation and robust performance capture, assisted by machine learning, will soon be possible with a single standard camera. With raytracing, game-engines start to blur the line between real-time and reality and volumetric video is just around the corner. These technological advances, together with more intuitive software tools, will make this style of filmmaking even more accessible.

Filmmaking is an ever-evolving artistic and technological pursuit to tell stories in new and compelling ways. The current Realtime Filmmaking and Virtual Production revolution is an important step along the way and has the potential to empower filmmakers to realise their creative vision beyond the flat screen.

Bild: Volker Kuchelmeister

VOLKER KUCHELMEISTER

is working as lead immersive designer and research fellow at the UNSW felt Experience and Empathy Lab (feel) in Paddington, New South Wales, Australia.

http://kuchelmeister.net

References:
[1] Project The Visit: http://kuchelmeister.net/portfolio/the-visit/
[2] Unity game-engine: https://unity.com
[3] Unreal game-engine: https://www.unrealengine.com
[4] Vicon high-end Motion Capture: https://www.vicon.com
[5] Optitrack high-end Motion Capture: https://optitrack.com
[6] Rokoko mid-range Motion Capture: https://www.rokoko.com/
[7] XSense mid-range Motion Capture: https://www.xsens.com
[8] Brekel MoCap and Facial Capture with Kinect: https://brekel.com
[9] Facial Action Coding System (FACS): https://imotions.com/blog/facial-action-coding-system/
[10] Polywink facial blendshapes on demand: https://www.polywink.com
[11] Facial Capture with iPhone X: http://www.bannaflak.com/face-cap/, https://mocap.reallusion.com/iclone-motion-live-mocap/iphone-live-face.html,
[12] OVR Oculus Lipsync: https://developer.oculus.com/documentation/audiosdk/latest/concepts/book-audio-ovrlipsync/
[13] Photogrammetry: Agisoft Metashape https://www.agisoft.com, Reality Capture https://www.capturingreality.com
[14] Project Fragmentation: http://kuchelmeister.net/portfolio/fragmentation/
[15] The Virtual Production Field Guide by Epic Games: https://www.unrealengine.com/en-US/blog/virtual-production-field-guide-a-new-resource-for-filmmakers