A WORKFLOW FOR NEXT GENERATION IMMERSIVE CONTENT

Nicolas Mollet[1], Tim Dillon[2], Fabien Danieau[1], François Le Clerc[1]

Technicolor[1], MPC[2]

ABSTRACT

We are nowadays experiencing the convergence of the Media Entertainment and Gaming industries thanks to the emergence of Virtual Reality. VR immerses users in an experience that gives them the illusion of reality and opens the door to numerous challenges and opportunities towards the definition of next-generation immersive media. On the one hand, immersive media are not standardized and not compatible over players and equipment. On the other hand, professional workflows for movie creation, previously designed for screen-experience, cannot be applied as such to new immersive experiences. But VR offers as well the opportunity to add other multi-modal feedback channels such as Haptics (sensorial), for which entirely new workflows need to be defined.

In this paper, we present a workflow to create, distribute and render immersive contents, involving immersive videos, haptic feedback, and interactivity. We focus on the locks and opportunities, and discuss real production examples.

INTRODUCTION – WORKFLOW OVERVIEW

Virtual Reality (VR), Video, and Video games are converging. Movies and games are getting closer, sharing techniques and contents. Immersive 360° videos have been produced for decades, and today anyone can shoot his own 360° video with a smartphone and a mirror kit. Similarly, Head Mounted Displays (HMD) have existed for a very long time, going back to the video game consoles of the 90’s (Nintendo virtual boy). Still, only now market and technology are simultaneously ready, through the emergence of new devices (the Oculus Rift, the Sony Morpheus, the HTC Vive or the Samsung Gear VR) and thanks to their quality of experience, their price, and the limits of the TV-screen experience. Finally, while « 4D cinema» (which consists in stimulating other human senses) has been available in dedicated theatres and movies, the « haptics » market at home is strongly growing.

The next generation immersive media can be defined as a set of multi-modal immersive experiences. It includes the stimulation of all of human senses: vision, sound, haptics, smell, proprioception, etc. It targets a perfect immersion in the experience, making it alive as real, and a perfect illusion for the user. Beyond the passive experience of watching a screen, the user is able to look around him/her and to interact with artistic content. It is not only about a two hour 360° videos, it may also be a 10 minute immersive artistic experience with adaptations and interactions with some characters or objects. We are at the very beginning of this evolution, still having tools and workflows either adapted to videos, videos+VFX, VR or video games. Mixing everything is a huge challenge. The tendency right now is to start with the audiovisual (AV) evolution, making it *visually* immersive. So the first step consists in integrating 360° omnistereo videos in the workflows, with obviously spatialized audio. But the following step, which will provide a true *Virtual Reality media* approach, involving real time rendering, adaptation, and interaction, requires deeper changes.

The typical workflow for video-streaming comprises three stages: production, distribution and rendering. Making it applicable to next generation immersive media raises a number of issues. The following sections describe those three stages, and illustrate their use with examples. Haptics is used as a transversal example of the workflow. We then present media produced by Technicolor and MPC taking into account this new workflow. We finally conclude and provide perspectives

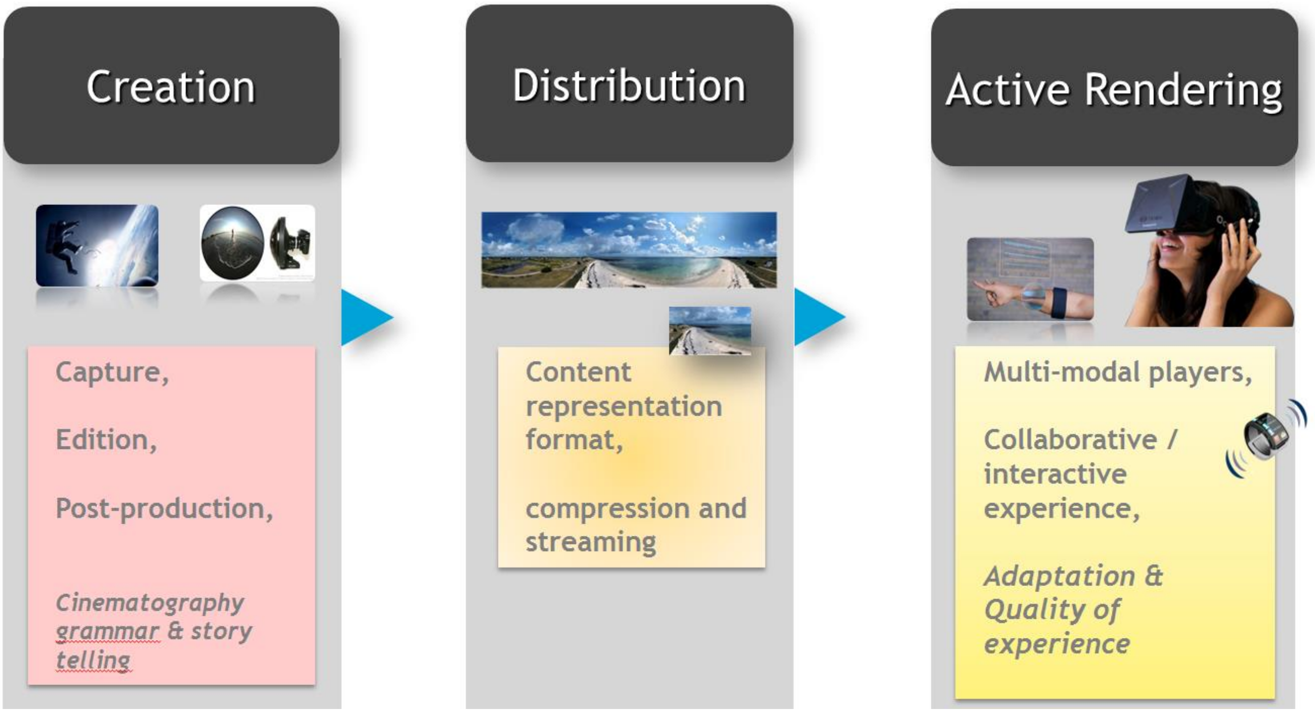
Creation
Capture,
Edition,
Post-production,
Cinematography grammar & story telling
Distribution
Content representation format,
compression and streaming
Active Rendering
Multi-modal players,
Collaborative / interactive experience,
Adaptation & Quality of experience

Figure 1 - workflow and challenges for next generation immersive multi-modal content

CREATION

The creation phase encompasses the techniques and tools to produce the new immersive content. We detail below the specific challenges for this phase.

Capture

The immense majority of cameras are made to acquire a planar image, in 2D or 3D. One of the future challenges here is to be able to go on new forms of video acquisitions. For example, the capture of spherical images, producing immersive 360° videos, has been known for decades. Solutions are now available for capturing omnistereo, and the workflows start to evolve towards that. But to go further on a VR approach, we will have to address the still open problem of free viewpoint videos. This calls for convenient, usable capture setups of reasonable size, and compatible with limited processing power. These solutions should provide a content allowing to move *inside* the *video*, around the acquired objects. This is the true intersection between video games, VR and videos, where real objects are modeled in 3D with their video texture. Approaches such as light fields could be involved here.

Creation tools

Contents are not limited to video, they could also be 3D CGI. These contents are currently created facing a screen. But VR offers the opportunity for artists to be immersed in their content, including at the creation phase. We can underline several recent announcements taking this approach, such as the ILMXLab, which proposes immersive rooms for artists to create and adapt the content, to be alive in it and feel it. How could this impact storytelling, how could it stimulate creativity, and finally, how to effectively create the content, the story, the interactions and so on? What would be the future desk for the artist is still an open question, where immersion, interactivity and collaboration are certainly important keywords.

Data

Captured or produced content needs to be directly edited and post-produced. The needs in terms of data are huge when we talk about immersive video and VR. UHD and HDR videos have a certain level of computational cost, while it is still done for a 45° field-of-view experience. Moreover, the VR community tends to agree that 8k resolution per eye, for a HMD context, would be the resolution target for a true VR experience. Natively supported resolutions of the current workflows and tools are far from this target. When we go on VR, it is easy to understand that tools which are only made for videos will have to evolve to cope with interactions for example, and will have to take more from the video games editing tools.

Cinematographic rules

immersive content provide a new user experience, meaning that the content should be thought as immersive at the creation stage. Directly watching a legacy content, such as an action movie, in a HMD does simply not work. The action sequences change the point of view every 2 seconds on average. Moreover, the point of view is selected to be seen at a certain distance, and optimized for a 45° field-of-view experience. Camera movements will make most people dizzy. Actually, thinking immersive implies rethinking all of the cinematographic rules, in order to bring to the user an immersive experience without getting him dizzy, without losing him, and on the contrary offering him new opportunities. As an example, the sound, or even haptic effects, may be used to drive the gaze attention of the final user. A dialog between two actors may be felt realistic, looking to the left and right, being with the characters. It actually opens new rooms of creation, where the artists will have completely novel ways of providing experiences.

Example: Haptics

Haptic feedback is traditionally used to highlight the physical events occurring in the audiovisual content (explosion, gun shots, etc.). But haptics may be considered as a complete medium in the sense that it may also convey emotion or semantic content. So the design of the haptic feedback is equally important to the design of AV content and should receive the same attention. Creating haptic effects means designing haptic feedback which will be felt by the user, independently from any haptic device. This is a hard task due to the complexity of data to be edited which are by nature heterogeneous: vibration, temperature, force-feedback, motion, etc. Thus the haptic designer has to take care about the sensitivity of the human body for each of these effects to be able to combine them in a proper way. But the real challenge comes from the edition of these effects in time (synchronization to the audiovisual content) and in space (location on the user's body).

To tackle this challenge we proposed a haptic editor embedding a model of the user (see Figure 2). For a chosen duration within the audiovisual content, the designer may select various effects (vibration, temperature, motion), edit their parameters (intensity, direction, etc.) and locate them on a 3D body model. This editor has been designed for touch screens to ease the manipulation of the 3D model. It also offer a playback feature which modifies the body model in real-time when the video is played. The designer thus has a preview of the haptic feedback. Creating haptic effects from scratch may be cumbersome. In the case of motion effects, translation and rotation have to be simultaneously edited. To ease the creation of such effects we explore the use of sensors. An inertial measurement unit (IMU) may capture the motion of an actor (2). The data are time-stamped, enabling the synchronization with the audiovisual content. Such a sensor is now accessible to the mass market and embedded in most of smart devices. Our editor takes advantage of these sensors and allows to directly record motion effects or import data from an IMU.

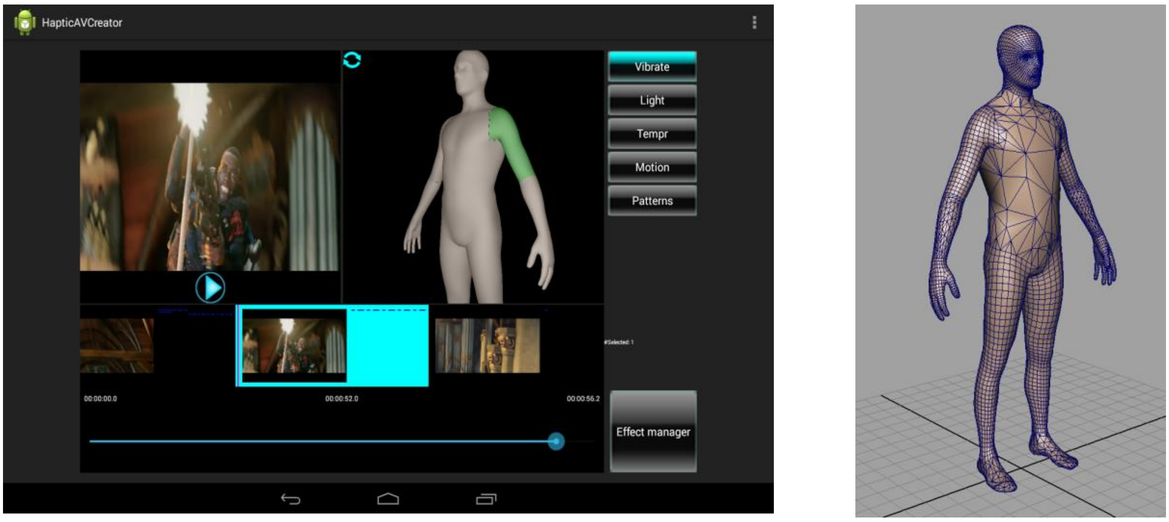

Figure 2 – (Left) Screenshot of the haptic editor. (Right) Example of haptic channel-adaptive mesh (much more resolution on the hands than on the torso)

DISTRIBUTION

As previously mentioned, the amount of data required for true VR experiences brings new challenges. The fact the user is looking in a certain direction is obviously an input to take into account. But it means that the streaming requires constant inputs from the user. If we simply broadcast everything, we will also have to find technical solutions to support very large videos. Moreover, the distribution and the final experience need to take the formats into account. The reality in the VR context is a multitude of formats (for example, more than 200 3D formats exist today), so definitely interoperability between the distribution (broadcast, physical support, etc.) and the rendering devices will be a very important issue. That is the case for 360° videos, but that is even worse if you consider a multi-modal immersion, as you have to mix with haptic devices from different manufacturers to provide other non-standard data.

Example: Haptics

Some pioneering contributions have demonstrated the feasibility of standardizing media with haptic effects (1). However, formats lack a description of the location of the effects directly addressing a part of the user's body. Besides, they do not take into account the human haptic perception. For example it is known that if two vibrating stimuli are too close on the skin, only one vibration will be felt. From such a description of haptic effects, useless information may be stored or transmitted to a haptic renderer. Moreover, a single update rate is specified in these formats while several locations on the user's body are stimulated and could have different temporal sensitivities. To tackle these issues, our system aims at producing a "perceptually-driven encoded haptic signal" from the high-level descriptions of haptic effects provided during the authoring stage. The key idea is to map those high-level effects on several underlying 3D body models which take into account the perceptual properties of a mean human being for the different haptic channels (vibration, temperature, pressure, etc.).

One body model may be defined for each haptic channel. Each body model is characterized by a set of vertices and faces with a variable spatial resolution (face size) depending on the user spatial sensitivity on the considered channel (see **Error! Reference source not found.** right). Each face also embeds some meta-information corresponding to some key properties related to the specific targeted haptic channel and location such as the required temporal resolution or the min and max admissible values.

Once the mapping has been done, a signal is generated for each model as a concatenation of a face identifier (relatively to the associated 3D model), a time-stamp (computed regarding the associated temporal resolution) and a normalized actual value (min/max mapped to 0/1). Only the active faces, where a change occurred, are considered. All those different signals (for each haptic channel) are then muxed into the "perceptually-driven encoded haptic signal" through a simple concatenation where each model is identified by a dedicated identifier. As a final step, the signal could be compressed to be transmitted or stored with classical techniques. This "perceptually-driven encoded haptic signal" natively takes into account the human perceptual properties (spatially and temporally) thanks to its model-based construction and may be highly compact and adaptive for the transmission. Indeed, in a streaming context, the different body models may be part of a dedicated standard (and should not be transmitted), and only the model and face identifiers as well as the timestamps and values should be transmitted.

ACTIVE RENDERING

The media consumption from the end user point of view is an awesome domain to address. This is where we can dream about new experiences, and where we can expect to have a huge impact on the future of entertainment at home, by providing new and incredible immersive experiences. The AV media are now consumed through TV, mobile phones, tablets, laptops, and projection. All of course are providing the classical paradigm of "screen" experience. The fact that HMDs are coming to a usable solution will make everyone discovering how true immersion is amazing. But the media consumption will not limit itself to this approach, it will require hardware solutions to feel immersed at home. It will include the HMD, obviously. But it raises social problems, where you want to be immersed with your family inside the content. In order to achieve this, the artist creating its content will have to precisely think about how to include the user and his direct environment in the content without breaking the immersion. If you are watching a movie with a particular color style through a HMD, you may want to be immersed and see your body in the content with a particular effect applied on it. It may be the same with your direct neighbors, that should be immersed naturally. Finally, HMD is not the only support to consider. The impact at home will be larger and with a longer term approach, placing an immersive solution as a central entertainment system at home, as the TV is today. It will be useful for telecommunication, video games, and media consumption. We can dream about amazing solutions to provide multi-modal and interactive experiences at home, and everything needs to be either industrialized or created.

Example: Haptics

The role of the haptic renderer is to command the haptic devices from the haptic effects designed independently from any devices. The challenge is to map the haptic effects located on the body model to the actual devices located on the user's body (potentially at different locations). Besides the devices have their own physical constraints which need to be respected by the haptic renderer. We propose here a first solution to perform such a mapping based on a human body model. The body model used for the design of haptic effects is shared to the haptic renderer. Haptic devices are located on this representation of the user's body. The body parts are defined within a hierarchical structure. Thus if one body part is stimulated, all the children are stimulated. For instance if the user's left arm is supposed to vibrate, the user's smartwatch on his left hand will vibrate if he has one.

More generally our system deals with the notion of haptic scalability. The system is aware of the location and capabilities of each device (vibration, temperature, motion, etc.), and it tries to generate haptic feedback as close as possible to the one designed by the content creator.

Example: Social lock, HMD facial replacement

We developed an Augmented Reality prototype system for recovering the full face appearance of a user wearing a HMD. Involving your family at home inside your immersive experience, or inviting remote friends, then becomes possible. Indeed, facial expressions are a key element of non-verbal communication, and are known to be conveyed mostly by the upper face part. The block diagram of Figure 3 below summarizes our technical approach. Our solution is user-specific. The general idea is to composite a warped 2D texture of the online user face appearance onto a static 3D model of the user’s head.

The 3D model is built offline from a set of photographs of the user taken from a variety of viewpoints, using an off-the-shelf 3D reconstruction software. The user does not wear the HMD and is requested to maintain the same neutral face expression in all photographs.

We also learn offline a set of representative appearances of the frontal view of the user’s face under a variety of expressions and visemes. To this purpose, we detect landmarks on a training dataset of user face images using a state-of-the art algorithm (3) and cluster the images based on the normalized layouts of landmark locations. The cluster representatives provide expression-specific face texture templates (Figure 3, top right). We leverage the tracking system built into the Oculus DK2 HMD in our prototype to obtain the location and angular pose of the HMD in every frame of the live video stream. Combined with camera calibration data, this information allows us to register the 2D face image with our 3D textured head model. In parallel, we apply landmark detection on the face image and, based on the layout of landmark locations, compute the reconstructed face texture from the closest expression-specific template. Finally, we composite this texture on the UV map of the 3D model and project the model onto the 2D live image in the direction of the detected pose to obtain the face reconstruction.

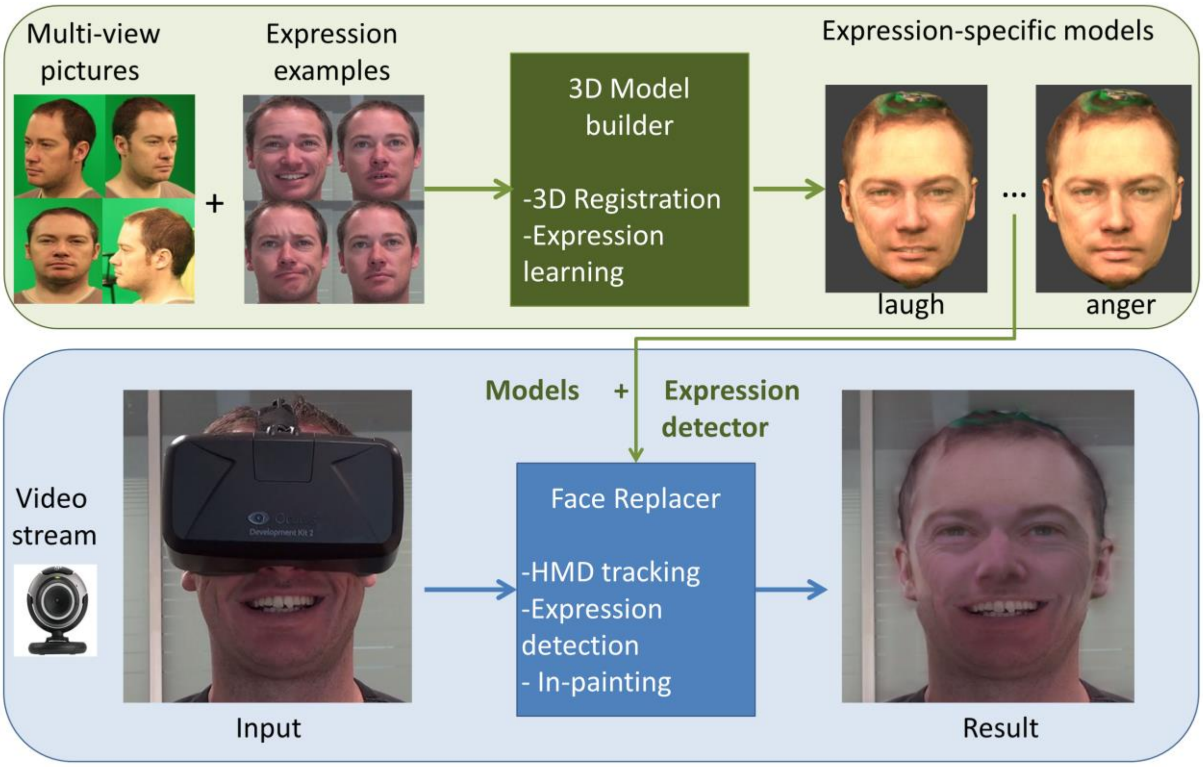

Figure 3: block diagram of the visual face reconstruction algorithm

IN USE – NEW IMMERSIVE MEDIA PRODUCED

360 Video

Catatonic was directed by Guy Shelmerdine for VRSE, works with MPC as partner for the VR post-production, which included stitching, stereo depths, VFX and titles. The experience was captured on location at a derelict mental hospital in Pasadena, CA with VRSE works' proprietary VR camera systems. Catatonic's production value and cinematography bridge the stark terror of a gripping horror film with the inescapable immersion of virtual reality. Participants are ushered into a custom-built wheelchair by live nurses in 1940s uniforms. The viewer is then be fitted with a Samsung Gear headset and headphones. Along with the 360º 3D immersion of VR, the patient will also feel jolts emanating from a ButtKicker™, a vibrating device built into the base of the wheelchair. Catatonic officially screened at SXSW in March 2015. Official Website: www.catatonic.co.

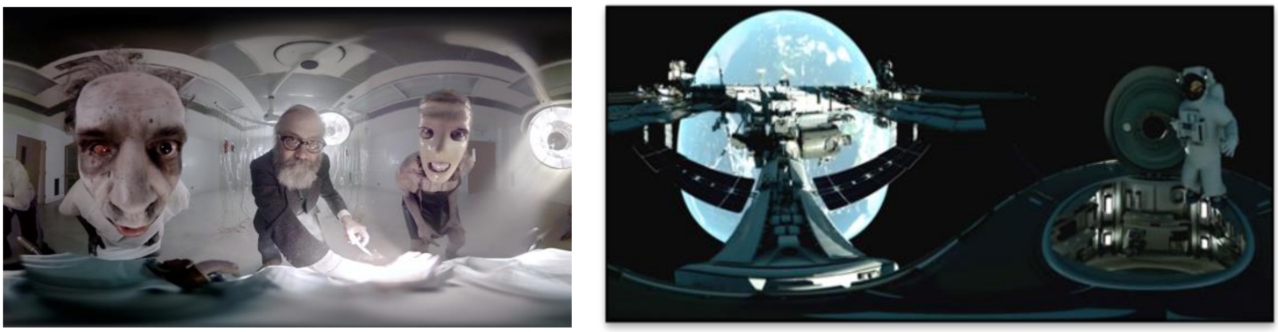

Figure 4: (left) Catatonic, (right) Orbit trailer

Orbit Trailer is a 360 AV+Haptic media. Our typical haptic setup was composed of a smartphone, a smartwatch and a cushion integrating four vibrators, placed on a chair to stimulate the user's back. Extra devices may be plugged and change the way the haptic rendering is performed. Regarding the design of those haptic effects, our editor was used to create vibration effects on the left and right parts of the user's body, as a high description level with a *Haptic description language for the artists*. For instance with Orbit, an asteroid comes from the right side of the screen and hits the spatial station of the left. This is represented by a vibration going from right to the left of the user. This “spatialized stereo” effect may be felt by the user if holding the smartwatch and the phone in each hand, depending on the user point of view orientation, and it is also rendered on the left and right parts of the cushion.

Real-time rendered content

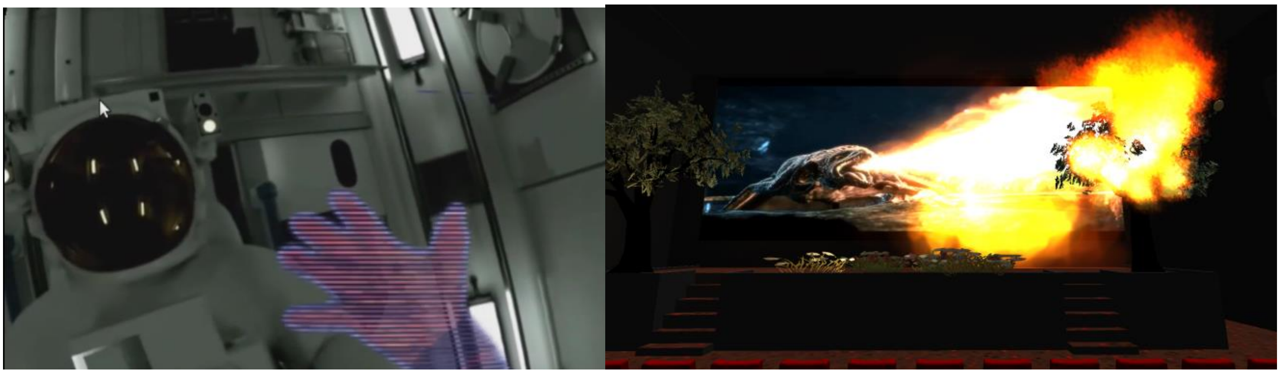

Figure 5: (left) Body insertion in the media, (right) content rendered in RT with the media

The place of the user in the experience is important, and the artists should take this into account to improve the immersion. In order for example to recover the user body inside the media, and not to break the immersion in terms of color/style/context, we insert the body of the user with a *moviestyle* effect embedded in the format. Figure 5 (left) shows the hand of the user with a hologram effect, adapted to an action taking place in space (orbit trailer). To achieve this, we equipped a HMD with video+depth cameras and segmented the user from them, then we applied the *moviestyle* effect on it.

As VR content is rendered in real-time, one of the first advantage compared to video is that the viewer will recover the parallax within the content. More generally, the content itself may be adapted to the user, to improve his experience, to participate in compensating cybersickness, or to provide interaction with parts of the media. The figure 5 (right) illustrate how a legacy movie can be completed with extra real time rendered content, getting the objects really out of the screen.

CONCLUSION AND PERSPECTIVES

The next generation immersive content will stimulate all of the human senses in order to create a full immersive experience. In this paper we have presented a workflow for such a media and we detailed its three main stages. We have first highlighted the issues relative to the creation which involve the needs of new authoring tools and capture devices, the handling of large amount of data and the necessity of rethinking cinematographic rules. In this context we propose a new editor for designing haptic effects based on a human body model. Then we detailed the stage of distribution and pointed out the lack of formats and standards for immersive videos and haptic effects. We also presented our current work on the formalization of haptic feedback. Finally the stage of active rendering was discussed. There is a need of devices to stimulate all of the human senses and algorithms to control them. But the addition of multiple devices should not break the social aspect of the cinematographic experience. In this direction we presented our haptic rendering system dealing with heterogeneous haptic configuration, and a system to hide HMDs during a shared experience by multiple users.

For future work, we will continue to address all of the presented locks to go to a complete VR workflow, and will validate it through the production of real media for the cinema industry

REFERENCES

1. Fabien Danieau, Anatole Lécuyer, Philippe Guillotel, Julien Fleureau, Nicolas Mollet, and Marc Christie. Enhancing Audiovisual Experience with Haptic Feedback: A Survey on HAV. IEEE TRANSACTIONS ON HAPTICS, VOL. 6, NO. 2. April 2013, pp. 193-205.

2. F. Danieau, J. Fleureau, A. Cabec, P. Kerbiriou, P. Guillotel, N. Mollet, M. Christie and A. Lécuyer. “A Framework for Enhancing Video Viewing Experience with Haptic Effects of Motion”. IEEE Haptics Symposium, pp. 541–546, 2012.

3. X. Burgos-Artizzu, P. Perona et P. Dollar, «Robust face landmark estimation under occlusion,» chez IEEE International Conference on Computer Vision, Sydney, 2013.