# AV Renderer Specification and Basic Characterisation of Audience Interaction

Deliverable D5.1.1

| | |
|---|---|
| FascinatE identifier: | Fascinate-D511-DTO-AV-Renderer-Spec-and-User-Interaction-v08.doc |
| Deliverable number: | D5.1.1 |
| Author(s) and company: | M. Borsum, J. Spille, A. Kochale (DTO);<br>E. Önnevall, G. Zoric (TII), J. Ruiz (UPC) |
| Internal reviewers: | O. Schreer (HHI), P. Rondão Alface, J.F. Macq (ALU) |
| Work package / task: | WP5 |
| Document status: | Final |
| Confidentiality: | Public |

| Version | Date | Reason of change |
|---|---|---|
| 1 | 2010-03-18 | Document created; initial draft |
| 2 | 2010-06-14 | Updated for review of chapter 5 |
| 3 | 2010-06-16 | Added lot of review comments |
| 4 | 2010-06-21 | Completed sections |
| 5 | 2010-06-28 | Final draft prior merge with reviewed chapter 4 |
| 6 | 2010-06-29 | Final draft for internal review |
| 7 | 2010-07-16 | Adapted following comments from internal review |
| 8 | 2010-07-29 | Final version |

© FascinatE consortium: all rights reserved

The work presented in this document was partially supported by the European Community under the 7th framework program for R&D.

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

This document contains material, which is the copyright of certain FascinatE consortium parties, and may not be reproduced or copied without permission. All FascinatE consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the FascinatE consortium as a whole, nor a certain party of the FascinatE consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

# Table of Contents

**1 Executive Summary ..... 1**

**2 Introduction ..... 2**

2.1 Purpose of this Document ..... 2

2.2 Scope of this Document ..... 2

2.3 Status of this Document ..... 2

2.4 Related Documents ..... 2

**3 Definition of AV Rendering Framework ..... 3**

3.1 Scope ..... 3

3.2 Common AV Scene Description Aspects ..... 4

*3.2.1 Modes for audio & video presentations ..... 4*

*3.2.2 Format of scene description ..... 5*

*3.2.3 Possible scene description models ..... 6*

3.3 Rendering Framework ..... 7

*3.3.1 Composer / device independent rendering ..... 7*

*3.3.2 Presenter / device aware rendering ..... 9*

**4 Processing Platform ..... 12**

4.1 General ..... 12

4.2 Guidelines for Platform Selection ..... 13

**5 Prevailing Large Audience Interaction ..... 17**

5.1 Introduction ..... 17

5.2 Mobile Audiences ..... 17

5.3 Domestic Audiences ..... 19

5.4 Crowds ..... 19

5.5 Conclusion ..... 22

**6 Gestural User Interaction ..... 24**

6.1 Definition of Set of Gestures ..... 24

6.2 Gesture database ..... 27

**7 Conclusions ..... 29**

**8 References ..... 30**

**9 Glossary ..... 32**

© FascinatE consortium: all rights reserved 

# 1 Executive Summary

To increase the level of user immersion of life events presented on interactive TVs, FascinatE has defined the goal to bring operations today performed in post production within the reach of user control. Furthermore, along with access to the panorama of the provided content the interactivity itself is supposed to become more user friendly and intuitive. To reach this within the FascinatE runtime usage studies have been initiated and implementations for gesture controlled user interfaces have been started.

In a first attempt this document defines processing platforms and rendering frameworks assisting this approach. Bringing media production process closer to the user without loosing image and sound quality necessitates sharing the processing burden across the media delivery and rendering work chain. This results in the need to identify scalable and powerful platforms that map on several usage scenarios and business models.

While a specific platform definition is not possible in the current state of the project, a set of required operations is defined following an introduced model for scene composition and presentation. Additionally, a structure to organise potential FascinatE platforms is suggested. Future work is indicated to include benchmarking several platform qualities.

Furthermore, a first study on audience interaction is presented and discussed. Due to the goal to consider various end user terminal platforms ongoing work is indicated to include data collection activities concerning specific devices such as mobile phones, TVs and public screens.

Finally, a first database of gestures is defined starting with a set of basic and intuitive signals to get commands across by using simple ways of body language. These gestures will be verified and assessed in the ongoing work and will establish a profound foundation to easier user/machine communication.

© FascinatE consortium: all rights reserved 

# 2 Introduction

## 2.1 Purpose of this Document

This document provides a definition of the AV renderer framework and a selection on a processing platform. It further reports on prevailing large audience interaction and gives a definition of a set of gestures to be recognised. These gestures are recorded as a database to support further development work.

## 2.2 Scope of this Document

To define the specific goals of FascinatE's rendering framework and processing platform in contrast to other projects this document provides an initial perspective. While mentioning the challenges of distributing the effort of rendering across the FascinatE work chain, the renderer itself will be treated as single entity in order to describe the required tasks more clearly and to concentrate on the expected operation. Further details on allocating certain rendering parts are expected within deliverable D1.4.1 for FascinatE's initial system specification.

It has to be noted that most statements within this documents represent a snapshot of understanding at the time of writing and are expected to be updated as soon as the requirements and use cases are set. This is especially true for chapter 3, 4 and 6 containing the elements for framework, platform and gesture database. Chapter 5 is considered to be a description initiating the required effort to prepare deliverable D5.3.1 (Requirements for the network interfaces and interactive systems usability) due in M23.

## 2.3 Status of this Document

This is the final version of D5.1.1.

## 2.4 Related Documents

Before reading this document it is recommended to be familiar with the following documents:

- D1.1.1 End user, production, hardware and networking requirements
- D2.1.1 Draft specification of generic data representation and coding scheme

© FascinatE consortium: all rights reserved 

# 3 Definition of AV Rendering Framework

This chapter defines an audiovisual rendering framework to be used for the FascinatE workflow in order to bring the captured and produced content onto displays and loudspeakers at the user environment.

## 3.1 Scope

The Audiovisual (AV) Rendering Framework defined in this chapter concentrates on topics relevant to accomplish the task of FascinatE goals for presenting media data. Hence, it does not discuss means for coding/decoding, media stream transmission and synchronisation. Furthermore, other system specific aspects that are covered by the generic system definition as system inherent like the system frame rate will be treated as state of the art as well.

The focus is to describe and define operations and means required to represent the media streams acquired and provided by the FascinatE media production on display and loudspeaker setups at the end-user location. In contrast to common media processing workflows the FascinatE renderer is supposed to be partitioned across media delivery networks and end user terminals following the selected business model, the available processing power and the available bandwidth of the access network.

For this purpose a **FascinatE Rendering Node** (FRN) (Figure 3-1) is defined as process that translates the FascinatE view of the real world for projection on a variety of displays and loudspeaker devices. This real world model is given as a layered scene providing media streams including descriptive metadata and additional scripting parameters. The scripting parameters can be provided by the production or preceding rendering stages as declarative metadata or a list of required operations to be performed within the rendering process.

Scripting can be understood as a way to perform program applications (like JavaScript), dynamically creating object trees (as in VRML) or static operation or parameter tables building or describing a layered scene. It should be noted that at the time of writing this document FascinatE scripting as mentioned above is understood as production list accomplishing certain operations or parameters and providing an additional media stream description potentially written in a markup language. Further definition of this production scripts is beyond the scope of this document and is expected to be detailed more advanced in D1.4.1 (Initial System Specification).

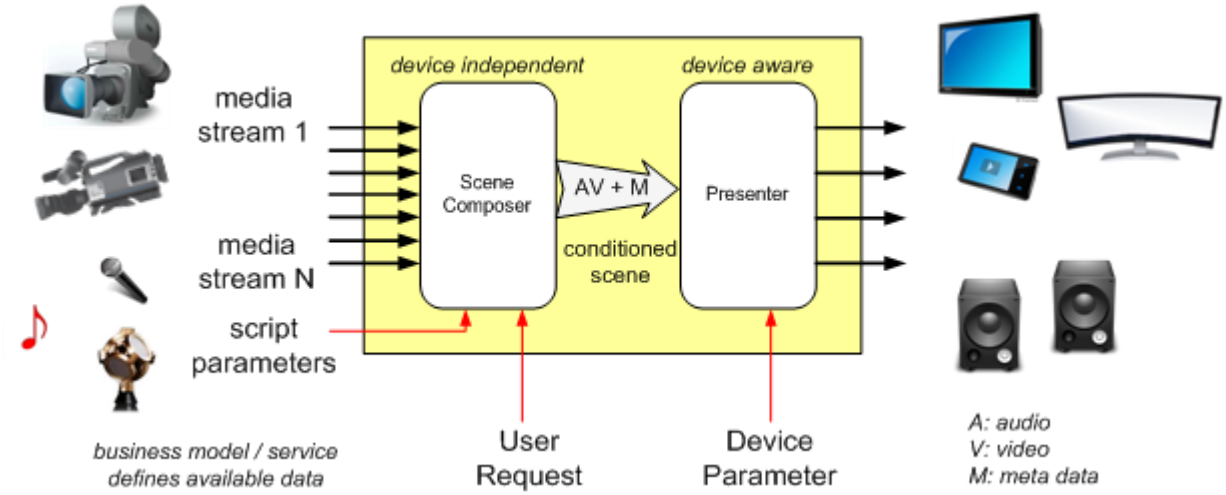



**Figure 3-1: FascinatE Rendering Node**

In the model the FRN is partitioned into a scene composer and a presenter. The streamed media data for rendering is provided by the production or by preceding rendering stages and is considered for the following discussion to be pre processed, demultiplexed and decoded, i.e. the FRN gets pure streaming data (audio / A, video / V, cf. to Figure 3-1) together with metadata (M).

To ensure encapsulated operation of FRNs certain assumptions on interfaces and corresponding data are made. It is assumed that the interfaces for streaming, scripting, user interaction and devices:

© FascinatE consortium: all rights reserved 

- Transmit media streams with negligible or small delays in order to overcome challenges caused by smooth user interaction response. This will require prefetching mechanisms and caches to be specified later, preferably during the runtime of this project.
- Provide unidirectional access to the respective blocks. Potential queries to the FRN to acquire properties and error states are treated as added feature to be described in a future version of this document.
- Reconfigures following parameters given by processed scripts, user requests (such as scene selections) and device parameters.
- Present data that does not need to be conditioned itself by decoding, retiming or transformation in a way that needs to reverse an operation of an earlier stage.

While all having composer and presenter components, FRNs have different complexity due to their location in the workflow and their respective availability of media data and script information, user request and device parameters. Such a **distributed rendering** (Figure 3-2) is supposed to offer format agnostic scaling across business model, processing power and network bandwidth. This is a significant difference to state of the art media production workflows.

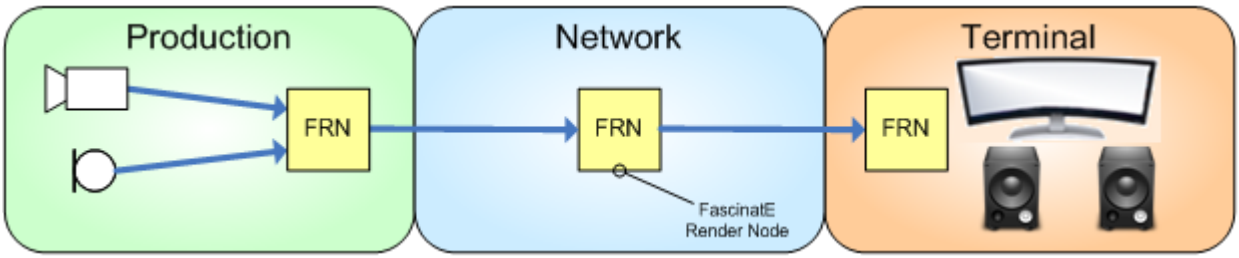



**Figure 3-2: Distributed rendering**

In current broadcast and media delivery scenarios it is common to execute all audio and video manipulating processing steps at the production side in a “post-production” stage. Processing steps identical for all customers or user groups will be executed for the FascinatE near the production site as well. These operations are most likely to be expected under control of the production itself.

To retain maximum image and sound quality, all device dependent processing steps are executed in the FascinatE processing chain as late as possible. Which processing steps are done at which part highly depends on the scenario and the business model. In the example of a cinema scenario it is affordable to even accomplish all the post-production processing at the end user side to get the highest possible quality level onto the screen or sound system. A mobile scenario demands most of the processing to be done ahead of the final media delivery to the terminal in order to reduce the processing workload of the mobile terminals.

This ‘process as late as possible’ principle is only valid for device aware processing. All device independent processing is identical for all customers and shall be done as early in the chain as possible, but will be defined by the selected business model.

For more details on partitioning the rendering process refer to chapter 3.3.

## 3.2 Common AV Scene Description Aspects

This chapter describes some common aspects of the scene description.

### *3.2.1 Modes for audio & video presentations*

The typical FascinatE panorama represents a view into the 3D real-world event. Thus camera and audio signal sources have 3D origins in the real-world.

**Video panorama**

To use this video panorama at the end terminal either some kind of surround display is needed or only a smaller part of the panorama can be shown on a flat display without visible distortion. On these flat displays the final view is rendered by perspective correction out of a FascinatE specific layered 3D scene

© FascinatE consortium: all rights reserved 

description. 3D in this context means only that 2D video planes are positioned in the 3D world coordinate system; there is no real 3D representation of the world (Figure 3-3a).

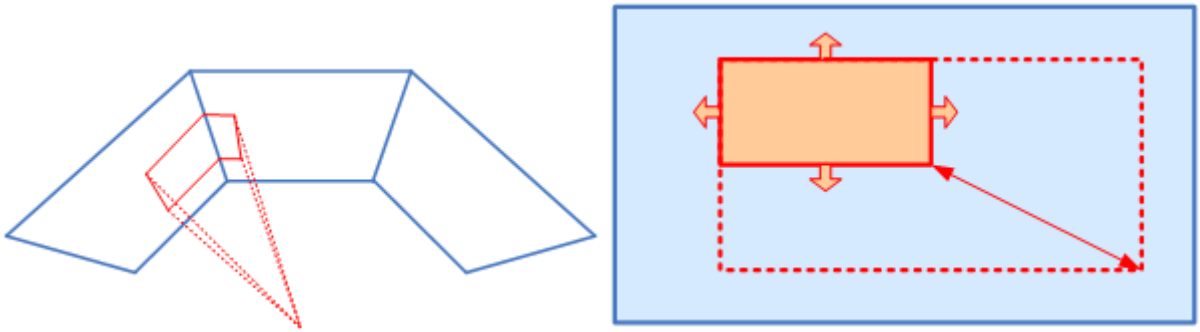


**Figure 3-3: FascinatE video panorama modes**
**a) standard 3D perspective projected b) pre-rendered panorama "FascinatE light"**

In some lower-end scenarios there might be only a pre-rendered flat panorama available for some end terminals. Because of the flat nature a head turn feature is impossible to implement. In this case a "FascinatE light" mode will be used, where the user can only select and zoom rectangular parts of this panorama. This kind of panorama can be composed early in the processing chain because it is independent from the end terminal (Figure 3-3b).

**Audio presentation**

The audio scene will represent a 3D audio sound field. This can be presented with a 3D or 2D Ambisonics system, with a more or less sparse loudspeaker setup depending on the Ambisonics order, or it can be presented with a 2D Wave Field Synthesis (WFS) system consisting of lines of loudspeakers very close to each other or a combination out of it. Furthermore, the audio scene can also be presented on typical stereo and surround setups e.g., 2.0, 5.1, 7.2 with less spatial information. Headphone or earphone presentations are also possible.

It is assumed that sound will be recorded with various microphones or microphone systems.

- Single sound source signals will be recorded with single microphones or will be estimated from other microphone signals using beam forming and additional location information. The source location will be measured directly or estimated using additional information. Later on, these sound signals with location coordinates will either be mapped to a loudspeaker (e.g., the center channel) or will be positioned in a WFS system [deVries, 2009]. Another possibility could be the encoding in Ambisonics signals, if the source signals should be presented with an Ambisonics speaker setup.
- Recorded sound fields, best used for ambience sound (reverberant or diffuse sound field), are encoded as Ambisonics description. This can either be presented directly by an Ambisonics setup or it has to be decomposed and mapped to single loudspeakers or WFS inputs.

### *3.2.2 Format of scene description*

The concrete format of the scene description will be defined later during the project lifetime. But some general requirements to fulfil the rendering tasks are described hereafter.

**Audio**

An audio scene in this context is defined as group of sound sources and/or of microphones in one location (room, stadium etc.) that captures the sound for the complete scene.

According to chapter 3.2.1, two approaches are possible to describe an audio scene.

- The first approach is to describe the sound field of the audio scene using a sound field description like Ambisonics. The necessary Ambisonics coefficients are taken from microphone arrays or are calculated for a synthetic sound scene. If a natural sound field is acquired no modifications of the audio scene are possible.
- The second approach is to describe all elements of an audio scene, i.e., the source signals, the positioning of the sources and if required also the acoustic environment of the audio scene.

© FascinatE consortium: all rights reserved 

- The elements of an audio scene may be estimated from the audio information if the parameters of the microphones (e.g., positioning, directivity) are known. However, the quality of this information highly depends on the microphone setup. This parametric approach allows modifications of the audio scene. A drawback is that so called dry sound sources are required, i.e., the sound source signals do not contain any spatial sound information. This spatial sound information (reverberation, echoes, etc.) are added as an extra sound when the audio scene is rendered.
- In practice, typical microphone signals are often wet. I.e., they contain the dry and acoustical signal components. However, they are treated like the dry signals as described above.

**Video**

To allow typical scene composition effects that the content consumer is familiar with from traditional broadcast TV, three different main categories build up the video scene (Figure 3-4).

1. Real-world (Figure 3-4 left): All camera images and also images composed or derived out of them, e.g. omnicam panorama stream, additional HD camera streams, pre-rendered panoramas.
2. Artificial, but real-world related (Figure 3-4 mid): Graphical information to be shown in the scene. They move together with the real-world video if the user changes the desired view, e.g. object highlighting, perimeter advertising boards, national/team flags during the hymn, distances and directions with circles and lines.
3. Artificial display related (Figure 3-4 right): Graphical information to be shown at distinct position on the screen. They hold their position on the screen, even if the user changes the desired view, e.g. station logo, time and scoreboard, statistical information like lap times, news ticker.

Both first categories have to be corrected in the 3D perspective based on the desired view of the distinct user to be mapped on the display. For the third category only a 2D mapping is necessary to adapt to the actual resolution and display capabilities. In the FRN these categories are processed one after the other, thus category 2 and 3 are building a transparent overlay to the real-world view.

An application at the end terminal can use the last two categories to construct its graphical user interface (GUI). With category 2 selectable objects or framing can be indicated, while category 3 is able to build up a traditional 2D GUI.

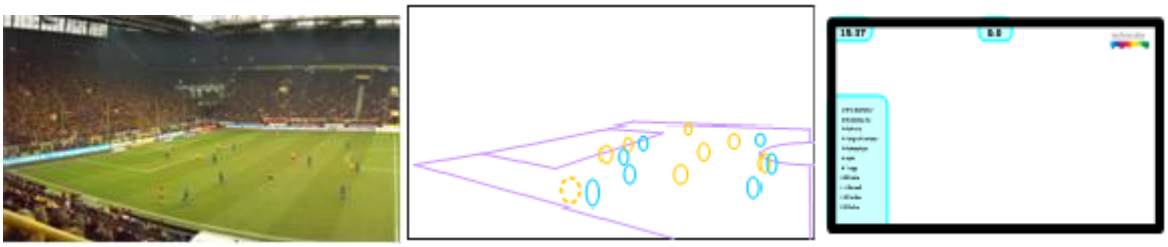


**Figure 3-4: Categories in scene composition**

### *3.2.3 Possible scene description models*

- Completely dynamic scene description with applications written in interpreted languages like Java / JavaScript is out of FascinatE scope. They would allow very customizable business models and tight control what the user can do and the way the service looks like. The drawback would be uncertain processing performance on inhomogeneous processing platforms. All FRNs would have to be designed with high performance margin to fulfil future applications.
- Dynamic processing graph for the scene description like MPEG BIFS. This would allow very complex and interactive services with the drawback of relative complex scene graph parsing. One feasible way would be to restrict the feature-set like in LASeR.
- More static, but updatable declarative scene description (like HTML/XML). This may limit the interactivity for some cases, but would allow a first straightforward implementation.

© FascinatE consortium: all rights reserved 

## 3.3 Rendering Framework

The general FRN as shown in Figure 3-1 contains several functions to accomplish the rendering task. These functions can be divided roughly into two categories:

The **composer** generates a data set representing the user selected scene. The available scene depends on the available media data with its descriptive metadata and a script to further describe the data and define specific operations to render a scene. The script can be provided by the production or from preceding rendering stages. Additionally, the scene selection requested by the user can be considered as another script configuring the composition stage. The compositor is meant to be device independent but scalable to the business model and processing needs.

The **presenter** finally renders the resulting media representation onto the end user display or loudspeaker set according to the provided device parameters. It has to be noted that the presenter is also responsible for mixing overlay information into the visual/audible media representation in order to allow for positioning of user interface widgets or content description information. Typically these widgets or fonts are terminal properties attached to a brand or service provider such as screensavers with branding logos.

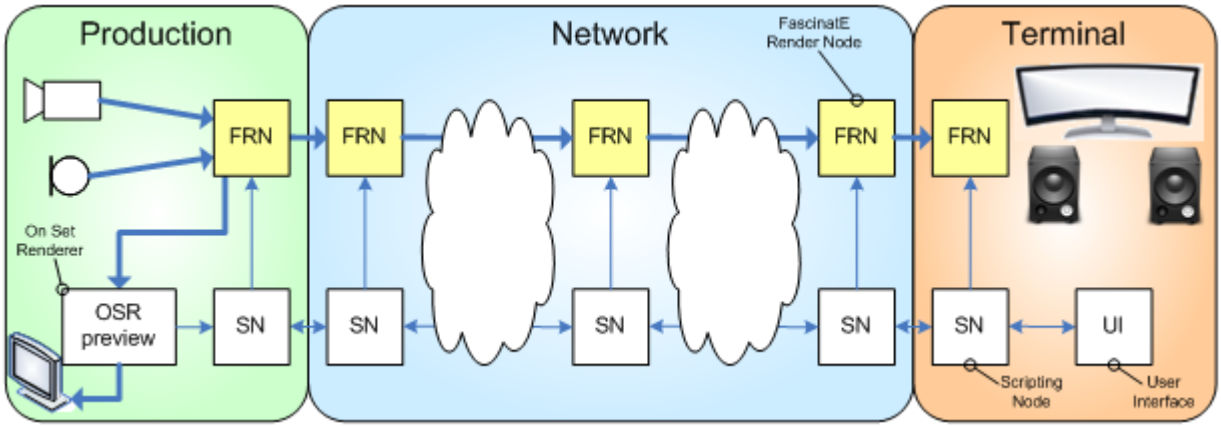



**Figure 3-5: Detailed distributed rendering in FascinatE – real world scenario**

As shown in Figure 3-5 there are different locations where FRNs can be placed between production, network and end terminal. Depending on the business model or the specific FascinatE implementation, the render functionality is spread over these nodes in different ways. There will be scenarios where individual nodes may contain no rendering at all and also some edge cases with all rendering in one of them. More information about this distributed rendering based on several scenarios is given in D1.1.1.

For example the processing intensive rendering is performed at the production side (see D1.1.1 Scenario 1 – production-centric delivery chain) and the device dependent operations at the terminal side. This will correspond to an FRN in the production employing just the compositor and another FRN at the terminal employing the presentation module.

In an example of a future full blown FascinatE work chain the FRN at the production could contain all image correction and meta-data production tasks, the first network FRN is directly connected to the production and maybe also located on-set. It does all the encoding, scaling and tiling for further network transmission. The middle network FRN implements the business model and reduces or enriches the FascinatE scene description as the user has paid for. The last FRN in the network is located near the end user and may implement some parts of the end terminal in mobile scenarios but also is meant to perform user related rendering. Finally, the FRN in the end terminal does all the decoding, presentation and handling the user interaction.

### *3.3.1 Composer / device independent rendering*

The composer block of the FRN shown in Figure 3-1 is responsible for all processing that is common for all or at least a distinct group of end terminals. For all these processing no information about capabilities of the end terminal or the desired view of the user is necessary. Therefore these functions will more likely be placed in FRNs that are located near the production site.

© FascinatE consortium: all rights reserved 

All data suppression that is done in the composition block (e.g. scaling or selection) is based on the scenario, the business model and bandwidth profiling of the delivery network.

**Composer for audio renderer**

Audio scene selections, rotation and remixing, controlled by the script parameters or user request, will be processed in the scene composer. The output of the scene composer is still end user device independent, like the original input material.

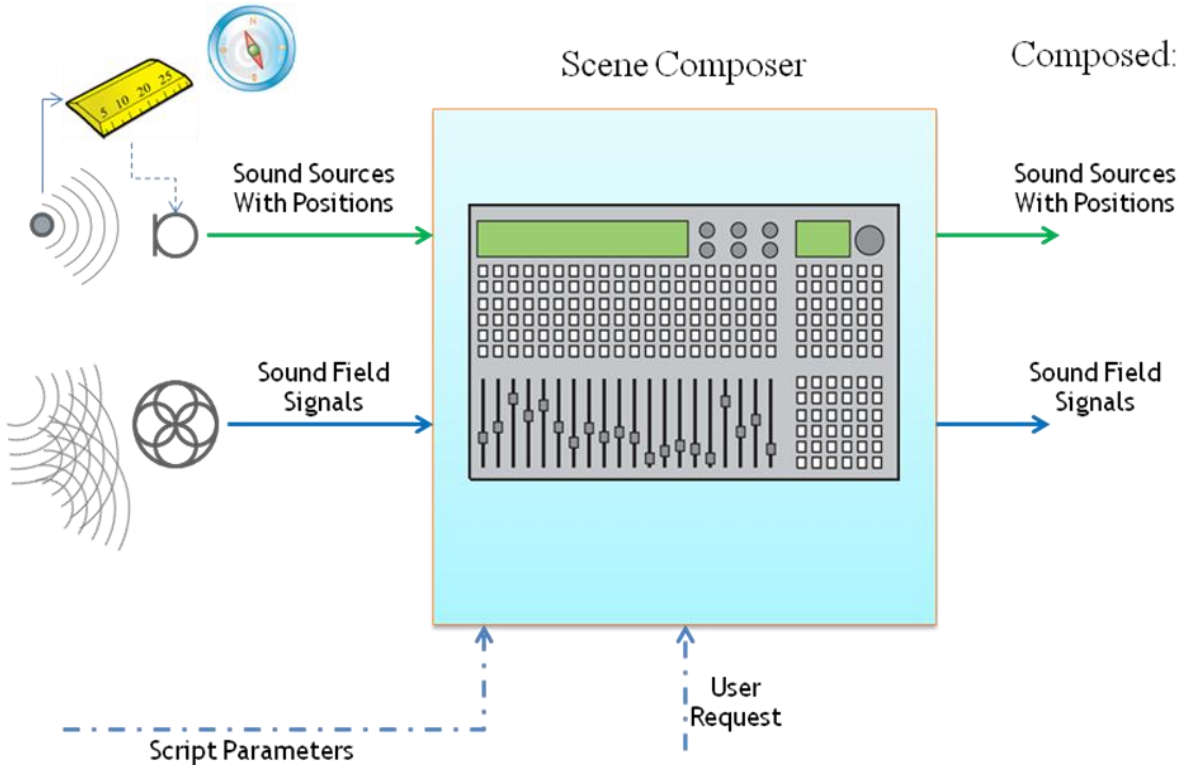



**Figure 3-6: Audio Composer**

The input contains either one or more sound source signals including positioning data and/or one or more sound field signals, also called Ambisonics signals, as described in the audio panorama section (section 3.2.1).

All input signals are processed by the scene composer either via script parameters or via user request.

With script parameters, the producer can select or manipulate audio signals in the FRN to give the user an optimal result. The user request enables the user to manipulate the audio scene, to his needs.

Depending on the script language, the implementation will be a script interpreter (e.g., JavaScript), dynamically created object tree (as in VRML) or a static list of operations. In the later case the audio composer can be seen as a remote controlled audio mixing console.

Typical operations would be:

- selection/mute
- mix/group
- adding jingles
- effects, like compression, delay, echoes, or reverb
- filtering (e.g., equalization, treble, or bass)
- update metadata (e.g., changing position metadata or rotating an Ambisonics field)

The output format will be the same as the input format, i.e., a collection of sound source and sound field signals. This enables the FascinatE system to cascade several audio composers from the production side to the end user terminal.

© FascinatE consortium: all rights reserved 

**Composer for video renderer**

Functionality that belongs to the composition:

- All post production tasks for image enhancement and correction; like removing lens distortion, adapting the colour and luminance,
- All post production tasks to allow the scripting engine to generate metadata and additional information layers, e.g. registration of all video signals to the FascinatE world coordinate system, object tracking and recognition, Region-of-Interest (RoI) definition and framing proposals for different screen aspect ratios. This potentially expands the composed FRN model to have also an output feed to the scripting engine. Details on this will be added in a future version of this document.
- Image scaling and tiling for better network utilization and bandwidth reduction.
- Recoding, e.g. from raw or lossless formats to typical broadcasting formats.
- Caching of all or parts of the scene for later use as on-demand service for time-shift or replays.
- Selection/reduction of image material based on the business model.
- Adding commercials and other additional data, even from other content providing sources.

All composition tasks are performed by modifying the FascinatE scene description. This means adding, removing or replacing video elements from the scene description

### *3.3.2 Presenter / device aware rendering*

The presenter block of the FRN shown in Figure 3-1 is responsible for all device and user selected viewport dependent rendering. Therefore these functions are typically located inside or at least close to the end terminal. To allow some parts of these to be rendered in the network the relevant device parameters and user viewport have to be shared to these rendering nodes. In the case of unavailability of the return channel all presentation rendering is located inside the end terminal.

**Presenter for audio rendering**

All audio transcoding is independent from the scene selection and heavily rely on the device parameters and will therefore be processed in the presenter part of the FRN.

© FascinatE consortium: all rights reserved 

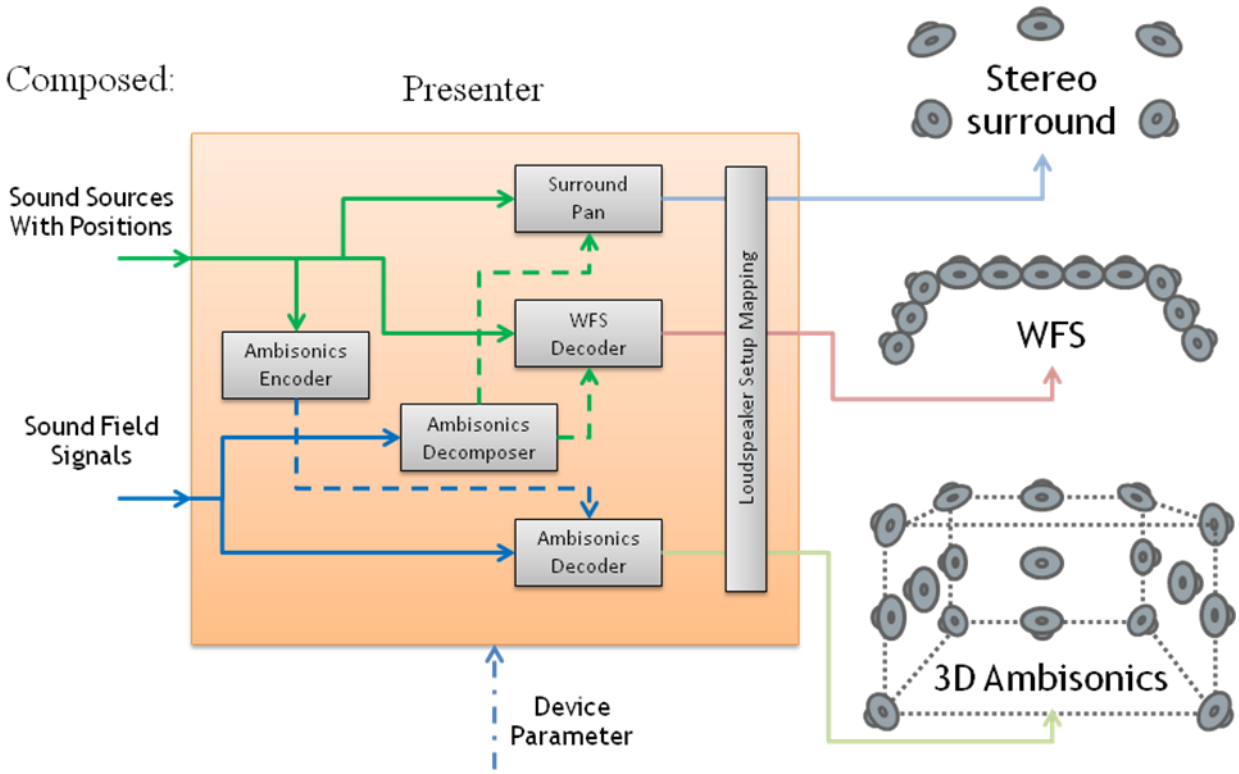



**Figure 3-7: Audio Presenter**

The inputs of the presenter are connected with the composer's outputs.

All input signals are processed by the presenter only depending on the device parameters.

As can be seen in Figure 3-7, for a simple **stereo/surround loudspeaker** setup the *sound source signals* with position information will be panned to the appropriate loudspeakers using a surround panning unit, like it is used in studio mixing consoles at the production site today. The output quality and the panning law, i.e., the weighting of the different loudspeakers, is device dependent and will not be specified in this document.

- *Sound field input signals* have to be either decoded by an Ambisonics decoder unit, optimized for an unsymmetrical loudspeaker circle, and have to be mapped onto the surround loudspeaker setup with the loudspeaker setup mapping unit
- or they have to be decomposed in a set of directional signals and panned as other sound source signals with the surround panning unit. This is known as Directional audio coding (DirAC) method from the literature. [Pulkki, 2007]

For a **WFS** system the *sound source signals* with position information will be directly processed by the WFS decoder. The *sound field input signals* can to be decomposed in a set of directional signals, e.g., a set of plane wave signals coming from various directions and can be treated as plane waves in the WFS decoder.

For a 2D or 3D **Ambisonics** setup the *sound field signals* can be directly decoded by the Ambisonics decoder unit. The *sound source signals* with position information have to be either encoded as Ambisonics signals first before they are processed by the Ambisonics decoder or they can be processed using the surround panning unit and mapped to the loudspeaker setup with the loudspeaker setup mapping unit. Panning a sound source to a 3D set of loudspeakers is known as vector base amplitude panning (VBAP) in the literature. [Pulkki, 1997]

© FascinatE consortium: all rights reserved

**Presenter for video rendering**

Functionality that belongs to the video presentation (chapter 3.2.2 for categories):

- Determination of visible elements/streams out of the scene description and thereby reduction of current frame scene graph by unused parts. This decision is mainly based on what can be seen with the user selected viewport, but also on some binary switches if the user enables/disables additional information layers.
- All perspective correction, stitching and blending of category 1 video elements from the scene description. The perspective to be currently used is defined by the device parameters of the connected displays and the viewport the user has selected.
- Generating artificial scene parts. This means perspective correction and alpha blending of category 2 elements from the scene description. This processing is also based on the user's viewport. If the user interface has enriched the scene description with GUI elements of category 2 these are rendered together with the normal elements.
- 2D-rendering, bitmap scaling and font rendering of all category 3 elements from the scene description. If the user interface has enriched the scene description with GUI elements of category 3 these are rendered together with the normal elements.

**Device parameters relevant for audio & video rendering:**

- Number and arrangement of speakers. Especially if a standard speaker setup such as stereo, 5.1, 3$^{rd}$ order Ambisonics or wave field is available.
- All speaker characteristics that have to be compensated, e.g. delay and frequency range for each single speaker.
- For high-end setups with Ambisonics or WFS also the real angle/position of all speakers.
- Number, type and arrangement of locally connected displays. Especially the angles and positioning of multiple displays to each other.
- Characteristics for each of the displays like resolution in pixels, maximum supported colour depth and possible frame rates.

© FascinatE consortium: all rights reserved 

# 4 Processing Platform

The objective of this chapter is to assemble guidelines and recommendations to select a processing platform for rendering FascinatE media data. It should be noted that at the time of writing this first version most assessments were based on assumptions extrapolated from state of the art equipment that processes FascinatE specific operations still in planning and development. This document can therefore only provide guiding hints.

## 4.1 General

Within the context of this chapter the **processing platform** is understood to be the technology of a rendering system's hardware and software that operates FascinatE specific applications. As the rendering process itself is supposed to be widely scalable, the provided guidelines have to be organised according a potential platform profile and implementation level. Examples are given whenever considered to be meaningful.

**Profiles** will classify the processing platform into feasible operations, i.e. if the operation can be performed by a platform class or not. **Levels** on the other hand, will group supported features such as parallel camera streams, number of audio channels, etc.

Following the conditions given in the requirements document D1.1.1, the processing platform has to be scaled according the established service model, the available rendering resources and the user selected terminal. Expressed as implementation complexity – as indicator for the functional density within a specific platform form factor - over processing performance, the variety of different processing stages utilized in a typical FascinatE workflow is depicted in Figure 4-1.

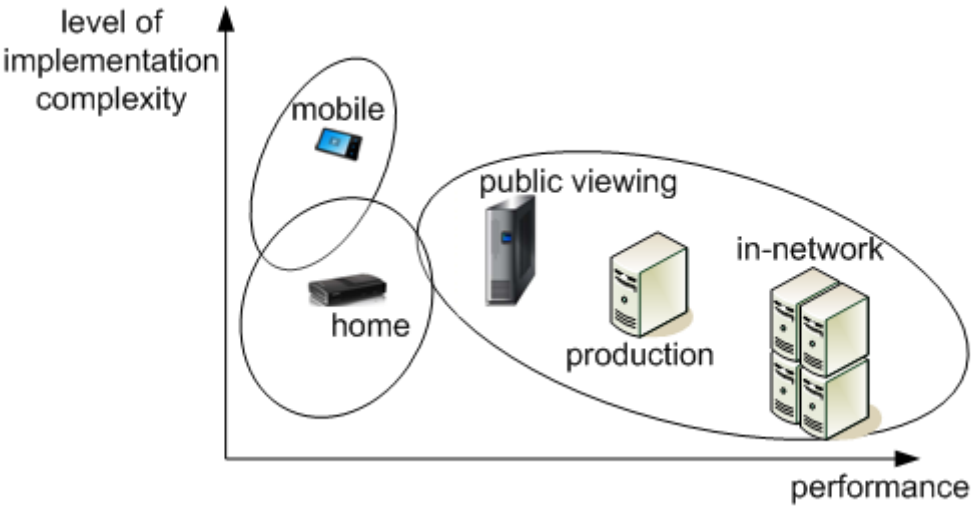



**Figure 4-1: FascinatE platform complexity diversification**

Basically these various platforms can be grouped to have **similar processing requirements** following their stage in the FascinatE workflow, i.e. related to their operation in production, in-network or as end-user-terminal. FascinatE's goal is to bring more operations traditionally performed in the production in the reach of control of end-users. This requires moving processing complexity up the content delivery and presentation chain.

Nevertheless, as long as the network infrastructure requires suppressing media content that is produced in order to overcome the access bandwidth bottleneck to the user terminal, it is expected that the most complex operations are performed off the production site in the network. This allows for rendering farms to fulfil needs for specific service models and usage groups.

Beyond such performance groups are certain **platform qualities** which need to be satisfied. These qualities scale by the required tasks for streaming media data and processing images or sound signals.

© FascinatE consortium: all rights reserved 

Potential platform qualities can be summarized as such for:

- Processing performance
- Access bandwidth or data throughput (input as well as output)
- Scalability of architecture (single core standalone or multi core rendering farm)
- Offered level of user interaction, i.e. interactivity
- Low power consumption, i.e. integration level
- Effort to support immersive life event experience

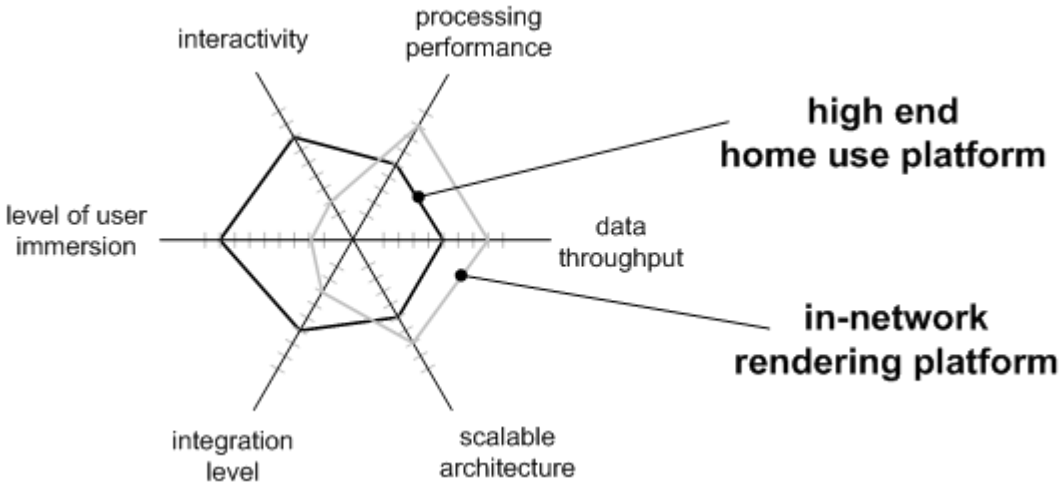



**Figure 4-2: Examples of weighted qualities per platform**

Processing platforms accomplish specific tasks in this context. Examples are Linux based media gateways as network extension in the home environment, high performance media servers acting as standalone rendering station for public viewing or thin client smart phones provided with pre rendered content from the FRN located in the network.

The qualities or characteristics mentioned above can now be **weighted for every platform** as shown in Figure 4-2 in order to assess the corresponding hardware architecture and software framework. This will require determining proper components picked from processor roadmaps and mature libraries for graphic, image and signal processing with the emphasis on open sources and standards to have market wide quality control. The content security challenge this imposes needs to be considered for conditional access, digital rights management and f.e. vulnerability of usage profiles like on every content access platform with advanced features.

It also has to be understood that FascinatE functionality needs to be leveraged by adopting an evolutionary approach that allows FascinatE tasks as added feature on existing platforms.

## 4.2 Guidelines for Platform Selection

While it is not possible to specify a generic FascinatE platform for every intermediate or final processing stage some basic assumptions can be gathered in order to survey the variety of feasible platforms.

Basically FascinatE is assumed to be an add-on for end-user-terminals, hence raising the bar for certain processing requirements. On the other hand it also can be understood that current terminals can feature FascinatE functions if the selected service model provides sufficient processing performance, bandwidth, etc. in preceding stages within the network or at the production site. This assists an evolutional introduction of the FascinatE service.

Intermediate rendering platforms for the production server and in-network renderer are considered as standalone or clustered systems scaled to the budget of producers and service providers.

Tasks of processing platforms can be divided into media stream access, rendering, system configuration and control and user interaction messaging and control.

© FascinatE consortium: all rights reserved 

These tasks are performed typically in functional blocks such as:

**Hardware related**

- Physical link to a network
- Central Processing Unit (CPU) – also hosting the system software, potentially with multiple cores
- Graphical Processing Unit (GPU) – commonly used for graphical overlays and games
- Digital Signal Processing (DSP) – like processing for audio data manipulation
- Storage components (secondary and primary)

**Software related**

- Low level driver (hardware abstraction)
- Operating system and/or middleware
- Application software providing also the graphical user interface and handling of user command ingest done for example by user gesture detection methods and processes.

To scale the processing platform according its position in the FascinatE distributed rendering chain, certain performance groups are defined in the following. As mentioned above, the scaling is done in profiles and levels.

**Profiles** are considered to express the ability for certain functions. Examples for this may be the function to change the visual perception just by selecting a certain camera cluster feed output (cluster hopping) or the ability to render the users own region of interest (RoI) with blending several camera layers and reframing the new perspective. An audio related example is the capability to perform a zoom on acoustic sound sources that compares to the visual zoom to enlarge an object.

| PROFILE | NAME | COMMENT |
|---|---|---|
| P1 | LOW | Simple mobile or home terminals |
| P2 | MAIN | Mainstream home terminal, high end smartphone |
| P3 | HIGH | High performance server, high end game station |

**Table 4-1: FascinatE platform profiles**

In contrast to that provide **levels** a reference about quality and quantity of the implemented functionality. These are number of audio channels, supported media streams or decoders or featured degree of interaction. Levels will be defined during the runtime of the project in order to address the feasible scalability of FascinatE platforms. For the time being separate levels for audio, video and interaction are considered to be meaningful (L.A, L.V, L.I).

While profiles indicate the programming complexity, levels are seen as hints to assess the respective processing performance.

Examples (with H = HIGH, M =MAIN, L=LOW):

- Public server P/L.A, L.V, L.I: `P/L [H / HHM]`
- Production/Network server P/L.A, L.V, L.I: `P/L [H / HHL]`

Following the business model and/or the selected service concept the various processing platforms will be balanced across the FascinatE work chain. Mainly this will influence the selected infrastructure and the required effort to spend resources in hardware and software. Roughly this can be expressed as thin or thick processing stage. Along with the profiles this leads to estimates about the expected level of integration and performance needs. A first estimation is summarized in Table 4-2.

It has to be noted that this is actually a more structured way of sorting rendering platforms compared to the rough sketch in Figure 4-1. The indicated thick in-network rendering platform can therefore also be understood as a rendering farm made of thin units. Furthermore, multiple thin units for the home scenario might well map onto distributed rendering farms that follows metropolitan area network like infrastructures with units for communities or buildings.

© FascinatE consortium: all rights reserved 

| | PRODUCTION | NETWORK | END-USER |
|---|---|---|---|
| MOBILE | thin | thick | thin – mid |
| HOME | thin | multiple thin | thin – mid |
| PUBLIC | thin | thin | mid |

**Table 4-2: Platform resource effort across FascinatE work chain**

To select appropriate hardware and software components a **benchmarking** of the performance of the implemented FascinatE processing has to be defined and performed. Available benchmarks typically focus on certain aspects of a standalone system, like on the 3D visual rendering performance expressed in achievable frames per seconds.

For FascinatE therefore an own collection of tools need to be defined to assess the audio- and video-processing speed and quality and the interactivity response as indicator for the level of activity of the system. Other FascinatE platform characteristics of interest here are given in chapter 4.1 as platform qualities.

While intrinsically format agnostic, the initial baseline for such benchmarks will be the specification of main parameters defined in the system specification document D 1.4.1. Examples for this may be the featured minimal data throughput or the selected sound sample resolution.

The outcome of this assessment indicates the choice needed from component **roadmaps** following the respective stage of the considered platform. This typically involves decisions on embedded computing or high end standalone components that balance tolerated power consumption with needed processing power.

For the time being only few hints can be formulated to overcome potential **bottlenecks** or processing challenges. A list of indications is provided in the following:

- Multiple video data streams per camera cluster might require for parallelisation beyond multithreading such as multi core operation. It might be meaningful to provide one core per camera cluster if those clusters contain more than one camera perspective.
- With image registration done at the first rendering node, following processing stages that need to correct the visual perspective resulting from user requests may have to deal with basic texture operations in a graphical processor. As the platform might need to deal with several high data rate video streams at once, it might be more critical to get the data from the CPU to the GPU. Such a requirement can be formulated as memory access bandwidth to transport parallel media streams. Current GPU roadmaps do not target this challenge sufficiently. GPU and CPU on the same die may render this issue less critical [Vuduc, 2010].
  More generalized this can be expressed as requirement of data transport between different rendering nodes.
- Processing accuracy and computational time to reach real time behaviour are further challenges to get covered. Examples for this are the selection of 64bit integer or floating point calculations needed for Ambisonics sound super positioning. Depending on the selected algorithm this may be more or less critical.
- The latency of the whole system needs to be minimized in the way that for certain user requests the system can react smoothly allowing a fluent navigation and swift acoustic or visual perspective change. This demands that in certain conditions the respective processing platform does not need to update internal data structures, i.e. necessitates caches. With latency in this context the delay between a request, like a perspective change, and the real data update is understood. It summarizes the required processing duration for the rendering steps of the processing platforms involved and the transmission of the data across a potential heterogeneous network.

© FascinatE consortium: all rights reserved 

**Summary**

At the time of writing this document, a selected processing platform is considered to have a scalable hardware architecture like any state of the art high performance PC platform or game station. To overcome potential bottlenecks regarding memory access and multi stream processing a multi core platform with multi channel memory access seems to be meaningful. It also requires a state of the art network access and a mainstream graphic processor with adequate high throughput data link towards the CPU.

As mentioned earlier it is further expected to rely on open sources and standards for the software framework. While LINUX as operating system and OpenGL as graphics library may be a natural choice, little can be stated so far on the required middleware and modules needed for implementing the algorithms for FascinatE processing.

© FascinatE consortium: all rights reserved 

# 5 Prevailing Large Audience Interaction

## 5.1 Introduction

When we are changing our ways of using media technology and in particular when we are designing for new ways to watch and interact with TV-screens, we need to account for the ways in which TV interaction is fitted into broader social contexts. In recent years the use of interactive technologies, like mobile phones and computer games, has made the public more used to advanced technological interaction. On the other hand watching TV together is an interaction that has been going on for decades. It is therefore important to study what people *actually* do in front of TV screens.

TII has mainly been focusing on public sports viewing during the first months of this project.

Through ethnographic studies in sports bars in Stockholm we have studied what people *do* through focusing on bodily actions like waving, pointing and touching, together with what people say in conversations, comments on what is happening on the screen, and outburst like screaming. It is during the peak events in sports viewing, like a goal, that a lot of "social action" takes place but definitely not all of it. In *Unpacking the Television: User Practices around a Changing Technology,* Louise Barkhuus and Barry Brown discuss "the *sociability of television*, and how TV, contrary to its image, is linked into the sociability of the household" [Barkhuus, 2009:15:9]. For FascinatE it is also crucial to study how the TV and other devices connected to a broadcast can effect and be affected of people's television content use. This section will develop a focus on users through studies in end-users environments.

## 5.2 Mobile Audiences

With mobile audience we refer to users of mobile phones or tablets, and similar small size devices. At the time of writing, TII has not done ethnographic studies on mobile audience. However, there exist several related studies on mobile TV (users). Next we present issues relevant for FascinatE.

It was shown, e.g. in [Harper, 2006], [Cui, 2007] that content is one of the key aspects for success of mobile TV. Special made-for-mobile content results in higher user satisfaction as reported by [Buchinger, 2009]. On the other hand, for a success, mobile TV needs to be easy and intuitive to use, not requiring much time for learning how to use it.

As suggested by authors in for example [Buchinger, 2009], [Knoche, 2007] or [Cui, 2007], some of the key factors for mobile TV from Quality of Experience (QoE) perspective are:

- **User (profile, motivation and needs).** Depending on the age, gender and culture, users are interested in different content (news, entertaining, sports…). Along with this goes personalization of the content, e.g. through user profiles. Generally, mobile TV is used to [Buchinger, 2009], [Cui, 2007]:
    - Kill time while waiting (longer or shorter breaks, commuting),
    - Stay up to date, i.e. with popular (sport) event,
    - Create a private sphere,
    - Feel less lonely (e.g. lunch break, if there is nobody to spend it with),
    - Relax,
    - Be entertained (in free time),
    - Own, share or exchange content,
    - Show novelty (a desire to be first),
    - Create or consume personalized content.
- **Mobile device.** Mobile device size, processor and battery constraints influence mobile TV experience. Another important and complex aspect is sound control, especially in public space.

© FascinatE consortium: all rights reserved 

In Repo et al. study, the participants expressed distress using sound in public [Repo, 2004]. This issue needs to be investigated further.

- **Context**. Viewing happens both individually and as shared experience (especially for young users during school breaks), but also as a way to separate from others, or for secret use (situations where watching TV is not suitable) [Cui, 2007]. Mobile phones allow private consumption because of short viewing distances and the viewing angle [Knoche, 2007]. Two main mobile TV use contexts are differentiated:
    - *Public space* (public transportation, waiting rooms, coffee shops etc.). It is used on the go to kill time, to be entertained or to be updated. According to Knoche, users are worried about becoming too absorbed in viewing, and thus distracted from other tasks while being on the move, e.g. missing trains [Knoche, 2004]. They also fear increased risks of accidents [Oksman, 2007]. Generally, users appreciated condensed information suitable for breaks and waiting periods (no longer than 10 minutes), and in situations when user is mobile (walking, cycling), audio is preferred [Oksman, 2007].
    - *Private space* (at home, at work, private car etc). It is usually used to relax, before going to sleep, but also in other situations because of no need for negotiation with other family members, ability to control where content is watched, comfort and privacy [Cui, 2007].
- **Content.** Most interesting genres for mobile TV include news, music, sport and live broadcast, similar as for traditional TV, but the content needs to be adjusted. Interesting for FascinatE is the question that authors ask in [Buchinger, 2009]: *How to compose made-for-mobile sports content, e.g. soccer match? On the one hand it is necessary to show the whole scenery of the match to understand what is going on; on the other hand it is necessary to show very small details like the ball to be able to follow the game.* At the same time authors in [Cesar, 2009] argue that *mobile TV is bigger than you think*, taking into consideration the viewing distance, i.e. viewing ratios and other parameters like resolution and frame rate. Also, in [Buchinger, 2009], [Oksman, 2007], the importance of user-generated content, audio and video sharing, for mobile TV is emphasized.
- **Technical performance.** Especially for live sport events, the immediacy of the content is very important, which is very well illustrated with example from [Buchinger, 2009]: *Consider the case that there is a significant time delay of several seconds between the transmission of standard TV and Mobile TV. An audience seeing the scoring of a goal in an important championship with noticeable delay is likely to be severely annoyed, even if this delay is as low as 15 seconds. In such a situation, viewers of the delayed presentation may hear cheering from their neighbours, but still they wait for the goal.* One of our participants used the same example when talking about real-time performance in the workshop we organized (more details about the workshop can be found in D1.1.1, section 3.2 Requirements). Also, response time for switching between channels needs at least be comparable with traditional TV [Cui, 2007].
- **Social connectivity.** In [Schatz, 2007] mobile social TV is defined: *It is a form of computer mediated communication which creates a joint mobile TV watching experience.* With the integration of collocated social and mobile TV, special care needs to be put on user's limited attention resources. Besides remote joint watching of mobile TV, it is also possible, with larger viewing distances, to share the screen with small groups, and it carries more social meaning than watching TV together in a living room since it requires invitation by the device owner [Cesar, 2009], and it is more intimate, cheek to cheek experience [Cui, 2007].

Work in [Knoche, 2004] investigated mobile users' needs and expectations of future multimedia services, among which were live sport events. According to it, following live events on mobile phones is interesting for users when they are not able to watch the event at home or in pubs on the standard TV. In the same study, participants expressed they would rather experience live sport content in a group.

Generally, in above mentioned studies, participants were concerned about several issues connected to mobile TV: lack of decent content, battery life (a threat to more important communication needs, [Knoche, 2007]), phone/screen size, too slow channel changing (up to 10 seconds) [Cui, 2007], expensive consumption of the content [Knoche, 2004].

Future data collection activities will focus on viewing live events on mobile phones - the aim is to investigate users' needs connected to open issues like content made-for-mobile, social connectivity,

© FascinatE consortium: all rights reserved 

interactivity, immersion and charging methods in order to give basis for a successful employment of mobile FascinatE services.

**Secondary screen**

Mobile audience also assumes people who use mobile device as a secondary screen for previewing, replaying or enhancing content, e.g. mobile phones or tablets are often used in the living room in combination with TV screens, or in a public arena with large TV screens. More on multiple screens and interplay can be found in Deliverable 1.1.1, subsection 3.1.1.

## 5.3 Domestic Audiences

As pointed out in Deliverable 1.1.1, subsection 3.1.1 on Interaction practices, the "TV is regarded as a lean-back medium with its users often ascribing a relaxing and passive attitude towards it" [Tsekleves 2009:202]. With that said it is still a highly social medium since people gather to watch, and day-to-day talks sometimes consist of subjects from the TV repertoire.

At the time of writing, we have organised a three-hour pilot observation with accompanying video- and audio material, of a soccer game watched (by five viewers) at a person's home. Although this material has not been properly analysed (aspects of interaction with the screen/content, gestures and other will be further analyzed) some initial points from the home observation can be made. The TV is an interactional piece of furniture and the content shown is promoting talk. The body movement is similar to the ones we have observed at sports bars such as a lot of pointing towards the screen and waving arms. The arm movements seem to be as important here as in the public (**Figure 5-1**) Another similar observation is viewers talking to each other while watching the screen.

**Figure 5-1: Use of arms for cheering**

We need to make additional investigations in people's homes and lives. We have discussed the possibility to follow-up that investigation with a study of a family and their interactions with the TV screen and other devices in a home environment. Some studies have already been done in this area, but very few uses classical fieldwork. We have for example work in [Barkhuus, 2009] that made in-depth interviews with 21 participants regarding their use of TiVo (a digital video recorder). The main interest would be the families' use of the TV screens and how this is connected to other device-use. How are screens and other technical equipments formulating families' perception of spare time?

## 5.4 Crowds

Our main research has been done on crowds and public screens, specifically in sports bars (Figure 5-2). In this section we will concentrate on FascinatE's connection to live sport.

© FascinatE consortium: all rights reserved

**Figure 5-2: A typical pointing (towards the screen) in a sports bar**

**Crowd heterogeneity**

The type of the crowd might have relevance for the interaction with the screens. During our first observations at a sports bar we saw that the crowd consisted of everything from couples to big groups of friends and on this occasion they were all watching different games. The sports crowd also seems to vary over the weekdays, which was particularly obvious at one of the sports bars, in the City Centre. On a Tuesday, the spectators and visitors consisted of a majority of groups of four to six men. The game broadcasted on the TV's was part of the football tournament Champions League. On Friday, the bar was not having as many visitors. On this occasion, the visitors also included lots of women. The game played was ice hockey (Sweden – Belarus, Olympics). This seems to be in line with studies in [Levy, 1989] on sport crowds on arenas, which distinguish between different kinds of sports crowds depending on the days of the week.

According to [O'Hara, 2008a] "[t]he relationship between potential audience members and what they are doing in particular locations bears on how they come to interact with the screens". Action taking place on a TV screen reflects how body language and eye contact dramatically change during a game. The movements sometimes go from concentrating almost motionless to waving arms and shaking heads. This could be viewed as a physical reaction to the understanding of the game. O'Hara's study has highlighted how players, through their interactions with the in-game elements and characters, both create and advance the narrative of the game [O'Hara, 2008b]. This is the basic of how experience is created and spread. The reaction and interaction with the screen could therefore be seen as the way the viewers of a live sport event create different narratives. Kenton O'Hara, Hazel Grian and John Williams although points to the danger of simplifying the descriptions of a crowd, i.e. to "homogenise the player and level of participation whereas in actuality there are many different levels of participation ranging from active and obsessive to relatively passive and casual" [O'Hara, 2008b:131]. They here focus on "game players", but we believe that it is applicable also when it comes to the design of interactive TV applications for example sport viewers.

We will now give you four analytical points that could be of importance for a user-friendly FascinatE.

**Rhythm and talk**

Sports have a rhythm or synchronisation involved that effect how the game is told by TV-producers and video mixers. The knowledge of this makes it easier to know when you can speak to not interrupt an exciting event. In our video material we have found that the language many times consist of agreements not arguments. With that said it is not always that people agree about what is happening on the screen but it seems like long debates or disagreements are rare. The viewers seem to need the same perceptions of the game not to interrupt an interesting event.

The public and social setting is important in forming a sort of context for viewers' actions. The sports bar environments stimulate a so-called *interrupted talk*. It is not necessarily a less personal one but it creates a focus on something more than the relationships between the members of the crowd. In a much generalized manner we could say that this kind of environment makes talk between some easier and between others harder. This interrupted talk is not always predictable since it is not only during the pauses that fans talk. It is somehow reflecting their engagement in the game. On the other hand, physical expressions like waving, pointing and utterances, happens frequently. This is what binds the experience together.

© FascinatE consortium: all rights reserved 

**Repeating and imitating**

The concept of the audience is imitating or repeating. The *audience wave* is one example when a whole arena is doing a live repeating motion and *applause* another one. From our video material we have found that viewers often repeat the same utterances as each other. In some examples there are also physical imitations to what is happening on the screen. One example is a man who starts to wobble back and forth to his left and right side when he is watching an exciting downhill ski competition. This could be seen as an immersive TV-experience.

**Synchronized cheering to screen and others**

As an audience it is important to time the cheering for a feeling of togetherness. This is also connected to what is happening on the screen. One example from our findings is when a person first celebrates to the screen and then turns to his friends. He also turns to other people in the crowd, not belonging to his group, and cheers (Figure 5-3). It is obvious that this is a social event, not only affecting the persons you are there with but also everyone at that specific place.

**Figure 5-3: Synchronized cheering (The person to the far right is cheering but none of the others at his table. He then looks at other people in the room)**

It also seems like interaction is taking place through speaking with the device. We have seen viewers shouting questions to the players (TV). The viewers are thereby interacting with the screen even if it doesn't "give" anything back. Reeves et al. argues that in situations when people are watching TV together, face-to-face interaction might be as important as their individual interaction with the screens i.e. "interaction affects and is affected by spectators" [Reeves, 2005:741]. Another way to interact with the TV is by leaning forward so you get a closer look on what is happening on the screen. This becomes a way to be a part of the action on the screen. In some ways this creates an immersive TV through an embodied experience.

**Appropriate cheering**

There are different ways of cheering and they are all affected by rules. At a filmed episode at a sports bar a man is cheering for a fall in a downhill skiing event. The others at the table are not having the same expression. One man in the company hardly looks at the cheering guy and the others just glance at him. The question is why they are cheering differently. One explanation could be that it is not good sportsmanship to cheer for a fall and not a victory. Even the commentators do a configuration of the cheering with their "It was not fun for her, of course, but she may enjoy the gold from yesterday". This becomes a discussion about what is appropriate cheering, which has something to do with appropriate body movements. How do you cheer? When do you not cheer? How long do you cheer and so on. The cheering is depending on the understanding of the sport and makes you more physically connected to what is happening on the TV. Gestures are therefore important to study and analyse when it comes to how fans create a common culture around TV watching.

**Gesture interaction with screen**

Although gestures are an important part of sports viewing, not all gestures are appropriate to be used as templates for FascinatE. If gestures are used to interact with the device they need to be quite different from the gestures that are produced through watching sports. It is also important to point out that gestures can be quite problematic and make the interaction with the screen harder. Figure 5-4 shows a

© FascinatE consortium: all rights reserved 

crowd interacting with the screen through “extreme waving”. On the big screen there is a quiz and you answer the questions by standing on a specific spot. Although this is somehow a part of the game it also demands a lot of actions from its viewers and participants.

**Figure 5-4: Crowd interacting with the public screen**

The main point is that the possible use of FascinatE through gestures should not enforce a new physical interaction between the end-user and its device but work as a strengthening for the general interaction.

## 5.5 Conclusion

In this section, we have covered basic characterisation of an audience in different settings based on literature review and on studies done so far. Key points are:

- Mobile audience:
  - Various contexts: on the go, public space (public transportation, coffee shops, waiting rooms etc), or private space (at home, at work, private car etc.)
  - Personalization of the content and specifically made-for-mobile content
  - Social connectivity
  - Mobile device characteristics; mobile phone as viewing and communicational device
- Domestic audience:
  - Social context (family and friends watching together)
  - One or multiple users with hierarchy of users, possible use of multiple screens
  - Relaxation
  - Use of gestures while watching sport content on TV

© FascinatE consortium: all rights reserved 

- Crowds:
    - Multiple users with one common screen and multiple personal screens.
    - Crowd heterogeneity
    - Social connectivity (repeating and imitating, synchronized cheering, gesture interaction with screen)

Usability, interaction design and interface assessment are covered in Deliverable 1.1.1 in subsection 3.2. Requirements.

With relation to FascinatE we see the next step as a step towards data collection activities concerning more specific use in devices (mobile phones, TVs, public screens) by accompanying users in their watching of TV content. This will give us better insight on how end-users could interact with FascinatE.

In June 2010 we have observed people watching the Royal wedding in Stockholm that was live on three large screens enabling the public to follow the wedding and the cortege. Next, we will analyze videos from the wedding. Moreover, we are planning to do further analysis of material from home environment with emphasise on interaction with the screen/content, gestures and other.

Considering crowds, no further user research is planned, but possibly some specific context (e.g. concert use case) to study in later stages of the project. With domestic audience, we will continue with a study of people and their interactions with the TV screen and other devices in a home environment. Future data collection activities will also focus on viewing TV content on mobile phones and users' interaction with it.

© FascinatE consortium: all rights reserved 

# 6 Gestural User Interaction

It is believed that gesture based user interaction will take a major part in future and innovative interactive systems. Therefore, it is proposed that the FascinatE system may be fully or partly controlled through a set of human gestures, which are associated to some of the main functionalities of the system. The objective of FascinatE's human gesture recognition system is to ease user interaction with the global system as much as possible. Control by gesture will not completely replace other control devices but will represent an alternative to traditional interactive methods (such as remote controls, tablet PCs, etc.) which is more innovative and user-friendly.

## 6.1 Definition of Set of Gestures

Some basic and intuitive gestures are proposed in this section. Such gestures are linked with the FascinatE interactive commands. These commands control the audio and video renderers and define the interface between them and the end user interactive system. The list of FascinatE interactive commands are:

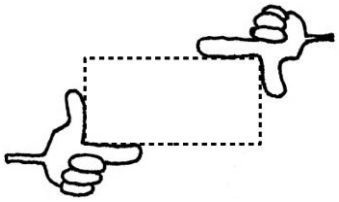
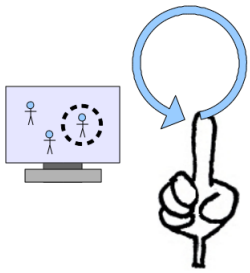
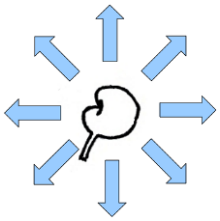
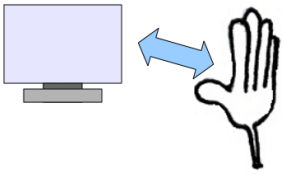
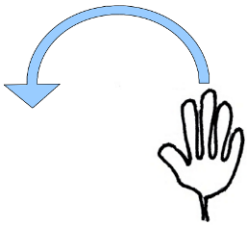
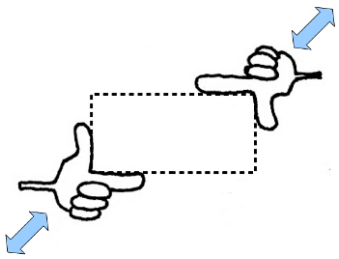
- *Decides framing:* Selecting a region of the screen as framing. This can correspond to a camera view (for instance selecting a camera from an omni-view) or a virtual view (such as doing zoom in a real camera view or adding two camera views together)
- *Selects objects for tracking:* Selecting an object that will be followed from now on (a specific player from a football match)
- *Scroll X-Y:* Navigating through menus or moving in a omni-camera view that occupies more than interface screen
- *Select / Back:* Used to select or de-select in menus, or navigating in the interface (for instance, going back to default camera view)
- *Zoom:* Zooming in camera views
- *Slow motion replays:* Asking for replays at user requested times
    - *Duration of replay:* The user can select the duration of the replay
- *Select windows (multiple cameras on screen):* Ability to create compositions in the screen with multiple cameras.
- *Separation foreground / background sound*
- *Set gain on dialogs:* Choosing the volume in dialogs

It is clear that this list may be altered in the future, adding other gestures or modifying the existing ones depending on user requirements and technology progress.

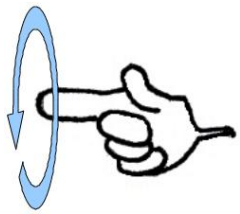
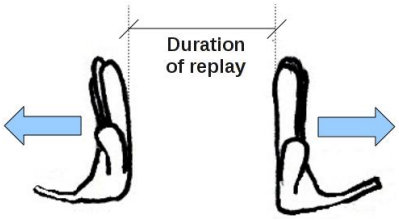
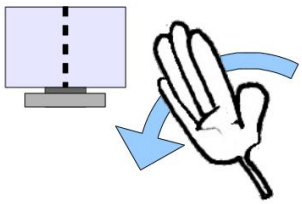
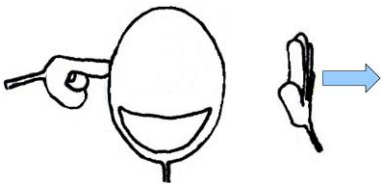
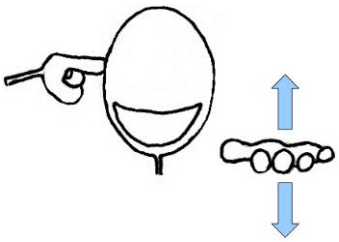
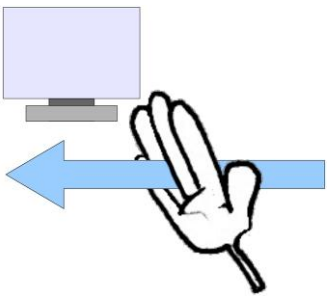
In order to obtain natural and user-friendly gestures, the above mentioned list has been forwarded to some members of our research team at UPC. Each member has chosen his/her favourite gesture to perform every functionality in the list. As a last step, all members have met and discussed the proposed gestures to finally obtain an agreed unified proposal which is presented in Table 6-1.

We would like to remark that the proposal in Table 6-1 should be taken as a first approach for FascinatE gestures. Feed-back with FascinatE partners, especially with TII, will surely lead to modifications and improvements of this suggestion. Section 3.2.2 in deliverable Deliverable 1.1.1 proposes an usability study and evaluation approach of the proposed FascinatE gestures.

© FascinatE consortium: all rights reserved 

| Action | Gesture | Comments |
|---|---|---|
| Decides framing |  | The opposite edges of the framing rectangle are set with each hand |
| Selects objects for tracking |  | Draw circle with finger to select the desired object on the screen |
| Scroll X-Y |  | Grab-move-release gesture |
| (Menu) Select |  | Open hand towards screen and backwards |
| (Menu) Back |  | Describe half circle with hand |
| Zoom |  | Same as framing but changing rectangle size |

© FascinatE consortium: all rights reserved

| Action | Gesture | Comments |
|---|---|---|
| Slow motion replays |  | Typical 'repeat' movement |
| Duration of replay |  | Control slider with separate hands |
| Select windows (multiple cameras on screen) |  | Sword-like motion (horizontal/vertical)<br>To select working window: Select gesture |
| Separation fg / bg sound |  | Touch ear + separate hand |
| Set gain on dialogs |  | Touch ear + raise/lower open hand |
| Change channel |  | Move open hand from left to right to switch forward and right to left to switch backwards (in figure) |

© FascinatE consortium: all rights reserved 

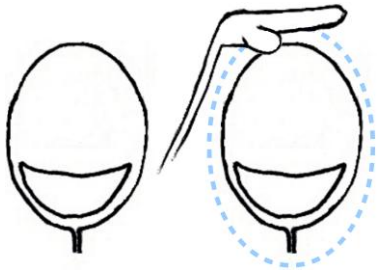
| Action | Gesture | Comments |
|---|---|---|
| Pass 'active user' token |  | The active user places one hand over another user's head to pass the activity token |

**Table 6-3: Proposal for basic FascinatE control gestures**

## 6.2 Gesture database

FascinatE gestures will be recorded using the home scenario UPC's set-up, which consists of a central Time-Of-Flight (TOF) SR4000 camera and two regular cameras at both sides (Figure 6-1).

The available SR4000 camera provides confident depth measurements between 0.5 – 5 meters. Such distance range is valid for FascinatE's home scenario, where users usually sit on a sofa about 2-3 meters away from the TV set.

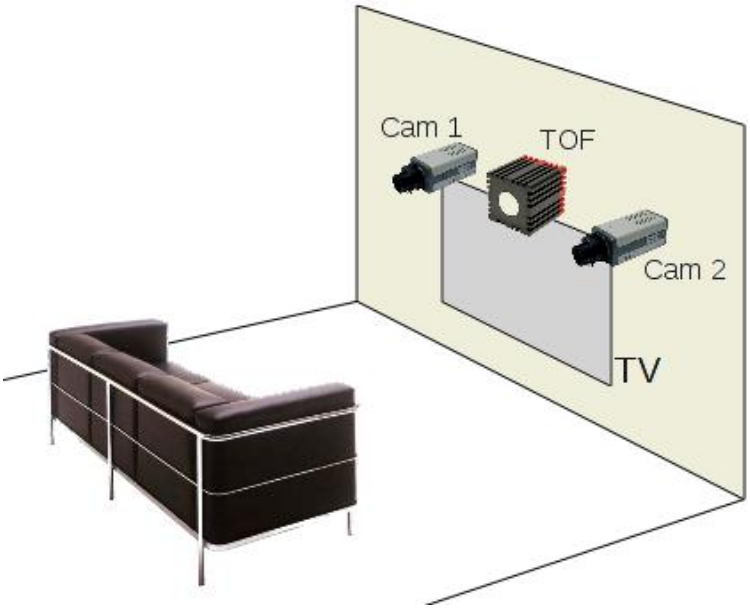



**Figure 6-1: FascinatE recording setup at UPC**

Being able to obtain a unique spatial reference for all the cameras is of great importance since one could map image zones detected on one camera onto the other cameras (i.e. hands and head of the active user). In order to do so, calibration of all cameras is under study. Calibration of a TOF camera is not straightforward due to its poor resolution and unknown camera model [Lindner, 2006; Kolb, 2009]. Furthermore, noise in depth measurements, coordinate transformation errors and numerical errors related to the camera API may strongly affect calibration. However, the first results appear to be satisfactory, as shown in Figure 6-2.

© FascinatE consortium: all rights reserved 

Projection errors observed on cameras 1 and 2 are due to TOF depth measurement errors. In Figure 6-3, the 3D points obtained with the TOF camera are shown. One may expect a flat reconstruction of the calibration plane, but the obtained point cloud shows a distorted calibration plane, which explains the errors observed in Figure 6-2.

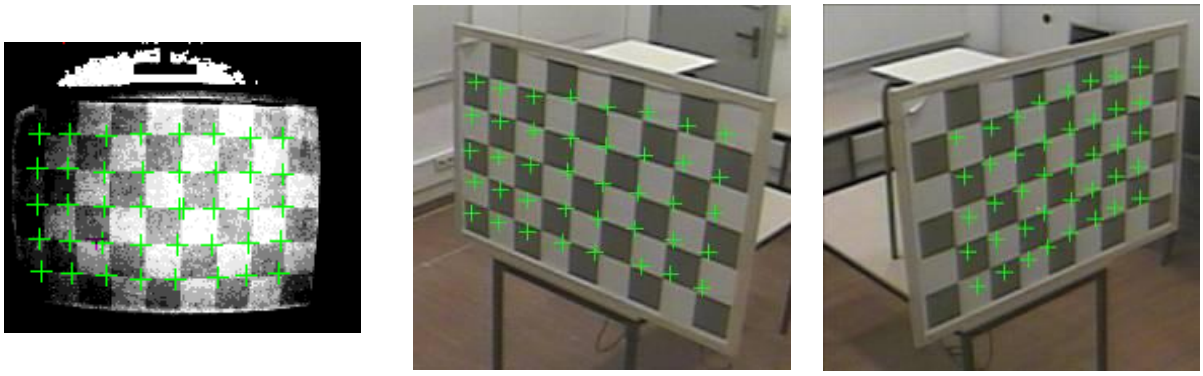


**Figure 6-2: First rough calibration results: TOF, camera 1 and camera 2 (left to right)**

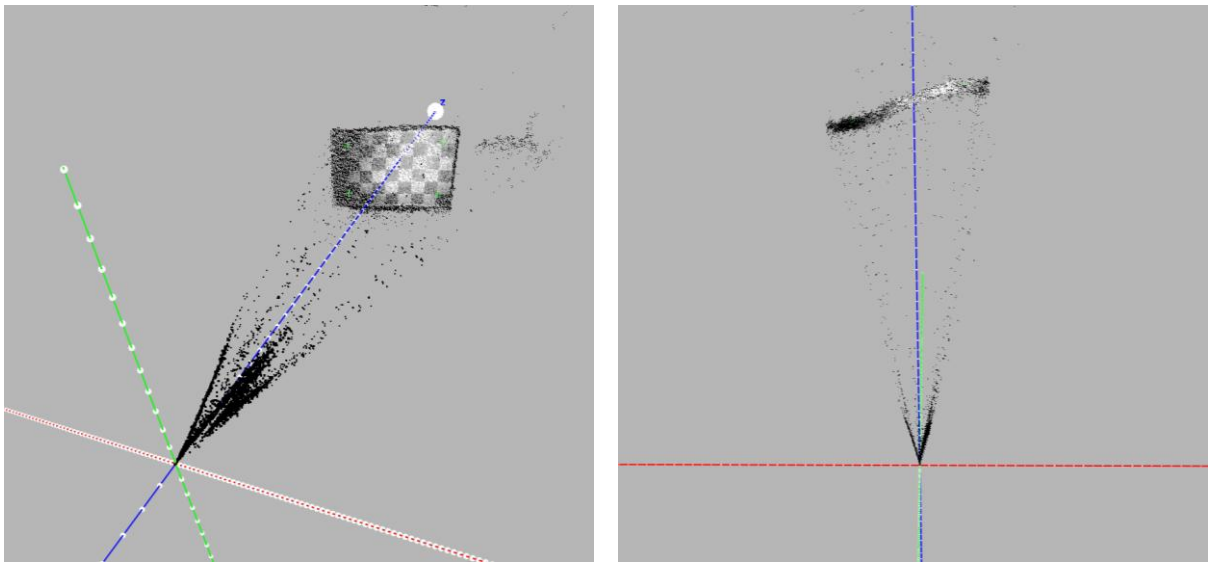


**Figure 6-3: 3D depth points obtained with TOF camera:**
**(a / left) Perspective view of the TOF captured 3D points.**
**(b / right) Top view, a plane is expected, the error of measurement is clearly visible.**

© FascinatE consortium: all rights reserved

# 7 Conclusions

Processing platforms and frameworks for FascinatE audiovisual rendering were explored in this document. With FascinatE's goal to bring the postproduction closer to the end user, it is shown that this requires a distributed rendering operation and hence a large scale of different processing platforms to be considered during the runtime of the project. Nevertheless, a structure to organize potential FascinatE platform is proposed that will be backed with future benchmarking results.

Furthermore, aspects of audience interaction were discussed, backed with studies from workshops and literature. Several fields for further enquiry were identified and will be investigated in the continuing projects runtime. This will be supported with first definitions of user interaction commands based on gestures that later will be used to build a comprehensive gesture database.

© FascinatE consortium: all rights reserved

# 8 References


[Barkhuus, 2009] Barkhuus, L. Brown, B. (2009) *Unpacking the Television: User Practices around a Changing Technology.* ACM Transactions of Computer-Human Interaction 16, 3 (sep. 2009), 1-22. ACM Press.

[Buchinger, 2009] Shelley Buchinger, Simone Kriglstein, Helmut Hlavacs, A comprehensive view on user studies: survey and open issues for mobile TV, Proceedings of the seventh European conference on European interactive television conference, 2009, Belgium

[Cesar, 2009] Cesar, P., Knoche, H., Bulterman, D.C.A (2009) From One to Many Boxes: Mobile Devices as Primary and Secondary Screens. In A.C. Roibás, A. Marcus, and R. Sala (eds.), Mobile TV: Customizing Content and Experiences. Heidelberg, Germany: Springer-Verlag (2009)

[Cui, 2007] Cui, Y., Chipchase, J. and Jung, Y. 2007. Personal TV: A Qualitative Study of Mobile TV Users. In Cesar, P., Chorianopoulos, K. and Jensen, J. F. (eds.) Interactive TV: A Shared Experience. Proceedings of 5th European Conference, EuroITV 2007. Springer, 195--204.

[deVries, 2009] Diemer de Vries (2009) *AES Monograph Wave Field Synthesis*, Audio Eingeneering Society inc., New York, ISBN No. 978-0-937803-68-4

[Harper, 2006] Richard Harper, Tim Regan, Mark Rouncefield, Taking hold of TV: learning from the literature, Proceedings of the 20th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: design: activities, artefacts and environments, November 20-24, 2006, Sydney, Australia

[Jumisko-Pyykkö, 2008] Satu Jumisko-Pyykkö, Mandy Weitzel , Dominik Strohmeier, Designing for user experience: what to expect from mobile 3d tv and video?, Proceeding of the 1st international conference on Designing interactive user experiences for TV and video, October 22-24, 2008, Silicon Valley, California, USA

[Knoche, 2007] H. Knoche and M.A. Sasse. Getting the Big Picture on Small Screens: Quality of Experience in Mobile TV. Idea Group, 2007.

[Knoche, 2004] H. Knoche and J. McCarthy. Mobile Users' Needs and Expectations of Future Multimedia Services, In Proc. of WWRF12, 2004.

[Kolb, 2009] A. Kolb and E. Barth and R. Koch and R. Larsen, Time-of-Flight Cameras in Computer Graphics, 2009

[Lindner, 2006] M. Lindner and A. Kolb, Lateral and depth calibration of pmd-distance sensors, ISVC, 2006, pp. 524-533.

[Levy, 1989] Levy, L. (1989) *A Study of Sports Crowds Behaviour: The Case of The Great Pumpkin Incident.* In Journal of Sports and Social Issues. 13 (2), pp. 69-91.

[O'Hara, 2008a] O'Hara, K. Glancy, M. And Robertshaw, S. (2008) *Understanding Collective Play in an Urban Screen Game.* Proceedings of the 2008 ACM conference on Computer supported cooperative work. ACM Press.

[O'Hara, 2008b] O'Hara, K. Grian, H. and Williams, J. (2008). *Participation, Collaboration and Spectatorship in an Alternate Reality Game.* Proceeding of the 20th Australasian Conference on Computer-Human Interaction. ACM, pp. 130-139.

[Oksman, 2007] Oksman, E. Noppari, A. Tammela, M. Mäkinen, and V. Ollikainen. Mobile TV in Everyday Life Contexts -- Individual Entertainment or Shared Experiences?, In Proc. 5th EuroITV, pages 215--225, 2007.

[Pulkki, 2007] Ville Pulkki, "Spatial Sound Reproduction with Directional Audio Coding", J. Audio Eng. Soc., Vol. 55, No. 6, 2007 June

[Pulkki, 1997] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," J. Audio Eng. Soc., vol. 45, pp. 456-466 (1997 June).


© FascinatE consortium: all rights reserved 

[Repo, 2004] Repo, P., Hyvonen, K., Pantzar, M and Timonen P. Users Inventing Ways To Enjoy New Mobile Services – The Case of Watching Mobile Videos, In Proceedings of the 37th Hawaii International Conference on System Sciences.

[Reeves, 2005] Reeves, S, Benford, S, O'Malley, C, and Fraser, M. (2005) *Designing the Spectators Experience.* Proceedings of Human Factors in Computing Systems. ACM, pp. 741-750.

[Schatz, 2007] Schatz, R., Wagner, S., Egger, S. and Jordan, N. 2007. Mobile TV Becomes Social -- Integrating Content with Communications, In Proceedings of the ITI 2007 Conference. June 25--28, 2007, Croatia.

[Tsekleves, 2009] Tsekleves, E. Whitman, R. Kondo, K and Hills, A. (2009) *Bringing the Television to other Media in the Home: An Ethnographic Study.* Proceedings of the 7th European Interactive Television Conference.

[Vuduc, 2010] Vuduc, R.; et al; On the Limits of GPU Acceleration; Paper presented at the 2nd USENIX Workshop on Hot Topics in Parallelism, June 14-15, 2010 US Berkeley,CA

© FascinatE consortium: all rights reserved 

# 9 Glossary

**Terms used in FascinatE**

| | |
|---|---|
| DirAC | Directional Audio Coding |
| FRN | A FascinatE Rendering Node is defined as a process that translates the FascinatE view of the real world for projection on a variety of displays and loudspeaker devices.GUI Graphical User Interface |
| HTML | HyperText Markup Language**;** provides a structure to documents such as paragraphs, lists and tables. Mainly used for internet web pages. |
| LASeR | Lightweight Application Scene Representation; Rich media standard to describe 2D vector graphics (aka ISO/IEC 14496-20 / MPEG-4 Part 20) |
| Loudspeaker | In the context of this document loudspeaker stands for a single loudspeaker or a loudspeaker system, like a 5.1-, Wave Field Synthesis - or Ambisonics setup. Loudspeakers could also be replaced by head- or earphones. In general loudspeaker stands for an electro acoustic sound transmitter system. |
| MPEG BIFS | Motion Picture Experts Group – Binary Format for Scenes; binary format for 2D and 3D audiovisual content, based on VRML (ISO/IEC 14496-11 / MPEG-4 Part 11) |
| RoI | Region of Interest |
| VBAP | Vector Base Amplitude Panning |
| VRML | Virtual Reality Modelling Language; text file format for representing 3D vector graphics. Popular to be used for describing objects and scenes. |
| XML | Extensible Markup Language; textual data format, used like HTML for documents but also for structuring any kind of arbitrary data |

**Partner Acronyms**

| | |
|---|---|
| ALU | Alcatel-Lucent Bell NV, BE |
| ARI | Arnold & Richter Cine Technik GMBH & Co Betriebs KG, DE |
| BBC | British Broadcasting Corporation |
| DTO | Deutsche Thomson OHG, DE |
| HHI | Heinrich Hertz Institut, Fraunhofer Gesellschaft zur Förderung der Angewandten Forschung e.V., DE |
| JRS | JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT |
| SES | Softeco Sismat S.P.A., IT |
| TII | The Interactive Institute, SE |
| TNO | Nederlandse Organisatie voor Toegapast Natuurwetenschappelijk Onderzoek – TNO, NL |
| UOS | The University of Salford, UK |
| UPC | Universitat Politècnica de Catalunya, ES |

© FascinatE consortium: all rights reserved