End User, Production and Hardware and Network Requirements

Deliverable D1.1.2

FascinatE identifier:	Fascinate-D112-UPC-Requirements-v07.docx
Deliverable number:	D1.1.2

Author(s) and company:	J. Ruiz-Hidalgo, J.R. Casas, X. Suau (UPC);
	A. Gibb (BBC); M.J. Prins (TNO); G. Zoric, A. Engström, M. Perry, E. Önnevall, O. Juhlin,
	P. Hannerfors (TII); J. Macq (ALU); O. Schreer (HHI);

Internal reviewers:	O.A. Niamut (TNO), G. Thomas (BBC)
Work package / task:	WP1
Document status:	Final
Confidentiality:	Public

Version	Date	Reason of change
1	2011-11-29	Initial input for second version of deliverable
2	2011-12-19	Merge updates from end-user perspective and Annex A from scripting
3	2012-01-15	Merge updates from production perspective
4	2012-01-30	Merge input from HHI, ALU, JRS and TII
5	2012-02-01	Merge input from TNO, updated conclusions
6	2012-02-10	Address comments from internal reviews
7	2012-02-14	Final version

Acknowledgement: The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248138.

Disclaimer: 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

Executive Summary

This document is an update of D1.1.1 and it defines the overall requirements that the FascinatE system should meet. The deliverable proposes three scenarios, depending on the configuration and functionality of the complete delivery chain, in which the possible FascinatE requirements are discussed:

- Scenario 1 (production-centric): All FascinatE functionality is provided by the production side and there is no computational load shifted to either the provider or the terminal.
- Scenario 2 (terminal-centric): A complete Layered Scene Representation (LSR), together with production scripts, are provided to the terminal which is responsible of rendering and presenting it to the end user.
- Scenario 3 (provider-centric): This can be interpreted as an intermediate step in the evolution of FascinatE technology. In this case, the LSR will be rendered to a format tailored to the delivery network and targeted terminal.

Together with the proposed scenarios, several use cases are defined to better understand the role of the FascinatE system in real-life situations. Based on the proposed scenarios and the level of interaction, this document describes the requirements and high-level functionality of the FascinatE system. These requirements are divided into three main parts: end-users, production teams and network infrastructure requirements. For each part, high-level requirements for the FascinatE system are listed.

In the case of end-user requirements, it covers issues that should be kept in mind when designing FascinatE based services.

In the case of production requirements, it was found that it is important to understand how to integrate FascinatE technology into existing technology and working practices, how existing production staff operates an automated script-based production system or, for instance, what tasks will production staff accept to be automated.

In the case of network provider requirements, it is shown that each of the three proposed scenarios comes with different requirements. Scenario 1 may be implemented with existing and deployed delivery networks. Scenario 2 puts strong requirements on the bandwidth of the delivery network and may only be introduced after significant advances in physical network technology and signal processing. Scenario 3 focuses on processing functionality. Within FascinatE, we consider this scenario the most relevant for innovations in the delivery network

Many details of the requirements discussed in this document will become clearer and better-defined as the project progresses. This deliverable has provided more details in many areas compared with the first version (D1.1.1), adding new use cases to further explore high-end panoramic environments and hybrid delivery networks, updating the rest of the use cases to the current state of the project, adding more granularity and a more comprehensible link to system requirements from the use cases, further analysing hardware requirements in end-user and network perspectives and, finally, incorporating a more detailed usability assessments for end-users. A third updated version of this document will be produced at the end of the project (D1.1.3, in Month 42).

1 Introduction

The objective of this deliverable is to define the overall requirements that the FascinatE system should meet. The requirements are defined from three different points of view: end-users, production teams and network infrastructure. In the first case, end-user requirements are established that will provide consumers with a novel and engaging experience in terms of the functionalities available from terminal devices. Production requirements are defined based on how production teams would expect to interact with the system. Finally, network requirements determine the expected capability of networks and processing hardware that the FascinatE system should be able to work on.

This is the second of three deliverables addressing requirements in the FascinatE project: a first version (D1.1.1) was produced in month 6 and a final requirements document (D1.1.3), informed by things learned during the duration of the project, will be issued at the end of the project in month 42. A summary of the changes included in this document with respect to D1.1.1 is:

- The scenarios proposed in section 2.1 are further clarified and explained
- Two new use cases have been added to further explore the requirements of a high-end panoramic environment (*Use case 2*) and a hybrid delivery network (*Use case 11*)
- Use cases have been adapted to the current state of the project and defined with more granularity. They have also been extended to link with possible system and hardware requirements to further bridge this document with the system specification of D1.4.2
- Hardware requirements have been further analysed and extended for end-user and network perspectives
- The state of the art in gesture recognition has been updated
- Usability requirements and assessments have been further specified and clarified
- The integration of the FascinatE project with the Vision Mixer has been extended and further revised
- Final conclusion have been updated to the current state of the project

This document is primarily designed to help members of the FascinatE consortium define the requirements of the system to be developed, and to provide a reference against which the achievements of the project can be judged. However, it will also be of more general interest outside the project, as it helps to explain what the project is trying to achieve and the technological environment in which the project is operating.

Although this deliverable looks at overall requirements rather than specific aspects of the FascinatE system, it is useful to refer to general aspects of the system so as to tailor the discussion to the planned developments in the project. The following assumptions about how the FascinatE system will operate and what it could provide should therefore be borne in mind:

- Audio and video will be captured using a selection of cameras and microphones. Specifically, there will be one or more fixed very wide-angle cameras (referred to as 'OMNICAMs'), multiple conventional broadcast cameras with the ability to pan, tilt and zoom, and microphones that may capture both the sound field at one or more points, and individual sound sources.
- A mechanism will be provided to combine these A/V sources into a Layered Scene Representation (LSR).
- It will be possible to produce a range of different views (or regions-of-interest) of the scene by selecting different viewpoints and fields-of-view, to suit different viewer preferences and device capabilities (e.g. making fields-of-view appropriate for the screen size of the device).
- The metadata describing how to create a particular view from the LSR is referred to as a 'script'. Scripts could be generated at the production side (e.g. analogous to the shot framing and selection decisions made by a cameraman and vision mixer (VM), or at the end-user side (e.g. by a user choosing the part of the scene they want to examine in detail), or some combination of the two.

1.1 Organization of the Document

In order to provide meaningful and correct requirements, it is important to understand the limitations and new functionalities provided in FascinatE. Three different scenarios can be envisaged, depending on the configuration and functionality provided by the complete delivery chain. Chapter 2 lists these three scenarios and provides possible use cases that can be realized within the scenarios. Note that, actually, the relation between scenarios and use cases is a loose one; a use case may be (partly) realized by multiple scenarios. In order to better organize the use cases within the proposed scenarios, the level of user interaction allowed in each scenario is also considered.

Based on the proposed scenarios and the level of interaction, this document describes the requirements and high-level functionality of the FascinatE system. Chapter 3 presents the requirements from the end-user perspective. First, the chapter covers issues that should be kept in mind when designing FascinatE based services in order to provide quality of experience as desired by users. Second, it gives interaction design guidelines for services based on FascinatE to provide a rich and user-friendly experience. Thirdly, it describes the usability assessment and the planned evaluation approach that will be followed in order to assess the fulfilment of the end-user requirements by the FascinatE system. Chapter 4 focuses on production requirements. It describes the restrictions placed on the system from the point of view of the production staff, workflows and systems. Chapter 5 presents the networking requirements. It details the appropriate requirements for the network role and, additionally, describes high-level network functionality. Finally, some conclusions are drawn in Chapter 6.

1.2 Related Documents

Before reading this document it is recommended that the reader is familiar with the following documents:

- *D1.1.1 End user, production and hardware and network requirements* is the first version of this document.
- *D1.4.2 Interim System Specification* defines the FascinatE system, explaining the functionality of each block and the interfaces defined to communicate between them.
- *D1.5.1 First System Integration* describes the status of the development of key modules in the system as of Month 21 (as shown at the demonstration at IBC 2011).
- *D2.1.1 Draft Specification of Generic Data Representation and Coding Scheme* defines the generic data representation for format-agnostic production and the type and structure of calibration data needed.
- D4.1.1a *Service Concepts, Business Models, Delivery Modes* identifies viable business models and associated services in the context of FascinatE.
- *D4.2.1 Capabilities of Current and Next-Generation Delivery Networks involved in FascinatE Services* provides an overview of the capabilities of current and Next-Generation delivery networks involved in FascinatE services.
- *D4.2.2 Delivery Network Reference Architecture* describes the first iteration of the detailed FascinatE delivery network.
- *D5.1.1 A/V Renderer Specification and Basic Characterisation of Audience Interaction* provides an initial overview of the possible interactions in the system and serves as a basis for defining requirements on user interaction metadata and terminal properties.
- *D5.3.1 Requirements for the network interfaces and interactive systems usability* details the requirements for the network interfaces of the terminal block, essentially between the Rendering functions, the delivery block and the User Control Nodes.

2 Scenarios and Use Cases

2.1 Scenarios

This section defines three different scenarios based on where the main processing or computational load is located in the delivery chain. In Table 1 in the next section, the three proposed scenarios are studied and related to the amount of interaction allowed to the end user in the FascinatE system. In FascinatE deliverable D4.1.1a, Chapter 5, the three scenarios are linked to value chain configurations and business models.

2.1.1 Scenario 1: Production-centric delivery chain

This scenario considers a current state-of-the-art delivery situation. It has its focus on innovations in the production domain. There is no FascinatE functionality in the network and terminal. In this case, the distribution of FascinatE content is tailored to a specific delivery format and one or more rendered views in the form of TV channels and/or media streams are presented to the user. Dedicated channels/streams exist for widescreen angle, zoom and region-of-interest (ROI) views. In this scenario, end-user interaction is limited to switching between channels and selecting streams. The degree of interaction allowed for current end-users is completely determined at the production side by the number of rendered views made available. Figure 1 shows a high-level functional architecture for the production-centric scenario. For a more detailed review of the elements presented in the Figure, see deliverable D5.1.1. Most of the computational load of the system resides in the production side.

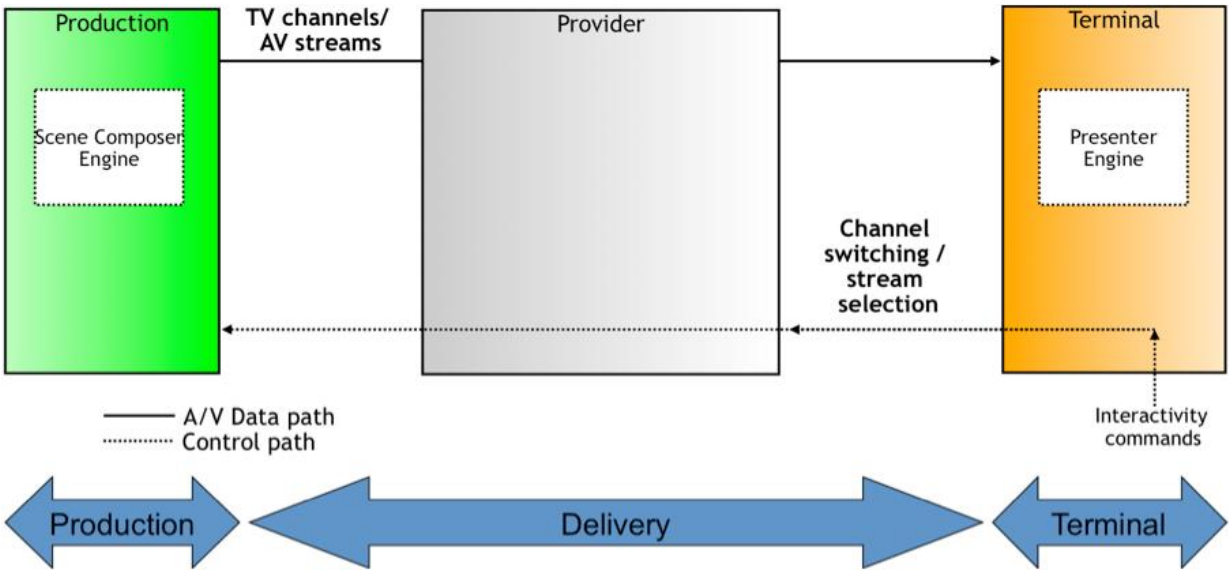


Figure 1: High-level architecture for the production-centric scenario

2.1.2 Scenario 2: Terminal-centric delivery chain

This scenario considers the final FascinatE evolution. It assumes an idealistic delivery network which allows for distribution of a full LSR, with the terminal receiving all the captured A/V streams. The LSR will be rendered by the terminal itself before presenting it to the user. This assumes that production scripts are sent towards the terminal, containing production-side knowledge that specifies the required processing steps. Figure 2 shows a high-level functional architecture for the terminal-centric scenario. Note that significant computational load has been moved to the terminal side for the processing of production generated scripts in response to user commands.

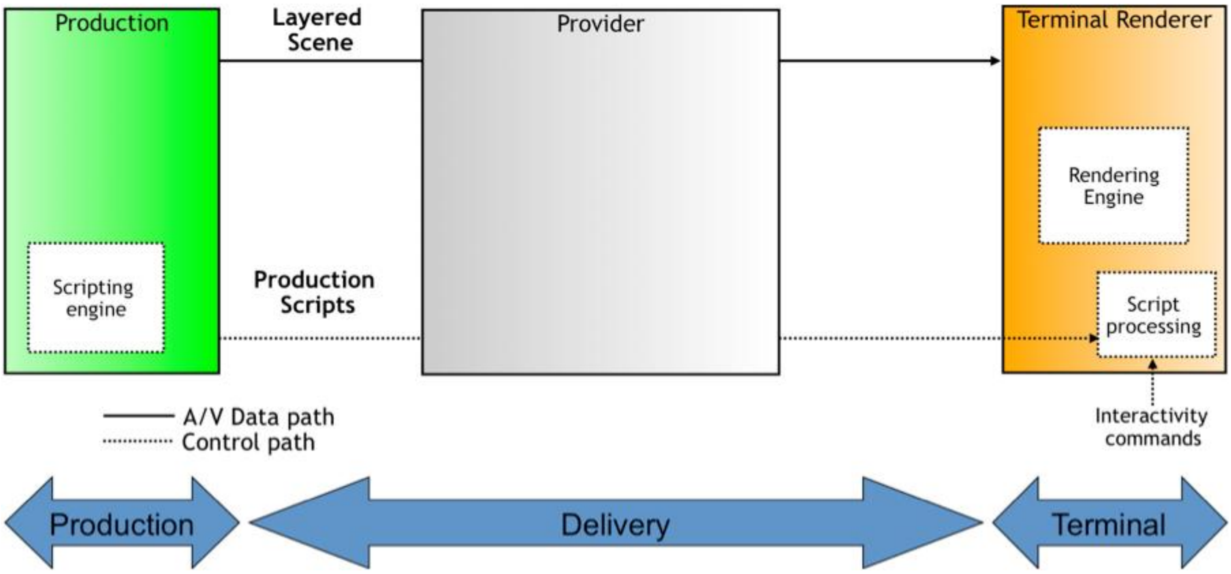


Figure 2: High-level architecture for the terminal-centric scenario

2.1.3 Scenario 3: Provider-centric delivery chain

This usage scenario highlights how the FascinatE technology will impact the way A/V media is delivered and will enable new types of service. Unlike in previous scenarios where the network is essentially seen as a bitpipe, we assume here that the delivery block contains a main processing function that renders the LSR of the A/V scene and processes the accompanying production scripts. The associated use cases described in section 2.2.3 will show examples of how a service provider can use the FascinatE delivery technologies to offer new type of interactive audio/video services to a large range of end-devices and end-user profiles.

On the one hand, this scenario can be interpreted as an intermediate step in the evolution of FascinatE technology, as it allows for limitations on the data rate that can be delivered to the end user, and on the processing power within the terminal. Within the delivery network, the LSR will be rendered to a format tailored to the delivery network and the requesting or targeted terminal. The script processing is also located in the delivery network and receives the interaction commands from the terminal side and the production script from the production side. Based on these inputs it can control the rendering function for providing the right view in the appropriate format to the terminal. Although this scenario requires additional functionality in the delivery network it saves bandwidth in the network without losing interactivity freedom compared to scenario 2.

On the other hand, the rationale to push more processing functions in the delivery block is not only based on short- or mid-term technical limitations, but also on business aspects. This scenario positions service providers as another potential class of users of the FascinatE technology. Here the term “service provider” is to be understood in a broad sense. It encompasses not just network and video service providers, but also local broadcasters or any other third-party which can benefit from the flexibility offered by the LSR to create new services: linear TV programmes, personalized, interactive services, etc.

Figure 3 shows a high-level functional architecture for the provider-centric scenario. In this case, most computational load is shifted to the provider side for the processing of both scripts generated from production and interactive commands from end users.

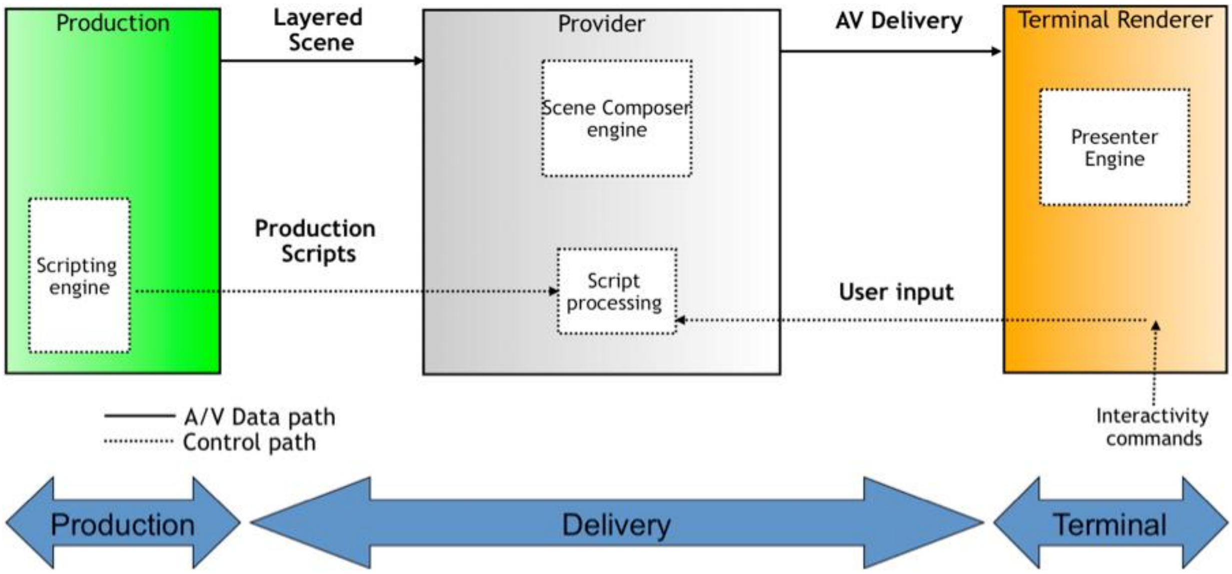


Figure 3: High-level architecture for the provider-centric scenario

2.2 Use Cases

This section details several possible use cases that can be realized in the scenarios described above. All use cases detailed in this section are focused from the end user, production and the network perspective. In order to be able to cover all different aspects of the possible use cases, the level of interaction of the end user is also considered. In this case, for each scenario, one can think of situations where a limited interactivity with the system is permitted or possible. Furthermore, situations where the end user has all the possibilities of the FascinatE interaction at their disposal can also be envisaged. The different levels of interactivity can be applied to all three scenarios.

Table 1 shows the relation between the level of interactivity presented to the end user and the proposed scenarios and how this interaction affects the production and network aspects of the system.

Scenario **End user interaction**	**Production-centric**	**Terminal-centric**	**Provider-centric**
No interaction. **State of the art production of a linear TV programme**	**Production**: Works as today. Extra tools provided by FascinatE allow for novel shots and audio **Network**: Only delivers linear video stream from production onwards.	**Production**: All production is automated. Content and scripts are transmitted to the terminal. However, all reproduction of content is fixed from the production or provider side but content is automatically adapted to terminal capabilities **Network**: Requires higher bandwidth to the user than the production-centric scenario	**Production**: Work load can be split between production gallery and provider gallery. Content is automatically adapted to terminal capabilities **Network**: Requires high bandwidth to production, but low bandwidth to the end user
Medium interaction. **User can choose between pre-defined streams of content**	**Production:** In order to generate multiple streams, more staff is required. Some of this work could be automated **Network:** Must deliver a large amount of different streams	**Production:** The case where all production is basically automated/supervised mark-up and script generation **Network:** Must deliver entire LSR to the end user	**Production:** The work of generating multiple streams could be split between a “skeleton” gallery at the production end, and provider galleries **Network:** Requirement between provider and network is larger than in the case above including capability for automatically selecting parts of the LSR
Full interaction. **FascinatE interactivity**	**Production:** Lots of scripting both automatic & supervised in gallery **Network** Interaction is limited to stream selection, a simulation of “full interaction” can only be realised by using delivering more channels/streams	**Production:** Production & provider galleries can cooperate on script creation and metadata generation **Network:** Must deliver entire LSR to the end user	**Production:** Production & provider galleries cooperate on script creation and metadata generation **Network:** Must deliver entire LSR to the end user including capability to render specific ROIs

Table 1: Overview of the proposed scenarios and degree of end user interactivity

The use cases presented in this section can be classified following the proposed organization. Figure 4 shows how the use cases detailed in next section can be classified depending on their level of interactivity and the scenario they can be realized. As seen in the Figure, several use cases can be included in different scenarios.

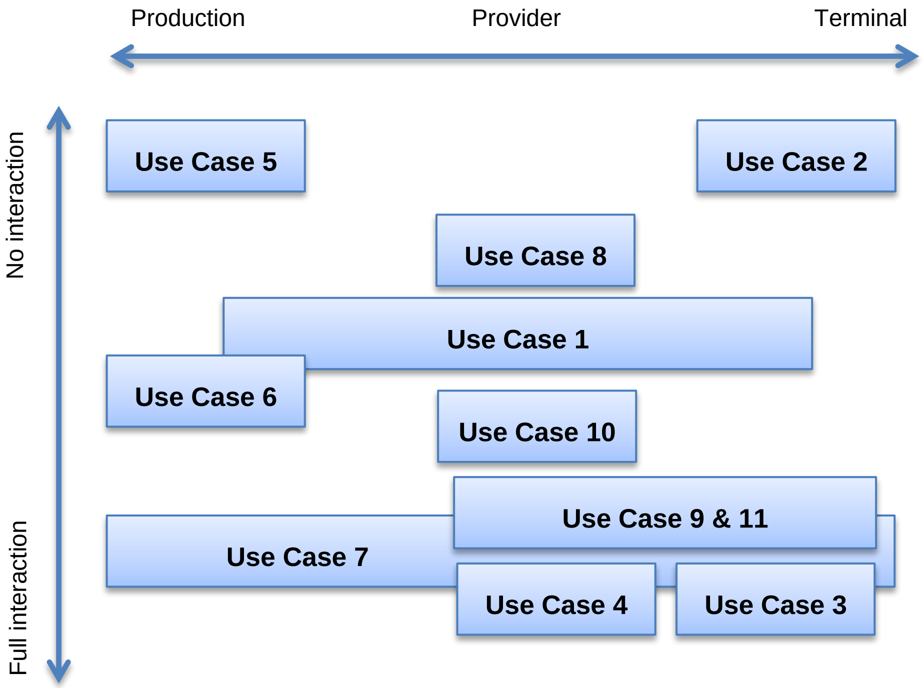
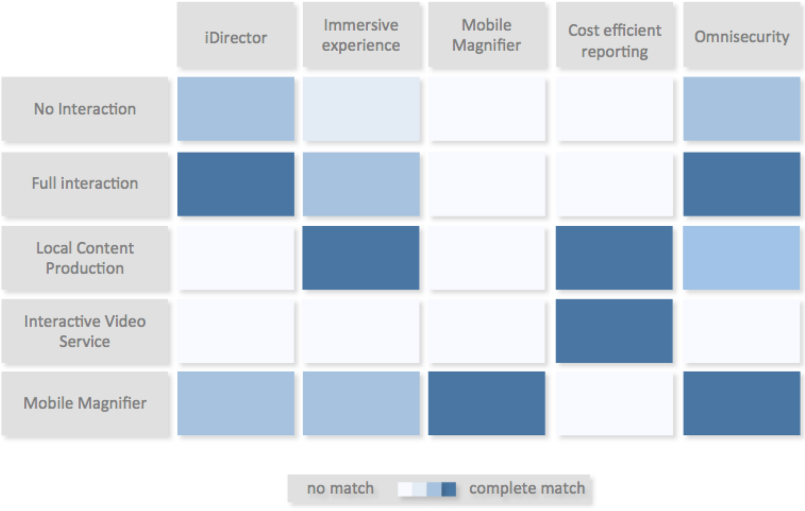


Figure 4: Classification of use cases

In FascinatE deliverable D4.1.1a, Chapter 4, a set of five service concepts is described, derived from business interests and based on the potential value that FascinatE can offer. Figure 5 shows the relation between the use cases described here and the service concepts described in D4.1.1a. The potential of each service concept in a given scenario is also considered in D4.1.1a.



	iDirector	Immersive experience	Mobile Magnifier	Cost efficient reporting	Omnisecurity
No Interaction	▒	░	□	□	▒
Full interaction	■	▒	□	□	■
Local Content Production	□	■	□	■	▒
Interactive Video Service	□	□	□	■	□
Mobile Magnifier	▒	▒	■	□	■

no match □░▒■ complete match

Figure 5: Relation between use cases and service concepts

2.2.1 End user perspective

End user perspective use cases describe real situations that can occur in the FascinatE system from the end user point of view. These use cases are centred around watching football and listening to an opera but can be applied to any other event such as concerts, standard TV programmes, or any other sport events.

***Use case 1*: No interaction in a home theatre environment**

> *John Smith arrives home late to watch his favourite football team Barcelona against Chelsea. The match has just started and he realises he just missed a goal as football players are already celebrating it. Even though several streams and channels are available to him*

showing the same football match from different angles, none of them is showing any replay he likes. A bit frustrated, he continues watching the match but, this time, he decides to select a wider view of the football field by using his remote. Later on, the system signals him by a small icon in the top-left side of the screen that a complementary channel is streaming a view automatically following his favourite player, who is playing a fantastic match. He quickly changes channels so he can follow him, displaying this dedicated player view on his TV set.

From this use case some system and hardware requirements can be foreseen:

- A communication interface between the end user and the terminal audio and video renderer is needed. It should be able to present the end user with information of current channels and views available to the user.
- The system must provide an input technique (remote control or gesture interface) to the user to be able to switch channels.
- A scripting engine is needed to present the terminal renderer with additional personalized channels that match the end user's profile.

***Use case 2*: No interaction in a high-end panoramic cinema environment**

John Smith is a big fan of the opera diva Anna Netrebko. Unfortunately, she does not perform in the small town where John lives. However, there is a premium event place in his town offering immersive experience in a cinema equipped with a 180° panoramic screen providing super resolution video quality and 3D sound. This cinema presents tonight a live transmission of a performance of Anna Netrebko at the arena in Verona. Arriving in the cinema, he is one out of thirty premium guests, enjoying the opera "La Traviata" with Anna Netrebko. He watches the opera in crystal clear image quality with 180° field of view. The sound system is based on wave field synthesis. He is able to watch the performance as he would have gotten the best and most expensive seat in the arena in Verona right in the middle in front of the stage. From time to time some close-up views of Anna Netrebko and other performers are shown in different parts of the panoramic screen. Thanks to this, he can perceive details of the show, facial expressions of the artists, which he never would see from any seat in the arena itself.

From this use case some system and hardware requirements can be foreseen:

- The panoramic video must be transmitted to the cinema via high-bandwidth network connection. The video inserts such as close-up and zoom in views are already composed on the production side. This is achieved using the accompanying broadcast cameras located next to the omni-directional video capturing device. The 3D audio is captured by a number of different microphones and the rendering is performed in real-time. The audio is streamed together with the video to the receiving end.
- The premium cinema theatre must be equipped with a high-quality projection system capable of displaying 7k x 2k resolution video. Furthermore, a wave field synthesis loudspeaker system is required for high-quality rendering of the 3D sound signal.

***Use case 3*: Full interaction terminal-centric**

John Smith is already late to watch his favourite football match of the season. He quickly turns on the TV set in his living room and starts enjoying a high-resolution video of the match together with a high quality surround sound of the commentators and ambient noise in the stadium. However, just in the middle of the game, his daughter Jane walks into the living room, grabs the remote control and changes channels just as his team was about to tie the match. By the time he is able to get the remote again and change back to the match, he misses the goal. He decides then to connect to the FascinatE interactive system. Automatically, the FascinatE system detects his presence and recognizes him as a user of the system, knowing that he prefers to interact using visual gesture recognition. Therefore, John is able to roll his hand backwards to request a replay. The system then shows him three possible camera views and a slide on screen to define the replay duration. John separates his hands for a custom selection of the replay duration and points at his preferred camera view. After some time enjoying the football match, his wife reminds him he has to finish some errands, so he gets his mobile phone out, selects the football channel and continues watching the football on the small screen while he walks out. Unfortunately, the *phone terminal is not powerful enough to provide all the functionalities his TV set does, so he selects a channel, which automatically focuses on players in his favourite team.*

From this use case some system and hardware requirements can be foreseen:

- The system must be simple (e.g. intuitive gestures) and must respond to user commands in a fast and reliable way (e.g. typically less than 200ms).
- The system should be non-intrusive, leaving the visual and audio channels as open as possible.
- The system should work on relatively small hardware (e.g. possible to locate in a living room). No more than a single computer / set-top box and one camera should be needed.
- A user profile with user preferences should be available. These preferences are used to select the default input technique to control the system and to obtain information of the end user's viewing preferences.
- The system must work under several illumination conditions (e.g. at night with no lights on) and under extreme use poses (e.g. standing up, sitting on the sofa, etc.). Depth cameras are, therefore, more suitable to be used for gesture recognition than normal colour cameras.
- Authentication is needed (e.g. face recognition) in order for the system to first: automatically recognise users and select preferences accordingly and second: allow the creation of a hierarchy of users to control the system.
- The system needs to provide free viewpoint selection and navigation capabilities. A set of gestures must be defined so they can allow the user to freely navigate the scene. Also, multimodal alternatives (e.g. touch based control with tablets) could be provided to allow navigation. In this case, gestures must be consistent across different options.
- The system must provide replay capabilities so the user can pause the current view, rewind and replay specific parts. In this case, storage capacity is needed to store in real-time key events and allow the possibility of playback to the end user.
- A scripting engine is needed to signal to the end user that different channels and regions of interest are available and to define them. A communication interface between the end user and the scripting engine is needed.
- The system should be able to render the scene on different end-devices (high definition TV, tablets, mobile phones, etc.).

***Use case 4*: Full interaction provider-centric**

John Smith walks into his living room and turns on his FascinatE TV set to watch the Champions League final match. While John is watching his daughter walks into the living room and tries to change the channel. However, the FascinatE system does not respond as he is the master user controlling the system at the moment. As his daughter, a little bit frustrated, stays in the room playing, John decides it would be nice to feel more immersed in the match. He increases the ambient noise by touching his ear, separating both foreground and background noise by raising the right hand. He now feels more like being in the stadium. Later on, he decides to change the view perspective and the FascinatE system suggests several options. On the right hand side of the screen, he chooses the panoramic view while on the left he lets the system following his favourite player plus the real time score of some other matches the system suggested him to follow. However, after a while, John decides that he should let his daughter control the system so he tells the FascinatE system to let his daughter control it. She is very happy to be able to select a different channel with cartoons. Meanwhile, John gets his tablet out and selects the same view he was watching before on the main TV. The network provider receives this request, and selects the optimal view to mimic the configuration John was using on his main TV.

From this use case, the system and hardware requirements foreseen are the same as the terminal-centric scenario listed in *Use Case 3*.

2.2.2 Production perspective

***Use case 5*: A FascinatE system used to create a normal, linear TV programme**

> *England are playing France in a friendly rugby match before the 2013 Six Nations gets started. The BBC is covering this match. They have an experimental FascinatE system in their production gallery. For the first time, the TV director can see the whole match by the centre line OMNICAM feed, which he has chosen to have displayed on the top four monitors in the gallery. The OMNICAM view shows the director a few useful pieces of information, such as which section of the panorama the virtual cameras are viewing, and the view from cameras one and two, which are being tracked against the OMNICAM image.*
>
> *An incident breaks out between two players, away from the ball. The nearest camera operators move quickly to cover the incident as it develops, but none of the conventional cameras captured the start of the incident. The FascinatE operator is able to provide a replay merging video from the OMNICAM and camera two which shows the start of the incident in low resolution. The operator can increase the resolution once camera two is on the incident.*
>
> *Watching at home, Barry and his housemate Didier immediately begin arguing about whose fault the incident was, before the OMNICAM-enhanced replay makes it clear who started it.*

From this use case some system and hardware requirements for the production side can be foreseen:

- Speed and artistic quality of clip production: A user interface needs to be provided that allows the FascinatE operator to quickly produce well-framed shots of a key event, including shot framing and dynamics resembling those which a real camera operator would provide, For the replay functionality described here, the production of these clips need not be fully real-time (the video is already non-live as this is an action replay) but it should be possible to produce a clip within about 10 seconds of the incident happening in order to be useful as an action replay. For analysis at half-time, a longer preparation time (of the order of 30-60 seconds) would be acceptable.
- Technical quality of clips: The video signal produced must be compatible with the rest of the broadcast infrastructure (e.g. 1080i 50Hz 16:9). The visual quality of the OMNICAM image should be comparable to the broadcast camera image in terms of colour fidelity, noise level, motion rendition, etc.
- Switching between broadcast and OMNICAM output: The system needs to provide a visually-acceptable way of switching to a broadcast camera feed (giving higher spatial resolution and possibly a more interesting viewpoint) once a relevant camera is covering the incident, taking into account the fact that there may be a parallax effects caused by there being a significant distance between the broadcast camera and the OMNICAM. This includes the ability to match the colour rendition of the OMNICAM to the broadcast cameras, and to choose shot framing and blending methods that allow the viewer to keep track of the objects of interest (e.g. players and ball) as the view switches. The operator will need a display system that allows him to see the views of an incident offered by multiple cameras, including the OMNICAM, to judge when to switch between cameras.
- Economics: The cost of installing and operating the system should be comparable to other enhancements that could be used for such a broadcast, such as a super slow-motion camera.

***Use case 6*: A linear TV programme with some interactivity**

> *The 2018 England Football World Cup is being covered by the BBC. The YouView-2 connected TV platform supports the FascinatE system. These internet-connected set top boxes are available nationwide, and FascinatE content is available to anyone with a fast enough internet connection. The BBC is producing coverage of the World Cup in FascinatE format, thanks to widespread uptake of this system.*
>
> *The production must cater for both conventional viewers without FascinatE-capable equipment, as well as those with. So the production team decide to use the FascinatE system to create normal TV coverage, and at the same time produce lots of rich metadata and scripts, which will allow viewers with FascinatE capable equipment to choose from a variety of different coverage, replays, and views of the games.*
>
> *The production gallery at Wembley Stadium is set up for FascinatE production. A special terminal shows the FascinatE operators the feeds from the OMNICAMs plus the overlaid high-resolution views. The FascinatE scripting operators watch their terminals marking up*

> *interesting events and tracking players, with the help of automated systems. The VM operator is at a normal vision mixing desk, which has virtual cameras coming into it as well as the conventional ones.*
>
> *At home, Didier is watching the England vs. France group stage match. Using his FascinatE set-top box he is watching a view of the game he has customised himself. On the upper half of his HDTV he is watching the panoramic view of the whole pitch. On the lower half of his TV he can see the main programme that is being produced, and also is following his favourite player, France's new striker.*

From this use case some system and hardware requirements for the production side can be foreseen:

- Speed and artistic quality of clip production: As in the previous use case, a user interface needs to be provided that allows the FascinatE scripting operators to produce scripts that describe well-framed shots of interesting parts of the scene, including shot framing and dynamics resembling those which a real camera operator would provide. However, the production of these scripts must be **fully real-time**. This is likely to require some automation, and ideally should allow one operator to control the production of multiple scripts.
- Production of metadata: where a large number of possible views are being produced, the operator must be able to associate metadata with each script that would be meaningful to an end user, to allow them to select clips based on a textual description (e.g. name of player) rather than just by seeing a visual 'thumbnail' of the clip.
- Transmission of data from the outside broadcast site: The audio and video signals needed to represent all the additional selected regions need to be sent from the outside broadcast site to the broadcaster's delivery systems. The limited bandwidth generally available from outside broadcast sites may require the content streams to be selected to provide only the required views without sending additional material that is not needed.

***Use case 7*: Fully interactive TV production**

> *GOBCOM are covering the final Rolling Stones concert using a FascinatE production system. They are providing video for the internal big screens, for mobile devices in the stadium, and streaming live to millions of viewers around the world. They are using script-based production to provide their many, many users with a customised view of the concert.*
>
> *The production gallery is full of staff working on FascinatE consoles. Some of them are supervising automated tracking programs, making sure that the systems are always selecting the best shots of each member of the band. Other members of the team are generating information about the views from different cameras and feeding this into the automatic script generation system.*
>
> *The big screen operator is using the input from the various cameras, both real and virtual, to create the backdrop to the show.*
>
> *Viewers at home can select their preferences for what kind of show they would like to watch, and the FascinatE system will build it automatically. They can select specific things to follow, like band members or the audience.*
>
> *People in the audience at the show can use the FascinatEURMobile service, as described in Use Case 10.*

From this use case some system and hardware requirements for the production side can be foreseen, in addition to those listed in *Use Case 6*:

- Multiple production users: The system must support simultaneous use by a number of production staff. This will require video, audio and script data to be available in real time across a number of production workstations.
- Range of screen sizes: The system needs to help operators produce content suitable for a wide range of screen sizes, for example by allowing a set of 'rules' to be specified which allow a given shot to be automatically adapted to a given screen size. This would allow shots customised for a range of screen sizes to be produced by a single operator (it would be uneconomic to have a separate operator producing content framed for each size of screen).

2.2.3 Provider perspective

***Use case 8*: Local content production**

> *July 2015. PBC, the national TV broadcaster of Palombia, is offering to its viewers the retransmission of the world championship of Athletics. For the first time in history, the 2-million Palombian audience has the opportunity to follow, live, each of their local athletics stars performing in the two national sports: discus and hammer throwing. Over the last years, PBC has had to face criticism that frequently, during such events, they did not sufficiently cover the performances of their national heroes, nor their physical preparation between the throws. In defence of PBC, the problem was simply that Palombian sportsmen were not a priority for the official production crew present at the event and therefore were not covered that much by the video feeds made available on the contribution links. This year however, the event is captured using the novel FascinatE acquisition technology and the A/V feeds are made available in the LSR. Thanks to the scripts that describe what content of interest is available in the panoramic view and at which AV fidelity, PBC is now able to locally and economically produce its own bouquet of TV channels covering the event that targets the local market and still is financially feasible. Those are simulcast on all the TV distribution platforms of the country and are produced in such a way that, anytime, the Palombian fans can follow each of their favourite athletes in action, beside other selected highlights and panoramic views of the stadium. Although this evolution has not required any change of consumer equipment, PBC is now able to offer to its audience an entire new experience of sport on TV, with content adapted to the local tastes and demands. Since the announcement of its new content offer, PBC has seen a twofold increase in their advertisement revenues.*

This use case can be deployed in several ways, each with its own specific requirements:

1. The most basic use case is to create one or several (say N) standard channels from the same LSR source. Each channel is clearly identified (e.g. via its name in an EPG) regarding the type of content/views it contains. For instance
   a. a channel can be associated to a physical camera location.
   b. Or it can be associated to a pre-rendered view, e.g. focusing on a certain type of actions, players (favourite team), etc.

   In this case, the end-user experiences a traditional lean-back TV service, where he can choose among a static selection of channels, whose content type is known by the end-user.
2. In a more advanced use case, the multiple views are not mapped to a static number of independent linear TV channels, but are proposed as part of a single TV service, with some coarse-grained interactivity for the end-user. For instance, some switching points are defined over time, where the end-user is given the option to move to another view among a predefined set of alternatives.

From this use case some system and hardware requirements can be foreseen. They are described in more detail in section 5.4.1:

- The creation of additional views requires some rendering functions after the capture stage.
- The way these views are created from the LSR content must either be controlled manually or controlled by production scripts that are ingested along with the LSR.
- The bandwidth capacity of the network must be dimensioned for the number of concurrent views made available.

***Use case 9*: Interactive Video service to any device**

> *June 2018. The Football World Cup is organized in Palombia. PaloCom, the major telecom operator in the country has seized this opportunity to launch a novel interactive video service on its IPTV platform. This will be available for premium live content, captured in the (now well-known) FascinatE A/V production format. The novelty today is that the main PaloCom video head-end is directly fed with the entire LSR, instead of traditional linear TV feeds. This LSR is made available by the national broadcaster PBC who is in charge of the A/V acquisition for this edition of the World Cup. (Thanks to a major increase in its financial resources over the last 3 years, PBC has been able to invest in a complete new set of FascinatE-ready acquisition equipment). This enables PaloCom to take advantage of the flexible LSR of the game to offer truly personalized and interactive video services. In addition*

> *to a set of pre-selected views (much like the bouquet offer which made the success of PBC three years ago), the interactive services allow PaloCom customers to fully select on-the-fly which portion of the stadium scene they want to see on their display. Interactive instant replays are announced for the next release of the service. Although the technical details are kept confidential, PaloCom has opted for a delivery solution where the interactivity requests are handled by the operator and fully rendered streams are transmitted to the end-customer device. Note that, Telombia, the main competitor of PaloCom, has recently introduced a very similar interactive service relying on high-end equipment at the end-customer premises to process the complete FascinatE LSR. Whereas Telombia serves only the 10% of their customer base having a 100Gbps fibre-connection, PaloCom is able to reach almost any device with virtually any access technology, from top-notch fibre down to the worst-case 50Mbps 6G mobile connections, requiring only a lightweight software installation.*

This use case can be deployed in several ways**.** From the end-user point of view, a full “interactive video service” should translate in a comparable experience for any type of end-device and network conditions: that is the ability to perform any panning and zooming operations in the content, being only constrained by what the audio/video sensors were able to record in the LSR. However, from the service provider and network operator point of view, the deployment of such a service can follow multiple roadmaps, essentially depending on the processing and access bandwidth conditions of its customer base:

1. [Telombia case] – *High Profile Terminals* – Full interactivity within the LSR can be offered only as a premium service. In that case, it is restricted to only customers with an actual access bandwidth superior to the full LSR and an end-device (SmartTV, set-top-box) that has the hardware capabilities for the reception and rendering of all A/V layers.
2. [Palocom - Home case] – *Main Profile Terminals* - it can be offered to any Home TV environment, with FascinatE-dedicated rendering hardware, but traditional residential access bandwidth. In this case, the network is required to transmit only the required portions of the LSR to the home terminal, so as to respect the bandwidth limit and fulfil the interactivity requests. An ad-hoc segmentation of the LSR must be defined, which determines the granularity at which the delivery mechanisms can be optimized
3. [Palocom - Any Device case] – All Terminal including *Low-Profile Terminals* – In addition to the previous case, the service can also be deployed to yet a larger range of devices, with limited assumptions on A/V processing capabilities. Assuming only the capability to decode a standard A/V stream at the resolution that matches the device’s display, processing requirements are therefore moved to the network side.

From this use case some system and hardware requirements can be foreseen. They are described in more detail in section 5.4.2:

- If the entire LSR can be delivered and processed by the terminal, a very high bandwidth (see section 5.2.2) is needed.
- If a traditional residential access bandwidth is assumed (see section 5.1.1), the network is required to transmit only the required portions of the LSR.
- If the A/V processing capabilities and access bandwidth is limited, all rendering operations are made in a network proxy.

***Use case 10*: Mobile magnifier**

> *In 2013 Jim is at the final concert of the Rolling Stones, and is listening entranced by his favourite music. In the stadium the concert is being recorded with a cluster of fixed cameras to a record high-resolution panoramic view, with additional detail being added to key areas of interest from the adjacent manned HD broadcast cameras. Jim uses his mobile to connect to the FascinatEURmobile service. After a connection has been established, his mobile initially shows the picture from its own camera on its screen. He points the phone camera to the stage and selects the drummer to be in the centre of his picture. He presses the OK button and the picture is replaced by a high quality close-up live stream of the drummer, as recorded by the camera system and repurposed for mobile usage. Jim can see the wrinkles of Charlie Watts and is very happy. He pans a bit by using his touch screen and watches Charlie do his thing. After a while he gets bored and selects Mick Jagger in the same manner. He presses the button ‘Follow me’ on his screen to make sure Mick will not walk out of the viewing frame on the mobile, as he jumps up and down on the stage. When the concert has finished, the FascinatEURmobile service informs Jim that an edited version of*

> *the concert is available. Jim watches it on his mobile during his trip home, just to enjoy the concert again. But he is disappointed with what the directors have selected for scene cuts and framing. So he activates the free navigation mode to get access to the whole 'database' and navigates freely (in both time and viewing window) to watch his favourite parts of the concert again.*

This use case is similar to subcase 3 of *Use Case 9*, as it relates to making FascinatE content and services available on mobile devices, such as tablets and smart phones. Additionally, the following system and hardware requirements can be foreseen:

- Network availability at a live event: this requires presence of a local wifi network or 4th generation mobile broadband network, for high-bandwidth connection at the event.
- Scalability: a delivery mechanism is required that scales to a large number of users, allowing for interactivity and adaptation of content to a variety of mobile devices. Such a delivery mechanism should be implemented in a managed or overlay delivery network, such as a CDN.
- Storage and caching: to provide on-demand and replay functionality, the delivery network must cache A/V segments that are delivered during content consumption, or store them for offline access after the event.
- Processing: given the variety in mobile devices, some low-profile terminals may still allow for some limited forms of processing, e.g. combining A/V segments at the terminal. Cloud-based components are required to handle high processing demands.
- On the production side, this use case requires functionality to relate the picture taken by the mobile devise, to the content captured by the camera cluster. Also, it requires feature and object tracking to create personalized views.

***Use case 11*: Hybrid Delivery**

> *John Smith arrives home late to watch his favourite football team Barcelona against Chelsea. The match has just started and he realises he just missed a goal as football players are already celebrating it. Even though several streams and channels are available to him from his IPTV service provider Palocom, showing the same football match from different angles, none of them is showing any replay he likes. The system signals him by a small icon in the top-left side of the screen that a complementary second screen service, called FascinaTwo, is available. This service allows him to receive a view automatically following his favourite player. John takes his complementary second screen device, a tablet, and starts the second screen application. With this application, John can receive specific views that have been captured and created by a third party. He can interactively navigate on his second screen, or he can use his second screen to control navigation on his primary screen.*

From this use case the following system and hardware requirements for the delivery network and mechanisms can be foreseen:

- Network access: this use case benefits from a hybrid delivery network, where primary content is delivered through regular service provider subscription, and second screen application and content is delivered via an over-the-top mobile broadband connection. The relation between the two services must be signalled on one or both of the networks.
- Interaction: a second screen application on a mobile interaction device, such as a tablet or smart phone, is required to connect to the additional services and content. The application should provide interactivity on the devices, as well as allowing the control of content on the primary device.

3 End User Perspective

This chapter discusses the requirements of the FascinatE system from the end user perspective. First, a study of similar systems currently available is presented. Next, a definition of the requirements available so far is extracted and, finally, a conclusion summarizes this section.

3.1 Background and Research Landscape

In this section, first systems that set the FascinatE project in the context of current state-of-the-art are briefly described. Next, qualities that FascinatE-based services should possess in order to provide quality of experience as desired by users are described and motivated through related work. Finally, gesture based interfaces are explained in more detail.

3.1.1 Technology coming up which reflects or has an impact on FascinatE technology

Systems visible to the end-user that have some FascinatE-like elements:

A hint of the possibilities offered by being able to extract portions of a very high resolution image for small displays may be seen in the 'HD View' work from Microsoft Research [Microsoft, 2010]. This uses 'gigapixel panoramas' and allows the user to interactively select a portion of the image to view. However, the images are stills, and there has been no significant work to our knowledge on using video to create images of anything like this resolution.

The first system that might go at least partly in the same direction as FascinatE is S.PORT from Sony, although this is still in a prototype stage. It has been installed in Arsenal's Emirates Stadium in London and allows Arsenal fans to watch replay, statistics and game scores on their PSP. Against this background it is also planned to capture the football game with two or more HD cameras and to stitch together a panoramic video in real-time using Sony's ZEGO processor technology. The user can then navigate within the panoramic view by interactively re-framing a small part of the scene and watch it on the PSP4.

Other systems:

- Quicktime VR - http://www.apple.com/quicktime/technologies/qtvr/ - still images that can include clickable hotspots; this has been around for many years.
- imLIVE - http://www.immersivemedia.com/markets/imLIVE/index.html - live streaming 360 degree video. The camera (http://www.immersivemedia.com/products/capture.html) looks neat, but 'only' captures 2400x1200 pixels and so is no good for zooming a long way into. They offer a complete end-to-end solution.
- Camargus – http://www.camargus.com/ - similar to the above, but with more cameras offering higher resolution.

In [MITLabs, 2010], the use of a tablet PC to pan around a scene being captured by three cameras is presented - one feeding a front monitor, and the others providing images for you to discover by 'looking around' through the hand-held device. This could be an interesting way of letting viewers 'browse' outside of the main image on a TV, making use of the panoramic video.

3.1.2 Navigation and interaction

A review of recent literature on navigation and interaction was included in D1.1.1 and for brevity is not reproduced here.

3.1.3 Gesture based interfaces

In some of the environments, it is believed that gesture-based (both touch based and visual-based) user interaction may take a major role in future and innovative systems. Many companies have recently been involved in developing interactive systems at different immersive levels. As stated in the previous section, **Immersion and liveness** (3.1.2), the main purpose of such systems is entertainment, being able to immerse the user in the event. Furthermore, some of them claim to be interesting for other applications like medical surgery or monitoring disabled persons.

Some of the commercial and technical characteristics of the recently proposed systems are listed and commented hereafter. As an example, camera setup, dictionary of gestures, system functions, userfriendliness, specific devices, etc. are some of the important issues to be taken into account when evaluating an interactive system.

Kinect: Commercialized by Microsoft Co., Kinect has caused a revolution in the field of player motion capture for video gaming. Devices such as the Wiimote or any other remote control system have become an old-fashioned version of gaming after Kinect was released in 2010 [Kinect, 2010].

Kinect is not a complete system itself, but a complement to the acclaimed Xbox360. More precisely, Kinect is composed of a microphone, an RGB camera and a depth camera, everything assembled in a 20cm bar. The RGB camera is mostly used for user (face) recognition, while the depth camera is Kinect’s crucial component which allows precise tracking and gesture recognition.

In order to develop Kinect’s depth camera, Microsoft has bought PrimeSense [PrimeSense, 2010], a company which had already excelled in the construction of depth cameras. An infra-red projector combined with a monochrome CMOS sensor allows Kinect to see the room in 3-D under any lighting conditions. How Kinect’s camera works has not been officially released, even though some reverse-engineering projects may give a clue on this topic [DIY,2011].

Kinect’s output has mainly been used to perform body tracking, the user being able to move freely to interact with a system. Kinect is based on a pose classification algorithm based on random-forests [Shotton, 2011], which delivers a precise pose estimation allowing the user interact with virtual elements on screen.

Microsoft has affirmed that they will not go below a latency of 0.1 seconds. However a lower latency would be desirable in some Kinect games applications.

Kinect’s output has had a very good acceptance in the Computer Vision community and many research projects have included Kinect in their research work.

HHI iPoint Presenter: HHI has developed a gesture-based interactive system for industry and medical surgery applications, as well as entertainment. HHI has based its prototype on a vertical dedicated camera which “sees” the user's hands. This way, iPoint can detect, track and interpret hand gestures so that the user may interact in real time with the system [iPoint, 2010].

With iPoint, one may manipulate virtual objects on screen and navigate through menus in real-time. A dictionary of gestures is also included in iPoint, which contains some basic navigation gestures such as zooming, selecting or rotating amongst others.

Extreme Reality XTR3D: Extreme Reality, in collaboration with Texas Instruments (TI) has developed a low cost gesture recognition system for mobile terminals. XTR3D uses a standard webcam as optical sensor, which drastically reduces the system's cost.

In their website, XTR3D demonstrates what they call ”touchless gesturing” with a mobile device – whereby users can control applications by simply pointing, clicking, dragging, and scrolling. Therefore, a small dictionary of gestures is to be recognized and classified [XTR3D, 2010]

XTR3D Human Device Interface claims to be cross-platform, being applied to TV gaming and animation. Capturing and tracking of the upper-body is also one of XTR3D features.

In conclusion, the most eye-catching and research-friendly system is Kinect, enabling full interactivity with the system by means of real-time full body tracking. The user may point at different places on the screen to navigate through menus, select applications and perform a large variety of movements which are captured and interpreted by the system.

One may appreciate that there exist few systems which provide device-less interactivity. Furthermore, only Kinect allows full interactivity, the others offering upper body or hand gesture recognition.

The number and type of camera is also an important point to be taken into account. FascinatE's scope does not envisage the use of a large number of cameras. Actually, UPC's setup for recordings consists of a central Kinect camera. Two lateral color cameras where envisaged before Kinect was released, but have been removed given Kinect’s RGB camera option. Thus, systems like Organic Motion Stage (10 cameras) [Organic, 2010] or HHI iPoint Presenter (special vertical camera) are not adapted to FascinatE's requirements.

The characteristics of the Kinect sensor are closest to what FascinatE aims to offer in terms of gesture recognition. However, both projects differ in some details. FascinatE does not need to track the full body precisely, but only some 'hot' body parts such as hands and head, even if rough tracking of the rest of the body will be helpful. FascinatE might require a gesture recognition system which works continuously, especially in the case of long periods of no movement of the active user (e.g. for the duration of a film). FascinatE's gesture recognition system should be capable of tracking and interpreting gestures after such long periods of inactivity, while Kinect always deals with highly active users.

A second level of tracking precision may include hand pose recognition and finger tracking [MITLabs, 2011]. Such feature should enable different ways of interaction through hand gesturing, combined with the above mentioned body gesturing.

An important point about the Kinect camera is the growing open-source community developing drivers and tools adapted to Kinect, such as OpenNI [OpenNI, 2011] and PointCloudLibrary [PCL, 2011].

Available and suitable technologies

The interest in vision-based action recognition has dramatically grown over the past years. Research on this topic has not been focused in a single direction but quite the opposite. A great variety of approaches and points-of-view are being proposed continually.

However, one may extract [Poppe, 2009] some steps in a gesture recognition system, which may facilitate the task of classifying such an enormous amount of research work. Such steps are:

- Feature Extraction
- *Tracking (if needed)*
- Action Recognition
- Classification

There is no limitation about the number, nature or complexity of the features to be extracted; nor about the action recognition algorithms to be used. Nevertheless, the chosen strategies should be consistent with the system requirements, specially those related with temporal constraints.

The FascinatE gesture recognition system aims to be a user-friendly interface, providing a full interactive experience which goes beyond the functions offered by typical remote control devices. Issues like real-time and system latency should meet user expectations, therefore temporal requirements of the system should not be underestimated.

Furthermore, FascinatE users will enjoy the system in a great variety of scenarios, with uncontrolled lighting (a scenario with no illumination is considered), partial user occlusions and many other unexpected artefacts.

Extracting features from an image or video sequence is the first important task of a gesture recognition system. In a similar way, some authors talk of 'image representation' referring to this first step. Indeed, it is just a matter of linguistics, since the objective remains the same: finding the characteristics (or features) which contain the important information for gesture recognition purposes.

According to Poppe [Poppe, 2009], feature extraction systems may be classified as either global or local representations. Global representations encode the region-of-interest of an image as a whole, dividing it into smaller zones through subsequent steps. The main drawback of such systems is that they are very sensitive to noise, partial occlusions and viewpoint variations. Therefore, they are less suitable for FascinatE.

On the other hand, local representations describe the observation as a collection of local descriptors or patches. These approaches are not subject to background subtraction and they behave better faced with changes in viewpoint and partial occlusions. Therefore, feature extraction algorithms using local representations may be particularly suitable for FascinatE, given the unconstrained nature of the environment in which the system needs to operate.

A short overview of some local-based feature extraction strategies and aspects are mentioned hereafter:

i) *Interest Points Detection (in space / time)*: Interest points are locations in space and time where sudden changes of movement occur in the video. Extending edge detection algorithms to 3D [Laptev 03], or using saliency and curvature operators [Willems, 2008] are only two examples of interest point extraction techniques.

ii) *Local Descriptors*: Image patches are summarized through a great variety of local descriptors. Local descriptors may contain a wide range of information, from 3D histograms [Laptev 08] to gradient and motion-flow operators [Dóllar, 2005], amongst many others.

iii) *Dimension-reduction algorithms*: A large number of high-dimension descriptors is usually obtained. Reducing the dimensionality of the problem is crucial. Some algorithms like PCA may be used. Patches and descriptors may be clustered to generate a codebook or bag-of-words.

iv) *Correlation between descriptors*: Descriptors may contain redundant information. Correlation between descriptors may help to reduce the amount of information representing the image, leading to a non redundant representation. Some common characteristics amongst descriptors are spatio-temporal co-occurrence [Scovanner, 2007; Savarese, 2008] or similar tracking features [Sun, 2009].

Generally speaking, local-based techniques trend to produce a large number of high-dimension interest points and descriptors. Reducing the dimensionality of the problem is crucial. Some algorithms like PCA may be used. Patches and descriptors may be clustered to generate a codebook or bag-of-words.

About Depth Range Cameras

Cameras which provide depth information have been widely studied and developed recently. We focus on the Kinect camera, given its great impact in recent research projects. Such cameras rely on the recognition of a structured light pattern, which shows how surfaces in the scene are oriented and at which depth are they located. After an internal filtering step, a depth map is delivered. Such an approach limits the measurement distance to, at most, 7 meters.

Recent conferences and journals on gesture recognition provide many references to strategies exploiting Kinect depth information. Given its ability to work without needing to rely on general scene illumination and the fact that depth-based information makes it easy to ignore more distant objects in the background, it seems to be an important research direction in the FascinatE context.

The tested Kinect camera provides two images per frame:

- *RGB image* : 1024x768 pixels
- *Depth image* : Depth map of 640x480 pixels (**Figure 6**):

These cameras offer a relatively good resolution of 640x480 pixels (VGA), much higher than previous depth cameras such as Mesa SR4000 [SR4000, 2010]. Furthermore, it allows image capture at about 30fps, which is a useful frame rate for real-time tracking applications. In addition, since the Kinect camera works with IR light, they are invariant to illumination changes, being able to make recordings in dark scenes.

The FascinatE project will study the use of the Kinect camera because depth information appears to be of great importance to detect gestures robustly. As an example, the most advanced commercial system nowadays, Xbox, also relies on depth information.

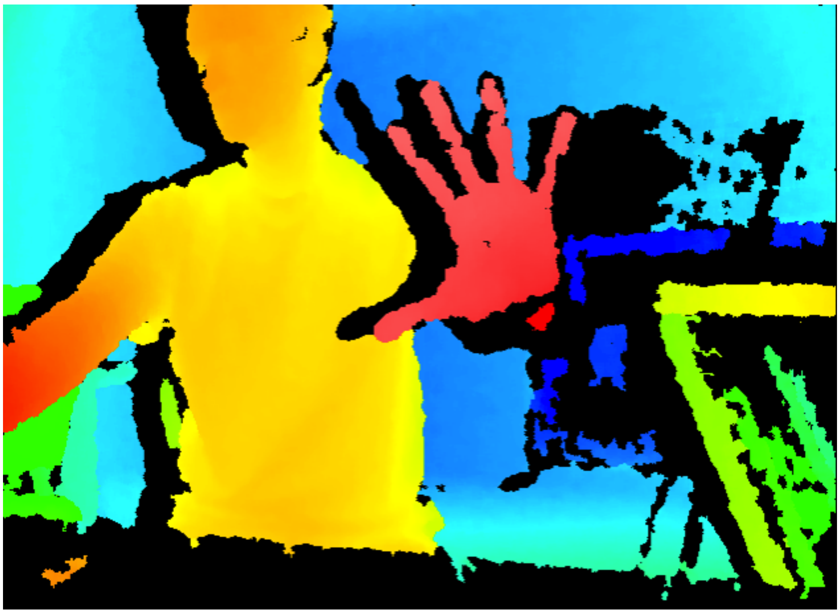


Figure 6: Kinect depth image

3.2 Requirements

This section aims to give interaction design guidelines for end-users using services based on FascinatE. It is more about user requirements' analysis, design process and testing cycle, and less about specification.

3.2.1 Interaction design

The design principle used in the end user interface design within the FascinatE project is to put the user, e.g. the viewer in the centre of the design aiming to make the user's interaction experience *as simple and intuitive* as possible.

It is therefore important to understand the quality expected by the user, both in terms of user experience from an interaction perspective, as well as content that FascinatE offers to users.

Questions to answer are:

- What should content include?
- How to access and manipulate content?
- What interaction methods are appropriate to the content?
- How to design intuitive interfaces allowing the end-user to engage?

To understand the potential needs of the end-users, it is also necessary to understand the functionality required by the system, i.e. to create a list of the (FascinatE) functionality requirements.
From the FascinatE project description and provided use cases, the main functional requirements are:

- Layered audio/video scene with cylindrical panorama
- Scripts supporting reconfiguration, object tracker updates and salient object/actions lists
- Advanced gesture based end-user interaction with the content
- Scalable delivery increasing diversity of end-users devices and network connectivity

Thus, the way users will interact with the system in various settings, will affect how the production of the content is carried out. As target users, both passive and active users are observed.

The remainder of this section starts with describing the planned approach. Next interaction mechanisms common for various terminals are stated and then environment-specific requirements are given. The section continues with the description of interactive commands for controlling the audio and video rendering and at the end gesture based interaction is explained in more details as an example of interactive commands.

Planned approach

Prior the actual start of the design of user interface, it is needed to do user experience work – analyse users and collect a list of user requirements. User requirements are based on the analysis of the potential users of the system and it is done through user studies.

In order to get requirements, in parallel with doing an overview of state of the art literature (including interactive TV, mobile TV, novel interaction techniques and similar), the following methods are being used:

- Workshops, brainstorming sessions and interviews with participants like e.g. designers, keen (live) TV viewers, sport fans, semi-pro producers, “normal” users and similar. Material that will be used includes storyboards, mock-ups, prototypes, etc. Study and usability report on user interfaces and demos defined and developed so far in the WP5 (video renderers and interface mock ups) can be found in D5.3.1, Section 4.
- Ethnographic studies of existing audience interaction to identify how, why and when people react and interact in natural settings. Results obtained so far as well as future studies are described in *D 5.1.1, Section 5*.

A pilot workshop (Figure 7) gave us a first impression of users' perspective on the use of FascinatE. Studies that came after helped us deepen understanding of users' viewing preferences as described next.

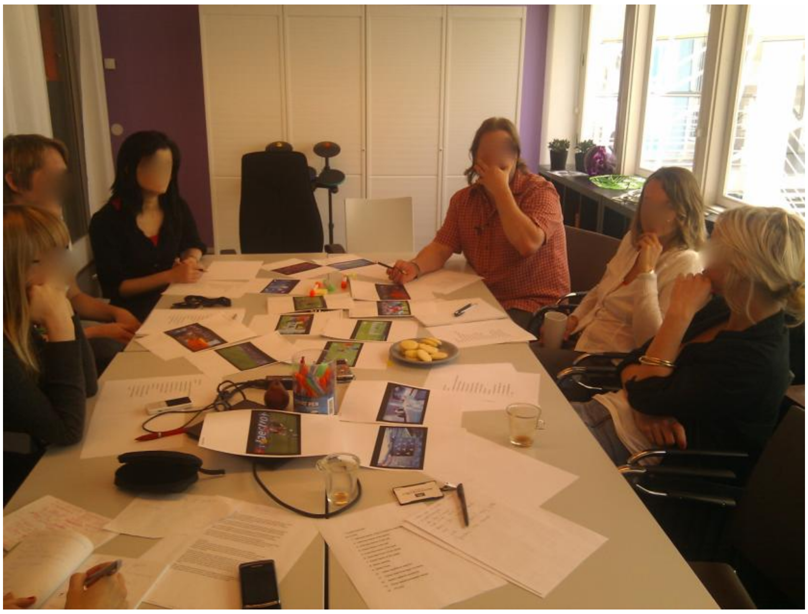


Figure 7: Snapshot from the pilot study

Common interaction mechanisms and metaphors across devices

A set of viewers' preferences is obtained by user analysis. In the iterative design process we are taking, prototyping and usability testing results are used to refine the end user requirements. When talking about viewing preferences, we consider both what users appreciate in TV viewing as it is today, and what novel TV services should offer to them to be at least comparable with the next generation interactive TV.

Here is the list of so far collected end user preferences i.e. design guidelines:

- **Level of interaction.** Passive and active use, design for various levels of interaction,
  - Relaxed exploration instead of information seeking (starting with familiar content and continuing with browsing of relevant items) [Chorianopoulos, 2008],
- **User profiles and pre-configuration** before the event (the idea: with more configuration before the event, less interaction will be needed during the event, resulting in more relaxed viewing of desired content),
  - Users' preferences of the (sport) event created with respect to event type, event venue, athletes' performance, nationality, team, or statistics.
- **Socially supported viewing.**
  - Semi-professional mode in which somebody else (e.g. amateur producer) produces the content for end user (a group),
  - Social navigation including following what others are viewing, sharing our own current view and rating of user profiles. For this, an interaction channel is needed, carefully designed to minimize user annoyances and distraction from the main TV content; latency is a critical issue, particularly for real-time communications.
  - Alarms from a system or other users, or subscribing to somebody else's view (friend or semi-professional) instead of active viewing). Viewers are worried about missing important moments while interacting with the system. One of the participants said: "*If you are your own producer, you know that you are going to miss something*".
  - Group-based video streaming rather than individual – it enables shared experience, but also scalability
- **Content presentation.**
  - Use of multiple screens,
  - Dynamic playlist of streams (sorted according to priorities) offered to the user depending on what was watched before, social trending information (e.g. stream "popularity" among your friends), user profile, location, or type of terminal.

  - Event as combination of multiple streams (“picture in picture”). The question is “*How many streams people can handle?*” – the answer from the studies was not more then 2-3 streams, depending how much movement is on each of them.
  - Possibility to interact with audio and video streams independently,
  - A need for an easily-accessible video stream with “the best view” – producer recommended.
  - A need for an easily-accessible panoramic view (overview picture).
  - Importance of replays: collection of replays (either as producer choice or from what others were viewing/replaying) as a video stream to offer.
  - Each stream needs to tell story (be narrative). Stream examples:
    1. Following person or object (e.g. player, ball, key actions)
    2. With focus on one of the participating teams/nationalities (e.g. audience cheering, main player etc.)
    3. "Promoting channel" (e.g. most important moments, result changes etc.)
    4. Collection of reruns (producer suggested and friends’ favourite)

Environment-specific requirements

The choice of an interaction technique, and a type of content depends on a specific setting; terminal properties and characteristics of the environment, i.e. context, need to be taken into consideration - each environment and/or terminal has its own features and limitations. In FascinatE, three main environments are differentiated: mobile, home and public. Next, we give the environment and terminal properties for each of them collected so far:

Mobile:

- When a mobile phone is used as the terminal, multitasking is required, mostly because of the communication requirements [Cui, 2007], e.g. watching TV and answering phone calls,
- Context, more details can be found in *D5.1.1, Section 5.2*:
  - On the go: single use, a pause and/or mute functionality is required, condensed information suitable for breaks and waiting periods (no longer than 10 minutes); in situations like walking or cycling, audio is preferred, immersion should be avoided,
  - Public space (public transportation, coffee shops, waiting rooms etc): one or multiple users (others invited by the owner of the device), audio use is limited,
  - Private space (at home, at work, private car etc.): one or multiple users (others invited by the owner of the device), privacy and control
- Importance of user-generated content, audio and video sharing [Buchinger, 2009], [Oksman, 2007],
- Terminal properties:
  - Screen size limitation [Buchinger, 2009], or not, if the viewing distance is taken into consideration [Cesar, 2010]
  - Battery life (a threat to more important communication needs [Knoche, 2007]),
  - The acceptability of the medium shot (with the greatest amount of detail) in the football video was less acceptable than the long and the very long shot at lower resolutions [Knoche, 2008],
  - Communication technologies: SMS/MMS, wireless, 3G/4G,
  - SIM card for end user identification,
  - Screen based text must not obscure action, but must be large enough to read, e.g. solution is to “swipe” to overlay text onto or off the screen,
  - *Potential interfaces*: keyboard, voice, stylus, gesture (e.g. touch screen – gesture recognition; swipe and pinch gestures),
  - *Sensor based interaction*: accelerometer (shake as hand gesture, tilt), RFID (Radio Frequency IDentification), magnetometer, camera (visual search),

Home:

- Social context is complex and varies over time (family and friends watching together); public shared space; negotiation with regards to the interface (remote control) – interface should be shared among the group, and immediately available and instantly shareable among the group [Vatavu, 2010],
- *Typical setting*: One or multiple users with hierarchy of users, possible use of multiple screens (mobile phones, tablets, or laptops). Typically in the living room (typically "lean back" interaction), but also in e.g. a (sport) pub ("lean forward" interaction)
- Use of secondary mobile screen which has more control compared to the first screen; if used to enhance TV viewing, it should display the same image at the same time as the main screen
- Terminal properties:
    - Authentification (e.g. face recognition), detection of location, and tracking of users (e.g. arms or hands)
    - Potential interfaces: gesture recognition (e.g. pointing), motion (e.g. Wii, - free space mouse or similar), tangible, voice;
    - *Examples*: Microsoft surface or interactive coffee table [Radu-Daniel, 2008] as a shared interface, sensor based (e.g. tangible cube device [Block, 2004]), physical mobile (phone) interaction: touching, pointing, scanning

Public:

- *Typical setting*: Multiple users with one common screen and multiple personal screens. Public viewing can be: (1) directly at a live event (e.g. concert, sport event, or festival), or (2) as live broadcast in cinemas, theatres, open spaces etc.
    - Use of multiple screens (e.g. mobile phone) to supplement content to the live broadcast/event (enhanced TV) – more control. Alternative screens should display the same image at the same time as the main screen.
- Terminal properties:
    - Possibility of "crowd" control – e.g. mobile interaction (web/SMS) or physical location/ actions of crowds

Interactive commands

Interactive commands control the audio and video rendering and define the interface between the terminal renderer and the end user interface. They are extracted from the functionalities requirements defined in this document. Interactive commands constitute base elements for designing the UI and they need to support the following:

1. **Switching between predefined video and audio streams, i.e. Region-of-Interest selection.** Those streams will be generated by the producer (or semi-professional) and offered to the end user, whereas it is possible to interact with audio and video stream independently. The choice (and number) of streams will depend on several factors, including capabilities of the viewing device, user preferences, feedback information and similar.
    - Each stream needs to tell story (be narrative).
2. **Navigation**, i.e. doing virtual camerawork in the panoramic picture (in navigation in the panoramic view from OMNICAM, tight zooming with full resolution is only possible in from the region(s) **viewed** by already existing pan/tilt/zoom cameras). Navigation assumes moving both in time (i.e. replays) and space. A more detailed description of some associated commands is available in D5.1.1.
3. **Audio manipulation**
    - Separation of foreground sound from background sound,
    - Setting gain of the speech component(s) of the audio.
    - Volume up/down/mute

Gesture based interaction

As an example, in a home oriented scenario, the visual gesture based interaction should include:

- Automatic user detection and identification
- Administration of users: power / normal
- Management of active user
- Only one user is able to control the system
- In multi-user scenarios, the first user is by default the active one. Other users become active when receiving the token.
- Reasonable latency of the system [Nielsen, 1993]:
    a. Visual feedback of some gestures (e.g. zooming or raising volume) must be fast enough to allow fluid interaction (e.g. less than 0.2 seconds)
    b. Other interactive commands (such as changing channels or dividing screen) are less restrictive in latency (e.g. less than 1 seconds)
- Gesture recognition limited to a predefined set of gestures

3.2.2 Usability assessment

The FascinatE system, and services based on it will be evaluated by testing on users in order to get direct input on how real users use the system.

Usability assessment in general focuses on measuring how a human-made product relates to its intended purpose; it discovers errors and areas of improvement by observing people using the product. Usability can be defined as the extent to which the system can be used by specified users to achieve specified goals in a specified context of use. Some of usability principles that need to be checked are:

- **Effectiveness.** Can you achieve what you want to?
- **Efficiency.** Can you do it without wasting effort?
- **Satisfaction.** Do you enjoy the process?

Factors to be included are e.g. suitability for the task, learnability, error tolerance etc.

Planned evaluation approach

Usability evaluation is *an assessment of the usability of a product, system, or interface.* Design process will include:

i) Integration of different features and functionalities in a gradual way, and

ii) Surveys of user feedback at each of the foreseen system demonstration

The results of the first test will be used as a control measurement, and all subsequent tests can then be compared with the control measurement to indicate improvement.

Suggested approaches:

i) **Heuristic evaluation**. It involves evaluators examining the interface and judging how it matches the known usability principles, i.e. "heuristics". The main goal is to identify any problems associated with the design of user interfaces.

ii) **Laboratory based usability studies**. It includes observing participatory users in semi-experimental, laboratory based environment, and possibly logging.

iii) **Naturalistic evaluation** – It includes observation of users in the specific context and natural environment, and possibly logging.

Heuristic evaluation

Usability heuristics used for heuristic evaluation are based on ten general principles for user interface design collected by Jakob Nielsen [Nielsen, 1994]:

1. *Visibility of system status* - The system should always keep users informed about what is going on, through appropriate feedback within reasonable time.
2. *Match between system and the real world* - The system should speak the users' language, with words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

3. *User control and freedom* - Users often choose system functions by mistake and will need a clearly marked "emergency exit" to leave the unwanted state without having to go through an extended dialogue. Support *undo* and *redo*.
4. *Consistency and standards* - Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform conventions.
5. *Error prevention* - Even better than good error messages is a careful design which prevents a problem from occurring in the first place. Either eliminate error-prone conditions or check for them and present users with a confirmation option before they commit to the action.
6. *Recognition rather than recall* - Minimize the user's memory load by making objects, actions, and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate.
7. *Flexibility and efficiency of use* - Accelerators -- unseen by the novice user -- may often speed up the interaction for the expert user such that the system can cater to both inexperienced and experienced users. Allow users to tailor frequent actions.
8. *Aesthetic and minimalist design* - Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility.
9. *Help users recognize, diagnose, and recover from errors* - Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution.
10. *Help and documentation* - Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large.

Prototype Testing

The aim of this method is to test a system on users as early, and as often, as possible. It is not necessary to have fully working systems, but rough and functional prototypes on which users perform certain tasks and an observer records the results. In the early stage of FascinatE evaluation, laboratory based testing is more appropriate. As the prototypes reach maturity, FascinatE will be also evaluated in the natural environment. Demos available at the moment are described in D5.3.1 and presented together with the corresponding usability reports.

FascinatE related usability requirements

Considering a specific characteristic of the FascinatE systems (live video interaction), the summary of the most important high-level requirements from the end-user perspective (both usability and usefulness) as learned from studies, literature and general interaction design practices, is listed next:

- Interaction needs to be non intrusive, by leaving the visual channel between user and the screen open
- Latency of the system needs to be low enough not to interfere with real-time content
- The user interface should:
    - c. Be simple, intuitive, consistent across similar control options
    - d. Posses a dimension of fun that makes interaction process captivating [Vatavu, 2010]
    - e. Be non - modal (only one mode is needed, so that each command has only one meaning).
- The form of input needs to be mapped to the intended output as directly as possible
- Support for multimodal input (choosing modality, e.g. Ability to “undo” simple actions without accessing deep menu structures is needed
- Feedback should be as soon as possible, and clear in its message - observability
- Design for interruption should not be time sensitive

3.3 Conclusion

In this section the requirements of the FascinatE system from the end user perspective have been discussed.

As suggested, and explained in Section 3.1.2, the following should be bore in mind when designing for FascinatE-based services:

- Interaction practices in various environments,
- Social component of TV watching,
- Use of multiple screens,
- Immersion and liveness,
- Virtual camerawork,
- User generated TV/video and
- Instant replay.

While there has already been some work done in the direction of gesture based interfaces, as covered in Section 3.1.3, there is still a lot to be done to understand what users would like to get from the system like FascinatE, and how they would like to use it (Section 3.2.2). After this initial testing has been carried out, it will be possible to create a full set of end-user requirements.

However, a first set of user requirements has been extracted based on current state-of-the-art literature and studies performed so far (Section 3.2.1):

- The user interface proposed by FascinatE should show the following properties: simple, intuitive, efficient, non intrusive, consistent and clear.
- There should be a reasonable latency of the system, at least comparable with today's systems.
- Users' viewing preferences should be kept as a very important goal. In general, TV viewers want to be entertained, get informed and relax.
- Three main environments are differentiated within FascinatE: mobile, home and public. Each environment could provide different levels of interaction depending on the terminal capabilities, the social context and the typical settings of the specific environment.

In some of the environments, it is believed that gesture-based (both touch based and visual-based) user interaction may take a major role in future and innovative systems. Therefore, the FascinatE system will be partly controlled through a set of visual human gestures (see deliverable D5.1.1) to provide a grater immersion and liveness experience. This control-by-gesture will not completely replace other control devices, but will provide an alternative to traditional interaction methods such as remote controls or PCs.

4 Production Perspective

4.1 Introduction

This section of this document describes the restrictions placed on the system from the point of view of the production staff, workflows and systems. We maintain the focus on production of coverage of live events. We consider:

- Technical requirements of the system hardware such as interfacing with existing systems,
- How people interact in existing production galleries and what these interactions achieve,
- New production techniques both required and enabled by new technologies introduced in the project.

The term "requirements" is a little misleading in the context of A/V content production under the FascinatE system. The factors which we discuss here might be better described as guidelines, or design considerations. Nothing in this section of the document places a quantitative restriction on the design or operation of the system, with the exception of some technical details. In this section we are trying to capture the ideas and impressions of production staff, such that we can create designs and prototype systems that are appealing. We describe ideas that production staff have talked about when presented with the very early outline of the system which we have today. As the project develops we anticipate that these guidelines will begin to mean different things. Some will have to be clarified, and some will have unexpected consequences.

4.2 Production Systems Today

The A/V content production system proposed by FascinatE is a paradigm shift beyond current TV programme production. However, most of the elements of which this new system is built have their roots in existing production methods. In addition, a discussion of production systems helps to provide a context for the discussion of production methods and roles under FascinatE.

4.2.1 Roles - people involved and their hierarchy

In this section we examine the roles of various staff responsible for a television production. This discussion is intended to be general, although we maintain our established focus on these roles in live events. In some sections the role deviates significantly from what is performed today. It is appropriate to group these tasks against approximately similar ones at this early stage of the project. Once we know more about what a FascinatE production gallery might look like, it will be more appropriate to suggest new, FascinatE-specific production roles.

Camera operator/remote cam op

A/V content production today uses a wide variety of camera operators. Alongside conventional manned broadcast cameras mounted on tripods or held by the operator, there are remotely operated cameras controlled by joystick, cameras on jibs and rails and fixed cameras, to name a few. The camera operator’s job is to provide the director with shots that he can use as part of his programme. Exactly what this means depends upon the type of programme being made. The following factors contribute to a "good shot":

- Content - what is in the shot,
- Static composition - The relationship in space and focus between the subjects in the shot, and also with the areas of the shot which are empty,
- Dynamic composition - how the motion of the shot relates to the motion of the subject. For example, tracking some racing athletes, or tracking a journalist walking across a shot,
- Angle - This is related to both content and composition.

In order to achieve this, a camera operator will have a wide range of controls on their camera, as well as more or less freedom of movement. In general there is a three-way trade-off between size (and hence freedom of movement), technical control, and cost. Although as with all things technological, there is more available for less every day.

In general, the highest quality pictures come from large cameras with large lenses. These are big and heavy, and so are difficult to move. Also, they have a certain amount of inertia which restricts rapid movement. This in turn restricts the kinds of dynamic moves which can be captured on screen.

Smaller cameras can also provide very high quality pictures, suitable for broadcast. A camera which a single operator can move around freely, or hold in one hand, is capable of producing video suitable for broadcasting. If a camera like this is used without some kind of mechanical support then the resulting video will shake and move noticeably.

Very small cameras are used in places where a camera operator could not get. Some examples:

- A Formula 1 driver’s view
- The view from the centre stump on a cricket pitch
- A close up view of a timid or dangerous wild animal

These cameras are usually fixed in one place with one view by an operator, and then left recording. Sometimes they can be controlled remotely, with more or less degrees of freedom.

The different types of shots and properties of different cameras are all used by the director to tell a story. It is the camera operator’s job to get the director the best footage they can, appropriate both to the subject and the type of camera they are operating.

Operating a camera is seen as a craft. Whilst the operator will have some knowledge and skill of the technical aspects of a camera, in particular those required to do his job, they will not generally be a technical expert. In all but the smallest of productions, the operator will rely on technical staff for most of the maintenance of the camera.

Camera operation is a skill which is learned by other staff. It is common for journalists to learn camera operation so that they can go into the field alone and produce video content to transmit back to a TV studio.

Director/vision mixer

In a TV production the director is responsible for executing the vision of himself and the producer. In conventional TV production this means the director will instruct the camera operators of which shots he requires them to take, will request graphical overlays on the screen, will decide when to cut to a replay, or to an advertisement break. The director also has control over the people who will be seen and/or heard as part of the production. In drama production, the director will be providing instructions to the actors. For sports and news, the director will often be providing information and instructions to the presenters and/or commentators.

Depending upon the size of the production, the director might delegate some or all of their responsibility to a team of staff.

The director has overall control and responsibility for everything which happens during the production. In some cases, the director will be operating the VM. Often the director will have very little direct control of any aspect of the production. The director will rely on the skills and experience of the various other staff and operators involved in making the programme. The director will expect the operators to follow their instructions immediately, as well as to use their own judgement if the director’s attention is elsewhere.

The Vision Mixer

Depending on the scale of production, the VM console might be operated either by the director or by a dedicated operator. The VM sees what is being captured by each of the cameras all the time. It is the VM who chooses which of these cameras is transmitted at any given moment. This is a very difficult job, and defines a lot of what the end user will consider to be "the programme." The VM will always be considering where their next shot, or next few shots, will be coming from.

The main tasks of the VM are to switch between cameras and sources i.e. to determine what will be shown in the broadcasted feed. The way in which a switch between two sources is made is called a transition. The most common transitions used are:

- Cut – The switch is made instantly.
- Dissolve – The two sources are mixed, one fading in, one fading out, superimposed.
- Fade – The source fades to or from a solid colour, usually black.
- Wipe – One source replaces another by following a two dimensional pattern.

The VM also controls if any graphical effects should be added to the broadcasted image, such as company logos or text. By switching between cameras (cutting) in different ways, the VM has a lot of control over the feel of the programme.

Note that the term VM is applied to both the operator and the console.

Audio engineer

The audio engineers are responsible for the capturing of suitable sounds to accompany the pictures for the programme. They will position microphones around the scene, then mix the signals from them together in one or more ways as required by the director.

The role of the audio engineer varies depending on the type of programme being made. Sports will often require the audio to alter to reflect what can be seen on screen. This means that the audio engineer will be actively involved throughout the production, constantly following the VM and the director's instruction. The audio engineer must have an appropriate mix ready to cut to when the VM cuts the video feed. In some sports there are mobile microphones which the audio engineer can rely upon to always produce suitable sounds (to accompany views of the game). These might be controlled by human operators, or attached to people involved with the game who are following the action, for example the rugby referee.

Music usually requires the audio mix to be fixed. In the case of pop music, this is because the band will often have their own engineers, who will create the bands "sound". For classical performances it is usually the goal of the director to emulate for the audience at home the experience that a member of the concert audience is having. Because of the very large number of microphones usually involved in a classical production, anything other than a fixed mix would probably be too complex to attempt to engineer.

Replay Operator

The ROs work as a team to provide the director with replays of key events. Especially during sports coverage, events can happen faster than live coverage can make sense of them. When one of these events happens, it is the RO's job to review the footage which has been captured and make suitable clips available to the director. When a replay is in progress, the RO is in control of the video being transmitted to the viewers. On almost any sports event there will be a team of ROs, with a master operator keeping track of what all the operators are doing.

The RO usually uses a device manufactured by EVS (brand name for replay/highlight video servers). Each EVS console usually work on four camera feeds at a time, displayed on a split screen setup. These camera feeds can then be dynamically selected and worked on two monitors in parallel (see Figure 8). This allows the operator to produce sequences of replay shots. The device stores video data on hard disk so as to be randomly accessible and navigable. The operator can review footage from either camera, scrubbing back and forwards at variable speed.

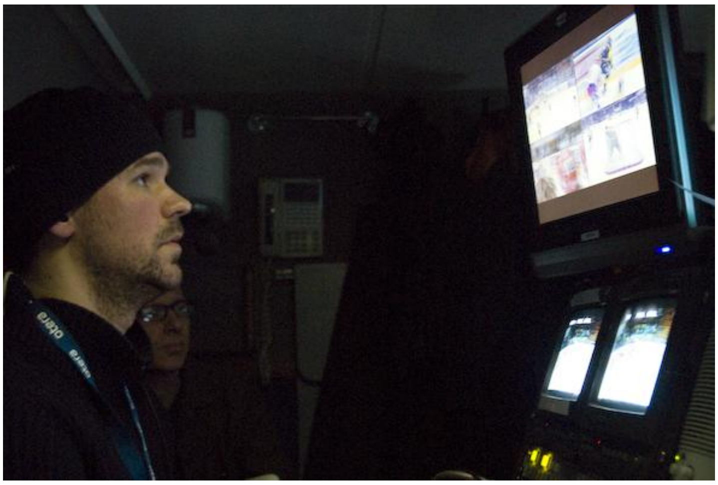


Figure 8: Replay operator on an EVS console

The job of the RO requires a diverse range of skills. Operating the console itself requires skills similar to both a VM and an editor. The operator is likely not to be operating under the direct control of the director, and so must make decisions about the quality of the shots they have independently. To that end they must understand whatever sport they are covering, and the job specific skill of how to clarify events particular to that sport.

More so than other roles in TV production, the RO must be capable of taking in and processing information very rapidly. They must comprehend things which they have seen on their cameras and be ready to act on them immediately. This part of their role makes ROs especially suitable as candidates for operating semi-automated metadata creation systems under FascinatE. These data could then be interpreted by the scripting engine to create useful scripts.

4.3 TV Production Technology Today

TV production today uses a range of technology. The FascinatE project will have to interface with some of it, and will replace other parts.

Camera

At a fundamental level, the job of the camera is to capture video. In order to do that effectively there are a number of additional features that a modern camera might be equipped with.

Most cameras are connected to a central system of video feeds. This allows the operator to see the views from other cameras. In the simplest system, the operator can view the output from the main VM. In more complex systems the operator can also view what other cameras can see.

Whatever video systems FascinatE uses, they must interface with existing broadcasting equipment using standard interfaces.

Video processing

The camera is connected to the “racks”. In a large outside broadcast or studio production this is physically separate from the camera: In another room, the gallery, or in the scanner truck for an outside broadcast. Here, there is often some control of the cameras operational parameters. Typically, those things which require some technical analysis, or which should be set once and then left alone, are controlled here. White and black level are examples of parameters which might be controlled by the racks. In HD production the focus is sometimes carried out by this operator.

Vision Mixer

The VM is a large and complex device. It accepts input from all the cameras, playback devices, and graphics systems. The operator can then mix them together in various ways to control the video part of the final programme output. The VM enables the operator to transition between different camera or pre-recorded inputs in different visual ways. The operator also has control over which graphics appear on the screen at what time.

The control surface of the VM appears as a large array of buttons, which often can light up. They are spread around the control surface in logical groups. Some of these buttons will control which video feeds are displayed on some of the monitors on the video wall. Most of the time these are fixed, but the vision mix operator will be able to control a small number. Figure 9 shows an example of a VM console in an outside broadcast vehicle.

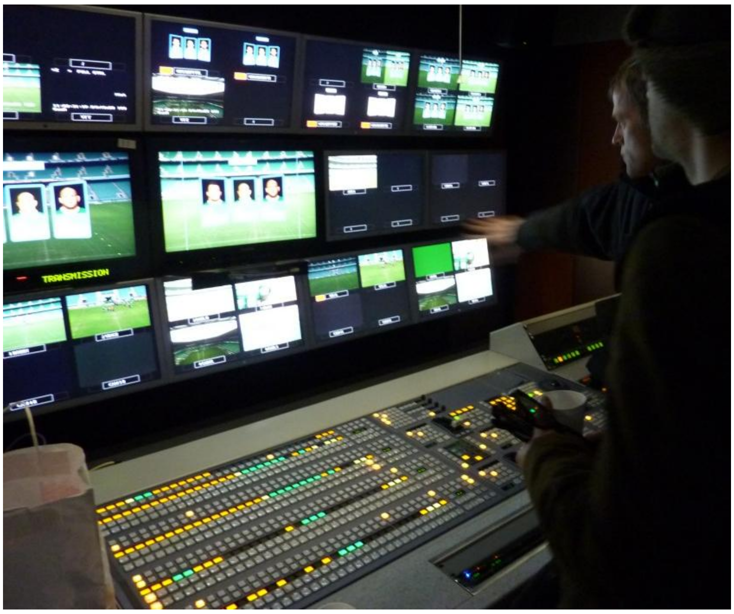


Figure 9: Production gallery and VM console

A VM is made up of different modules, depending of its size and the scope of its use. The ones described here are some of the more essential. While technically not a part of the mixing console, monitors are included as well, since they are more or less essential for the use of the console.

VM modules are described next:

- **Monitors.** A monitor is a display device usually displaying either the signal of an input to the mixing console or the output of a bus. It can be either a hardware TV or monitor dedicated to one source, or a virtual monitor on a larger display device. A common setup is to have a monitor showing each of the inputs to the video mixer and two monitors showing the output of the program and preview bus respectively.
- **Bus.** A bus is basically a switch with multiple inputs and a single output. Some buses allow for multiple inputs to be combined, whilst some only allow one input to be switched through at the time. In a video mixer, a bus generally takes the form of a row of buttons, each representing an input such as a camera feed. By pressing the button of a certain input on the bus, that input is switched to the output of the bus. Usually the buttons of the bus have a tally light so as to indicate which input is active.
- **Programme bus.** The purpose of the programme bus is to select which input should be sent to the main output of the mixer, i.e. the feed that should be broadcast. The inputs of the programme bus typically consist of live camera feeds and other video resources ready for broadcast. The programme bus has a dedicated monitor logically called the programme monitor, on which the crew can see what is being broadcast at the moment.
- **Preview bus.** The preview bus works in the same way as the program bus, with one exception. The output is sent to a dedicated monitor (the “preview monitor”) and is not broadcasted.
- **Key bus.** The purpose of the key bus is to be able to add different graphical effects to the output of the mixer, e.g. text or company logos.
- **Transition module.** The transition module controls how the transition between the output of the preview bus and the program bus is performed. Firstly the type of transition is selected. The next step is to initiate the transition. This could be done automatically by pressing a button (the “auto trans” button, which performs the transition in a pre-set time interval, or by controlling it manually by using a slider bar. When a transition is made, the source that was active on the preview bus is now active on the program bus and vice versa.

Usage. There are a number of ways to use the VM. When performing cuts between sources, the VM could press the buttons on the program bus directly or selecting them on the preview bus and using the cut button on the transition module. If a transition other than a cut is desired, the VM selects the source on the preview bus, then selects what type of transition is desired and if some graphical effect should be used in it on the transition module, and uses either the auto trans button or the fader bar to perform the transition.

For the FascinatE project, the most important aspect of the operation of a VM is that the system is closed, and the data about cuts and so on are not readily available. We need some way to extract this data and use it to drive scripts.

Replay machine

Instant replay involves replaying video footage of events very soon after they have occurred, usually in breaks from the event's 'action'. Instant replay is an important component of contemporary 'live' television. It supports visual interpretations of real-time broadcast events by showing how event-critical incidents have unfolded, and is a common element of live TV that viewers expect to have. This recorded footage needs to be cut into the live footage so that it does not disrupt the on-going action. The key design issue for this area in FascinatE lies in users being able to select and search for relevant and topical content to make sense of the action; it is likely that this aspect will also contribute to the users' experience of televisual 'liveness'.

The ability to create instant replay material in the production of contemporary television relies on the use of non-linear (tapeless) media, which allows 'random access' to stored video footage. Video and audio material is captured to a storage device, which allows recorded footage to be searched, segmented, re-sequenced and played back. In live sport that involves the use of multi-camera recordings, these systems allow programme editors to cut into the live broadcast to show recorded footage from cameras that were not initially selected for broadcast, allowing the use of multiple angles on action taking place during the game and at different playback speeds. The role of the instant replay operator (RO) is to act as an editor, assessing and selecting sequences very rapidly as soon as they occur to create material that can be cut into the live footage when possible or appropriate. These operators are not just technical operators, skilled at working with the video to produce content when requested – they need to be highly attentive to the developing game in producing relevant and timely footage. The RO controls a specialised tapeless recording machine (usually an "EVS" as explained above). Highlight "clips" are stored on a computer hard drive for quick recall and maximum flexibility. Highlights can be grouped together into playlists for replay packages supporting storylines or specific players during a game. These recording machines frequently take in two cameras at a time, constantly recording, and can output two channels at a time. Highlights received while playback occurs can be clipped out of the buffer and added to an ever-expanding library of clips. Networked together, two EVS machines can share and playback each other's clips and playlists. More basic video recorders can only clip and playback highlights, but cannot create playlists [SportsTV, 2010].

In live sports production, the major part of the production studio is taken up by workstations for the VM, the producer, the script and the graphics operator, all facing a video gallery. This video gallery displays all the visual resources the VM has at hand; manned and unmanned cameras placed around the arena, two monitors showing the RO's work and one display for graphics overlays. Usually close to the video gallery is the RO's workstation. The VM and the RO can communicate verbally and hear the commentators on loudspeakers inside the studio. The VM is directly audible to the commentators via an intercom headset, while the RO can speak back to the commentators by pressing a button to activate the intercom. In all, a large team (typically 15-35 people), including camera operators, sound and image engineers collaborate to produce the live broadcast.

Events in the material can be accessed instantly as they occur, and individual sequences can be edited into playlists to provide multiple camera angles on a situation. At this point, and in the same way as with the live cameras, these replay image sequences can be selected and cut into the broadcast feed by the VM. The RO's work involves the continuous identification of potentially interesting situations in the game. When such a situation takes place, they typically examine the footage to examine which camera captured a suitable view of the situation by rewinding the video that had just been stored on the server. They would then select one (or more) video streams that showed this situation. On locating this they will set an 'in-point' to the selected feed and then typically wait for directions from the VM. If the VM, who relies on the RO to have done just this, calls for footage, the RO prepares to roll the sequence on command. If no such call is made, the sequence may be stored in a video bank for later use.

Visually, the replay unit drives a monitor showing multiple camera feeds. This setup records multiple live camera feeds continuously throughout the game, and enables the operator to go back in time to any of their camera feeds, search within the video and edit short sequences to be replayed. Typically, the following key functionalities and their corresponding interface controls are available to the RO: 1) a camera selection interface allowing the operator to select from multiple live or recorded camera feeds;

2) video jog wheel used for searching within the stored video; 3) a playback control lever (controlling playback speeds); and 4) a video bank for storing clips for later access, individually or as playlists.

A number of techniques are utilised by the RO in creating timely and meaningful replay footage:

- *Temporal coordination through media threading*: at the same time as searching through logged data, the RO listens to the on-going audio commentary, using this as a resource to check the live video feeds on occasions where they talk about possible replayable topics.
- *Tracing historical references backwards in time*: the RO can use the live camera's image of the current visual action as indicators of the actions that had occurred previously in the game, and gradually doing 'detective work' by going backwards and forwards in the recorded footage to help make sense of the logged media.
- *Distributed and parallel search*: by allowing the others in the team to know the RO is undertaking a search, the production team can simultaneously search and make sense of important events for replay, and thus cover more visual material in the brief time available.
- *Synchronising production with game time*: Replay production is oriented towards game time in that it allows the production team to fill gaps in game play. The intermittent structure of game time, and especially the pauses in play, provides opportunities to focus more on editing and less on the live action. This is because it is unlikely that any new game action will emerge that is appropriate to use for replays during this time.
- *Narrative formats supporting replay production*: The live feed of video provided by the camera operators during game intermissions is helpful for the RO, even though he may not use this material in the edited version. It is useful because the narrative format changes outside of game time. At this point, the camera operators switch from following the action to showing what had happened. This switch in narrative formats has two consequences for the replay producer. First, it provides the RO with time to search and edit their material. Second, it provides them with a bridge between the narration of the game in between the actual situation and the replay of it, allowing replay material to be meaningfully inserted into the live footage.

Users expect to have access to instant replay in near real-time. Accessing this replay information quickly is not easy, and professional instant ROs replay on a number of mechanisms to do this work. These include using the on-going commentary and the use of current live images in interpreting what had happened prior to this event, as well as distributed/social search, and the use of game time and narrative formats when inserting replay content into real-time content. Professional operators use multiple screens to do this work, and whilst we might not expect amateurs to have access to a gallery of screens, they might use secondary screens (e.g. laptops, iPads, touch screen phones) to review visual content before displaying on their primary screens, or simply to use these secondary screens for viewing instant replay sequences. The OMNICAM in FascinatE offers a unique benefit to instant replay – close up footage from a number of zoom levels can be accessed relatively easily, either professionally or by viewers. What is clear is that FascinatE without a form of instant replay will provide a very impoverished form of live TV experience.

Segments of visual material viewed by others might be tagged, in a similar way to Amazon.com ("recommender system"), i.e. "people who looked at this segment with your personalised interests, also looked at this instant replay footage". There might also be backchannels designed for this that would allow topic-specific discussions between users when trying to make sense of which moments in time and camera angles would be best used in instant replay. It would also be useful to make known pauses in the game time more visible to users so that they could make use of this time to do their instant replay production. Similarly, dealing with replay production could be a tricky problem when the commentators are viewing a different zoom and angle to the viewers – so finding a way to couple these more formally might be a useful design goal. Given that scripts may be used to generate framing, these might also be used retrospectively by viewers to automatically cut to a different perspective to show significant game features that have just occurred. Some work on such automatic event recognition in sport has been published already [Wang, 2004].

Studies of instant replay production will continue during the FascinatE project, looking at this in different forms of live TV, and seeing how it is used the production process, both to support individuals in producing and accessing instant replay material.

Audio

Audio capture technology is not as involved as video capture. It normally does not require as many people to operate. A number of microphones will be placed around an event. A few may be manned, but most are unmanned and fixed in place. Some processing happens to the signals arriving from these microphones. Some processing is fixed before an event, other aspects are changed dynamically as the event changes. As a general rule, as little as possible is changed in the audio once an event in running.

The microphone signals are passed into a mixing desk where the engineer adjusts the relative levels, amongst other things. Like the VM, audio mixers do not all allow the export of data about the settings of their controls. This will provide a challenge to the FascinatE system.

The channels which are at times used for broadcasting in a conventional system are still used when they are off air. Audio channels are often used for communications between physically separate sites. The FascinatE system must take care to enable these functions to continue, whilst not allowing them to be transmitted to the viewer.

4.4 New Production Roles under FascinatE

Under the FascinatE system the roles of the various people outlined above will change. We use the Production use cases outlined in Section 2.2.2 to illustrate the departures from current methods which the FascinatE system will allow, or require. The three use cases are:

- *Use case 5* – A FascinatE system is used to create a conventional linear TV programme. The OMNICAM, and the various automatic scripting abilities of the FascinatE system are used by the professional production staff to create more informative replays and more exciting views of the game. The viewers at home tube to one channel, and have no interactive control.
- *Use case 6* – A linear TV programme is created by the FascinatE production gallery. Instead of being delivered to the home as a single video stream, it is instead delivered as a LSR and a variety of scripts. The user can choose either to watch the directed content, or can navigate freely around the content which is available. The user has little guidance as to what might be interesting to watch, away from the main programme. Content is lightly curated so that unsuitable material is not transmitted.
- *Use case 7* – The production staff put most of their efforts into creating scripts. They rely on the auto-assisted VM to take care of the basics of tracking the game whilst no incidents are happening. The staff spends most of their time creating replay scripts, and supervising the automatic generation of live scripts. The user has a choice of a wide variety of “curated content” which they can choose to view as the match progresses.

We now consider how existing roles in production might change under these different use-cases.

Director under *Use case 5*

Under this use case the output of the production team is a linear TV programme. Of all the production staff it is the director who is most focussed on the output of the team. Therefore his job remains very similar. He will have additional tools to work with. The tools developed by the FascinatE project must integrate with existing production hardware if we are to realise this use case, or one like it.

In this use case, the director could benefit from an OMNICAM view. This was highlighted in discussions with production staff. In sports production the team will usually sit in a scanner (a lorry) watching the input from the cameras. These cameras do not present a coherent picture of what is going on at an event. It is only through skill and experience that directors can translate this view into a coherent understanding of what is happening on the pitch. An indication on the OMNICAM view of the location of cameras and other key parts of the production was also thought to be useful.

Director under *Use case 6*

Under this use case, the role of the director is similar. The main function of the production team is still to produce a high-quality programme, and the director is still in charge of this team. As a live programme is being made, the director will perform more or less the same job, with a few of the enhanced tools at his disposal as described in the previous section. Still, the director will have some influence over which regions of interest should be targeted with scripts providing alternate views.

Director under *Use case 7*

Here the situation for the director is very different. In this use case we have no “main” programme, which in the other use cases has the director’s entire attention. In this case the job of the director will include a more extensive “off-line” component, discussing their expectations with staff before a programme begins. The director might use the outputs of the various scripting systems to create a preferred view. This might provide a similar experience to watching the directed version of the show, as in use cases 5 and 6. They might also provide the final decision about which cameras are transmitted.

In the earlier section describing the role of the director, their role was described as “The person with responsibility for realising their and the producers vision for the programme”. They will maintain this role under this FascinatE production system, but exactly what that will mean is not clear. Perhaps in a real sports game there will always be more regions of interest than can be tracked? Then the director would be responsible for deciding which groups of areas of interest should be the focus of the coverage.

Vision Mixer

The three use-cases present different levels of control and automation over the VM’s control over how individual camera feeds are assembled into a live broadcast. These range from absolute control in *Use case 5* to very little control in *Use case 7*.

Vision Mixer under *Use case 5*

Here, the role of the VM is largely likely to remain unchanged in terms of the image-related expectations from them, although the methods by which they interact with camera content (virtual and live camera inputs) are likely to differ substantially from the ways that they currently perform their work. There are a number of different ways of interaction (and levels of control) that can be anticipated, from one in which all cameras, including virtual and live camera inputs, are fed into an existing gallery and operated on through a conventional VM, to one in which they have direct control of the OMNICAM output.

Assuming that the mixer has access to virtual cameras well as remote cameras, they will be able to select complementary footage from either real or virtual cameras of the same event. They will also be able to script (or request scripts from others) dynamically, and access existing scripts to select automated following of particular forms of action by the virtual cameras. It is unlikely that scripted footage will be automatically selected for transmission, but that this will be ‘flagged’ as potentially interesting footage for transmission, perhaps by placing footage deemed to be relevant inside a ‘preview’ screen in the image gallery. They will perform cuts between live and replay footage and between cameras as before. Where commentators are involved, commentators will see the same footage as all of the viewers (in addition to any additional information) and will comment on the footage that viewers can see. They may make comments that affect the shots selected for transmission by the VM.

Vision Mixer under *Use case 6*

As for the director, the VM is operating as normal in *Use case 6*. The high-quality main programme must still be made, and managing the VM console will take up most of this operator’s time. Depending upon the production, the VM might also be responsible for checking the suitability of feeds from other cameras for transmission. This must be done by someone, to ensure that the end user only has access to content that they might wish the viewer to select between. All camera operators will spend some time capturing material which is not suitable for transmission, for whatever reason.

Vision Mixer under *Use case 7*

In *Use case 7*, the VM is essentially the end user/s. The users will access the footage over a mobile terminal, FascinatE enabled TV set, or public screen. They may be accessing a broadcast feed either from the production gallery, or may be able to select different views into the game from different cameras (real, virtual or OMNICAM). Of all of the use cases, the FascinatE enabled TV set offers the most flexibility, and it is conceivable that the users might be able to operate a sophisticated remote control that replicated the functionality of the Vision Mix operator. It is also envisaged that mobile terminals are likely to be used both independently and in concert with other viewing media. This need not mean that they are controlled differently in both scenarios, but that they allow users/mixers to access different content streams.

Scripting is likely to be very difficult in this third use case, but users/mixers may be able to set up preconfigured user settings (e.g. ‘prefer the red team’, ‘select views that show the goal). One way to enhance scripting with information gathered from viewers would be to use a ‘like’ option (accessed by a button on a mobile terminal, the remote control on a TV set, and perhaps by volume on a large screen):

this might allow users to be more likely to see similar types of footage in the future, to tag footage and to see previous tagged sequences that other people also ‘liked’, or to see the game from the same viewpoint as people who also ‘liked’ particular settings.

The small amount of screen real-estate on mobile terminals and the low ability to control multiparticipant viewing at public screens means that the ‘default’ view is likely to be one most commonly viewed on these kinds of display. Other shot types selected are likely to be deviations from this, and selected from a much smaller, and edited, set of alternatives.

In the production gallery we have a fairly homogeneous group of people, all of whom share some of the skills of a director, VM, and RO. These operators would be generating scripts by creating packages of replay content to illustrate certain interesting events, or by following certain regions of interest. It might be possible to have these operators supervising semi-automated coverage of the game whilst no incidents are happening.

The three scenarios present different distribution of the control over how individual camera feeds are assembled into a programme. Aside from this distribution of mixing control, the change in roles for camera operators is largely dependent on the level of integration with the regular production workflow.

Camera Operator under *Use case 5*

In the first use case, FascinatE is integrated into a standard production workflow. Framing selections from the LSR are fed into the video gallery as virtual cameras, alongside manned cameras, replay footage etc. In this case, camera operators at the live event will

- maintain their working roles, providing live coverage. They will continue to choose shots based upon a combination of their own experience and the director’s instructions. They are complemented or partially replaced by remote FascinatE camera operators.
- likely need to acquire skills of combining their footage with scripted virtual cameras, in order to provide enhanced resolution images for key areas of interest. This depends on reliable real-time registration between images from manned and virtual cameras. Support for this also need to be built in the system, e.g. in the form of communication channels for requesting framing by manned cameras, or by adjusting FascinatE scripts to fit the nearest available manned camera.

Camera Operator under *Use case 6*

In the case where a FascinatE gallery produces one main output, but offers a number of alternative views:

- one or more FascinatE operators would be offering footage to a VM or master operator. In smaller, low budget productions these roles could be merged into a single master operator performing the tasks of controlling a set of scripted virtual cameras and mixing their output into an edited program.
- manned camera operators could work individually to provide additional viewpoints and close-up footage, feeding into the main FascinatE unit.
- manned cameras could also provide the master operator with complementary high resolution feeds to support virtual cameras, as described above.
- Operators might have to indicate whether they had a shot worth seeing or not, to decide whether the view from a particular camera is included in the optional content.

Remote/virtual camera operators could be working one or more virtual cameras within FascinatE. Operating one camera, in the most manual setup, could be similar to traditional camerawork, using e.g. a remote camera control interface to manually pan, tilt and zoom within the interface. But input devices could also include any combination of touch, gesture and physical interfaces. Operating several virtual cameras would most likely be more of a monitoring role; attending to multiple semi-auto scripts, manipulating them dynamically when needed, and offering them up to the VM or master operator.

Camera Operator under *Use case 7*

In *Use case 7*, the virtual camera operators would be producing video feeds in a similar manner to the use-cases above, but the fact that the feeds are produced with a scripting engine and a more distributed production chain may have some implications. The direction of the production can be separate from the VM or master operator role on location. There may be a designated director with no responsibility for mixing. In a setting where no director is present, camera operation tasks may instead be negotiated beforehand in greater detail, or a communication backchannel could support requests to the camera operators from the scripting engine, either automatically or through remote script operators.

Manned cameras could also be generating metadata for the scripting engine to use. This would not necessarily affect the role of the camera operators themselves, but would increase demands on real-time image recognition and registration between individual cameras and the system. Aligning manned cameras with virtual cameras to provide greater detail would most likely have to be done through automatically adjusting virtual cameras to manned ones, rather than through requests, as the mixing and assembly of images becomes more distributed and automated.

In all three use cases, the relationship and ratio between manned and virtual camera operators depends to some extent to how we see them working collaboratively to produce close-up footage of events, on the technical limitations in doing this within the panoramic image alone, and on the technology for juxtaposing manned and virtual cameras in real time to produce detailed shots. Early input from producers also indicated that they may depend on the scale of production. Producers saw great potential in using FascinatE to enable low-end productions with primarily remote camerawork, but they were hesitant to replace close-up camerawork in larger productions (up to 30 manned cameras) with footage they feared would be flatter and less detailed due to the optical differences between a fixed OMNICAM and a single manned camera.

4.4.1 Audio Engineer

As the number of video programmes output by the studio increases, so the number of audio mixes required increases. Automated mixing will require some degree of human support and interaction. It is likely that more audio engineers will be required to provide a sufficient number of different mixes.

Audio Engineer under *Use case 5*

As with the other roles, the audio engineer’s job changes very little in *Use case 5*. The output linear programme is the audio engineer’s main focus. As long as the production gallery is just making one TV programme, then the audio engineer will mix audio appropriate to that programme.

As with video, some of the automated or semi-automated tools which the FascinatE projects develop might either provide the audio engineer with more tools which they can use to create content, or might also alleviate some more straightforward tasks.

Audio Engineer under *Use case 6*

Use case 6 introduces metadata and scripts. The FascinatE studio must pass scripts to the user. In the same way that the VM creates scripts instead of a programme, so the main audio mix is passed down the transmission chain as scripts containing mixing parameters. Information about groups and processing is passed down too. From a technical point of view, this implies automated audio mixing at the user end based on control parameters for the mixing desk. Extracting these parameters will require a carefully designed API that must be implemented by a desk in order for it to be FascinatE compatible. A mixing desk would be required to record all its operational parameters and output them, as well as responding to control from outside.

Use case 6 also includes some alternative views that a user could select. These views will likely require audio mixes. It was highlighted as part of our discussions with Production staff that it is very difficult to get an audio mix right. So we expect that this use case would require more audio engineers, to mix, or at least supervise the automated mixing, of audio to accompany the various video feeds. Alternatively, we could provide a single audio mix for the whole programme. Some users will listen to the commentary of sports on the radio, whilst watching the TV pictures with no sound. We might seek to create a more interactive version of this experience.

Audio Engineer under *Use case 7*

Sound is mixed in an automated way based on scripts and metadata in *Use case 7*. Those scripts and metadata have to come from somewhere! At the moment, automatic matching of sounds to sources is at or beyond the state of the art. So we expect that some people will be involved in generating the metadata and scripts of which audio should be associated with which video pictures, and how they should be mixed. There are two candidate approaches; using staff to create semantic metadata about the subject of a microphone signal, where it is and so on, or using staff to create audio mixes suitable to the video streams being transmitted. In discussions with production staff it was pointed out that mixing correctly is very difficult to do right, and is easy to spot if it is wrong. This implies that fully-automated intelligent mixing based on semantic metadata is a bad idea. Instead, we should use operators to create mixing metadata. These metadata can provide a basic mix which we might then tweak in the renderer.

The attention of an experienced audio engineer is taken up completely with mixing the right audio for a scene, in applications where audio mixing is done one the fly, such as in sports.

In our curated content environment, it only makes sense to mix audio for video that will be leaving the studio. There might also be several views of the same subject which would require similar audio mixes. Metadata associated with the pictures should be sufficient to do this kind of loose grouping. But what happens when we have views of the same scene from an opposite viewpoint?

In discussion with production staff it was suggested that one audio engineer might be able to handle the adjustment of more than one audio stream associated with more than one set of pictures. Still, for one engineer to look after more than a few would be very difficult.

The basic situation we envision is to have a handful of audio engineers each providing rough mixes for a small number of different video feeds. Exactly what is automated and what must be done by humans will be defined by this project as we explore that area.

4.4.2 Replay operator

Replay operator under *Use case 5*

Under *Use case 5*, we can imagine two different ways that the RO’s job could go. Either they see the usual set of cameras, and have fixed captured video to work with, or they have complete access to the data from the OMNICAM, including high-resolution sections as captured by the OMNICAMs associated cameras.

In the first of these examples, the job of the RO is more or less the same. They will be watching their two feeds, and when incidents occur, they will be prepared to play back some of their shots should the director wish it. The view from the OMNICAM would probably help with the ROs comprehension of what was going on at a sports event. So they should be in a position where they can see the OMNICAM feed(s).

In the second (and more interesting) of these scenarios, we imagine that the OMNICAM data is available in full to the RO. This presents a significant departure from the existing situation. It allows the RO to compose their own camera shot based on the data coming from the OMNICAM, and in the associated LSR. For example, if one of the camera operators has missed part of a developing incident, the RO can fill in the necessary footage from the OMNICAM.

One way this could be made to work would be for the RO to see the registered camera views overlaid on the OMNICAM, then to be able to adjust the motion of the view so that the interesting incident was not missed.

Alternatively, we could allow the RO free access to the OMNICAM data to compose their own shots. This would be useful where the conventional cameras have missed an incident altogether. This would be useful in situations where there are not enough cameras to cover a large area at once. For instance, in motor racing, there is often only one camera at a corner which covers both the entrance and exit to that corner. The same area covered by an OMNICAM would be able to view both the entrance and exit at once.

Of course, it is likely that the view of the scene from the OMNICAM is at a lower resolution than the view from the camera it is replacing. It remains to be seen whether this will provide a service which is compelling to the user.

Replay operator under *Use case 6*

The same two possibilities exist within this use case as described under *Use case 5*, above. So from the point of view of making a basic programme, the RO must still perform the same role. That is, they must still provide replays on demand to the director. However, these replays will be delivered to the end user as metadata or FascinatE scripts. This immediately opens up the opportunity for the end user to watch curated content in the form of the directed replays whenever they like (assuming some video/audio caching either at the user terminal or at the edge of the network.) We can expand on this idea to consider the possibility that the ROs as a team could deliver even more replays and reviews of interesting parts of the event. These could then be delivered as FascinatE scripts to the user terminal, and form part an even richer selection of curated content.

However, in the earlier parts of this section on production requirements we have frequently referred to the need to create scripts and metadata about what can be seen in the various views on a scene. This is where the ROs job changes. Instead of playing back replays at the request of the director, the replays could be generated more or less continually. These are then passed into the FascinatE system as scripts which can be picked up either by the director or by an interactive home user.

Replay operator in *Use case 7*

Under *Use case 7* the whole production system is very different. As discussed above, most of the key roles change significantly compared to conventional TV production. The role of the RO changes the most. In a conventional TV production, the job of the RO is to extract sections of video from the recent past relevant to an incident or event. The purpose of these sections of video is to help the viewer's understanding of what went on during that event.

It is important to separate out the RO's job from the purpose of what he achieves. The purpose of the replay is to improve the viewer's understanding of some event that happened too fast to be understood in a single viewing, or that needs to be seen from several different viewpoints.. This is achieved by replaying relevant clips of that event. One of the goals of the FascinatE project is to create a system that is both clever enough, and contains enough metadata, that a replay could happen in an automatic (or nearly automatic) way. Creating the clever system is not part of the production requirements. Generating the scene information and metadata which enables that clever system to direct automatic replays is the responsibility of the production side. It should also be pointed out here that we anticipate FascinatE to be able to do far more than automated replays with the scene information and metadata which is transmitted with the audio and video,

In order to achieve this goal of a clever system capable of automatically generated replays it is critical to generate as much information about the content of all the different video feeds as possible. The RO (or someone with a similar set of skills) is likely to be the main source of manually generated information about footage being captured. Instead of operating an EVS replay machine, an operator would work at a terminal which is used for capturing metadata about what is going on in a given camera shot. This is likely to be semi-automated. In order to implement the features which we have described elsewhere in this document, we need to generate metadata about:

- which people and events are in which shots;
- which cameras have the best view on a particular person or event;
- what the main focus of a particular shot is.

This list is far from exhaustive. As the design of the scripting and rendering engines progresses we will find out what kinds and quantities of data are necessary to make high quality script-based footage. This in turn will inform what these operators have to do under *Use case 7*.

4.5 List of Production Requirements

This section summarises the production requirements of FascinatE. As was mentioned in the introduction, some of these requirements may change as the project evolves. Some of them refer to specific applications, which may or may not be built.

Hardware:

- Whatever video, audio and communications systems FascinatE uses, it must interface with existing broadcast systems through standard interfaces.
- Where a specific FascinatE production interface is not being used, an easy way of extracting data on camera selection is required from the video mixer as well as the individual microphone feeds from the audio mixer.
- New hardware developed for FascinatE must take up a comparable amount of room to existing broadcast kit within outside broadcast vehicles.

Control:

- A means should be provided to select a view or views from the OMNICAM and present it to the VM as a standard video signal.
- A means should be provided to give the director some influence over which regions of interest should be targeted with scripts providing alternate views.
- A means should be provided to ensure that the end user only has access to the content that the director might wish the viewer to select between.

- A means should be provided to generate scripts by creating packages of replay content to illustrate certain interesting events, or by following certain regions of interest. It might be possible to have these operators supervising semi-automated coverage of the game whilst no incidents are happening. This is likely to require the following:
    - A means should be provided to recall parts of the OMNICAM view from the recent past, so as to allow replay metadata to be added.
- A means should be provided to display the whole view from the OMNICAM or OMNICAMs from the production gallery. In addition, certain useful metadata such as camera and microphone positions should feature on this view.
- A means should be provided to enable (basic) single-person coverage of a live event, such as a lower-league football match.
- A means should be provided to enable the director alone to create a “Preferred view” in *Use case 7*, using the output of the semi-automated scripting system.
- A means should be provided for the FascinatE operator(s) to indicate which of the views which are available in the panorama are to be used for any given virtual camera shot. However the renderer is not compelled to abide by his choice.
- A means must be provided for the production staff to add metadata and information to video and audio content in real time (such as the name of a football player being followed). This method may be automated or semi-automated, supervised or unsupervised. Some aspects of this could make use of existing approaches or practices in the production of metadata for clips, e.g. in taking information such as names of players that are superimposed as graphics or captions. However, such metadata cannot easily be extracted from current production tools, so for the purposes of demonstrating a FascinatE prototype it may be easier to provide a facility to enter such data directly.
- A means should be provided for reviewing the end-user view which the system produces on a few candidate devices from the gallery.

5 Networking Perspective

The delivery network takes a central part within the FascinatE environment. In general, the delivery network allows for distribution of content originating from the production domain towards end-user terminals. Also, the delivery network may provide an interaction channel that allows users to give input for e.g. selection and navigation purposes.

This chapter considers the FascinatE requirements from a networking perspective. For each of the three scenarios described in Section 2, requiments appropriate for the network role are discussed in Section 5.2. Additionally, high-level network functionality is described in Section 5.5.

5.1 Network Capacity Today

5.1.1 *Fixed delivery networks*

There has been a strong increase over the last years in broadband penetration in Europe. Broadband access means access via a ADSL, Cable or FTTx network where the bandwidth can vary from 256kbps to over 30Mbps. In Figure 10 is shown the penetration rate in the 27 European countries. The weighted average is around 26%.

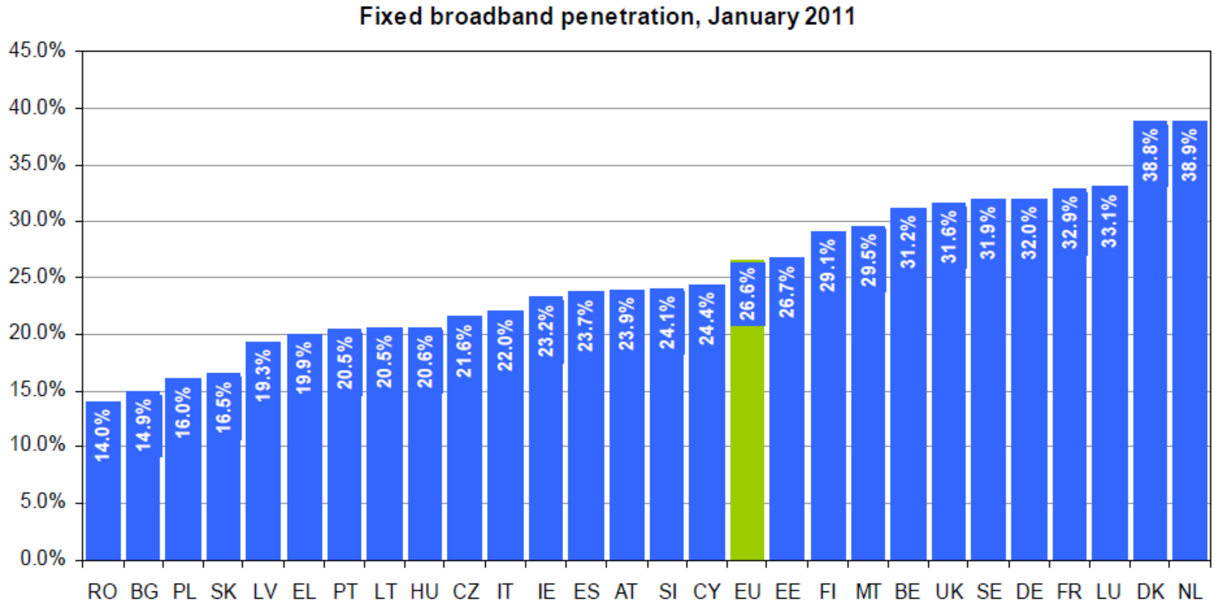


Figure 10: Broadband access penetration rate in the EU [EU-DAS, 2011]

The average download speed of broadband subscriptions in the EU has greatly improved between 2004 and 2011. At the end of 2011 85% of EU broadband subscriptions are estimated to be associated to nominal speeds above 2 MB/s. [EU-DAS, 2011]. Today about two percent does have FTTH access.

The European Commission has stated that in 2020 broadband access of over 30Mbps must be available for all EU citizens; half of them should have access to broadband links over 100Mbps in that year.

5.1.2 *Mobile delivery networks*

Much technical development is taking place in the area of mobile access links with new technologies as LTE. An overview of these technologies is given in Table 2.

Acces type	Access network technology	Theoretical limit	Practical (download speed users can get now)	Expected in 3-5 years
Unicast	HSDPA/HSUPA (usually abbreviated together as HSPA)	*7,2 Mbit* 14.4 Mbit/s DL 5,76 Mbit/s UL	0,7-1.4 Mbit/s (DL) 0.28-0,7 Mbit/s (UL)	1-3 Mbit/s
Unicast	UMTS	384 Kbit/s	220-320 Kbit/s	See HSPA
Unicast	Wifi (801.11G)	54 Mbit/s	10 - 25 Mbit/s	10 - 25 Mbit/s
Unicast	Wifi (801.11N)	150 Mbit/s / 300 Mbit/s	20 - 80 Mbit/s	20-80Mbit/s
Unicast	WiMax	70 Mbit/s	1-2 Mbit/s	<20Mbps
Unicast	LTE	326.4 Mbit/s (for 4x4 antennas)	1-5 Mbit/s	<20 Mbps
Unicast	LTE Advanced	Up to 1 Gbit/s	In development	
Unicast	802.16M	Up to 1 Gbit/s	In development	
Broadcast	DVB-T	24 Mbit/s (total)	24 Mbit/s (total)	
Broadcast	DVB-T2	40 Mbit/s (total)	40 Mbit/s (total)	

Table 2: Overview of the technologies involved in mobile access links

3G/HSPA networks have wide coverage today in Europe. In some countries LTE networks have been rolled out now however it is unsure on what speed LTE subscriber penetration will take place in Europe.

Also it is not yet clear what speeds can be reached with the new LTE technologies. It has yet to be determined what download and upload speeds can be expected in normal real life circumstances. In Figure 10 is shown the mobile broadband penetration rate in the 27 European countries. The weighted average is around 26%.

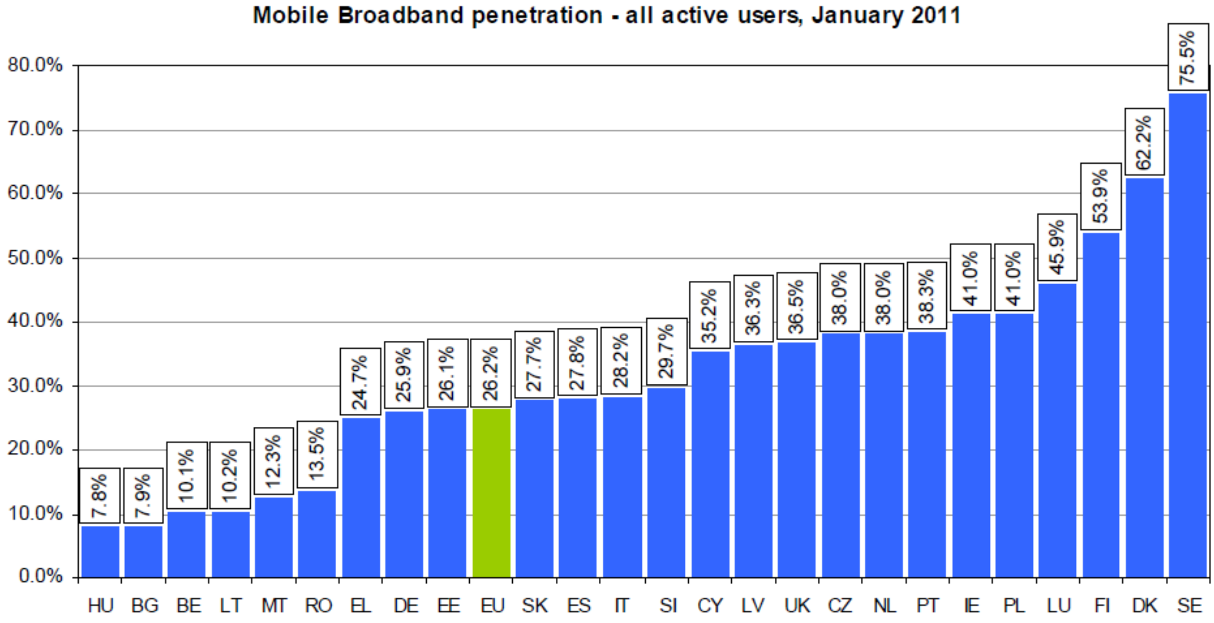


Figure 11: Mobile broadband penetration in EU [EU-DAS, 2011]

5.1.3 Core and contribution networks

Production network can make use of (bundled) 1Gbps and 10Gbps fibre access technologies. New standards for speeds of 100Gbps are in development. Availability depends on the local facilities.

However over the last few years on most (sport) event and studio locations has been invested in extensive fibre networks.

5.2 Delivery Network Requirements

The functional, system and hardware requirements on the delivery network include the following categories:

- **Bandwidth**

  The services provided by FascinatE require a significant increase in bandwidth compared to existing TV or video services. Bandwidth requirements are largely related to the distribution of the video signals, in either production format, intermediate format or a specific pre-rendered view. Other data flows that are transmitted over the delivery network are audio signals, scripting metadata and the user input and commands.

- **Latency**

  Latency covers the delays that are introduced in the FascinatE delivery network. Latency can be an important constraint for live video services, e.g. a soccer match and services involving (user) interactivity. Four types of latency can be distinguished:

  - End-to-end service delay. This relates to the time difference between the recording of audio and video at the recording location and the presentation of these signals to the user by the terminal. For live events the allowed delay may be in the order of seconds, while the allowed delay for on-site terminals may require real-time behaviour, e.g. when offering a mobile view to visitors during a music concert. The allowed end-to-end service delay is scenario and use case dependent.

    Note: the end-to-end service delay is the sum of the time for session setup, T_{SS}, and the delivery network latency, T_{DNL}:

    $T_{SS} + T_{DNL}$

  - Session setup delay. This delay is related to the time between a user requesting a media stream and the presentation of the content on a screen or through speakers. For certain delivery modes (i.e. unicast and multicast delivery), a terminal must request the A/V media before transmission of the media starts, introducing session setup delay. For broadcast delivery modes the media streams will already be transmitted to the terminal when a user requests a media stream, so the time between user input and content presentation is minimal.

  - Delivery network latency. This relates to the amount of delay the delivery network introduces for e.g. the transport, routing, conversion and rendering of media streams that are provided by the production domain and delivered to a terminal. In other words, the time difference between the ingestion of A/V data into the delivery network and the reception of this data by the terminal.

  - Responsiveness. This relates to the time between a command input and the generation of the result of that command. For interactive scenarios, i.e. where an end-user controls what is being displayed via user commands, responsiveness will be a dominant factor. Acceptable delay values depend on the scenario, use case, but possibly also on the type of Terminal. Acceptable responsiveness values are to be determined during the course of this project.

- **Formats and codecs**

  These requirements relate to the information being transmitted by the delivery network. For the production domain, the requirements for formats and codecs are specified in D2.1.1. The appropriate formats and codecs for delivery are for further study. The requirements for scripting formats are specified in Section 5.5.

- **Type of transport**

  The types of transport relate to delivery modes and delivery types supported by the delivery network. The delivery modes are: unicast, broadcast and multicast. The delivery types are: unidirectional, bidirectional, unidirectional with separate feedback channel.

- **Interactivity**

Relates to end-user interactivity with the FascinatE system, where interactivity is defined as user commands or terminal responses.

- **Processing**

  Some use cases, particularly in the service provider-centric scope, put high requirements on the processing resources which are expected in the delivery network. FascinatE considers new functionality to cope with these requirements. Some functions are natural evolutions of the ones that can be found today (packetization, filtering, routing mechanisms) while others can be seen as more disruptive, for instance functions that perform A/V processing (e.g. A/V coding, rendering). Such functional requirements and functions are treated in more details in Section 5.5.

- **Service and deployments requirements**

  These requirements relate to service deployment in an operation situation. E.g. Service Discovery and Selection (SD&S) information, subscription and billing information, Quality of Service (QoS) and so on. These requirements are out of scope of this document.

5.3 Requirements for Production and Terminal-centric (Scenarios 1 and 2) Use Cases

5.3.1 *Requirements for scenario 1*

Scenario 1, and the associated *Use case 1*, resembles the delivery of TV channels or video streams in current delivery networks. However, instead of offering one TV channel per TV station (i.e. where one view of the content is being offered), multiple views of the same content are generated and delivered to the terminal. These views are determined at the production stage. User interactivity is limited to session setup (if and when required by the underlying network type) and the selection of one of the available views by the end-user. From a delivery network perspective this means that current delivery networks can be used, as long as they meet the bandwidth and latency requirements and provide the appropriate functionality for service discovery and selection. Specifically, the family of non-IP DVB networks, IPTV networks, optical fiber and GPON networks and several mobile variants may be used in this scenario.

In Scenario 1, up to two types of data are transmitted over the network:

- A/V content in the form of a (finite) set of rendered views (i.e. TV channels or video streams).
- Interaction commands for session setup and modification (i.e. channel switch / stream selection). These commands are optional for broadcast delivery modes, where interactivity is local.

5.3.2 *Requirements for scenario 2*

In scenario 2, the LSR shall be transmitted in its entirety to a terminal, where views will be rendered and presented to the end-user. The end-user has full control of what is being displayed on the terminal screen. This scenario requires an idealistic network infrastructure providing a very high bandwidth to deliver a LSR. Also, the network should not increase the response time experienced by the user. Such end-to-end requirements are not expected to be feasible for residential video services in the near future as they would require an extremely disruptive change in the network infrastructure. However, for some specific use cases, it is reasonable to design a dedicated infrastructure able to support this scenario. Within FascinatE, these requirements are relevant for the situation where the OMNICAM is directly connected to the terminal FascinatE Rendering Node, as described in *Use case 2*.

Bandwidth

As specified in D2.1.1, a LSR consists of one or more camera clusters. For a camera cluster the data rates are given in Table 3.

	HDCAM	HDRCAM	SLOMOCAM	HDR OMNICAM
Spatial resolution	1920x1080	1920x1080	1920x1080	7k x 2k
chroma sampling	4:2:2	4:2:2	4:2:2	4:2:2
dynamic range	10bit	16bit	10bit	16bit
Temporal resolution	i60	i60	i180	i60
Bitrate (Gbps)	1.2	2.0	3.7	13.4

Table 3: Raw data rates and other information on the different camera types in a camera cluster.

Based on the numbers, the total amount of bandwidth required to transmit a camera cluster is given by the following formulas, assuming a camera cluster contains one OMNICAM:

$$BW_{cameracluster} = 1 \text{ x OMNICAM} + N \text{ x HDRCAM} + O \text{ x SLOMOCAM} + P * \text{HDCAM}$$
$$= 1 \text{ x } 13.4 + N \text{ x } 2.0 + O \text{ x } 3.7 + 1.2 * \text{HD Gbit/s}$$

where N is the number of HDR cameras, O the number of SLOWMO cameras and P the number of HD cameras. This assumes that a Camera Cluster consists of exactly one OMNICAM.
The total BW requirement for an entire LSR is then given as:

$$BW_{layeredscene} = H \text{ x } BW_{cameracluster} + I \text{ x } BW_{audio} + BW_{layeredscenemetadata}$$

where H is the number of camera clusters, I is the number of audio microphones.

The delivery network shall provide a bandwidth $BW_{layeredscene}$ to provide the delivery of one LSR. Production shall support a network interface to the delivery network providing a throughput of at least $BW_{layeredscene.}$ The terminal shall support one network interface or a combination of interfaces that provide a throughput of at least $BW_{layeredscene.}$ The total bitrate of the LSR, $BW_{layeredscene}$, will essentially depend on the composition of the camera clusters covering the video scene. Assuming that a camera cluster is made of an OMNICAM and 2 to 4 additional satellite cameras (either with HDR, Slow Motion or HD capabilities), the total raw video bitrate per camera cluster is expected to be in the range of 15 to 30 Gbit/s. The number of camera clusters will highly depend on very ad hoc production requirements. For some events it is reasonable to expect that several camera clusters will be required to cover the scene, thus leading to raw video bitrates in the order of magnitudes of 100 Gbit/s.

Video compression can naturally lower these bitrate requirements. Lossless compression usually provides bitrate reduction by a factor of less than 10. Lossy video compression can lead to lower bitrate. But since compression artefacts in a camera cluster can propagate through the subsequent rendering processes, high fidelity requirements (to be later quantified in the course of the project) shall be imposed for the compression of the LSR, at least at the production side. Therefore, even if lossy compression of the LSR is allowed, the compression ratio is expected to stay below two orders of magnitudes in any case, thus still yielding a total bitrate in the order of a few Gbit/s. Deliverable 2.1.1 reports intermediate compression ratios ranging from 1 ½ to 4.

Since the audio bandwidth requirements are substantially lower than the video bandwidth requirements we have not paid further attention on this as they are of less concern.

Latency

The end-to-end service delay shall be determined by the latency originating in the terminal. In other words, the delivery network shall not or have a minimal contribution to the end-to-end latency, such that $T_{DNL}=0$.or $T_{DNL}\approx 0$.

Formats and codecs

The A/V media shall be provided by production to the delivery network in a LSR format as specified in D2.1.1. The production scripts shall be provided by production to delivery in a production script and will be specified in deliverables D1.4.1 and D1.4.2. The A/V media shall be offered by the delivery network to the terminal in a LSR format as specified in D2.1.1. The production scripts shall be offered by the delivery network to the terminal in a production script format as will be specified in deliverables D1.4.1 and D1.4.2.

Type of transport

The delivery network shall support multiple distribution technologies. It shall make use of existing distribution technologies where possible and where these networks meet the bandwidth and latency requirements as stated above. Among the supported delivery modes are: broadcast via uni-directional networks and multicast and unicast via IP-based networks. Combinations of different network technologies shall be supported, e.g. a broadcast network for media delivery combined with a broadband IP access network as a feedback channel. The capabilities of existing delivery networks are further studied in D4.2.1.

Interactivity

For delivery modes requiring a session setup, the delivery network shall at least support interactivity for session setup. All other interactivity is located in the terminal.

5.4 Requirements for Network-centric (Scenario 3) Use Cases

Scenario 3 has a focus on the delivery network. Here, the delivery network provides the necessary functionality to adapt a LSR coming from the production domain to one or more views suitable for the end-user terminal.

We detail hereafter the functional requirements for the four use cases described in section 2.2.3.

5.4.1 Requirements for Use Case 8

This use case focuses on coarse levels of interactivity, where the end user is allowed to select predefined portions of the captured content. The requirements will differ depending on the granularity of the interactive offer. Recall that we have distinguished 3 sub-use cases:

1. Predefined channels:
    a. Either directly mapped to a predefined production cameras and audio mix.
    b. Or mapped to a view that can be generated from post-processing the captured LSR content.
2. Interactive Playmaps: the user is offered more complex navigation choices where many switching points are defined over time and space. The number of possible combinations does not make it practical to offer each possible combination with its own channel identifier.

The following requirements are derived

- The creation of additional views requires some rendering functions after the capture stage.
    - In sub-case 1.b : they are best located either as part of the production process, or at the service provider's video ingest facilities (e.g. to repurpose content for a regional audience). Rendering such views closer to the end-user end (e.g. in a network proxy or in the terminal itself) is also an option, but would lead globally to a waste of processing resources, as many rendering functions in the overall system would turn out to perform the exact same operations.
    - In sub-case 2 : when evolving to a very high number of navigations scenarios, it quickly becomes impossible to render each possibility upfront, as the ingest processing and the bandwidth capabilities of TV broadcast networks cannot indefinitely scale. Although some parts of the content can be pre-render at the production or at a service-provider's ingest point, the final rendering of content is best located close to the end-user. This has the advantage of distributing the processing load for personalized rendering and minimizing the round-trip delay of the interactivity loop.
- The way these views are created from the LSR content must either be controlled manually (but such a post-production role is rather unexpected for a TV service provider) or controlled by production scripts that are ingested along with the LSR and that describe how the relevant views can be created for the various contexts (user's preferred type of views, team, etc…).
- The bandwidth capacity of the network must be dimensioned for the number of concurrent views made available. Bandwidth constraints can therefore be an element when deciding whether a service is to be deployed as in subcase 1 or subcase 2. Bandwidth limitations may also require adapting the fidelity at which the content is packaged for delivery.
- Some mapping must be performed between the available views and the channel streams. For subcase [2], the switching points can be explicitly reflected into this mapping into a playmap. In this case, the delivery mechanisms can more efficiently filter the relevant views for each group of users

5.4.2 Requirements for Use Case 9

This use case focuses on supporting the finer-grain level of interactivity foreseen in FascinatE, where an end-user can freely navigate around the content. Recall that we have distinguished three situations that differ in terms of the access bandwidth and terminal capabilities that a service provider can assume regarding its end users.

1. *High Profile Terminal* : this assumes that the whole LSR can be delivered and processed by the terminal. For the delivery, this essentially requires the development of very high bandwidth (see section 5.2.2) transmission technologies, which is not a research topic in this project. Note that similar requirements apply for the transmission of the full LSR to theatre set-up (where the full set of A/V data is rendered without user interaction), like described in Use Case 2.
2. *Main-profile terminal*: this assumes traditional residential access bandwidth (see section 5.1.1). In this case, the network is required to transmit only the required portions of the LSR to the home terminal, so as to respect the bandwidth limit and fulfil the interactivity request.
    a. Therefore, as in the previous use case, an ad-hoc segmentation of the LSR must be defined, which determines the granularity at which the delivery mechanisms can be optimized.
    b. Unlike in *Use case 8*, no predefined sequence of the segments is available for steering the delivery of segments. However, in order to optimize the bandwidth and delay performance of the delivery of partial LSR data to the renderer, we still require the availability of delivery scripts that can instruct on the relevance of each segment over time and space.
    c. In addition, a dedicated interface is required with the terminal so as to ensure on-time delivery to the rendering functions of the parts of the LSR required to respond to the continuous interactivity requests..
3. *Low-profile terminals*: this corresponds to the situation with lowest assumptions on A/V processing capabilities and access bandwidth. This requires performing all rendering operations in a network proxy, so that only A/V data ready to be sent to display and speakers are sent to the end-device.

5.4.3 Requirements for Use case 10

This use case focuses on supporting the delivery of a FascinatE service to mobile devices of consumers at a live vent. Such devices are either low-profile terminals, with lowest assumptions on A/V processing capabilities and access bandwidth, or m*ain-profile terminals* with less restricted assumptions on A/V processing capabilities and access bandwidth. The following requirements apply:

- Network availability at a live event: this requires presence of a local wifi network or 4th generation mobile broadband network, for high-bandwidth connection at the event.
- Scalability: a delivery mechanism is required that scales to a large number of users, allowing for interactivity and adaptation of content to a variety of mobile devices. Such a delivery mechanism should be implemented in a managed or overlay delivery network, such as a CDN.
- Storage and caching: to provide on-demand and replay functionality, the delivery network must cache A/V segments that are delivered during content consumption, or store them for offline access after the event.
- Processing: given the variety in mobile devices, some low-profile terminals may still allow for some limited forms of processing, e.g. combining A/V segments at the terminal. Cloud-based components are required to handle high processing demands.
- On the production side, this use case requires functionality to relate the picture taken by the mobile device, to the content captured by the camera cluster. Also, it requires feature and object tracking to create personalized views.

5.4.4 Requirements for Use case 11

This use case focuses on hybrid delivery of a FascinatE service to a home situation with multiple screens, including a *primary screen* (e.g. Connected TV, set-top box) and a complementary *second screen* (e.g. a tablet or smartphone); for the second, low A/V processing capabilities and access bandwidth are assumed, whereas for the primary screen, less restricted assumptions on A/V processing capabilities and access bandwidth are assumed. The following requirements apply:

- Network access: this use case benefits from a hybrid delivery network, where primary content is delivered through regular service provider subscription, and second screen application and content is delivered via an over-the-top mobile broadband connection. The relation between the two services must be signalled on one or both of the networks.
- Interaction: a second screen application on a mobile interaction device, such as a tablet or smart phone, is required to connect to the additional services and content. The application should provide interactivity on the devices, as well as allowing the control of content on the primary device.

5.5 Delivery Network Functionality

In this section an overview is given of functionality that is required in the delivery network, in order to support the most demanding aspect of the selected scenario. At the highest functional level (Figure 3), the scenario is characterized by the fact that

- the delivery block has to support two fundamentally different A/V representation formats at its input and output.
- the delivery block has to intercept and process the user input.

As is described below, these observations imply the presence of at least one function for ingesting the layered A/V format and modifying it and at least one function processing and responding to the interactivity requests. The need of more specific functions (such as A/V adaptation, rendering, etc.) can already be stated at this point in time, but the functional architecture described hereafter is merely used for illustrative purposes at this stage of the project. Formal discussions on how high-level functions can be spread over the delivery network and how lower-level functions can be mapped to them will be addressed in D1.4.1. The delivery network is described in detail in FascinatE deliverable D4.2.2a. Its interfaces towards other functional blocks in the overall FascinatE system is described in deliverable D1.4.2.

A key realisation in the development of the delivery network architecture is the fact that inside the delivery network, i.e. between AV ingest and AV proxy, chunks of data are requested. These chunks of data are referred to as **tiles** and they relate to a specific spatial region of a video frame. In most cases, tiles are grouped for a certain time period, in which case they are called **segments**. The particular grouping can be dependent on the transport protocol used, but globally, the FascinatE delivery network is aimed at delivery of tiled and segmented content. Hence, the updated delivery network architecture consists of aggregated functional elements for such segmented delivery. The three main high-level blocks in the delivery network are listed below

1. **A/V Ingest**, including a *Terminal-agnostic Fascinate Rendering Node (FRN)* that performs video rendering operations that are beneficial for a large range of terminal conditions; a *Content Segmentation* function recast the LSR content into FascinatE media delivery units that are suitable for network encapsulation and further transport functions, and a *Segment Transport Server* that regroups all the functions required to initialize the actual delivery through the network infrastructure.
2. **A/V Proxy**, including a *Terminal-dependent FRN* that performs video rendering operations that are required for a specific terminal and a specific set of user requests; a *Content Re-assembly* function to remove the tile granularity of the received segments and deliver to the FRN the requested portions of the LSR, and a *Segment Transport Client* that regroups all the functions required to terminate the segment transport and signals upstream the request for segments.
3. **A/V Relay**, a set of intermediate transport nodes that can aggregate and/or relay segment requests at the transport protocol control level, and also serve as demarcation points between delivery modes for the downstream A/V flows (e.g. multicast vs. unicast or push vs. pull).

In **Table 4** below, we analyse how the requirements as derived from use cases 8-11 can be fulfilled with the delivery network functions as described in D4.2.2a and D1.4.2:

Use Case	Requirement	Network Functional Block	Gap Analysis
8	Need for Rendering functions that prepare the offered mix of views	Ingest Terminal-Agnostic FRN Proxy or Client – Terminal-Dependent FRN	Quantification arguments to optimize the location of rendering functions will be analyzed in D4.2.2b
8	Scripting required to specify what views can be generated from LSR	Interface with Production Scripting Engine (ITF-02 in D1.4.2)	A first description of information required from Production Scripts for delivery functions is described in D3.1.2. First prototyping of the interface and results are expected in D3.2.1b.
8, 9	Accommodate bandwidth constraints	Ingest Segmentation Blocks allow the delivery functions to transport content at various granularities and representation fidelity (e.g. resolution, frame rate, compression level, …)	First set-up and studies have been done for some configuration of multi-resolution and multi-rate encoding as reported in D4.4.1. Further performance proof-points of the approach will be given in D4.4.2
8	Expose content – channel mapping to delivery functions	Ingest Delivery Scripting Engine	First DSE prototyping will be documented in D3.2.1b
9	On time delivery of partial LSR data to terminal renderer	Interface with FRN (ITF-09 in D1.4.2)	A first description of the interface requirements is further detailed in D5.3.1. First prototyping of the interface will be reported in D5.1.3.
9	Optimize delivery performance based on Segment Relevance information	Proxy Delivery Scripting Engine	First performance results of an offline prototype were reported in D4.4.1. Real-time implementation will be documented in D4.5.1
9	Rendering functions performed in the network	A/V Proxy FRN and UCN functions	A demonstrator of a Network proxy for interactive video is available. Performance Improvements will be reported in D4.4.2
10	Connectivity at event	None; provisioned by underlying mobile transport network	Depends on mobile broadband deployments at event locations.

10	Scalable delivery of content towards many mobile devices	Content Segmentation, Segment Transport, Content Re-assembly	A demonstrator of a scalable delivery mechanism for interactive video is available and was described in D4.4.1. Performance Improvements and additional implementation will be reported in D4.4.2
10	Storage and caching of audio-visual data in a CDN	A/V Relay	First results for cloud/CDN-based storage will be reported in D4.5.1
10	Production-side content analysis	Production-side Content Analysis and Scripting Engine	Current content analysis and scripting components do not operate in real-time and only work on specific content (e.g. football).
11	Hybrid network access	Partly provisioned by underlying hybrid transport network, i.e. IPTV and wifi. Segment Transport is involved in network abstraction layer to separate A/V streams.	A first description of a hybrid delivery mechanism for interactive video planned for in D4.4.2. The implementation will be described in D4.5.2.
11	Second screen interaction	A/V proxy FRN and UCN functions, distributed over the network and terminal.	A demonstrator for second-screen interaction is available and described in D1.5.1. An implementation for hybrid delivery will be described in D1.5.2

Table 4: Gap analysis for network-related use cases

6 Conclusions

This document has described the overall requirements that the FascinatE system should meet in order to fulfil the needs of end-users, production teams and network infrastructure. The document is structured in three parts detailing the requirements from each of these three perspectives: end-user, production and network. It has proposed three different scenarios depending on the configuration and functionality provided by the complete delivery chain:

- In scenario 1, production-centric, all the FascinatE functionality is provided by the production side and there is no computational load shifted to either the provider or the terminal.
- Scenario 2, terminal-centric, assumes that a complete LSR, together with production scripts, are provided to the terminal which is responsible of rendering and presenting it to the end-user.
- The scenario 3, provider-centric, can be interpreted as an intermediate step in the evolution of FascinatE technology. In this case, the LSR will be rendered to a format tailored to the delivery network and targeted terminal.

In the case of end-user requirements, it has covered issues that should be kept in mind when designing FascinatE based services in order to provide a high quality of experience as desired by users. Also, design guidelines to provide a rich and user-friendly experience to FascinatE services have been detailed. From the discussion, several key points may be extracted:

- In all proposed scenarios the interactivity offered to the end user can vary but scenarios 2 and 3 have the potential of providing a higher level of interaction in a more natural way.
- The user interface proposed by FascinatE should show the following properties: simple, intuitive, efficient, non intrusive, consistent and clear.
- There should be a reasonable latency of the system. For instance, the visual feedback of some gestures (e.g. zooming or raising volume) must be fast enough to allow fluid interaction. On the other hand, other interactive commands (such as changing channels or dividing screen) are less restrictive in latency.
- Users' viewing preferences should be kept as a very important goal. In general, TV viewers want to be entertained, get informed and relax.
- Three main environments are differentiated within FascinatE: mobile, home and public. Each environment could provide different levels of interaction depending on the terminal capabilities, the social context and the typical settings of the specific environment.

Finally, for the end-user requirements, the usability assessments and the planned evaluation approach has been discussed in order to assess the fulfilment of the end-user requirements by the FascinatE system.

In the case of production requirements, there may be several distinct areas of challenges to be met by the FascinatE project:

- How to integrate FascinatE technology into existing technology and working practices?
- How do existing production staff operate an automated script-based production system?
- What tasks will production staff accept to be automated?

In particular, the following questions are key to the FascinatE system:

- What is the role of a director in an automated, script-based production?
- What will the OMNICAM be used for, and how will the operator(s) do this?
- What amount of trade off between quality and automation is acceptable in audio and video production?
- How will production staff generate scripts and metadata about video and audio content? How many people will this take?

Finally, in the case of network requirements, both requirements and some needed functionality from a network perspective have been considered. It becomes clear that each of the three proposed scenarios comes with different requirements. Scenario 1 may be implemented with existing and deployed delivery networks. Scenario 2 puts strong requirements on the bandwidth of the delivery network and may only be introduced after significant advances in physical network technology and signal processing. Scenario

3 focuses on processing functionality. Within FascinatE, we consider this scenario the most relevant for innovations in the delivery network. A summary of the requirements needed by the delivery network are:

- The need for rendering functions in various parts of the network, e.g to prepare the offered mix of views
- Scripting engines required to specify what views can be generated from the LSR
- Ability to accommodate to bandwidth constraints
- Expose content – channel mapping to delivery functions
- On time delivery of partial LSR data to terminal renderer
- Optimize delivery performance based on Segment Relevance information
- Scalable delivery of content towards many different devices
- Storage and caching of audio-visual data in a CDN or cloud
- Production-side content analysis
- Hybrid network access
- Support of channels to provide a second screen interaction

As some details of the requirements discussed in this document become clearer and better-defined as the project progresses, a third updated version of this document will be produced at the end of the project (D1.1.3, in Month 42).

7 References

[Barkhuus, 2009] Barkhuus, L. and Browns, 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.

[Anderson, 2006] C. Anderson, The long tail: Why the future of business is selling less of more. ISBN-10: 1401302378. Hyperion. 2006

[Auslander, 2006] P. Auslander, Liveness: Performance in a mediatized culture. ISBN-10 0415773539. Routledge, 2006

[Ball, 2007] R. Ball, C. North, D.A. Bowman, Move to improve: Promoting physical navigation to increase user performance with large displays. In ACM CHI Conference, 191-200, 2007

[Block, 2004] Block, F., Schmidt, A., Villar, N., & Gellersen, H.-W., Towards a playful user interface for home entertainment systems, in Proceedings of the European Symposium on Ambient Intelligence 2004, Springer, 207–217, (2004).

[Bowers, 2001] J. Bowers, Crossing the line: A field study of inhabited television. In Behaviour & Information Technology, Vol. 20, No. 2, 127-140, 2001

[Cesar , 2008] Cesar P., Bulterman, D.C.A., and Jansen, A.J., Usages of the Secondary Screen in an Interactive Television Environment: Control, Enrich, Share, and Transfer Television Content., in Proceedings of EuroITV, 2008.

[Chorianopoulos, 2007] K. Chorianopoulos, Content-enriched communication supporting the social uses of TV, Journal of the Communications Network 6 (2007), no. 1, 23--30.

[Chorianopoulos, 2008] Chorianopoulos, K. (2008) User Interface Design Principles for Interactive Television Applications. International Journal of Human-Computer Interaction. 24, 6. Taylor & Francis. 556—573.

[Cooper, 2008] William Cooper, The interactive television user experience so far, Proceeding of the 1st international conference on Designing interactive user experiences for TV and video, October 22-24, 2008, Silicon Valley, California, USA

[Coppens, 2004] Coppens, T., Trappeniers, L., and Godon, M. 2004. AmigoTV: Towards a social TV experience. In Proceedings of the EuroITV 2004 Conference (Brighton, UK).

[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.

[DIY, 2011] DIY Kinect Hacking, http://ladyada.net/learn/diykinect

[Dóllar, 2005] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, Serge Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05), Beijing, China, October 2005, pp. 65–72.

[Engström , 2010] Engström, A., Juhlin, O., Perry, M. and Broth, M. (2010). Temporal hybridity: Mixing live video footage with instant replay in real time. Proceedings of ACM CHI 2010, Atlanta, GA.

[EU-DAS, 2011] European Commission, Information Society, Digital Agenda Scoreboard 2011

[Geerts, 2009] David Geerts , Dirk De Grooff, Supporting the social uses of television: sociability heuristics for social tv, Proceedings of the 27th international conference on Human factors in computing systems, April 04-09, 2009, Boston, MA, USA.

[Hoeben, 2006] A. Hoeben, J.P. Stappers, Taking clues from the world outside: Navigating interactive panoramas. 2006

[iPoint, 2010] Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (HHI), iPoint Presenter. *http://www.hhi.fraunhofer.de/en/departments/interactive-media-human-factors/overview/ipoint-presenter/*

[Juhlin, 2010] O. Juhlin, A. Engström, E. Reponen, Mobile broadcasting – the whats and hows of live video as a social medium. Accepted to ACM MobileHCI Conference. 2010

[Kinect, 2010] Microsoft Natal Project / Kinect, h*ttp://www.xbox.com/en-US/community/events/e3/kinect.htm*

[Knoche, 2008] Knoche, H., McCarthy, J., Sasse, M. A. (2008) How low can you go? The effect of low resolutions on shot types. In Personalized and Mobile Digital TV Applications in Springer Multimedia Tools and Applications Series.

[Kolb, 2010] A. Kolb, E. Barth, R. Koch and R. Larsen, Time-of-Flight Cameras in Computer Graphics, in: Computer Graphics Forum, Volume 29 Issue 1, pp. 141-159.

[Laptev, 2003] Ivan Laptev, Tony Lindeberg, Space–time interest points, in: Proceedings of the International Conference on Computer Vision (ICCV'03), vol. 1, Nice, France, October 2003, pp. 432–439.

[Laptev, 2008] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, Benjamin Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[Lee, 2007] Jackie Lee, C. -H., Chang, C., Chung, H., Dickie, C., & Selker, T., Emotionally reactive television, in Proceedings of ACM IUI'07, pp. 329–332, (2007).

[Microsoft, 2010] *http://research.microsoft.com/en-us/um/redmond/groups/ivm/hdview/*

[MITLabs, 2010] *http://www.engadget.com/2010/04/09/mit-media-labs-surround-vision-brings-virtual-reality-to-a-tabl?icid=sphere_blogsmith_inpage_engadget*

[MITLabs, 2011] ROS MIT Demos, *http://www.ros.org/wiki/mit-ros-pkg/KinectDemos*

[Nakatoh, 2007] Nakatoh, Y., Kuwano, H., Kanamori, T., & Hoshimi, M., Speech recognition interface system for digital TV control, Acoustical Science and Technology, 28(3), pp. 165–171, (2007).

[Nielsen, 1993] Jakob Nielsen, Usability Engineering, Morgan Kaufmann, San Francisco, 1993. ISBN 0-12-518406-9

[Nielsen, 1994] Jakob Nielsen and Robert L. Mack, Usability Inspection Methods, John Wiley & Sons, New York, NY, 1994. ISBN 0-471-01877-5

[O'Hara , 2009] O'Hara, K. and Glancy, M. (2009) Watching in Public: understanding audience interaction with Big Screen TV in urban spaces. In Social Interactive Television: Immersive Shared Experiences and Perspectives. Cesar, P. Geerts, D. and Chorianopoulos, K. (Eds.). London: IGI Global.

[OpenNI, 2011] OpenNI, *http://openni.org/*

[Organic, 2010] Organic Motion Stage. http://www.inition.co.uk/inition/product.php?URL_=product_mocaptrack_organicmotion_stage

[PCL, 2011] Point Cloud Library, *http://pointclouds.org*

[Perry, 2010] Perry, M., Engström, A., Juhlin, O. and Broth, M. (unpublished) “EVS... now!” Socially segueing instant replay into live video. Conditionally accepted to Journal: Visual Studies.

[Poppe, 2009] Ronald Poppe, A survey on vision-based human action recognition, Image and Vision Computing, Volume 28, Issue 6, June 2010, Pages 976-990

[PrimeSense, 2010] PrimeSense, the Natal Project camera. *http://www.primesense.com*

[Ranjan, 2007] A. Ranjan, J.P. Birnholtz, R. Balakrishnan, Dynamic shared visual spaces: Experimenting with automatic camera control in a remote repair task. In ACM CHI Conference, 1177-1186, 2007

[Savarese, 2008] Silvio Savarese, Andrey DelPozo, Juan Carlos Niebles, Li Fei-Fei, Spatial–temporal correlatons for unsupervised action classification, in: Proceedings of the Workshop on Applications of Computer Vision (WACV’08), Copper Mountain, CO, January 2008, pp. 1–8.

[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.

[Scovanner, 2007] Paul Scovanner, Saad Ali, Mubarak Shah, A 3-dimensional SIFT descriptor and its application to action recognition, in: Proceedings of the International Conference on Multimedia (MultiMedia’07), Augsburg, Germany, September 2007, pp. 357–360.

[Shirky, 2009] C. Shirky, Here comes everybody: The power of organizing without organizations. 2008

[Shotton, 2011] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake, Real-Time Human Pose Recognition in Parts from a Single Depth Image, in CVPR, IEEE, June 2011

[SportsTV, 2010] *http://www.sportstvproduction.net/?page_id=8*

[SR4000, 2010] SwissRanger SR4000 Overview, *http://www.mesa-imaging.ch/prodview4k.php*

[Sun, 2001] X. Sun, J. Foote, D. Kimber, S. Manjunath, Panoramic video capturing and compressed domain virtual camera control. In ACM Multimedia Conference, 329-338, 2001

[Sun, 2009] Ju Sun, Xiao Wu, Shuicheng Yan, Loong-Fah Cheong, Tat-Seng Chua, Jintao Li, Hierarchical spatio-temporal context modeling for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR’09), Miami, FL, June 2009, pp. 1–8.

[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.

[TA2, 2010] TA2, Together anywhere, Together anytime, *http://ta2-project.eu*

[Vatavu, 2008] Radu-Daniel Vatavu, Stefan-Gheorghe Pentiuc, Interactive Coffee Tables: Interfacing TV within an Intuitive, Fun and Shared Experience, EuroITV 2008: the 6th European Interactive TV Conference, Salzburg, Austria, July 2008, LNCS 5066, pp. 183-187, Springer (2008).

[Vatavu, 2010] Radu-Daniel Vatavu, Cretivity in Interactive TV: Personalize, Share, and Invent Interfaces, In Mobile TV: Customizing Content and Experience, A Marcus et al. (eds.), Springer-Verlag, 2010.

[Wang, 2004] Wang, J., Xu, C., Chng, E., Wah, K., and Tian, Q. 2004. Automatic replay generation for soccer video broadcasting. In Proceedings of the 12th Annual ACM international Conference on Multimedia (New York, NY, USA, October 10 - 16, 2004). MULTIMEDIA '04. ACM, New York, NY, 32-39.

[Wang, 2009] Zhenchen Wang, Stefan Poslad, Charalampos Z. Patrikakis, Alan Pearmain, Personalised Live Sports Event Viewing on Mobile Devices, ubicomm, pp.59-64, 2009 Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2009.

[Wang, 2009] Zhenchen Wang, Stefan Poslad, Charalampos Z. Patrikakis, Alan Pearmain, Personalised Live Sports Event Viewing on Mobile Devices, ubicomm, pp.59-64, 2009 Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2009.

[Willems, 2008] Geert Willems, Tinne Tuytelaars, Luc J. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Proceedings of the European Conference on Computer Vision (ECCV’08) – part 2, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 650–663 (Number 5303).

[XTR3D, 2010] Extreme Reality XTR3D. *http://www.xtr3d.com/*

[Zhu, 2010] D. Zhu, T. Gedeon, K. Taylor, Natural interaction enhanced remote camera control for teleoperation. In ACM CHI Conference, 3229-3234, 2010

8 Glossary

Terms used within the FascinatE project, sorted alphabetically.

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

Other acronyms:

A/V	Audio and Visual
ADSL	Asymmetric Digital Subscriber Line
API	Application Programming Interface
CDN	Content Delivery Networks
CMOS	Complementary Metal–Oxide–Semiconductor
DSE	Delivery Scripting Engine
DVB-T	Digital Video Broadcasting - Terrestrial
EPG	Electronic Program Guide
EVS	Brand name for replay/highlight video servers
FTTH	Fiber To The Home
FTTx	Fiber To The x
GPON	ITU-T G.984 Passive Optical Network Standard
HD	High Definition
HDR	High Dynamic Range
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access
HSUPA	High Speed Uplink Packet Access
IBC	International Broadcast Convention
IPTV	Internet Protocol TeleVision
LSR	Layered Scene Representation
LTE	3GPP Long Term Evolution
MMS	Multimedia Message Service

OMNICAM	Omni-directional camera for ultra-high resolution panoramic video capture
PCA	Principal Component Analysis
PSE	Production Scripting Engine
RGB	Red Green Blue
RO	Replay Operator
ROI	Region Of Interest
SMS	Short Message Service
VGA	Video Graphics Array
UI	User Interface
UMTS	Universal Mobile Telecommunications System
VM	Vision Mixer