End User, Production and Hardware and Networking Requirements

Deliverable D1.1.1

FascinatE identifier:	FascinatE-D111-UPC-Requirements-v08.doc
Deliverable number:	D1.1.1
Author(s) and company:	J. Ruiz Hidalgo, J.R. Casas, X. Suau (UPC); A. Gibb (BBC), O.A. Niamut, M.J. Prins (TNO); G. Zoric, A. Engström, M. Perry, E. Önnevall, O. Juhlin (TII); J. Macq (ALU)
Internal reviewers:	G.A. Thomas (BBC), A. Havekes, (TNO)

Work package / task:	WP1
Document status:	Final
Confidentiality:	Public

Version	Date	Reason of change
1	2010-03-10	Document created, initial input
2	2010-06-23	First version with user and production perspective
3	2010-06-24	Figures and reference formatting
4	2010-06-25	Added network perspective
5	2010-06-29	Modifed introduction chapter to add scenario/interaction grid
6	2010-07-08	Added conclusions chapter
7	2010-07-08	Small typos and corrections
8	2010-07-29	Changes after internal reviews

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

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

Table of Contents

1 Executive Summary

This document 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, where to develop and study the possible FascinatE requirements:

- 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, 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): It can be interpreted as an intermediate step in the evolution of FascinatE technology. In this case, the layered scene 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. 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.
- In some of the environments, it is believed that gesture-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 human gestures (see deliverable D5.1.1). This control-by-gesture will not completely replace other control devices, but will provide an alternative to traditional interaction methods such as remote controls, PCs, or touch screens.

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

In the case of network provider requirements, 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. In this case, each scenario and use case described in Chapter 1 may require a different set of functions:

- In both scenario 1 (production-centric) and scenario 2 (terminal-centric), the delivery network itself does not provide any of the above mentioned functionality. See the sections on production and end-user interaction, respectively, for high-level functional architectures in those domains.
- The functionality for scenario 3 (network-centric) consists of all functions given in the high-level functional architecture or a subset of these functions.
- Concrete usage of FascinatE technology will blend aspects of the three scenarios.

Many details of the requirements discussed in this document will become clearer and better-defined as the project progresses and will be verified when demonstrators are trialled. An updated requirements document will therefore be produced at the end of the project (D1.1.2, July 2013).

2 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 to 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 first of two deliverables addressing requirements in the FascinatE project: a final requirements document, informed by things learned during the project, will be issued at the end of the project (July 2013).

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 when 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 what is termed a 'layered scene' description.
- 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 layered scene is referred o 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), 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.

Further background information on the project can be found on the project's website: http://www.fascinate-project.eu.

2.1 Organisation of this 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. 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 hardware and

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.

3 Scenarios and Use Cases

3.1 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 section 3.2, the three proposed scenarios are studied and related to the amount of interaction allowed to the end user in the FascinatE system

3.1.1 Scenario 1: Production-centric delivery chain

This scenario considers a current state-of-the-art delivery scenario. 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.

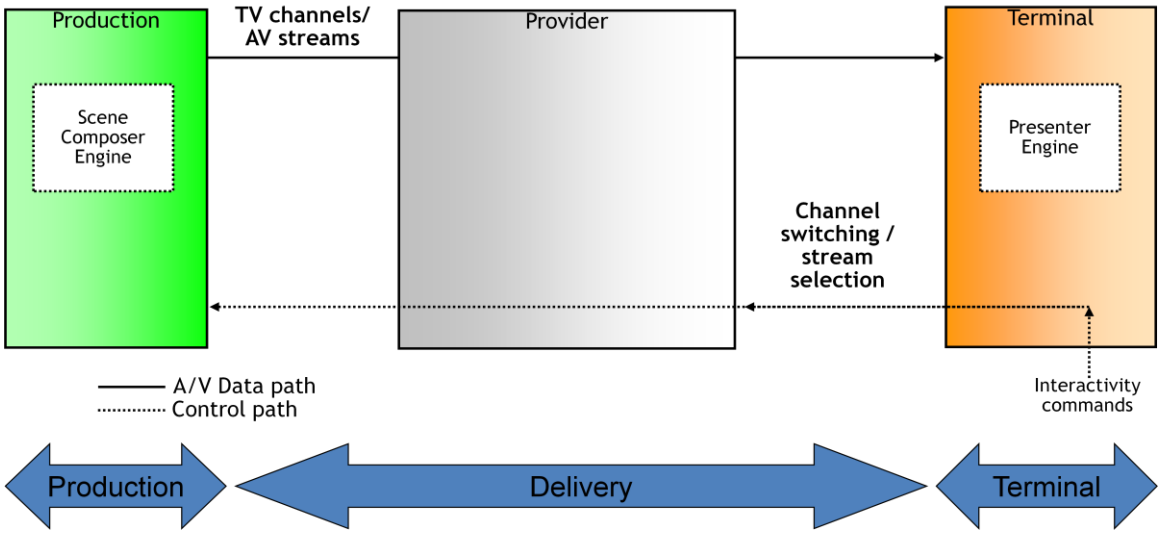


Figure 1: High-level architecture for the production-centric scenario

3.1.2 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 layered scene[1], with the terminal receiving all the captured A/V streams. The layered scene 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 specify 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.

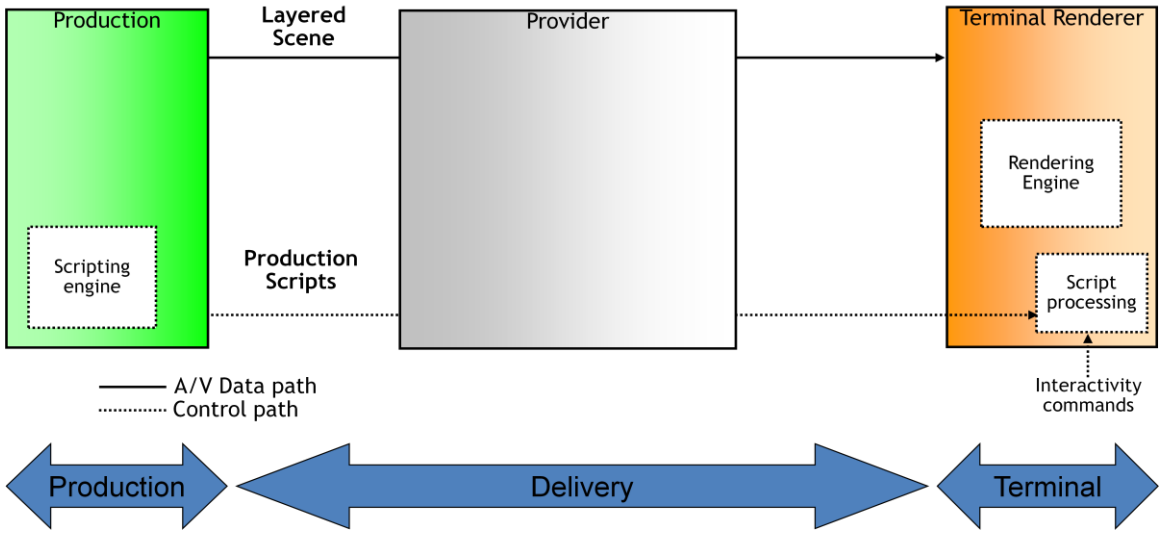


Figure 2: High-level architecture for the terminal-centric scenario

3.1.3 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 layered representation of the A/V scene and processes the accompanying production scripts.

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 layered scene 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 this 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 layered A/V content to create new services: linear TV programmes, personalized, interactive services, etc.

[1] The layered scene is defined in D2.1.1. A layered scene presentation contains a number of video and audio signals (ultra-high resolution images, images from panning/tilting/zooming cameras, omnicam images, audio signals from surround microphones) with the metadata describing their relative geometrical and photometric alignment.

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.

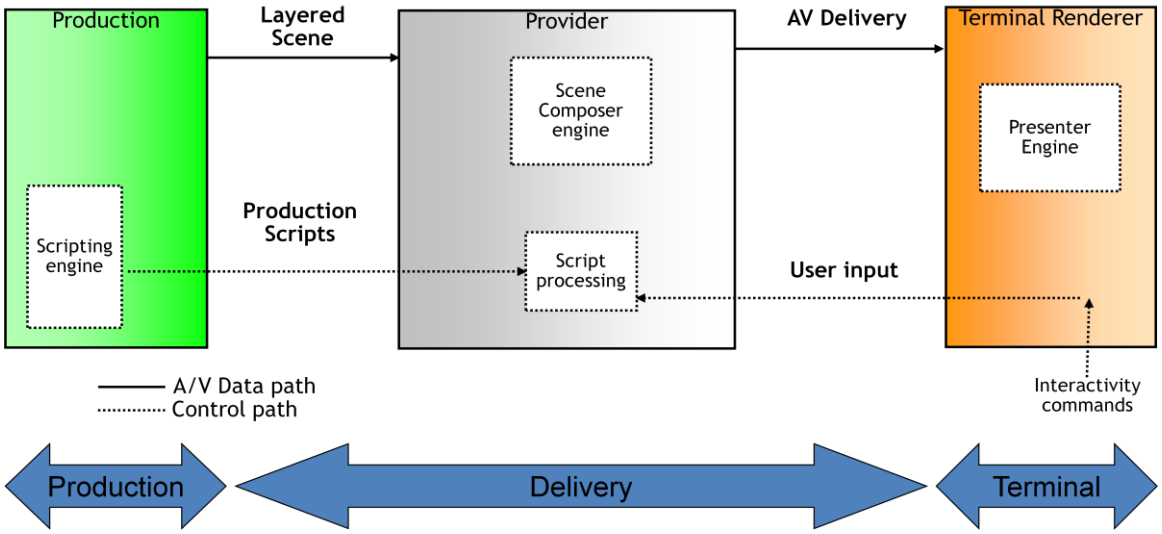


Figure 3: High-level architecture for the provider-centric scenario

3.2 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 and content and scripts are transmitted to the terminal. However, all reproduction of content is fixed from the production or provider side **Network**: Requires higher bandwidth to the user than the production-centric scenario	**Production**: Work load can be split between production gallery and provider gallery **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 are required. Some of this work could be automated **Network:** Must deliver lots of different streams, both AV and, maybe a small subset of scripts to include limited functionality to the end user	**Production:** The case where all production is basically automated/supervised mark-up and script generation **Network:** Must deliver entire layered scene representation 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
Full interaction. **FascinatE interactivity**	**Production:** Lots of scripting both automatic & supervised in gallery **Network** Must deliver the entire layered scene representation to the end user	**Production:** Production & provider galleries can cooperate on script creation and metadata generation **Network:** Must deliver entire layered scene representation to the end user	**Production:** Production & provider galleries cooperate on script creation and metadata generation **Network:** Must deliver entire layered scene representation to the end user

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.

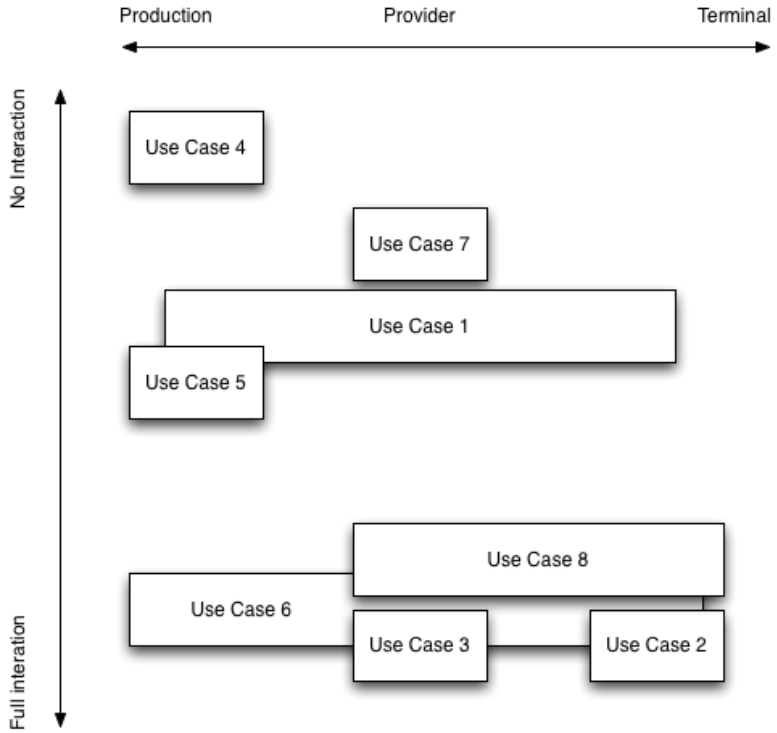


Figure 4: Classification of use cases

3.2.1 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 but can be applied to any other event such as concerts, standard TV programmes, or any other sport events.

Use case 1: No interaction

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. 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, which is playing a fantastic match, he quickly changes channels so he can follow him displaying this dedicated player view on his TV set.

Use case 2: 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 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.

Use case 3: 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 while 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 and 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 he selected to follow. However, after a while, John decides that he should let his daughter control the system so he passes the token on to his daughter by placing his hand over her head. This token indicates who is controlling the system and has preference over any other user trying to give commands to the system. She is very happy to be able to select a different channel with cartoons. Meanwhile, John gets his tablet out, selects the same view he was using before in the main TV. The network provider receives this, and selects the optimal view to mimic the configuration John was using in his main TV.

3.2.2 Production perspective

Use case 4: A FascinatE system used to create a normal, linear TV programme

England and France 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.

Use case 5: A linear TV programme with some interactivity

The 2018 England Football World Cup is being covered by the BBC. The Canvas-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 omnicam 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 vision mixer operator is at a normal vision mixer console, but 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.

Use case 6: 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 9.

3.2.3 Provider perspective

Use case 7: 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 AV feeds are made available in the so-called layered scene representation. 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.

Use case 8: 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 layered scene representation, instead of traditional linear TV feeds. This layered scene 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 into a fully new set of FascinatE-ready acquisition equipment). This enables PaloCom to take advantage of the flexible layered scene representation 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 layered scene. Whereas Telombia serves only the 10% of their customer basis 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.

Use case 9: 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.

4 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, conclusion summarizes this section.

4.1 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 posses 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 details.

4.1.1 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 , has been around for ages.
- 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' does 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

In [MITLabs, 2010], presents the use of a tablet PC to pan around a scene being captured by three cameras - 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.

4.1.2 Navigation and interaction

This section covers issues that should be kept in mind when designing FascinatE based services in order to provide quality of experience as desired by users. In order to make a clear presentation, each of the topics is formatted as:

a. Definition
b. Motivation/justification
c. Literature review and original data if any
d. Design relevant summary for FascinatE
e. Design sketch
f. Conclusions/future work

Interaction practices:

a. By Interaction practices, we mean the ways in which people behave towards the TV set in various environments, and more broadly, how they engage, in a variety of ways, with TV

content. This will examine group interaction around technology and designing for groups such as families, sports clubs, crowds etc. FascinatE aims to achieve more interaction with the broadcast through different devices such as TV and mobile phones. It is therefore important that the team understands current practices for two reasons: i) to avoid a misfit between our technology and real life, ii) to inform design and generate new design ideas.

b. As pointed out in the FascinatE proposal the use of media has changed rapidly the last few years through development of mobile phones and the growth of social media platforms. On the other hand, a TV is still seen as traditional technical equipment compared to computers. Through ethnographic studies of practices for watching TV we have started to outline general thoughts about FascinatE and the end-user interaction.

c. In Tsekleves, Whitman, Kondo and Hills *Bringing the Television to other Media in the Home: An Ethnographic Study* one of their informant makes a comparison between TV and computer: ”You can’t fall asleep on the Internet, can you?” [Tsekleves, 2009]. This reflects the importance of studying interaction practices for specific devices. The TV is to a large extent connected to relaxation in everyday life. The interaction is however connected to the content you are watching. Watching sport differs in a quite obvious way from watching for example a romantic comedy. Where the former demands presence and common behaviours like clapping and pointing, the second example can give a more laid-back experience.

One important point that Tsekleves et al. make is that people today are used to “multiple media related tasks”. It is common to use your mobile phone while watching TV or jumping between screens (TV/Computer/TV/Mobile). ” TV (…) is also typically situated prominently in a shared social space, such as the living room or kitchen” [Tsekleves, 2009]. This makes the TV a social device just in line with Barkhuus and Brown’s study *Unpacking the Television: User Practices around a Changing Technology*. According to Barkhuus and Brown the watching of TV creates a common “routine” and also effects the “shared experience” together with the social interaction between friends and colleagues [Barkhuus, 2009].

The use of public screens varies to a large extent depending on people’s movements in the setting. O’Hara et al. speak of this as the “city rhythms” [O’Hara, 2009:258]. People therefore interact with the screen in relation to their other commitments (walking to work, eating lunch, going home from work etc.) as well as in relation to other people. According to O’Hara et al. “watching in a larger crowd provides an important sense of *belonging* and *community*. (…) We see the importance of watching as a crowd, too, in sporting events, where a sense of competitiveness and being part of a larger group, transformed the atmosphere of viewing” [O’Hara, 2009:260]. This *sense* could be one of the reasons why people watch sport in a pub and not at home.

d/e. We have concentrated on four different studies found in the literature concerning TV watching as an introduction to interaction practices. All these works have in common the main topic, but they are focusing on different aspects for the interaction: Multiple screens [Tsekleves, 2009], TiVo [Barkhuus, 2009] and Big screens [O’Hara, 2009].

According to Tsekleves et al. ”the level of interactivity is not only limited by the potential of the technology, namely display mechanisms, hardware and interaction models but also by the user’s willingness to interact” [Tsekleves, 2009:202]. The viewers are therefore the main objects for the success of FascinatE. Although Tsekleves et al. focuses on the users “willingness to interact”, one big obstacle concerning use of technology is the difficulties of using the product. Therefore the interaction technique has to be simple and easy-to-use.

Barkhuus and Brown’s concerns are how new techniques have changed the experience and planning of TV watching. According to them “the active choosing of what content is to be watched” through downloading and recording your own films ”more resembles other types of media consumption such as reading, listening to music, or going to the cinema” [Barkhuus, 2009:15:3]. According to Barkhuus and Brown ”viewers do not simply watch TV, but search for, obtain, share, collect, and discuss television”. In that sense watching sports is not only about seeing but also interacting with the content and sharing this with others watching.

One point in O’Hara and Glancys *Watching in Public: understanding audience interaction with Big Screen TV in urban spaces* is the importance of audio to attract people to public/big screens. People already have a habit of using personal screens (such as mobile phones and TVs in home). It seemed although that people in O’Hara’s and Glancy’s study used the public screen to construct their daily routines. “[O]ne person who worked in an office nearby the

Manchester screen had shifted the time of her lunch break so that she could eat her lunch while watching the lunchtime episode of Neighbours".

It is also important to be aware of the location of the screens and how this is affecting the interaction. This might be applicable to both the public and the personal TV screens. In two of these studies the TV was placed in the main room (living room, kitchen) and this definitely affected the interaction with the screen. The screens therefore have important social aspects independent of the content.

f. To be able to develop an advanced but simple product for TV watching we need to find new ways of watching TV, with a combination of the media knowledge and use that we have today. So how do mobile phones, TVs and public screens relate and in what way are they a separately viewing experience?

Social Interactive TV:

a. This covers relevant current research in interactive TV, except research on future remote controls (which are covered in Section 3.1.3 Gesture based interfaces) and interaction practices (also covered in the above paragraph). As watching TV has always been a social experience, that aspect should be included as well. By using today's communications technologies, the viewers can be connected together, even if they are not co-present in the same physical setting. The Internet has supported social interaction around viewing, and this is a key driver for this topic.

b. Co-viewing practices are important for us to understand in FascinatE, because FascinatE is a type of interactive TV and at the very least has to offer services comparable to next generation systems. There might be an intrinsic problem for co-viewing in FascinatE in the sense that the freedom of perspective for viewing content may lead to different viewers viewing different images – this could make social interaction (synchronous or synchronously) more problematic.

c. Interactivity in interactive TV (iTV) spans from very low to high level in which viewers affect the program being watched. Although the technology has been available for more then a decade, there is still much to be done to improve the user experience. One of the key issues that needs to be resolved in iTV is how to interact with and navigate through the broadcast content. Most of the solutions available so far have been based on advanced remote controls (including coloured or functional buttons) and hierarchical menus (cursor or numerical navigation), which are often too complex and slow. However, with technological advances, new possibilities for user interfaces and interaction open, which might lead to even more clumsy design if users' current practices and expectations would not be considered [William, 2008]. Several representative interaction techniques, showing new research directions, are described below, whereas (more) gesture based interfaces are covered in section 3.1.3.

Speech remote control is used in [Nakatoh, 2007] to control a digital TV – to change between channels or to perform category search. Still, for simple actions like volume changing, the button input is still used.

An interactive coffee table [Radu-Daniel, 2008] is used to control the TV set using shared wide-area interface via simple hand movements across the video-sensitive surface of the table which may be performed by any of the viewers at any time. In doing so, the need for negotiation is avoided, and the interface is immediately available for all the participants.

A tangible cube [Block, 2004], with embedded gravity sensing and wireless communication capabilities, is used as an input device for playful changing between different TV-channel. On a TV screen, a 3D graphical representation of the cube is shown, with a TV stream being rendered on each face of the cube.

User interface for personalized live sports viewing on mobile devices in [Zhenchen, 2009] consists of viewing and navigational parts. Accelerometer sensors, which are incorporated in current mobile devices, are used to switch between viewing screen and navigational menu – it is only necessary to shake such a mobile device.

HiTV is an Emotionally-Reactive TV system [Jackie, 2007] where a digitally augmented soft ball is used as affect-input interface that can amplify TV programme's video/audio signals. It transforms the original video and audio into effects that intrigue and fulfil people's emotional expectation, and thus gives certain characteristics of social responses.

Social interactive TV is already an ongoing academic research and commercial programme. Using this technology, we can see what our friends are watching (e.g. through Facebook status:

"Mark is watching Big Brother"), tether the TV content together, or chat (audio/text) about the programme onscreen within the TV image. Off-channel media discussions, over the phone, in SMS conversations, in online forums or chat windows is for many people an integral part of TV viewing.

In [Schatz, 2007] the role of broadcast is redefined: *instead of being a plain consumable, TV content serves as conduit of social interaction i.e. socializing around the content might be more important than the content itself.* In the same work, it is further stated: *Social TV aims to provide multiple remote viewers with a joint watching experience*.

Many existing systems offer instant messaging capabilities like messages and status information. An example is AmigoTV [Coppens, 2004] - a prototype implementation that combines broadcast television with rich communication, offering social experience through voice communication or emotions and avatars

Social interactive TV has also been observed from a theoretical aspect. In [Chorianopolus, 2007], taxonomy of the social aspects of television based on presence and type of communication is given, while in [Geerts, 2009] sociability heuristics for evaluating social TV is presented.

Social interaction has also been observed outside TV context, e.g. in the project Together anywhere, Together anytime [TA2, 2010].

d. It is clear that interaction with FascinatE needs to be carefully designed. What might not be that obvious, is on what level the technology should support social communication around TV. Relevant scenarios include both synchronous and asynchronous communication, with both co-located and distance viewers, whereas activities assumed around it are e.g. sharing common conversational elements, social filtering, choosing programmes or controlling the content. However, social communication should not distract users from the TV programme, and its use should be simple and non intrusive.

e. One of the reasons that we might want to understand social practices around watching TV in FascinatE is in using 'social network' technology in the generation of tagged content and scripts to allow a shared perspective of common media content. This might cover remote viewing, chat, social filtering, interface metaphors. Possibly people may publish their viewpoints (to social networks?) and other people may subscribe to this (possible business model?).

f. Within the FascinatE project, studies will continue in order to understand what would be the best design practice for interacting with the FascinatE system. Considering social aspects, no primary user research will take place, but possibly some interaction mechanisms will be prototyped and evaluated in later stages of the project.

Multiple screens and interplay:

a. We no longer live in a single screen world, and content on these screens may be interrelated. For example, in a public arena, large TV screens may be used in combination with mobile phone screens, in the living room, TV screens are used in combination with tablets and mobile phones.

b. Mobile devices, computers, laptops, public screens and multiple TV screens proliferate, and provide both more screen surface and multiple views into related TV content. The use of multiple screens will change how we experience and access content, both individually and socially. In different formats/device form factors this content may be used to support interpretation and user experience with this content – this is FascinatE's technical contribution. We need to understand how these devices will be used together, and what values this opportunity will hold for users.

c. Although the idea of simultaneously using multiple media devices is not new, there are not many studies that investigate the use of multiple screens and their interplay in TV viewing. In [Cesar, 2008] four major uses of secondary screens (handheld devices) in an interactive digital television environment are identified: control, enrich, share and transfer television content (to take away a copy of the television content). Control enables the decoupling of the television stream, optional enhanced content, and television controls. Enriching media content might be done by, for example, including personalized media overlays such as an audio commentary.

d/e. Secondary screens (e.g. laptops, iPads, touch screen phones) provide technology to handle collaborative watching (remote and/or co-present). Examples of FascinatE relevant use of multiple screens are: use of small devices to provide more detail for individual users on the

content seen on the large screens (enhancing viewing), personal rewinding/instant replays, reviewing of visual content before displaying on primary screens, sharing of common conversational elements, social networking around live broadcast, or for the end user production of content for the larger screen by accessing and editing content via scripts.

f. Within the FascinatE project, further studies will be carried out of people watching TV on multiple devices, both in public and home environments.

Immersion and 'liveness':

a. 'Immersion' in FascinatE is formulated around embodied physical presence within the surrounding media technology, i.e. surround sound and a large screen in the home or a curved panoramic cinema display. But the concept could be expanded to also involve *engagement* in the viewer experience by other means. This is closely related to the *liveness* of the event, which can be conveyed more or less successfully through the way it is produced and displayed at end user terminals.

b. Producing the sensation of being there at a live event while viewing it remotely will include, but is not exclusively about embodied physical presence within the surrounding screens and audio. Other topics include the role of narrative in maintaining live qualia, e.g. balancing shots of audience with focal action, interaction between mediated (i.e. TV) activity and present activities (i.e. what is going on around the viewer), and the role of audio in generating the viewer experience. This mainly has relevance to the home and public scenarios of FascinatE and is not likely to apply strictly mobile use.

Television and the Internet are converging, both in hardware and in user behaviour. With the proliferation of broadband Internet and web services capable of aggregation of information and media, users are now actively customizing their media intake on the Internet to fit their particular interests (see previous section). Combined with the trend of more and more of these sources being updated in real time – twitter feeds, live sports reporting and betting, live web cameras etc – there are even more such layers of live content feeding into the experience of watching a live transmission of an event.

c. [Auslander, 2006] discusses the common assumption that live qualities in events and performances are inherent features that are not compatible with mediatized formats such as television. He argues that, on the contrary, the feeling of a live experience in e.g. a theatre play is often produced and even enhanced and saturated by media, through making references to narrative formats from television or by the use of screens and recorded audio. This can be seen in the same way in the events to be captured by the FascinatE system – concerts, live sports etc – in the way they are enhanced with non-live or near-live media such as 'jumbotron' screens displaying replay sequences and close-ups in sports arenas or close ups of performers projected on backdrops at live concerts [Perry, 2010]. Hoeben and Stappers [Hoeben, 2006] argue that subtle physical and physiological properties play an important role in how we perceive and interact with screens and images in general, and with environments displayed on panoramic images in particular. They introduce balancing high dynamic range and managing unnatural depth of field in panoramic images, among other things, as means of making the viewing experience more natural and immersive.

d. Immersion in the viewing of a live event is a strong feature in FascinatE, and is clearly related to the embodied physical presence within the screen and audio space in the end user setting. But it also involves engagement with the live event as such, and this could be designed for within the features already planned for in the system, both at the production end and at the end user terminals.

e. Live aspects should be kept in mind when designing features such as automated script selection, narrative formats, replay functionalities and intuitive navigation within the panoramic image at both the production sites and at the user end. Second screens and parallel live information flows also come into play in scenarios involving experiencing a live event through multiple media sources.

f. Further studies will be made on how broadcast live events are edited and produced, as well as on how users perceive and interact with live and non-live media while viewing live events.

Virtual camerawork:

a. Virtual camerawork includes ways of framing live video content within the larger panoramic image, for display locally in the case of end users, and for output to a vision mixer or broadcast

to end users in the case of a FascinatE operator. It further includes mechanisms for dynamically controlling and tracking this frame within the panoramic image, analogous to pan-tilt-zoom operation of a manned camera at a live event.

b. At the user end, FascinatE represents an opportunity to interact with the entire scene of a broadcast live event. This interaction will to a large extent involve selecting and mixing between automated or remotely produced scripts. But it may also involve manipulating those scripts and manually controlling framing and navigation within the panorama. The means for this interaction then becomes an integral part of the viewing experience.

At the production end, manually or semi-automatically controlled “virtual cameras” will likely be working side by side with automated scripts and manned cameras. Automated scripts may need some means of manual adjustment. The control of these virtual cameras needs to be as accurate and responsive as a manned camera, which sets high demands on their remote control operation and interaction models. The requirements on the production end may be different and more critical than those on the user end, where interaction through devices at hand or gestural interaction may be preferred to professional hardware.

c. Embodied interaction, arguably, promotes the better use of humans’ physical embodied resources such as motor memory, peripheral vision, optical flow, focal attention, and spatial memory to enhance the experience, understanding, or performance of the user [Ball, 2007]. Such models, as well as touch interfaces and state of the art models for remote controls should be further explored in order to find plausible ways for end users to navigate within the panoramic image and manipulate automated scripts generated elsewhere in the system.

There are several systems developed for remote control of cameras to frame and display shared workspaces, teleconferences, lectures, etc. [Sun, 2001; Ranjan, 2007]. Earlier experiments with navigation in panoramic images have used combinations of image recognition, tracking, stabilisation and filtering to model a professional camera operator’s pan-tilt-zoom operations. But they generally address much less dynamic events than those envisioned in FascinatE; lectures, monitoring tasks and collaborative work in controlled settings. A study by Bowers [Bowers, 2001] highlights the need for the director or vision mixer to influence and direct the framing of regions of interest by remote camera operators. Commercial systems are available for remote control of cameras, in both surveillance systems and in web based solutions using software controlling remote web cameras.

d/e. The end user side will be informed by our parallel work on secondary screens (e.g. laptops, iPads, touch screen phones) as well as future remote controls, gestural and embodied interaction. Investigations into professional remote control hardware (pan-tilt-zoom controllers using joysticks, sliders and levers) may be a suitable starting point for exploring navigation and operation models for the production end.

f. Further research needs to be carried out into remote control that meets the demands of high accuracy, speed, responsiveness and smooth operation similar to that of professional camerawork. This will also include investigating high end commercial systems for remote camera operation.

Interaction in user-generated TV/video:

a. Users are engaging with live media in a multitude of ways online, both on computers and on mobile devices. Behaviours familiar from search and information retrieval online are transferred to rich media like audio and video, and new forms of aggregation and consumption emerge as services and end terminals become more powerful.

b. Television viewing and internet use are converging. Although TV viewing in the home, one of the FascinatE scenarios, will largely remain a lean-back activity as we think of it today, it is also likely to be influenced in parts by evolving and more active user behaviours of internet use as the difference between a TV set and a computer screen becomes less clear. There is also an increase in live media online as high speed broadband lowers delivery barriers.

On the user end, we see users taking an active part in finding and compiling information and media from a diversity of sources. This has been an ongoing trend since the wider penetration of the Internet, and has come to involve rich media such as video in the past decade with the proliferation of broadband Internet and so-called web 2.0 services. Users have come to expect the ability to customize their media consumption both manually through search and through

aggregators and mash-up services combining video, audio, text, map data etc, often in real time.

c. People are looking for social media broadcasts, niche material and unique customised sources that are not provided by traditional media sources. This has been made possible by dramatically lowered cost of production and distribution, which has in turn broken down professional categories in media production [Shirky, 2009]. Amateur and semi-professional producers now exist side by side with high end content providers. People are adopting production tools and methods to produce live media with mobile phones [Juhlin, 2010] and other tools at hand, and broadcasting to friends and wider audiences. These niche productions also find their audiences, who consume them in addition to more mainstream media [Anderson, 2006].

d/e. Allowing for interaction with the live video image in various stages along the production chain, from the production gallery to the end user terminals, is a key feature in FascinatE. This could be leveraged in allowing for production teams outside of the main production site to assemble scripts and footage into broadcasts customized for specific target audiences. These production teams could be both in-house partners and independent semi-professionals, given adequate support and access to the raw video data and control features. Tools could include mixing functionalities, selection of a variety of pre-produced camera scripts, access to metadata and tools for adjusting scripts within some given constraints. This naturally raises interesting possibilities along with concerns for production quality on the content provider's part. If production tools would be used externally, they would need to be carefully designed so that they meet the demands of the content providers and delivery networks, while being manageable by independent producers.

f. Future work would first need to involve discussions within the project regarding the roles of such external producers and the set of features that would be appropriate. It would further include how tools and production methods could be designed and transferred to semi-professional or non-professional hardware.

Instant replay:

a. Instant replay involves replaying video footage of events very soon after they have occurred, usually in breaks from the event's 'action'. Replays may be shown in slow motion, or from multiple camera angles. Most instant replay sequences show close-up action. Less frequently, wide angle instant replays are also used when explaining strategy, and also in concert with telestration (still or moving images that are supplemented with a light pen).

b. 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 ongoing 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'.

c. 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, resequenced and played back. In live sport that involves the use of multicamera 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 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.

In live sports production, the major part of the production studio is taken up by workstations for the vision mixer (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 replay operator's work and one display for graphics overlays. Usually close to the video gallery is the replay operator's workstation. The VM and the replay operator (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 vision mixer. The replay operator'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 vision mixer. If the vision mixer, who relies on the replay operator to have done just this, calls for footage, the replay operator 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 replay operator 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.

d. Users expect to have access to instant replay in near real-time. Accessing this replay information quickly is not easy, and professional instant replay operators replay on a number of mechanisms to do this work. These include using the ongoing 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.

e. 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].

f. 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.

4.1.3 Gesture based interfaces

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, user-friendliness, specific devices, etc. are some of the important issues to be taken into account when evaluating an interactive system.

- *Natal / Kinect Project*: Commercialized by Microsoft Co., Natal is intended to be a revolution in the video gaming field. It is believed that, after Kinect, devices such as the Wiimote or any other remote control system will become an old-fashioned version of gaming after Natal release (expected around 2010) [Kinect, 2010].

  Natal is not a complete system itself, but a complement to the acclaimed Xbox360. More precisely, Natal 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 Natal's crucial component which allows precise tracking and gesture recognition.

  In order to develop Natal'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 Project Natal to see the room in 3-D under any lighting conditions. How PrimeSense's camera works is not clear yet, due to industrial secrecy.

  Natal performs full-body tracking, the user being able to move freely to interact with the system. Nevertheless, Natal is not really based on gesture recognition but on a tracking precise enough to allow the user interact with virtual elements on screen. Therefore, there might not exist a dictionary of gestures.

  Microsoft has affirmed that they will not go below a latency of 0.1 seconds. However, in the first public shows and meetings, one may appreciate a higher latency in some applications.

- *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 system is Microsoft Natal / Kinect, even though it has not been released yet. Natal provides 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 deviceless interactivity. Furthermore, only Natal 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 great amount of cameras. Actually, UPC's setup for recordings will consist of a central TOF (Time Of Flight) camera and two lateral regular cameras. Thus, systems like Organic Motion Stage (10 cameras) [Organic, 2010] or HHI iPoint Presenter (special vertical camera) are not adapted to FascinatE's requirements.

In conclusion, Natal Project characteristics are the most similar ones 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. duration of a film). FascinatE's gesture recognition system should be capable of tracking and interpreting gestures after such long periods of inactivity, while Natal always deals with highly active users.

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 Time-Of-Flight Cameras

Cameras which provide depth information have been widely studied and developed recently. We focus on Time-Of-Flight (TOF) cameras, specially on the SwissRanger SR4000 [SR4000, 2010]. Such cameras rely on the Time-Of-Flight principle, which measures the back and forth travelling time of a light beam starting at the camera and rebounding on the desired object at distance Z. SwissRanger cameras work by modulating the outgoing IR beam with an RF carrier, then measuring the phase shift of that carrier on the receiver side. Such an approach limits the measurement distance to, at most, 10 meters.

Surveys and reviews of recent research works on gesture recognition provide few references to strategies exploiting TOF depth information. However, given their ability to work without needing to rely on general scene illuminantion 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. Moreover, combining TOF cameras with regular cameras to take advantage of TOF depth and stereo information at the same time is also an interesting and challenging issue.

The tested SR4000 camera provides three images per frame (Figure 5):

- *Amplitude image* : Amount of IR light received per pixel
- *Depth image* : Time-Of-Flight calculated depth
- *Confidence image* : Value which indicates how reliable the depth measurement is (from 0 = worst measurement to 7 = best measurement)

TOF cameras offer a relatively poor resolution of 176x144 pixels (QCIF). On the other hand, they allow recordings at about 25fps, which is a useful frame rate for real-time tracking applications. In addition,

since TOF cameras work with IR light, they are invariant to illumination changes, being able to make recordings in dark scenes.

The main drawback of TOF cameras is the erroneous measurement due to multiple reflections in the scene. Such multipath receptions may deform objects in the scene. Furthermore, object texture is also of great importance, some materials being extremely absorbent (dark and matt) and others very reflecting (mirrors, floor tiles). As an example, the chairs in Figure 5 offer poor measurements (low confidence values) because of scattering of IR light due to chair's texture. In addition, other TOF inherent errors such as systematic depth errors and motion artefacts may appear [Kolb 10]. IR light is modulated with a sinusoidal signal in order to be able to measure a difference of phase to calculate depth information. Because the theoretically required sinusoidal signal is not achievable in practice, the measured depth does not reflect the true distance but contains a systematic error. The systematic error is typically in the range of 5 cm, after any bias in the distance error has been removed. Motion artefacts are due to the TOF integration process, which involves the processing of 4 different measures to capture one single frame. If an object is moving, its contours may not coincide in such frames, resulting in a depth estimation error.

The FascinatE gesture recognition system will be equipped with a TOF camera because depth information appears to be of great importance to detect gestures robustly. As an example, the most advanced commercial system nowadays, Natal Project, now kown as Kinect, also relies on depth information.

Figure 5: TOF images. From left to right: Amplitude, Depth and Confidence

4.2 Requirements

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

4.2.1 Interaction design

It is 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 methods are appropriate to the content?
- How to design intuitive interfaces allowing user to engage?

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

Remaining 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

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 will be 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.
- 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 gave us a first impression of users' perspective on the use of FascinatE. Interesting points that need to be explored further are:

- Semi-professional mode in which somebody else (e.g. amateur producer) produces the content for end user (a group),
- 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),
- Social navigation including following what others are viewing, sharing our own current view and rating of user profiles,
- Collection of re-runs (either as producer choice or from what others were viewing/replaying) as a video stream to offer,
- Importance of replays - possibility to choose which moments we want to see again (as it is now, users are often unsatisfied with the editor's choice),
- Users might be 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*". One possible solution is the use of alarms from system or other users, and another is subscribing to somebody else's view (friend or semi-professional) instead of active viewing (going back to the first point).

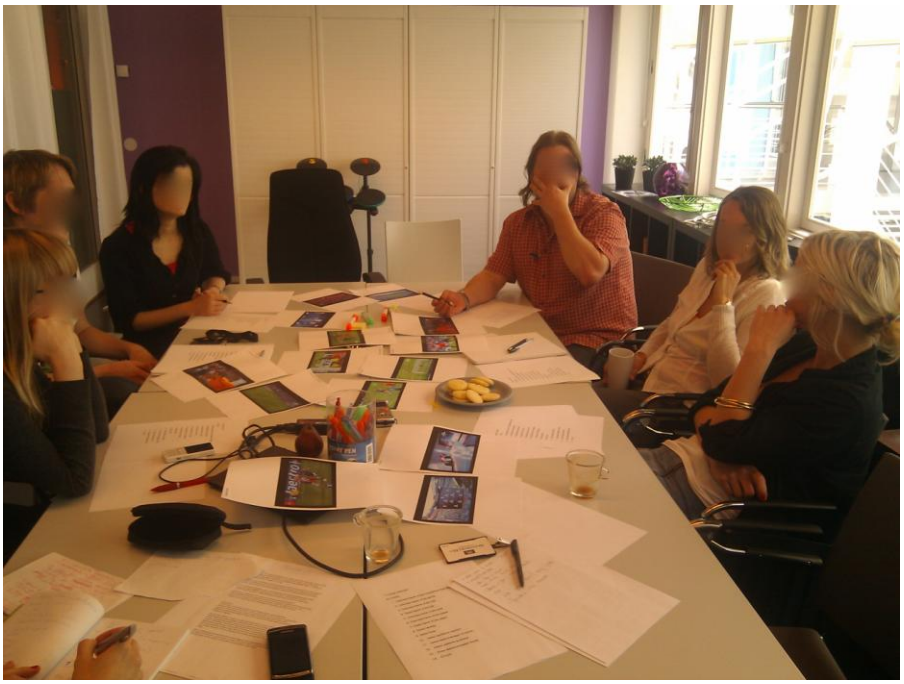


Figure 6: Snapshot from the pilot study

Common interaction mechanisms and metaphors across devices

This section covers high level requirements from the end-user perspective (both usability and usefulness), as learned from initial studies, literature and general interaction design practices.

The user interface in a FascinatE-based service should posses following qualities:

- Simple (e.g. using gestures, speech, actions),
- Intuitive (e.g. using familiar objects),
- Efficient; should not require learning, memorising or recalling commands [Vatavu, 2010] (e.g. multifunctional remotes),
- Should posses a dimension of fun that makes interaction process captivating [Vatavu, 2010],
- Multimodal (choosing modality, e.g. gestures, speech, or tactile, according to the content and environment),
- Non intrusive, by leaving the visual channel between user and the screen open,
- Consistent across similar control options,
- Map the form of input to the intended output as directly as possible,
- Ability to “undo” simple actions without accessing deep menu structures,
- Observability – feedback should be as soon as possible, and clear in its message,
- Design for interruption – not time sensitive,
- Non - modal (only one mode is needed, so that each command has only one meaning).

Users’ viewing preferences learned so far are stated next (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):

- TV viewers want to be entertained, get informed and relax; also they want to share information, discuss content, i.e. be social,
- Relaxed exploration instead of information seeking (starting with familiar content and continuing with browsing of relevant items) [Chorianopoulos, 2008],
- Group-based video streaming rather then individual – it enables shared experience, but also scalability

- Social TV: 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

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 cannot 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; snipe 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 them and the end user interface. They need to support the following:

1. **Switching between predefined video and audio streams.** 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. So far collected guidelines include:

- 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.
- Users' preferences of the (sport) event created with respect to event type, event venue, athletes' performance, nationality, team, or statistics.
- Event as combination of multiple streams ("picture in picture"). The question is "*How many streams people can handle?*" – the answer from the pilot study, not more then 2-3 streams, depending how much movement is on each of them.
- 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),

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 moving/zooming/panning cameras). Navigation assumes moving both in time and space. Some associated commands, and their descriptions are (a more detailed description is available in D5.1.1):

- Deciding on framing,
- Selecting objects,
- Scrolling X-Y,
- Select / Back,
- Zooming in/out,
- Slow motion replays,
- Duration of replay,
- Selecting windows (picture in picture),
- Separation foreground sound from background sound,
- Setting gain on the speech component(s) of the audio.

Gesture based interaction

It will 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:
  - Visual feedback of some gestures (e.g. zooming or raising volume) must be fast enough to allow fluid interaction
  - Other interactive commands (such as changing channels or dividing screen) are less restrictive in latency.
- Gesture recognition limited to a predefined set of gestures

4.2.2 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. 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

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.

4.3 Conclusion

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

As suggested, and motivated, in Section 3.1.2, the following should be beard 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 those testing will be done, 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 so far performed studies (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 user interaction may take a major role in future and innovative systems. Therefore, the FascinatE system will be partly controlled through a set of human gestures (see deliverable D5.1.1). This control-by-gesture will not completely replace other control devices, but will provide an alternative to traditional interaction methods such as remote controls, PCs, or touch screens.

5 Production Perspective

5.1 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 AV 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.

5.2 Production Systems Today

The AV 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.

5.2.1 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

AV 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 vision mixer. 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.

Depending on the scale of production, the vision mixer console might be operated either by the director or by a dedicated operator. The vision mixer sees what is being captured by each of the cameras all the time. It is the vision mixer 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 vision mixer will always be considering where their next shot, or next few shots, will be coming from. By switching between cameras (cutting) in different ways, the vision mixer has a lot of control over the feel of the programme.

Note that the term "vision mixer" 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 vision mixer and

the director's instruction. The audio engineer must have an appropriate mix ready to cut to when the vision mixers 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 Replay operators 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 replay operator'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 replay operator is in control of the video being transmitted to the viewers. On almost any sports event there will be a team of replay operators, with a master operator keeping track of what all the operators are doing.

The replay operator uses a device manufactured by EVS (The abbreviation has no useful formal definition) Each EVS console usually controls two cameras. 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.

The job of the replay operator requires a diverse range of skills. Operating the console itself requires skills similar to both a vision mixer 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 EVS operator 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 EVS operators 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.

5.3 Technology involved

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

5.3.1 TV production technology today

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 vision mixer. 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 vision mixer 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 vision mixer 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 vision mixer 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 7 shows an example of a vision mixer console in an outside broadcast vehicle.

Figure 7: Production gallery and vision mixer console

For the FascinatE project, the most important aspect of the operation of a vision mixer 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

The EVS operator controls a specialised tapeless recording machine. 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. EVS operators 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 MAXS or MAVS can only clip and playback highlights, but cannot create playlists [SportsTV, 2010].

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 vision mixer, 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.

5.4 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 1 to illustrate the departures from current methods which the FascinatE system will allow, or require. The three use-cases are:

- Use-case 1 – 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 2 – 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 layered scene representation 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 3 – The production staff put most of their efforts into creating scripts. They rely on the auto-assisted vision mixer to take care of the basics of tracking the game whilst no incidents are happening. The staff spend 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 1

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 2

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 3

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 1 and 2. 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 vision mixer's control over how individual camera feeds are assembled into a live broadcast. These range from absolute control in use-case 1 to very little control in use-case 3.

Vision Mixer under use-case 1

Here, the role of the vision mixer 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 vision mixer, 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 vision mixer.

Vision Mixer under use-case 2

As for the director, the vision mixer is operating as normal in use-case 2. The high-quality main programme must still be made, and managing the vision mixer console will take up most of this operator's time. Depending upon the production, the vision mixer 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 3

In use-case 3, the vision mixer 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, vision mixer, and replay operator. 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 1

In the first use-case, FascinatE is integrated into a standard production workflow. Framing selections from the layered scene representation 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 2

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 vision mixer 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 vision mixer or master operator.

Camera Operator under use-case 3

In use-case 3, 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 vision mixer 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.

5.4.1 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 1

As with the other roles, the audio engineer's job changes very little in use-case 1. 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 2

Use-case 2 introduces metadata and scripts. The FascinatE studio must pass scripts to the user. In the same way that the vision mixer 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 2 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 3

Sound is mixed in an automated way based on scripts and metadata in use-case 3. 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.

5.4.2 Replay operator

Replay operator under use-case 1

Under use-case 1, we can imagine two different ways that the replay operator'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 replay operator 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 replay operators 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 replay operator. This presents a significant departure from the existing situation. It allows the replay operator to compose their own camera shot based on the data coming from the omnicam, and in the associated Layered Scene Representation. For example, if one of the camera operators has missed part of a developing incident, the replay operator can fill in the necessary footage from the omnicam.

One way this could be made to work would be for the replay operator 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 replay operator 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 2

The same two possibilities exist within this use case as described under use-case 1, above. So from the point of view of making a basic programme, the replay operator 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 replay operators 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 replay operators 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 under use-case 3

Under use-case 3 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 replay operator changes the most. In a conventional TV production, the job of the replay operator 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 Replay operator'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. 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 replay operator (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 3.

5.5 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.
- An easy, standardised way of extracting information from the vision mixer is required.
- An easy, standardised way of extracting information from the audio mixer is required.
- 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 vision mixer 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 3, using the output of the semi-automated scripting system.
- A means should be provided for the omnicam 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. This method may be automated or semi-automated, supervised or unsupervised.
- A means should be provided for reviewing the end-user view which the system produces on a few candidate devices from the gallery.

6 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.

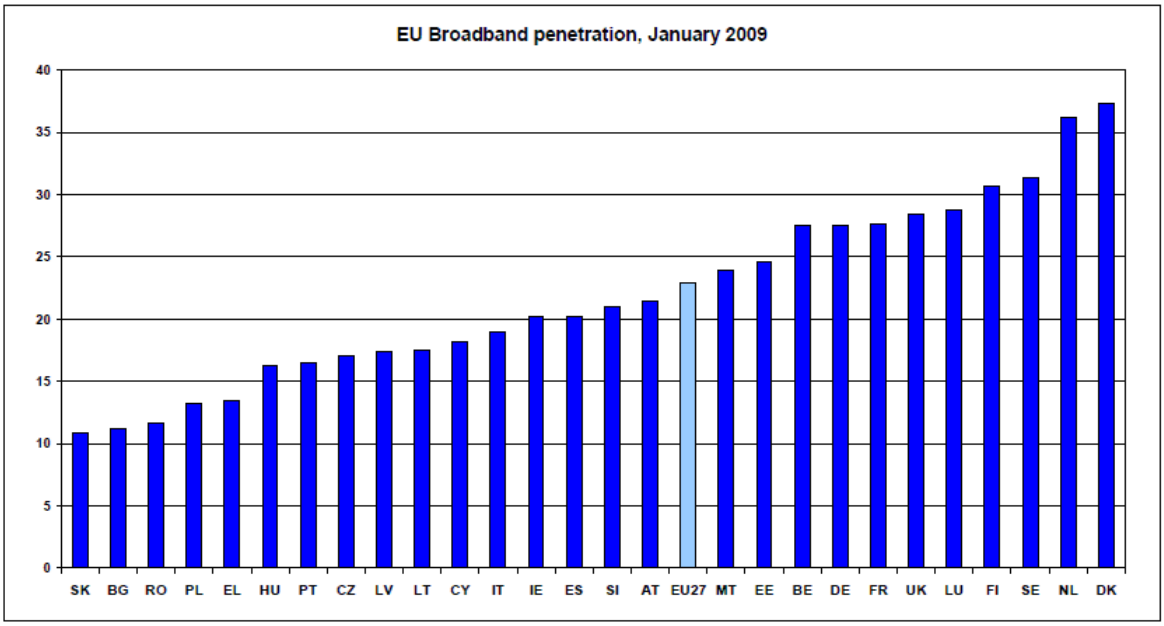
This chapter considers the FascinatE requirements from a networking perspective. For each of the three scenarios described in Section 2, requiments appropriate for the network role are discussed in Section 5.2. Additionally, high-level network functionality is described in Section 5.3.

Note: Processing hardware requirements are currently left blank (intentionally), as these depend highly on the expected terminal capabilities, the selected A/V formats and production capabilities. At the time of creation of this document version, these capabilities and selection of A/V formats were still unclear. The hardware requirements will be elaborated on in a future version of this document.

6.1 Network Capacity Today

6.1.1 Fixed delivery networks

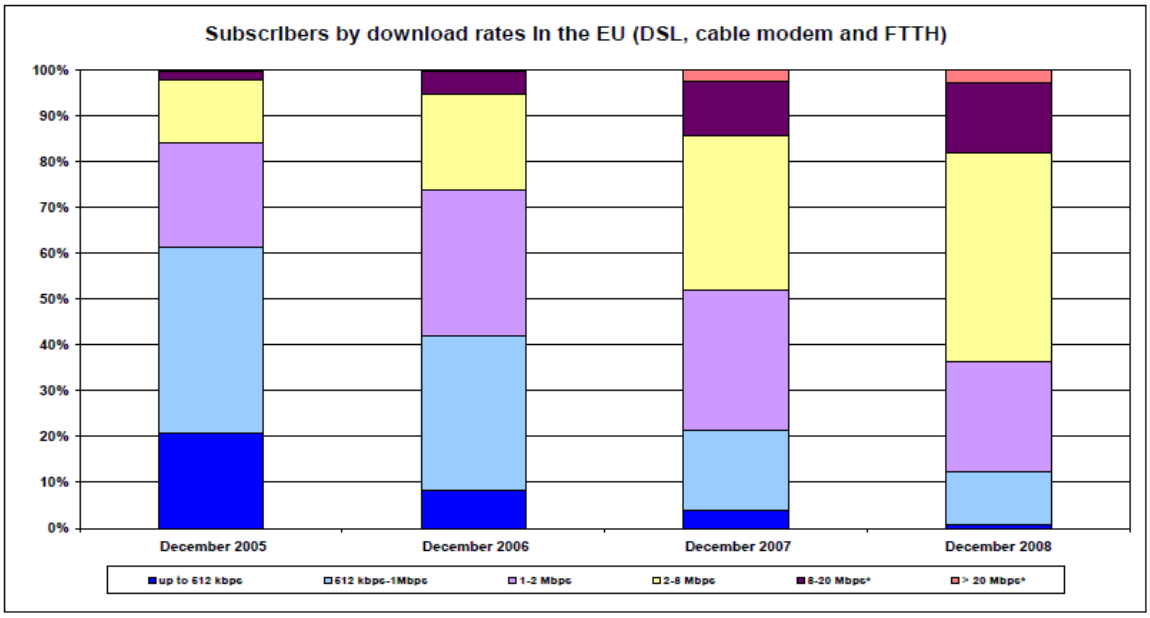
There has been a strong increase over the last years in broadband penetration in Europe. Broadband access means access via a ADSL, Cable or FTTx network where the bandwidth can vary from 256kbps to over 30Mbps. In Figure 8 is shown the penetration rate in the 27 European countries. The weighted average is around 23%.



Source: Communication Committee (CoCom)

Figure 8: Broadband access penetration rate in the EU

The average download speed of broadband subscriptions in the EU has greatly improved between 2004 and 2008. At the end of 2008 75% of EU broadband subscriptions are estimated to be associated to nominal speeds above 2 MB/s. Figure 9 shows that less than 5% did have access above the 30Mb/s. Today about one percent does have FTTH access.



Source: IDATE. 2008 data refer to 8-30 Mbps and >30 Mbps access lines

Figure 9: Subscribers by download rates in EU

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.

6.1.2 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 abreviated 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.

6.1.3 Core 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.

6.2 Delivery Network Requirements

The delivery network requirements 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 5.3.

- **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 5.3.

- **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.

6.2.1 Requirements for scenario 1

Scenario 1, the production-centric delivery chain, 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.

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.

Table 3 provides an overview of network technologies that we consider relevant for use in scenario 1.

Network type	Examples	Delivery Mode	Typical bandwidth	Typical latency
Digital TV broadcast	DVB-$C/T_1/T_2/S_1/S_2$	Broadcast	4-20 Mbps per TV Channel	250 ms
Multicast IP	Managed IPTV	Multicast	8-20 Mbps per TV Channel	2-10 ms In managed networks this value will be lower or less variable
Unicast IP	Best effort / internet Managed IPTV	Unicast	2 Mbps – 40 Mbps This depends on access line bandwidth	2-20 ms In managed networks this value will be lower / less variable
Mobile broadcast	DVB-H	Broadcast	3 – 10 Mbps	25 – 500 ms
Mobile IP	3G/LTE	Unicast	384 kbps – 2 Mbps / 100 Mbps	30 - 50 ms
Wireless IP	Wifi	(Broadcast) (Multicast) Unicast	11 - 150 Mbps	To be determined in D4.2.1.
Fibre IP	GPON, EPON	Broadcast Multicast Unicast	1 – 10 Gbps	To be determined in D4.2.1.

Table 3: Overview of network technologies relevant for scenario 1

Bandwidth

For scenario 1 a (limited) set of audiovisual views shall be simulcasted (i.e. offered at the same time) as rendered AV data. Production should make sure that each view is encoded at a pre-specified resolution and bitrate, that corresponds to most terminal display and processing capabilities. The network should be able to offer the required bandwidth for the requested streams. As a baseline we assume a bandwidth of BW_{view} provided by production, where the video is formatted as a regular HD or SD channel, and the audio is formatted as a regular stereo or multichannel audio representation. This leads to the following bandwidth requirements:

The production interface to the delivery network shall provide a bandwidth of at least $BW_{simulcast}$=N x BW_{view}, where N is the number of simultaneously offered views

The delivery network shall provide a bandwidth of at least $BW_{simulcast.}$

The terminal interface to the delivery network shall support a bandwidth of at least $BW_{simulcast}$ for networks providing broadcast delivery modes.

The terminal interface to the delivery network shall support a bandwidth of at least

$BW_{unicast} = BW_{view}$ for unicast and multicast delivery modes.

The range of bitrates for the compressed views (i.e. the bitrates of the media streams transmitted over the network) depends on several other factors:

- The terminal profiles and capabilities (terminal restrictions)
- The required fidelity, frame rate and resolution for media playback (quality requirements).
- The selection of audio and video coding formats (coding and transport requirements / restrictions).

As these factors are yet to be determined or specified, the bandwidth requirements can currently not be further specified, although the following remarks can be made:

- For unicast and multicast delivery modes the media streams can be transmitted independently over the access network. In a best-case scenario when an end-user switches views, the

transmission of the previous view is stopped (thereby freeing bandwidth resources) before the transmission of the new view is started (which then again uses the bandwidth resources). In a worst-case, both previous and new views are transmitted simultaneously for a limited amount of time.

- For a mobile terminal (with limited screen size) transmitting a panorama does not make a lot of sense. So whether all views are relevant for the target terminal remains to be seen.

Latency

The delivery network latency shall not exceed the delivery network latency of state of the art media delivery networks. In terms of responsiveness, the requirements are similar to channel/stream switching as in traditional TV and video services. ETSI TS 102 034 specifies solutions for fast channel changes that may be applicable. Latencies of existing delivery networks are further studied in D4.2.1

Formats and codecs

The AV coding and delivery formats for rendered views that are transported via the delivery network shall be consistent with those used in the existing delivery networks listed in Table 3. These include MPEG2-TS and direct RTP encapsulation, as well as HTTP-based streaming. MPEG2 and H.264 video coding is assumed, and MPEG-1 layer 2, HE-AAC or Dolby Digital audio coding. See ETSI TS 102 005 for an overview of common audio and video codecs for digital TV and video services.

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 that are yet to be determined. Furthermore, 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 session setup feedback channel, or a hybrid delivery setup where a panorama view is delivered via a broadcast channel and zoom views are transmitted via unicast IP.

Interactivity

For delivery modes requiring a session setup, the delivery network shall at least support interactivity for service discovery and selection, and content fetching via channel switching or stream selection. This relates to well-known session negotiation and session setup techniques, which are out of scope of this document.

6.2.2 Requirements for scenario 2

In scenario 2, the layered scene 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 layered scene. 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 the requirements described hereafter.

Bandwidth

As specified in D2.1.1, a layered scene representation consists of one or more camera clusters. For a camera cluster the data rates are given in Table 4.

	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 4: 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:

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

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 layered scene is then given as:

$$BW_{layeredscene} = H \times BW_{cameracluster} + I \times BW_{audio} + BW_{layeredscenemetada}$$

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 layered scene. 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 layered scene, $BW_{layeredscene}$, 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, SLOWMO 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 layered scene representation, at least at the production side. Therefore, even if lossy compression of the layered scene 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 AV media shall be provided by production to the delivery network in a layered scene 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 AV media shall be offered by the delivery network to the terminal in a layered scene 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.

6.2.3 Requirements for scenario 3

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

Bandwidth

Production shall support a network interface to deliver a complete single layered scene to the delivery network. It must be able to carry the maximum set of data generated within the production domain. As a baseline we assume a bandwidth of BW_{view} requested by the terminal, where the video is formatted as a regular HD or SD channel, and the audio is formatted as a regular stereo or multichannel audio representation.

The terminal interface to the delivery network shall support a bandwidth of at least $BW_{simulcast}$ for networks providing broadcast delivery modes.

The terminal interface to the delivery network shall support a bandwidth of at least

$BW_{unicast} = BW_{view}$ for unicast and multicast delivery modes.

The delivery network may support a variety of bandwidths, depending on its topology and functional architecture.

Latency

The latency requirements for scenario 3 are to be determined, e.g. by user studies and by analysing the capabilities of existing networks.

Formats and codecs

The AV media shall be provided by production to the delivery network in a layered scene 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 delivery scripts shall be offered by the delivery network to the terminal in a delivery script format as will be specified in deliverables D1.4.1 and D1.4.2. The AV delivery formats for rendered views are yet to be determined.

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

The script processing function in the delivery network should provide an interface to the terminal for receiving interactivity commands.

Processing

The requirements on processing resources are highly dependent on the use case considered, and in particular on the level of personalization and interactivity allowed.

- For example, in the case of local content production (see Section 2.2, use case 7), A/V processing resources can be centralized and the produced streams can then be delivered using typical networking resources (data forwarding, control …).
- In the case of more personalized and interactive services (see Section 2.2, use cases 8 and 9), the total A/V processing resources in the system need to scale with the amount of end-user requests. Also, the level of interactivity and the associated responsiveness requirements shall impose specific requirements on how close to the end-user processing functions (e.g. interactivity control and A/V rendering) needs to be located. Depending on the system design choices to be specified in D1.4.1, lower-level requirements on specific networking functionalities will be derived and lead to more detailed specification on the delivery network architecture to be described in D4.2.2.

Requirements on the scalability of the delivery network (e.g. the number of concurrent streams or concurrent script processing requests) will be provided in D1.1.2.

6.3 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.

6.3.1 High-level functional architecture

In Figure 10, a high-level diagram of the functional architecture of the delivery network is given. This diagram provides a reference of the functions that can be used in the delivery network. Each scenario and use case may require a different set of functions.

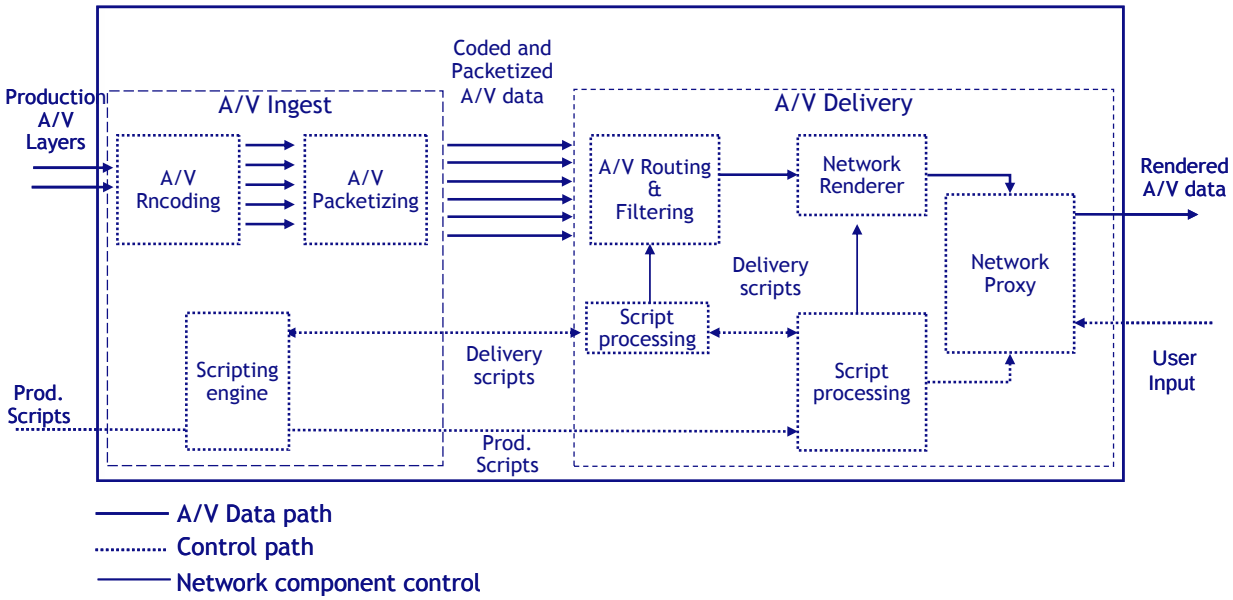


Figure 10: high-level diagram of the delivery network functional architecture

The high-level diagram consists of the following information:

1. High-level functionality. This reflects a logical grouping of functions that are likely to be combined to provide specific roles in the FascinatE delivery network.
    a. *A/V Ingest*. The A/V ingest is the interface between production and the delivery network, and is comparable to a traditional TV Head-end. The A/V ingest is responsible for ingesting production A/V layers into the delivery network.
    b. *A/V Delivery*. The A/V delivery is the network part where the flexible selection or composition of A/V streams or layers in the delivery network takes place, to allow for efficient distribution of the A/V media.
2. Information flows at the interfaces of these high level functional groups:
    a. At the production interface, input flows that are processed by the A/V Ingest consist of:
        i. *Production A/V layers*: They consist of the A/V data originating from production, in combination with the associated metadata describing the layered scene, as defined in D2.1.1

ii. *Production scripts*: These scripts are transmitted from production into the delivery network towards the location where the network renderer function is located.

b. At the terminal interface, information flows consist of:
   i. *User input and terminal response:* Interactivity commands as provided by an end user from the terminal to control a certain view of the layered scene.
   ii. *Rendered A/V data* (audiovisual data that provides a specific view of the layered scene representation) or *Pre-rendered A/V data* (audiovisual data that has been processed in the delivery network, e.g. filtered, but has not been rendered into a view) is output by the A/V Delivery block.

c. Between the A/V Ingest and Delivery, additional intermediate information flows may include:
   i. *Coded and Packetized A/V data*: It consists of a layered scene representation in a delivery distribution format.

d. *Delivery scripts*: They include scripting information that need to be exposed to the network elements. They are transmitted towards the last script processing function in the network

As stated before, a more complete description of the functional blocks of the system will be provided in Deliverable 1.4.1.

6.3.2 High-level functional architecture and scenarios

Each scenario and use case described in Chapter 1 may require a different set of functions. In both scenario 1 (production-centric) and scenario 2 (terminal-centric), the delivery network itself does not provide any of the above-mentioned functionality. The functionality for scenario 3 (provider-centric) consists of all functions given in the high-level functional architecture or a subset of these functions. Also, concrete usage of FascinatE technology will blend aspects of the three scenarios.

7 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 layered scene representation, 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 layered scene 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. The main questions to answer are:

- What should be included as content?
- How to access and manipulate content?
- What methods are appropriate to the content?
- How to design intuitive interfaces allowing users to engage?

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.
- In some of the environments, it is believed that gesture-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 human gestures (see deliverable D5.1.1). This control-by-gesture will not completely replace other control devices, but will provide an alternative to traditional interaction methods such as remote controls, PCs, or touch screens.

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. In this case, each scenario and use case described in Chapter 1 may require a different set of functions:

- In both scenario 1 (production-centric) and scenario 2 (terminal-centric), the delivery network itself does not provide any of the above-mentioned functionality. See the sections on production and end-user interaction, respectively, for high-level functional architectures in those domains.
- The functionality for scenario 3 (network-centric) consists of all functions given in the high-level functional architecture or a subset of these functions.
- Concrete usage of FascinatE technology will blend aspects of the three scenarios.

We propose to take into account all these questions and conclusions when designing the final FascinatE system architecture. Deliverables D1.4.1 and D1.4.2 will take this information in order to provide a complete system definition.

Many details of the requirements discussed in this document will become clearer and better-defined as the project progresses and will be verified when demonstrators are trialled. An updated requirements document will therefore be produced at the end of the project (D1.1.2, in Month 42).

8 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.

[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.

[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*

[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).

[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.

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

[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

[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

9 Glossary

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

