IBC Special Issue, September 2011

FascinatE Newsletter

http://www.fascinate-project.eu/

Ref. Ares(2011)1005901 - 22/09/2011

Inside this issue:

Welcome from the Project Coordinator

I am happy to introduce you to FascinatE! Since February 2010 our consortium consisting of eleven partners spread over Europe has been working to implement our idea of the ultra high resolution interactive television service of the future. A full system comprising appropriate capturing and analysis technology, networking components and various terminal devices is being implemented. The capturing side uses as base an ultra high-definition panorama augmented with additional cameras and also 3D and ambient audio. This information is compiled into a layered scene representation together with metadata. The networking components will be able to interpret the layered scene representation and will adapt the content depending on the type of service or the capabilities of the target device. In respect to terminal devices the whole range from high resolution, immersive displays for a bigger audience, home environments with TV sets down to mobile devices used by individuals are covered. Interaction methods tailored to the device, e.g. hand gestures for big devices and touch gestures for the smaller ones are being investigated.

During our first test shoot we had the opportunity to capture a full range of content of the English Premier League soccer game Chelsea vs. Wolverhampton Wanderers (article in first issue of the newsletter).

We will showcase our results at different stages of the project. The first demonstrations will be given here at IBC 2011.

In order to be kept up-to-date on the developments of the project please visit www.fascinate-project.eu or follow us on twitter "@Fascinate_Prjct".

Georg Thallinger,

Project Coordinator

The FascinatE Consortium

Special points of interest:

- Demonstrations highlighting key progress in project
- Real time stitching of OmniCam high resolution panoramic video
- Object based 3D audio system
- Gesture control of user defined scene
- Networks for real time video navigation

The FascinatE consortium at our kick off meeting in Graz, February 2010

Live Ultra-High Resolution Panoramic Video

In the advanced format-agnostic production framework developed by FascinatE, real-time acquisition of ultra-high definition panoramic video is essential because this enables either the production side or the end user to select interesting viewing directions independently of what the camera operator shot. The omni-directional camera developed by Fraunhofer HHI uses six HD cameras mounted on a mirror-rig to capture an ultra-high definition 180° panoramic video with a resolution of 7000 x 1920 pixels. The major challenge here is blending and stitching these six camera views to form one single panoramic image. Even in panoramic video, the viewer is quite sensitive to incorrect stitching and blending if moving objects pass the border between views. Hence, a set of geometrical and photometric corrections have to be applied to each camera view as well as the careful blending of image borders. Fraunhofer HHI is presenting its Real-time Stitching Engine capturing six HD cameras at 25 FPS and stitching them together to a panoramic video at 7k x 2k pixel resolution. The real-time panoramic video can be observed at a large screen next to the omni-directional camera.

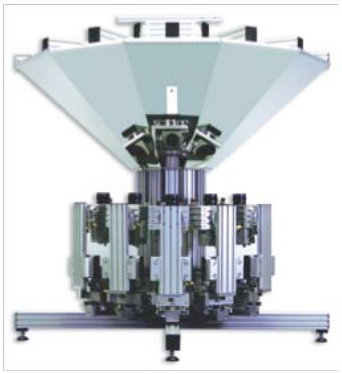

Figure 1. 2D OMNICAM

3D OmniCam

The Fraunhofer Heinrich Hertz Institute HHI has developed a scalable, mirror-based, multi-camera rig that can be used for capturing immersive high-resolution 3D video panoramas. With its special mechanical and optical features it enables an optimal arrangement of multiple HD stereo cameras that solves the fundamental dilemma between parallax-free stitching of video panoramas on the one hand and the parallax needed for 3D stereo reproduction on the other. The rig is scalable in increments of 24 degrees and supports acquisition of live 3D video panoramas of up to 360 degrees with a maximal resolution of about 15.000 × 2.000 pel for each stereo view. The prototype of this new 3D camera can be seen at the FascinatE booth.

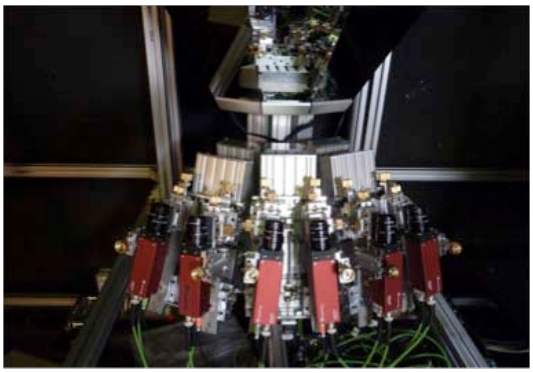

Figure 2. 3D OMNICAM

Person Tracking & Production Scripting

Person Tracking

Joanneum Research presents a demo for person detection and tracking in high-resolution panoramic video streams, obtained from a panoramic camera stitching video streams from six HD tiles. The tracking algorithm has to detect and track persons over six static and rectified HD image-sequences from the OmniCam. Instead of using the ultra-high definition image, each video tile is separately analysed by different workstations to enable real-time analysis. The AV content analysis uses a CUDA accelerated feature point tracker, a blob detector, and a CUDA HOG person detector, which are used for region tracking in each of the tiles before fusing the results for the entire panorama. The results of the person and blob detector for each image of the different image sequences yield the regions of detected persons for further processing. Furthermore, person IDs are linked to the appropriate combined regions with their corresponding feature points. The tracking system is demonstrated on a single PC with appropriate graphics board, processing a full HD stream. Results are shown in figure 3.

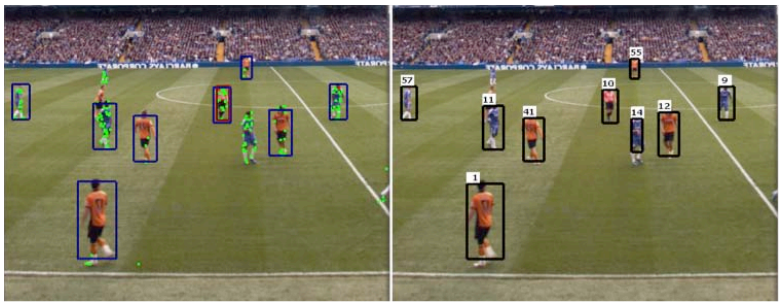

Figure 3 – The left image shows detected person regions and tracked feature points. The resulting tracked persons with their IDs are shown on the right.

Production Scripting Engine

The Production Scripting Engine (PSE) is responsible for decision making on content selection. The key feature is to automatically select a suitable area within the OmniCam panorama image, in addition to cuts between different broadcast cameras. Selection behaviour is based on pragmatic (cover most interesting actions) and cinematographic (ensure basic aesthetic principles) rules. In some cases, this is not fully automatic but involves a human in the loop, a production team member deciding between prepared options. The PSE is a distributed component with at least one instance at the production site and one at the terminal end. The output of the PSE is called “script”, which consists of a combination of content selection options and decisions, renderer instructions, user interface options etc. Scripts are passed to subsequent PSE components from the production site towards the terminal, where final instructions are given to a device-specific renderer.

Creating a format agnostic interactive broadcast experience poses some interesting challenges to the partners from Technicolor and The University of Salford who are responsible for the audio aspects of FascinatE. Of chief importance is the need to record the given audio scene in such a way that the content can be rendered on any reproduction system at the user end and can update depending on the dynamic viewing point. This demands a paradigm shift from how audio has been traditionally recorded for broadcast. Instead of broadcasting to match a specific hardware set up such as stereo, 5.1, 7.1 etc we adopt an object orientated approach which can be reproduced on any system. The audio scene is considered to be made up of a set of audio objects (point sources with a specific location) and an ambient sound field contribution. The challenge at the recording side therefore is to record the sound field as well as the content and location of the audio objects at the scene. This often involves using different or adapted recording techniques to what is considered standard practice in the broadcast industry. Ideally each sound source would be individually close miked and tracked in space, however in many cases (such as the first FascinatE test shoot at a football match) this is not possible and the content and position of the audio objects needs to be derived by processing the signals from several microphones near to the sources. The ambient sound field can also be recorded in such a way that it can be updated to match a given viewing position for example, using ambisonic microphones such as the Eigenmike® or the SoundField® microphone which record the 3 dimensional sound field at a given point. With audio objects and sound field accurately recorded it is possible to encode these sources in various sound field representations such as high order ambisonics (HOA) or wave field synthesis (WFS) which can in turn be decoded for any reproduction system from stereo (e.g. on mobile devices) to true 3D sound using HOA with height in large public installations. As the user pans around the visual scene it is possible to both rotate and translate this sound field to match the new viewing position based on camera pan and zoom. On the rendering side, it is important that the audio updates accurately with the updating view and that it matches the user preferences based on a combination of production choices and user input. FascinatE bridges the gap between passive viewer and active participant scenarios; current television broadcasts could be considered as passive viewing, where the audio remains stationary regardless of the camera position; conversely active participant viewing is more akin to a video game scenario where the audio updates completely with the viewing position. Of interest for FascinatE is which of these viewing paradigms the user subscribes to when navigating round the scene. Future work will therefore be centred on not only recording the audio scene such that the content is format agnostic but also on determining how best to render the audio to match user preferences.

"Instead of broadcasting to match a specific hardware set up such as stereo, 5.1, 7.1 etc we adopt an object orientated approach which can be reproduced on any system"

Figure 4. The Eigenmike® is used by FascinatE for high order ambisonics

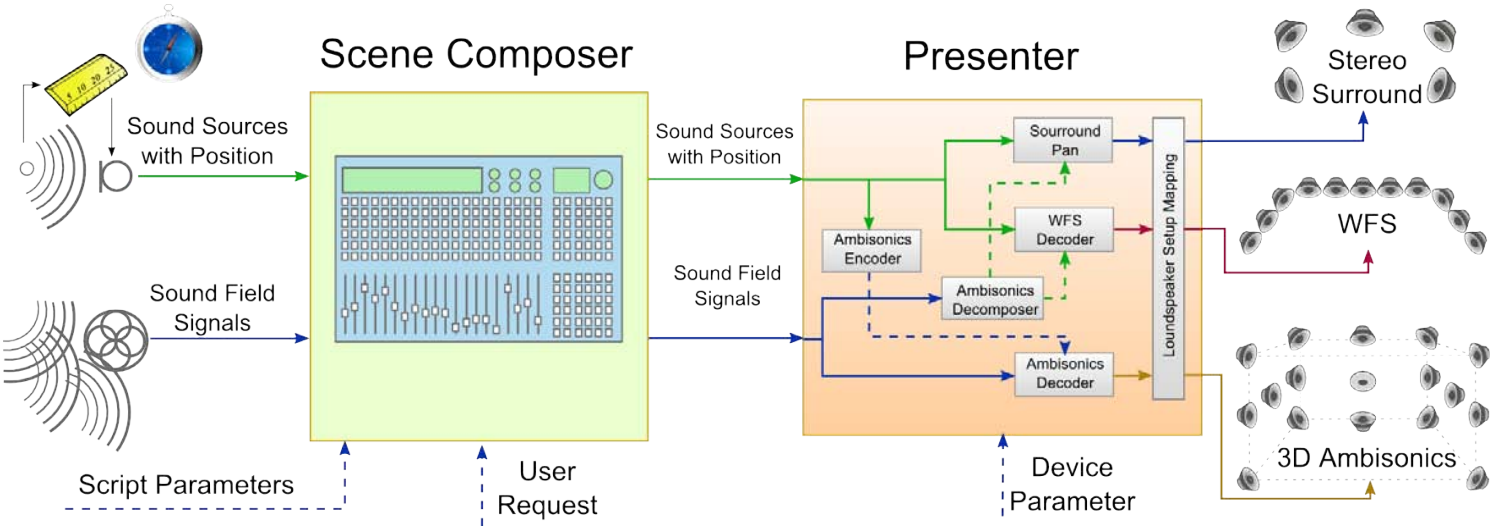

Figure 5. System diagram for FascinatE audio

FascinatE Rendering Node with ROI Zoom and Gesture Control

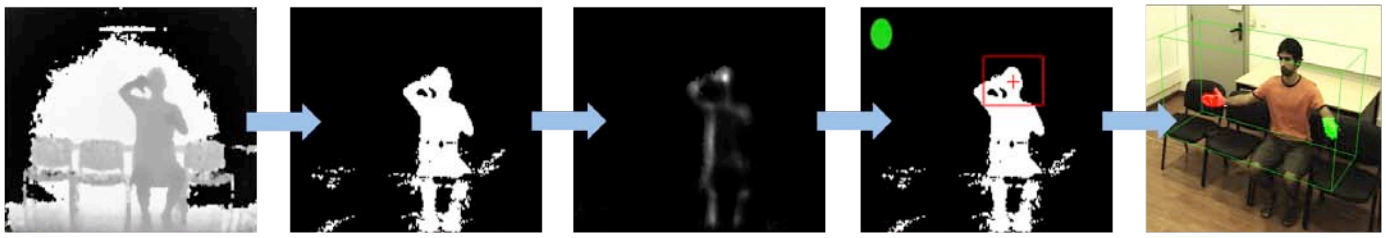

Figure 6: Head position estimation example

A first terminal prototype, shown in the Fascinate stand at IBC 2011, demonstrates the capability of navigating within content captured by FascinatE sensors, such as panoramic videos. The demonstrator employs gesture recognition to simplify the interaction between the terminal and the end user. It is also focused in a home scenario where the end user interacts with the rendered content on a high-definition TV set.

The Universitat Politècnica de Catalunya (UPC) developed for this purpose a fast and robust head and hand tracking algorithm using depth information from a range sensor, allowing interactive and immersive applications. This functionality is used to control a real time rendering platform developed by Technicolor. This platform is configurable by scripts and provides Virtual Camera navigation with pan, tilt and zoom commands.

"An applied XML based scripting mechanism controls and scales visual rendering performed on camera clusters offering multiple regions of interest"

In order to interpret user gestures as means to navigate within a panorama, hands and heads are tracked by exploiting depth estimation. This process includes modeling templates for heads and calculating an elliptical matching score. The template is resized depending on the distance the person is placed. For a given search zone a matching score provides head position probabilities and confidence values for position estimations.

For tracking the hands to understand the performed gestures, a workspace is defined as a 3D box, placed in relation to the detected head position. Within this 3D box, hands are detected by merging and filtering samples with similar size and depth information.

Finally, an empirical law relating the area of a surface in the image with its real world counterpart is obtained. A distinction of open or closed hands is obtained by segmenting the area of the detected hand. An example of all these steps is shown in Figure 6.

The variety of available end terminals require nowadays a format agnostic production to prepare the content best suited to all. FascinatE terminals and services will supply interactive, personalized visual perspective to enrich the user experience. Content navigation like pan, tilt and zoom allows the user a real immersive experience beyond simple channel switching. The scalable architecture of the rendering platform developed for FascinatE allows applications of different target terminals such as home theaters or smart phones.

An applied XML based scripting mechanism controls and scales visual rendering performed on camera clusters offering multiple regions of interest (see Figure 7). This supports automation of workflows and optimization of delivery channels. The visual rendering of such layered scenes into personalized perspectives on end user screens are performed by transformation from the circular panorama onto flat surfaces (fig 8). Additional effort is spend to place graphical elements for user information in relation to the selected region of interest and the display surface used for presentation.

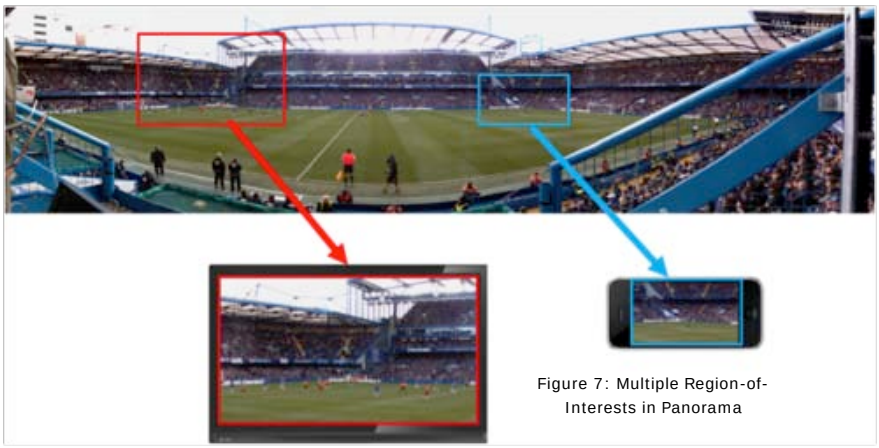

Figure 7: Multiple Region-of-Interests in Panorama

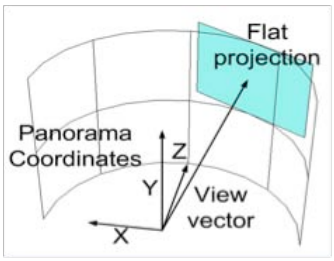

Figure 8 rendering of personalized perspectives by transformation from the circular panorama onto flat surfaces

In conclusion, the demonstrator presented at IBC 2011 is able to perform a fast (68fps) and robust hand and head tracking with an error of less than 6cm. The resulting smooth hand trajectories can be used for further gesture classification and analysis. This technology is applied to a real time capable terminal platform for pan, tilt and zoom navigation within a panoramic scene. An easy personalization by gestures is complemented by scripting support offering perspective options such as prepared region of interests.

The transmission of the FascinatE layered scene representation represents a major challenge for a delivery network, essentially in terms of bandwidth and processing requirements. As an example, the live delivery of the current FascinatE audio/video test material requires an uncompressed data rate of around 16Gbps. FascinatE aims to deliver immersive video services to a large range of terminals from high end audio/video set-ups with fibre connectivity, to low-powered mobile devices. To deliver an immersive and interactive media experience to any device in a scalable manner, the project has focused so far on the development of Audio/Video Proxies. Their role is to perform some of the media processing tasks on behalf of a terminal so as to reduce the processing and bandwidth requirements for the terminal hardware.

The two following prototypes are demonstrated at IBC 2011:

1. Network Proxy for Real-Time Video Navigation:

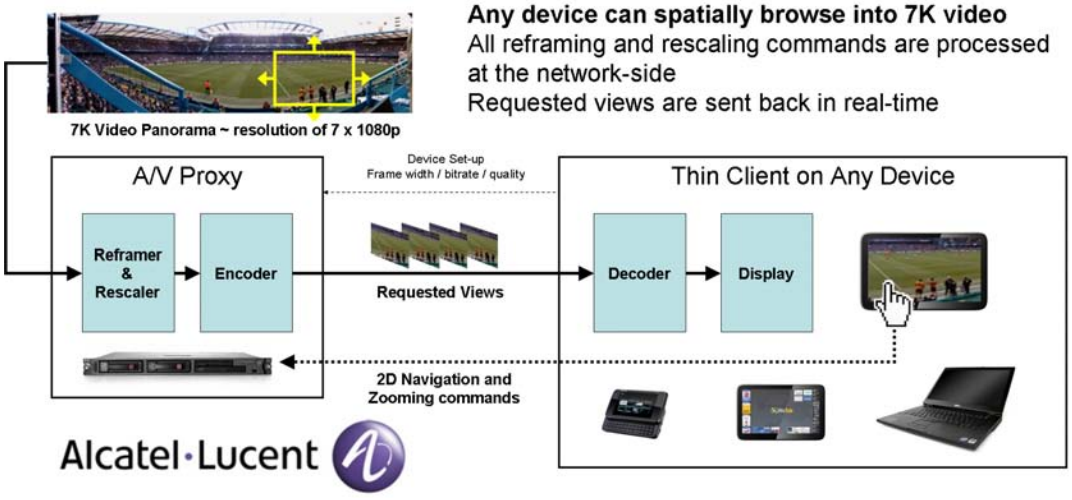

Figure 9 Overview of the Network Proxy for Real-Time Video Navigation

We focus in this demonstrator on a specific case where the proxy is able to process in real-time end-user requests for navigating a very high resolution video panorama (Figure 9).

The proxy has access to a 7k x 2k video panorama and sends a reframed and compressed video to each client device at an appropriate resolution and bitrate. In this demonstrator, the end-user can directly navigate the 7k panoramic video using a tablet or a mobile phone equipped with a touchscreen. The user commands are translated into a stream of 2D translation and zooming commands that are sent upstream to the proxy. The corresponding reframing, rescaling and coding processes are executed in real-time for each client device. The proxy then delivers a compressed video stream containing the requested views, which only requires a standard decoding step before display. With this approach, Ultra-HD content can be watched interactively in a natural manner, even on a low-power and small-display device.

2. Spatial segmented delivery of immersive media:

Spatial segmentation is used as a method to efficiently deliver parts of the 7k x 2k video panorama to devices which are not capable of displaying the entire resolution at once, such as smartphones and tablets. The general concept behind spatial segmentation is to spatially split each video frame into several tiles. The video frames corresponding to the various tiles are encoded independently and stored separate as a new video stream, or spatial segment. We focus in this demonstrator on the case where the A/V proxy only requests a subset of segments, based on the ROI selected by the user(s) for which it performs the spatial segment re-assembly. In the prototype, spatial segments are transported using a protocol similar to HTTP adaptive streaming. The spatial segments are then reassembled by the proxy. The navigation can be controlled on a mobile device, such as a tablet or smart phone. Functionality is further increased by using multiple resolution layers, which allow for smoother zooming (Figure 10).

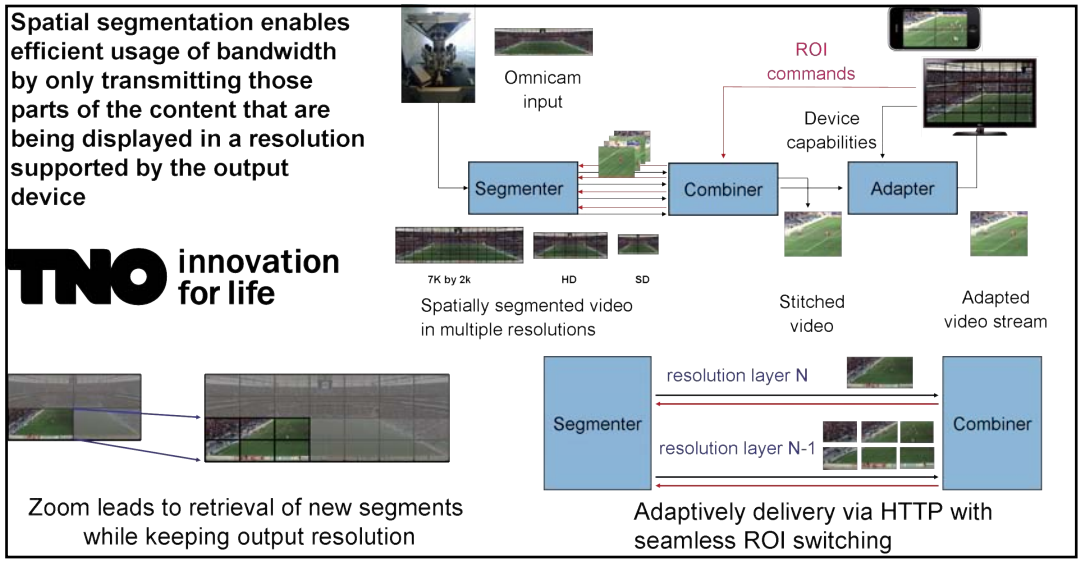

Figure 10 Spatial segmented delivery of immersive media - Overview of the Network Proxy for Real-Time Video Navigation

ALEXA M – the new capturing device for the OmniCam

The Alexa camera gained an overwhelming reputation in motion picture and broadcast productions for its outstanding quality and ease of operation. The OmniCam would greatly benefit from the quality of this device; but a much smaller form factor is required. Hence we reduced weight and size of the Alexa camera. At IBC the first working prototypes of Alexa M will be presented. In this modular camera version the sensor head is separated form the electronics back end, connected through an ultra fast fibre connection. The sensor head will be integrated into the next generation of the OmniCam for even more fascinating and immersive panoramic pictures.

Figure 11. The Arri Alexa M

A 3D model of the test bed scenario

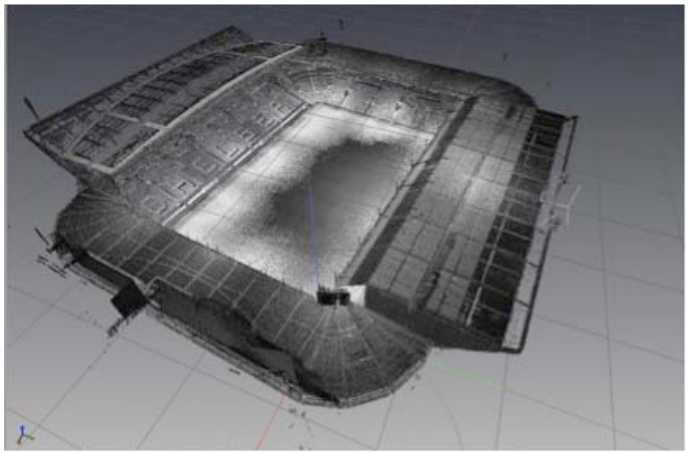

Figure 12. Laser scanned model of Stamford Bridge Stadium

3D laser scanning has proven its usefulness in many applications like civil engineering, architecture and archaeology. In the FascinatE system it is essential to know the precise 3D coordinates of all cameras and microphones. When the viewer of a FascinatE event selects an area of interest within the scenario, the displayed images and the presented audio signals should both focus on the same spot or area. A 3D laser scanner not only measures those coordinates, but also generates a static 3D model of the whole scenario enabling easier means of camera calibration and matching.

Use of broadcast cameras to support the Omnicam

It is impractical to obtain a tightly-zoomed high-definition close-up simply by selecting a window from the Omnicam image. For example, to obtain the same resolution as is available from an HD broadcast camera with a horizontal lens angle of 5 degrees would need the 180-degree panorama to have a horizontal resolution of about 70k pixels (36 times that of an HD image). It therefore makes sense to use conventional broadcast cameras to provide close-ups of key areas of the scene, as used in conventional coverage. To allow the user to zoom smoothly from a wide shot from the Omnicam into a region-of-interest covered by a broadcast camera, it is necessary to ensure that the images from the two cameras can be matched - both spatially and in terms of colorimetry. Figure 13 compares an image from an HD broadcast camera (right) to the corresponding portion of the Omnicam image (left), after background alignment and colour matching, showing the potential gain in resolution (BBC).

Figure 13 compares an image from an HD broadcast camera (right) to the corresponding portion of the Omnicam image (left), after background alignment and colour matching, showing the potential gain in resolution.

The FascinatE Project and Consortium

FascinatE is an EU-funded project involving a group of 11 partners from across Europe. FascinatE stands for: Format-Agnostic SCript-based INterAcTive Experience and is looking at broadcasting live events to give the viewer a more interactive experience no matter what device they are using the view the broadcast.

The FascinatE project is developing a system to allow end-users to interactively view and navigate around an ultra-high resolution video panorama showing a live event, with the accompanying audio automatically changing to match the selected view. The output will be adapted to their particular kind of device, covering anything from a mobile handset to an immersive panoramic display. At the production side, this requires the development of new audio and video capture systems, and scripting systems to control the shot framing options presented to the viewer. Intelligent networks with processing components will be needed to repurpose the content to suit different device types and framing selections, and user terminals supporting innovative interaction methods will be needed to allow viewers to control and display the content.

Contact Details and Project Office

Georg Thallinger
JOANNEUM RESEARCH
Graz, Austria
georg.thallinger@joanneum.at

Newsletter Editor

Ben Shirley
University of Salford
Salford, UK
b.g.shirley@salford.ac.uk

FascinatE is funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248138