ICoSOLE: IMMERSIVE COVERAGE OF SPATIALLY OUTSPREAD LIVE EVENTS

R. Bauwens[2], R. Grandl[3], D. Marston[5], M. Matton[2], C. Pike[5], M. Wijnants[4], W. Bailer[1], G. Thallinger[1]

[1]JOANNEUM RESEARCH, AT; [2]VRT, BE; [3]Bitmovin, AT; [4]iMinds, BE, [5]BBC, UK

ABSTRACT

ICoSOLE researches and develops approaches for the integration of content from professional and consumer capture devices, including omnidirectional and multi-source equipment, in order to provide an immersive media coverage of live events spread-out over larger regions.

1. PROJECT OVERVIEW

ICoSOLE[1] aims at developing a platform that enables users to experience live events which are spatially spread out, such as festivals (e.g. Dranouter in Belgium, Glastonbury in the UK), parades, marathons or bike races, in an immersive way by combining high-quality spatial video and audio and user generated content. The project develops a platform for a context-adapted hybrid broadcast-Internet service, providing efficient tools for capture, production and distribution of audiovisual content captured by a heterogeneous set of devices spread over the event site.

The approach uses a variety of sensors, ranging from mobile consumer devices over professional broadcast capture equipment to panoramic and/or free-viewpoint video and spatial audio. Methods for streaming live high-quality audiovisual content from mobile capture devices to content acquisition, processing and editing services will be developed.

In order to combine the heterogeneous capture sources, ICoSOLE researches and develops approaches for integration of content from professional and consumer capture devices, including mobile (and moving) sensors, based on metadata and content analysis. Methods for fusing visual and audio information into a format agnostic data representation will be developed, which enable rendering video and audio for virtual viewer/listener positions.

ICoSOLE develops efficient tools for media production professionals to select, configure and review the content sources being used. These tools capture, extract and annotate metadata during the production process and integrate this metadata throughout the entire production chain to the end user. Content is provided via broadcast, enhanced by additional content transported via broadband and novel interaction possibilities for second screen and web consumption. The content is also provided in an adapted form to mobile devices.

[1] The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 610370. ICoSOLE ("Immersive Coverage of Spatially Outspread Live Events", http://www.icosole.eu).

2. CONTENT CAPTURE

The capture of video, audio and metadata with professional equipment, as well as mobile consumer devices is addressed. Content generated by user devices, like smart-phones or tablets, is analysed in terms of quality before it is forwarded to a central storage.

2.1 Professional Content Capture

A proof-of-concept novel multi-angle high resolution panoramic video capture system has been developed and demonstrated to the public. The system captures an event from several vantage points, by means of high resolution panoramic or omni-directional camera heads, controlled from a single or multiple computers. The novel capture system concept aims at reducing capture costs and will be suited in the context of web broadcasting and the bottom end of the TV broadcasting market. It will be equally well suited for delayed broadcasting, as well as for live remote broadcasting.

2.2 User Generated Content Capture

A range of mobile consumer devices (tablets, smart-phones, etc.) has been analysed in terms of their recording performance as well as storage and transmission capabilities. It was investigated how metadata gathered by build-in sensors correlate to the quality of captured video content.

In order to ensure the quality of the output, content which does not reach the desired quality levels is discarded. Algorithms for sharpness, noise and over-/underexposure detection taking into account the limitations of mobile devices have been implemented. Apps for capturing content along with sensor data, including the functionality for visual quality analysis, have been developed. Immediate feedback based on the evaluation is given to the user, as it can be seen

978-1-4799-7082-7/15/$31.00 ©2015 IEEE

in Figure 1. The predefined thresholds can be specified by the applications configuration functionality and will be refined on findings from our field-trials.
The apps also provide means to upload the captured audiovisual content together with metadata. Especially in live scenarios difficulties can arise from congested or low capacity networks. A trade-off between the number of users simultaneously uploading content and streaming quality has to be made. Different common streaming approaches from UGC devices, including RTSP and RTMP, have been evaluated and first implementations for initial experiments have been developed. A promising novel upload approach, which runs on top of the well-known HTTP protocol, will also be considered: The captured essence, including metadata, is stored in short segments of a few seconds long, locally on the device, which runs a webserver. Clients (other users or a processing unit on site) will be able to fetch the segments via HTTP GET requests in parallel.

3. CONTENT PROCESSING AND AUTHORING

Preliminary studies have revealed that for the execution of different algorithms, a scalable generic audio processing platform will be advantageous. Such a platform is being developed and will be used as part of the ICoSOLE system, as well as the open source GStreamer framework. Associated algorithms include accelerated audio processing, e.g. for advanced real-time audio signal analysis, by using OpenCL based acceleration technologies.
Visual matching approaches, for example using compact feature descriptors, have been surveyed for their use to improve localisation and orientation of user-contributed content, refining information captured from sensors. First work on the analysis of UGC audio recordings has been performed, as well as preliminary work on content selection rules.
Software libraries have been implemented that cover general audio processing handling object-based audio. They also include the initial development of a binaural production system. These libraries have been developed for use within the BBC IP Studio framework that will allow streaming of object-based audio in a production environment.

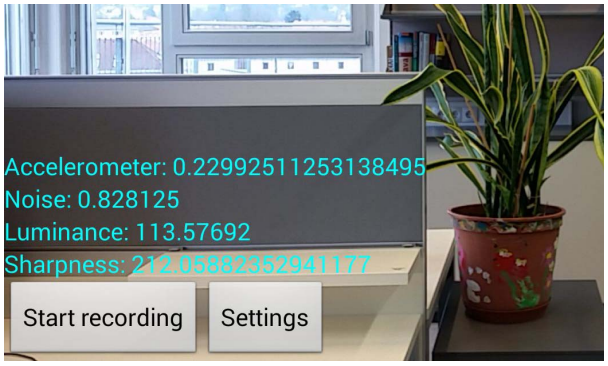

Figure 1: Screenshot of the Android capture application.

Based on the use cases of ICoSOLE, key concepts have been identified for live editing resulting in requirements for a control user-interface suitable to deal with ICoSOLE's live setup. The challenge here is the processing of huge amounts of user generated content, which is crucial for a successful and flexible live edit.

4. CONTENT DISTRIBUTION AND CONSUMPTION

A first prototype implementation of a new media streaming and processing pipeline, named *TOABeam*, has started. The pipeline allows demultiplexing and decoding media essences from files and network streams, as well as encoding and multiplexing of uncompressed media to destination files and streams. After researching several media frameworks, the GStreamer framework has been chosen, as it provides a solid foundation for the developments.
Preliminary work on playback technologies for different mobile platforms, e.g. 3D audio rendering technologies has been done. A suitable concept will most likely use conventional headphones in conjunction with the binaural rendering technology, which will feature an immersive 3D sound experience. Web browser based audio rendering tools (using the WebAudio API) for binaural audio, including vector base amplitude panning (VBAP) rendering, have been developed.
Interaction with and navigation of the event is enhanced by developments called the Wall of Moments and Venue Explorer. The Wall of Moments provides the user with a mosaic of user generated video streams from friends' smartphones from which they can choose the content they are interested in. Venue Explorer provides a navigable view of the whole event's venue and allows the user to zoom in and select the stage or activity of interest. Many of the underlying concepts with the Wall of Moments and Venue Explorer have been researched in [3] on festival viewer experiences.
Some of these developments are described in more detail in the following sub-sections.

4.1 Fast MPEG-DASH Stream Switching

In digital media streaming, there is typically a delay associated with switching or *zapping* from one media signal to another. This startup latency is primarily caused by the need to buffer significant amounts of media data at client side in order to make the media playback resilient to small disruptions in the network throughput. A secondary, less severe source of startup delay is to be found in the compressed nature of the digital media content: the received content first needs to be decompressed before it can be rendered. Slow startup can be frustrating for the end-user, in the sense that a typical user will want to be able to switch between individual media streams as swiftly as possible. In effect, unresponsive zapping behavior could have the very

undesirable consequence that the user loses his interest in the content.

To accelerate the time-to-first-picture or -sample in MPEG-DASH-based distribution environments, a content-prefetching framework has been implemented [2]. Besides fetching MPEG-DASH segments pertaining to the *active* streams – by which we refer to the streams that are currently being rendered by the client – the framework makes sure that (a subset of) the other existing elementary streams are available locally before they actually become needed for rendering purposes. In other words, pre-fetched content is not actually rendered by the client until the consumer effectively switches to the corresponding stream. As at this stage the desired content is already available locally (typically at an intermediate quality level though), the stream switch will be near-instantaneous. Of course, this benefit comes at the cost of either increased bandwidth consumption or quality reduction of the active stream(s) at unmodified bandwidth usage. As such, the total amount of streams to pre-fetch simultaneously is an important design decision that typically will need to be made on a per-application basis. Also note that the efficiency of the pre-fetching framework can be optimized by resorting to content recommendation algorithms to decide on the subset of content feeds that are most likely to be zapped to by the user.

4.2 WebRTC-powered Remote ODV Rendering

A WebRTC-powered remote rendering platform for omni-directional video (ODV) or so-called 360° video has been developed and validated [3]. The immersive traits of ODV footage are maximized when the footage is projected correctly (e.g., rendered on the "inside" of a 3D sphere or cylinder). This projection task can be computationally complex and in addition mandates the availability of 3D rendering functionality at client side. Especially in Web browser environments, access to graphics APIs is not universally supported across platforms. The remote rendering platform therefore offloads the 3D graphics responsibilities involved in ODV playback to a back-end server. Client-side interactions with regard to ODV viewport manipulation are relayed to the server, where they are translated into appropriate 3D scene modification instructions. Subsequently, only the currently visible spatial segment of the ODV scene (i.e., the current ODV viewport) is transported to the client using RTP embedded in a WebRTC session. Besides reducing the computational overhead and eliminating 3D graphics processing requirements at client side, the remote rendering platform hence also improves the bandwidth efficiency by removing the need to stream entire ODV frames. Experimental results have indicated that the latency introduced by the remote rendering platform is negligible; the propagation delay remains the dominant latency factor, as is the case in traditional streaming. As a result, delay reduction measures that are applicable in traditional streaming (e.g., caching) are directly exploitable in the proposed remote rendering solution as well.

4.3 The Wall of Moments

Immersivity of content (consumption) can be achieved in many ways. As well as immersive audio and video, the sensation of immersivity can also be enhanced by connecting a remote user to the event in a personalized way. This can be done by showing user generated content (UGC) videos of friends that are at the event, and are sharing "Moments" they experience. The main advantage is that they increase the feeling of being there for a remote user. The major drawback is the poor audio (and/or video) quality of UGC at large events.

The Wall of Moments intends to tackle this problem, by smoothly combining UGC and high-quality professional content. When new UGC content becomes available, videos are aligned with the general event timeline by synchronising them with professional time-coded content using visual or audio matching technologies. Figure 2 shows the main screen of the Wall of Moments prototype. UGC videos are shown and played back in a personalized browseable wall.

When selecting a particular moment, it is played back in full screen as shown in Figure 3. It also includes synchronized professional content (picture-in-picture) and relevant social media interactions. When the UGC moment is finished, the prototype automatically continues to play the professional content. On the top of Figure 3, a timeline is present showing different available moments synchronised with the main professional content. Different adaptive video streaming client implementations have been evaluated having a DASH based playout in mind. In order to test a variety of clients also some preparatory work on mobile clients, including possible hardware and software requirements, has been made.

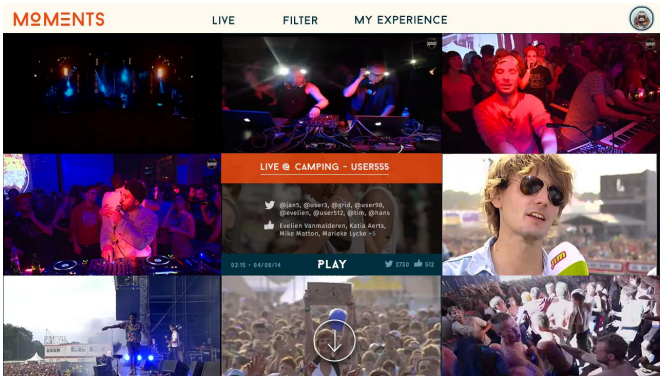

Figure 2: Wall of Moments.

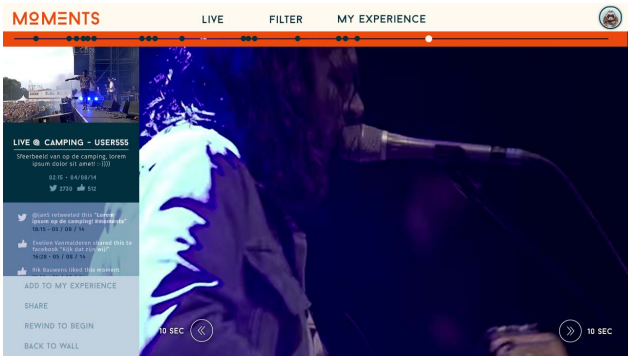

Figure 3: Synchronized content.

A HTML5/JS based player has been developed with homogeneous styling, which gives a consistent "look and feel" among different platforms, as well as an intuitive user interface for interaction possibilities.

4.4 Venue Explorer

The Venue Explorer [4] is designed as a content navigation interface for large events with several simultaneous sub-events. The aim is to enhance the feeling of presence when viewing an event remotely by giving the context of the spatial and temporal relationship between events and allowing interactive audiovisual exploration of the scene, as opposed to list-based navigation of event highlights. It is based on a visual overview of the event, which can either be a map or live video. The view can be navigated with pan and zoom controls.

Venue Explorer is a web application built using HTML5 features. Pan and zoom of video is achieved with DASH adaptive streaming, where an HD resolution video tile is selected from a 4k overview shot to achieve the best resolution for the current zoom level. A dynamic sound scape is created based on the available content in the current viewport, using the Web Audio API. Graphical overlays are used to provide information about the events to the user. A scene description is used to dynamically update content and metadata parameters for the event scene. The Venue Explorer design can also be applied to omnidirectional video and audio sources.

4.5 Immersive Object-Based Audio

3D spatial audio technologies are applied in the ICoSOLE project to enhance immersion. The object-based audio concept is applied, where elements of the audio scene are distributed separately with metadata that describes how these objects fit into the scene. This metadata can vary over time. A rendering process is then applied on the user-side to reconstruct this scene. This allows a single content source to be distributed to heterogeneous user devices, such as a tablet device with headphones or a smart TV with a connected sound bar. The rendering process is aware of the sound reproduction system and so can adapt the scene rendering to most accurately recreate the intended experience. The object-based approach brings other potential benefits, such as user personalization, interactivity, and responsive adaptation to the environment (such as environmentally-aware dynamic range compression [5]).

Rendering of 3D spatial audio for loudspeakers and headphones from an object-based representation has been implemented in JavaScript (using the Web Audio API) and C++. This will allow web browsers and native applications to integrate 3D audio.

Dynamic binaural synthesis from head-related transfer function (HRTF) data, using real-time input from a head orientation tracker, is used to create immersive headphone rendering [6]. Generalised amplitude panning for 3D loudspeaker arrays is achieved using the vector-base amplitude panning (VBAP) technique [7].

5. CONCLUSION AND FUTURE WORK

The project is currently finishing its first development cycle and will test the technology at a music festival in Summer 2015. The feedback gathered in this field trial will be the input for developing and improved system with more comprehensive functionality.

6. REFERENCES

[1] P. Quax, M. Wijnants, G. Rovelo Ruiz, W. Lamotte, J. Claes and J-F. Macq, “An Optimized Adaptive Streaming Framework for Interactive Immersive Video Experiences”, to appear in *Proc. IEEE BMSB*, 2015.

[2] P. Quax, J. Liesenborgs, A. Barzan, M. Croonen, W. Lamotte, B. Vankeirsbilck, B. Dhoedt, T. Kimpe, K. Pattyn and M. McLin, “Remote Rendering Solutions Using Web Technologies”, in *MMTAP*, February 2015, pp. 1-28.

[3] R. Velt, S. Benford, S. Reeves, M. Evans, M. Glancy, P. Stenton, “Towards an Extended Festival Viewing Experience”, in *Proceedings of TVX 2015,* 2015.

[4] M. Paradis, B. Gregory-Clarke and F. Melchior, “VenueExplorer, Object-Based Interactive Audio for Live Events” in *Proceedings of the Web Audio Conference*, 2015.

[5] A. Mason, M. Paradis, “Adaptive, Personalised "In Browser" Audio Compression” in *Proc. of the Web Audio Conference*, 2015.

[6] Jot, J.-M., Larcher, V., & Warusfel, O. “Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony”. In *Proceedings of 98th Convention of the Audio Engineering Society*, 1995.

[7] Pulkki, V. “Virtual Sound Source Positioning Using Vector Base Amplitude Panning”. Journal of the Audio Engineering Society, 45(6), 456–466, 1997.