2016

MIXED REALITY TECHNOLOGIES FOR IMMERSIVE INTERACTIVE BROADCAST

O. Schreer[1], W. Waizenegger[1], W. Fernando[2], H. Kodikara Arachchi[2],
A. Oehme[3], A. Smolic[4], B. Yargicoglu[5], A. Akman[5], U. Curjel[6]

[1]Fraunhofer HHI, Germany; [2]University of Surrey, UK; [3]HFC Human-Factors-Consult, Germany; [4]Disney Research Zurich, Switzerland; [5]Argela, Turkey; [6]SRF, Switzerland

ABSTRACT

Up until now, TV has been a one-to-many proposition apart from a few exceptions. The TV Stations produced and packaged their shows and the consumers had to tune in at a specific time to watch their favourite show. However, new technologies are changing the way we watch and produce television programs. For example, viewers often use second screen applications and are engaged in lively discussions via social media channels while watching TV. Nevertheless, immediate live interaction with broadcast media is still not possible. In this paper, the latest results of the European funded project ACTION-TV, which is developing novel forms of user interaction based on advanced Computer Vision and Mixed-Reality technologies, are presented. The aim of this research project is to let viewers actively participate in pre-produced live-action television shows. This expands the horizon of the interactive television concept towards engaging television shows. The paper explains the concept, challenges and solutions resulting in a first prototype real-time demonstrator for novel interactive TV services.

INTRODUCTION

ACTION-TV [1] proposes an innovative mode of user interaction for broadcasting to relax the rigid and passive nature of present broadcasting ecosystems. It has two key aims: (i) a group of users can take part in TV shows providing a sense of immersion into the show and seamless natural engagement with the content; (ii) users are encouraged to use TV shows as a means of social engagement while keeping themselves and their talents more visible across social circles. These aims will be achieved by developing an advanced digital media access and delivery platform that enables augmenting traditional audio-visual broadcasts with novel interactivity elements. In Figure 1, the ACTION-TV media cloud is depicted, where different social groups can actively participate in a TV show. By focusing on a game show use case, home viewers can take part in a quiz duel and a dancing competition, while their performance is captured and mixed with a live-action broadcast stream and distributed to their social community. This is achieved by 3D capture of the home viewer and insertion of visual appealing 3D model into the scene as a “virtual candidate”. An artistic sketch of the concept is depicted in Figure 2.

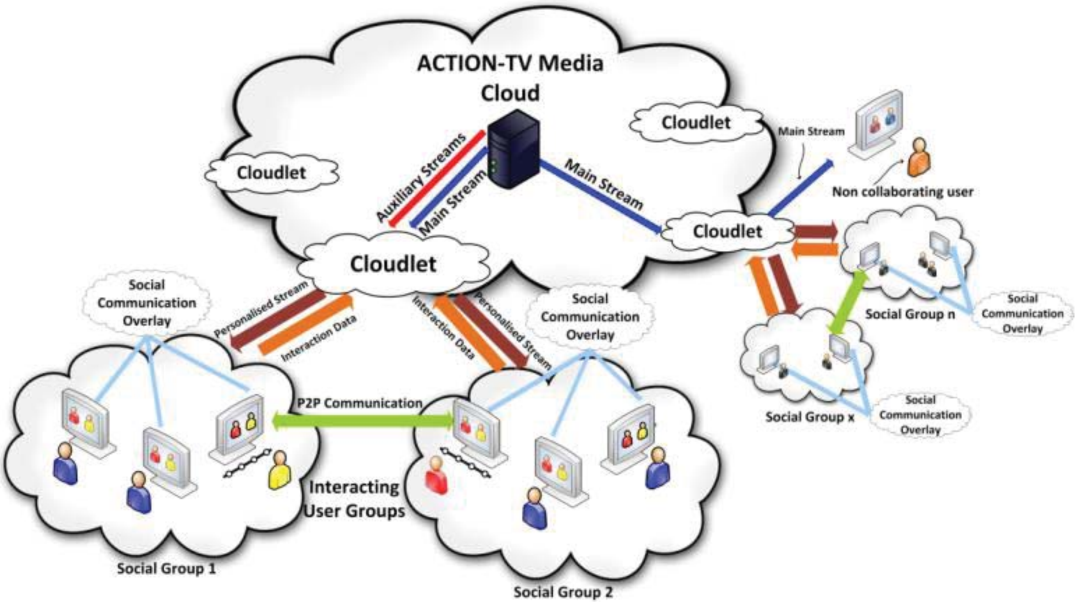

Figure 1 – ACTION-TV media cloud

The ability to interact with a program has been used to engage viewers and increase interest (and revenues) in television shows. The easiest form of interaction is a telephone conversation. Thanks to the internet, it's now easy for people to not only be heard but also appear on television (skype, streaming etc.). Furthermore, state-of-the-art virtual reality technologies have been reported in user engagement in television shows. In particular, HUMANOID [2] and ReV-TV [3] projects proposed to use avatar animation technology to bring in members of the audience into TV shows. The collaborative research project REVERIE [4] developed technologies for personalizing the avatar to resemble the actual participant's appearance through photorealistic human modelling. However, these technologies have been developed for live television shows where an actual two way conversation is possible. The possibility of interacting with a pre-recorded program is still very limited. ACTION-TV addresses this issue.

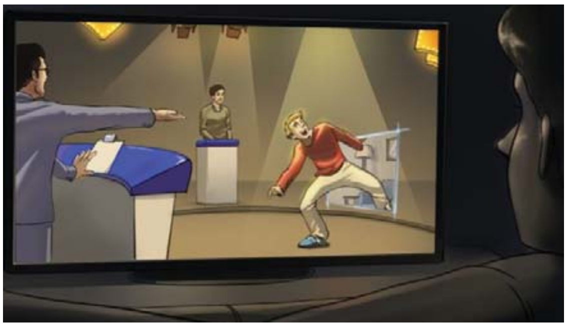

Figure 2 – Artistic sketch of the concept

The broadcast industry is moving from an SDI workflow towards an IP based workflow. ACTION-TV capitalizes on the IPTV workflow and has developed technologies which will change the way the users can interact with television programs.

1. 3D Capture and real-time rendering of the "home candidate"

2. Ability to realistically merge a "home candidate" in a pre-recorded television program

3. Ability for a "home candidate" to interact in real time (answer questions, dance) and be a part of a pre-recorded television show.

4. Ability for a social group to watch "their" home candidate in the program and communicate amongst one another via text or video.

ACTION-TV allows a home viewer to participate as a 3D avatar in a television studio. The interactive viewer moves from a passive state to an active one. This form of interaction is new and not yet seen in current television programming. The pre-recorded content is no longer tied to a specific time slot. A specific group can see the “transmission” at a time of their choice. The television experience becomes a big personalized game. Overall, this new concept leads to several technological challenges such as real-time 3D video processing, mixing various video streams and cloud-based computing. In the next sections, the ACTION-TV demonstrator, the system architecture, the content production workflow, the user model generation, real-time transmission over media clouds and user evaluation results for quality of experience based on a prototype demonstrator are presented.

ACTION-TV DEMONSTRATOR

The ACTION-TV demonstrator shows, how an active collaborating user at home can interact with a pre-taped television program. The Action-Duel (game show) use case was developed for several reasons. Most viewers know how game shows work and the basic concept is easily understandable. A host asks questions and the contestants (in the studio and at home) must choose the correct answer. In this scenario, a studio candidate plays against an interactive collaborating user (home candidate). The interactive collaborating user sits in his living room and is captured with a suitable camera setup. Additionally a tablet or computer is needed to input the correct answer. The ACTION-TV system adapts the captured home user to the camera shots and angles used in the pre-taped program. This allows the candidate to be shown up close or in a two shot, and merged in a realistic way with the pre-taped studio segments. The static nature of the format, the contestants don’t move around much in the studio and are usually shown from the waist up, allows us to concentrate on a realistic merging of the avatar with the television studio. During the quiz round, the social group can interact via text and video to comment on the progress of the home candidate. It’s fun to see one’s friends on TV, and to be able to communicate with the other members of one’s social group. The second half of the Action-Duel scenario is a dance round, which is more complex to realise than the quiz round. Here, two active collaborating users are virtually placed into the television studio. A professional choreographer and dancer demonstrates a dance routine. Both candidates have the chance to show how well they can follow the routine. Again, the social group can watch the “dance off” and interact via text and video. One of the challenges is to allow the active collaborating user at home to “see” the choreographer as if they were in the studio. The dance round requires a full body capture process with quick movements. The dancers at home are then virtually placed in the studio and dance along with the choreographer. The Action-Duel scenario is designed to showcase the interaction possibilities that the ACTION-TV technology can provide. The future use of the technology is not limited to the use case shown. The use case for the technology were developed and presented to inspire producers or content creators to come up with their own new and innovative forms of interactive programming based on the ACTION-TV technology.

SYSTEM ARCHITECTURE

The end-to-end ACTION-TV system consists of a hybrid cloud and terminal based architecture taking computational demand and delay constraints into account. The ACTION-TV system is built upon five main modules. The **Content Producer** uploads the pre-produced distribution package, which consists of a main stream, a set of auxiliary streams, and related metadata to the **Main Server** in the cloud. At the pre-scheduled time, the Main Server streams content to the broadcast audience. It also delivers auxiliary content for collaborating social groups, when it is requested. The **Cloudlet Server** is a cloud data centre, which is located closer to the users from networking point of view [5]. The Cloudlet Server performs computing tasks to produce the personalized media stream for passive-collaborating users of a particular user group based on the active-collaborating users' actions and to deliver the respective video streams to the specific users. The location of the data centre can be determined dynamically based on the geographic distribution of the members of the social group so that the communication delay is minimized [6]. The home domain hosts the **User Terminal** and **2nd Screen**. The User Terminal acts as set-top-box for *Non-collaborating users,* or receives a personalized stream generated by the Cloudlet Server for *Passive-collaborating users*. *Active-collaborating users* receive the personalized content directly generated at the user terminal in order to minimise delay. When there is more than one active-collaborating user in a social group, P2P links are established in order to pass information between them without noticeable delay. The User Terminal also captures user actions and distributes them to the relevant modules for further processing. It analyses multi-view video received from sensing devices for generating reduced mesh model of the user, which is used for inserting users into the personalized video stream. The 2nd Screen is available for collaborating users during the show. It will be used for interactive functionalities such as answering questions in a quiz show and voting. The proposed ACTION-TV signal processing architecture is shown in Figure 3.

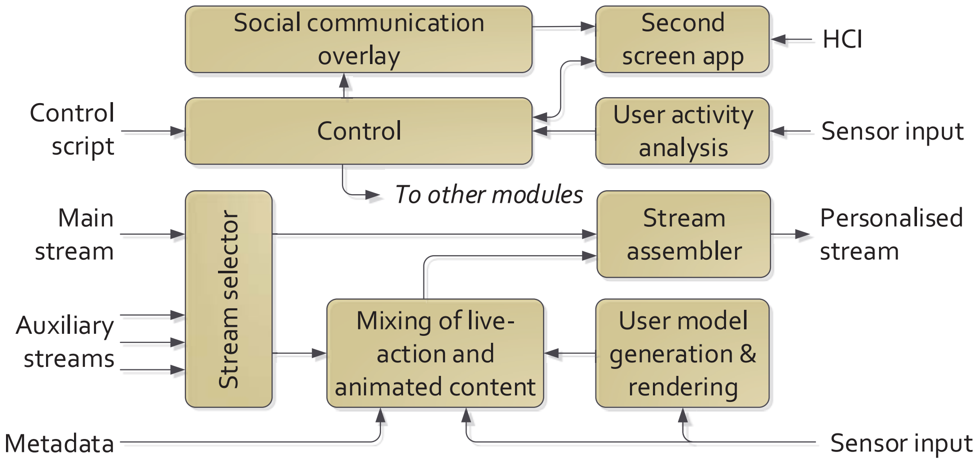

Figure 3 – Simplified functional block diagram

The control module orchestrates the interaction session by signalling the other modules to perform their functions at the correct time and in the correct sequence. The XML based control script describes the sequence of actions associated with the show. This includes the start and end times of rendering the user into the scene, type of user activity to be looked for, correct stream associated with each potential user activity, place and camera angle for rendering the user, etc. The control module, decodes the control script and issues control messages to relevant modules to manage the delivery and personalization functionalities. The social communication overlay functionality has a client-server based architecture. Server side resides in the main server and serves the conferencing rooms for all social groups. The client side resides in the 2nd screen display, which runs on a web browser, and lets the user join the conference session that is dedicated to its social group. The user activity analysis determines the type of activity performed by the user during a time period and checks whether it agrees with the expected activity, as indicated in the control script [7]. For example, if the activity required for a particular time interval is a dance act, the user activity analysis should determine whether the user performs the requested dance act. The feedback is used by the control unit to decide the subsequent video clip to be displayed for the user.

The stream selector performs the actions of a program editor. It switches between the main and auxiliary streams at interaction points according to the instructions received from the control module. Some of the selected clips are fed to the mixing module, which mixes live action and animated content, while the others go directly to the stream assembler. The stream assembler splices clips coming from two different paths to produce the personalized video. The mixing function is typically performed in the pixel domain. Therefore, any video clips, where the user is inserted, have to be decoded. However, the stream selector and assembler can operate in the compressed domain, as state-of-the-art video coding standards provide efficient methods for stream splitting and splicing operations [8].

CONTENT PRODUCTION WORKFLOW

Similar to conventional TV productions, the production of ACTION-TV content involves several phases resulting in the ACTION-TV distribution package used for broadcasting. In the **Pre-Production** phase, all camera shootings are planned. This includes the preparation of shooting scripts or story boards, assembling a production crew (director, camera men, gaffer, assistants, etc.), casting contestants or actors, choosing a location for the shooting and coordinating all aspects relevant for the production. In the **Production and Live Editing** phase, the AV content is recorded according to the shooting script or story board. In contrast to a conventional TV production, it is necessary to additionally produce auxiliary streams, and capturing all relevant data, which will facilitate the subsequent post-production phases. Auxiliary streams represent AV content, which will be presented to ACTION-TV users for interaction and engagement. The recording of AV content is conducted simultaneously with several conventional broadcast cameras. All cameras are gen-locked to easily determine frame correspondences in post-production. Content for the main stream and auxiliary streams is recorded from several camera angles, and a preliminary version of the main stream is created live during recoding by cutting between different cameras. In the final offline editing, the camera angle for some shots is adjusted to improve the pace of the presented content in general and to align the pace of the main and auxiliary streams. In addition, the length of the auxiliary streams is adapted to enable a coherent switching from auxiliary streams to the main stream or between auxiliary streams. The next step is the **Control Script Generation**. A control script is executed by a story manager, which orchestrates the distribution system to enable ACTION-TV users to interact and engage in auxiliary video streams while maintaining the logical coherence of the story. It contains information on semantically important events, intervals when externally triggered events are expected, and definitions of update and output functions, which update the story state and trigger the execution of an action, respectively. This information is provided by a story author or operator, who has an understanding of the overall story and the causal relationships between main and auxiliary streams. Figure 4 shows an illustration of the temporal order of the main and personalized auxiliary streams as they could be seen by an ACTION-TV user or his/her social group, when experiencing the quiz show round of the ACTION-TV demonstration scenario, Action Duel.

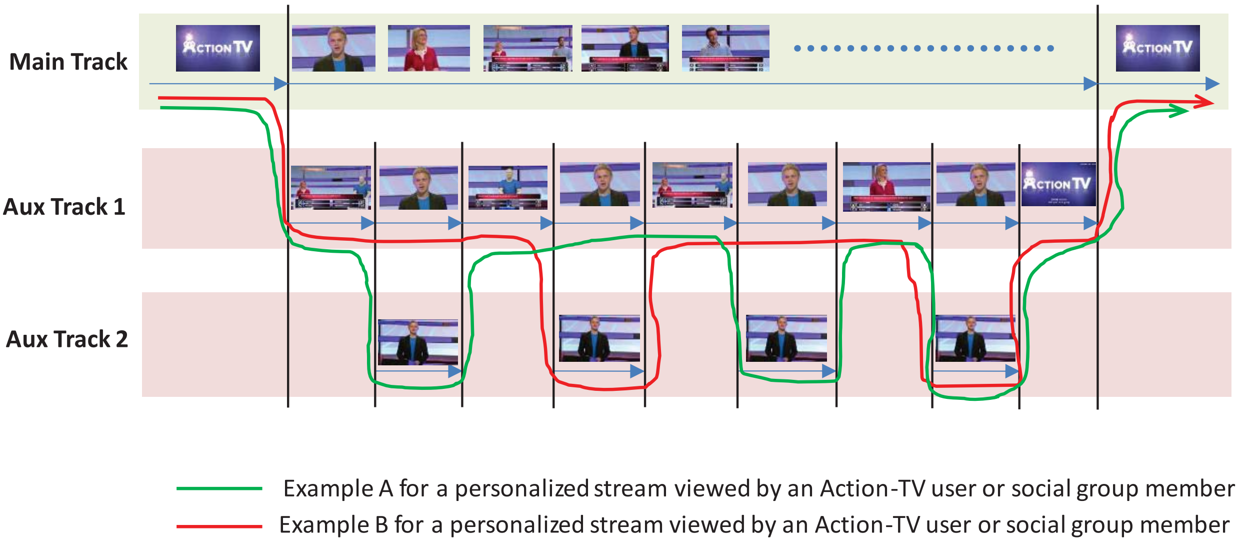
Main Track
Aux Track 1
Aux Track 2
Example A for a personalized stream viewed by an Action-TV user or social group member
Example B for a personalized stream viewed by an Action-TV user or social group member

Figure 4 – Personalized auxiliary streams shown to an active user or social group member.

Aux tracks presented in the figure contain personalized streams temporally ordered, while black bars indicate at which point in time a switch to a personalized auxiliary stream of another track may happen. After each question, the host comments the answers given by the contestants, which is initiated by a dedicated event. Which comments are given, depends on the action message emitted by the output function. In this round, the output function makes use of the current score stored in the story state to decide which action message to emit, while the story state is mainly updated based on the correctness of the answers provided by the contestants. After answering the four questions, Action Duel continues by switching to the main track, which presents the main stream as it is already seen by conventional TV viewers.

Metadata Generation is required to personalize auxiliary streams, i.e. to integrate the visual appearance of ACTION-TV users and their performances in auxiliary AV content. The appearance of the ACTION-TV user is represented by a 3D avatar which, in dependence of the ACTION-TV story, could be a fictional 3D character preselected by the user or it could be a photorealistic 3D model. The ACTION-TV user is then captured in the home environment (e.g. with a motion capture sensor or a multi camera system) and replicated by the avatar who is integrated in the personalized auxiliary stream. The appearance and performance parameters are transmitted to the cloud, where the personalization of auxiliary streams is conducted with the associated metadata. The metadata contains information to allow for plausible rendering of the avatar or 3D model in the AV content. This is geometric information (intrinsic and extrinsic camera parameters used for shooting the AV content, pose of the ground plane in 3D space), occlusion information to place the avatar also behind foreground objects and information about light sources to cast shadows. Therefore, the scene geometry and the viewing perspective of the broadcast camera need to be known. By using available features in the scene such as the ACTION-TV logo, position and orientation can be estimated to render a virtual candidate from the correct perspective (see Figure 5).

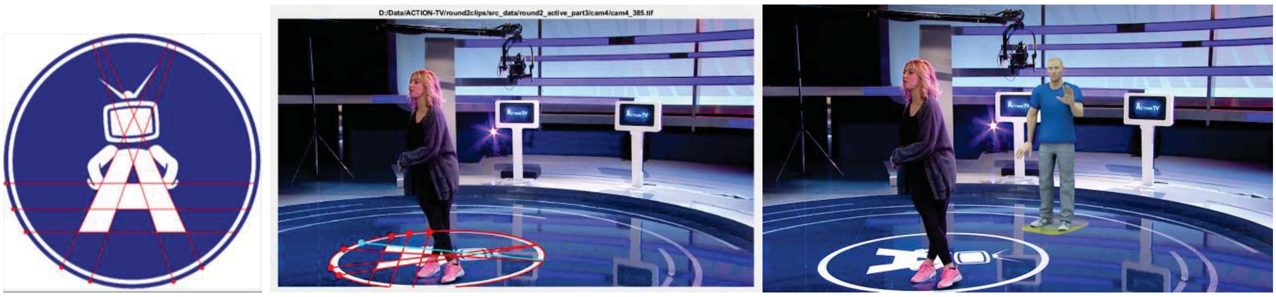
D:/Data/ACTION-TV/round2clips/src_data/round2_active_part3/cam4/cam4_385.tif
ACTION TV
ACTION TV

Figure 5 – ACTION-TV logo (left), detected features in the camera view (middle) and perspective correct rendering of virtual candidate (right)

Finally, the **Packaging** of all the generated data, i.e. the main and auxiliary streams, the control script and the metadata, is performed for distribution. To efficiently represent AV files they are encoded with HEVC, while the control script and metadata, which are natively represented in XML format, are binarized.

USER MODEL GENERATION AND REAL TIME RENDERING

The aim is to capture the home user with multiple cameras, create a photo-realistic 3D model and render it into the broadcast stream of the TV studio. Figure 6 shows the camera setup as well as the rendering result of the first prototype system.

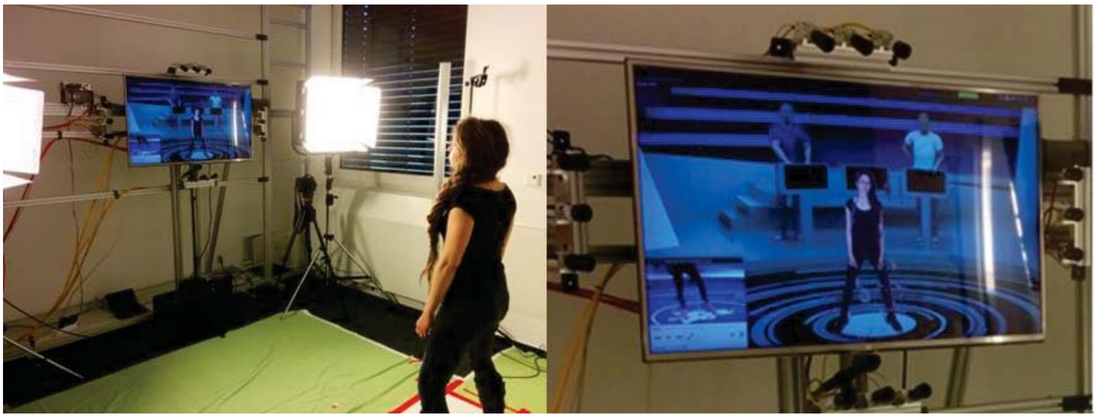

Figure 6 – Live capture of home user with multi camera setup (left) and rendering of the virtual candidate in the TV studio (right)

In the first phase of the project, a 3D model was estimated based on multi-view video analysis and point cloud fusion between different trifocal camera systems [9]. In the second phase, a photo-realistic 3D model will serve for visualization that will be adopted to the user's personal outlook (e.g. face, clothes) and animated by marker less motion capture.

REAL-TIME TRANSMISSION OVER MEDIA CLOUDS

In order to bring ACTION-TV concept alive, novel content delivery architecture is developed which considers various challenges. There exist stringent network requirements such as real-time transmission, low delay, high-quality video and synchronized receipt of media streams. The intense real-time rendering in the delivery chain creates additional computational delay putting pressure on the minimal latency requirement. Moreover, differing user types and collaborating groups requires different stream delivery policies.

Besides, dynamic streaming and switching of streams within a show is needed to provide ACTION-TV's novel user interaction capability. Finally there is a scalability problem posed by the large number of geographically distributed clients who are expected to access live media content simultaneously.

To meet these challenging media networking requirements, ACTION-TV architecture exploits both the media cloud architecture and P2P like direct communication, which will take place among interacting peers. This architecture integrates cloud computing, media processing and delivery strategies all together. By using the media cloud architecture, the system provides a scalable infrastructure enabling optimization of available virtualized cloud resources in real-time intense content processing and delivery jointly. To cope with stringent low latency processing and network delivery requirements, media clouds will be optimally placed as serving cloudlets through a resource allocation procedure, which determines the optimal data centre allocation (to individual 'social groups') that maximizes the group QoS of the end-users adhering to the ACTION-TV scenario. With the placement of distributed media clouds and optimum selection of serving cloudlets, the system will reduce the network latency by becoming closer to geographically scattered user terminals and will be able to provide synchronization of collaborating users to achieve joint content consumption. To accommodate for different user types and interaction points within a show, a game logic and a corresponding network control messaging framework has been designed. Through this framework, cloudlet servers and user terminals can be configured per user type and social group. Based on live user interaction analysis, the game logic and the framework is able to react to the users actions and switch over to the corresponding media streams. As an example, in a quiz show the transmitted media streams change according to whether the user answers each question correctly or not. With advanced user interaction analysis, a user who does not dance in a dance duel can also be detected and relevant media streaming can be delivered to motivate the user to start interacting with the show.

Finally, it is necessary to minimize the action-to-reaction delay perceived by the active-collaborating users. Hence, their own local rendered stream as well as the other active-collaborating users' local rendered stream must be available at each other's user terminals as soon as practically possible. To achieve this, a peer-to-peer connection structure was designed between these active-collaborating users. By providing P2P like direct but very low latency transmission of locally rendered streams among interacting peers, the system is able to immerse users into the same show content at right interaction points, which makes the system more realistic to the viewers of the show.

USER EVALUATION AND USER EXPERIENCE

During the project, several user tests were carried out in order to assess participants' perceived User Experience (UX). In an early stage of the project, it was investigated how users would like to be represented in an ACTION-TV show. Several visual representations in various resolutions have been used for the tests. By using marker less tracking technology, these models have been animated in real-time by the user. The major result was that more realistic representation of users has been preferred against representation by artificial characters. In addition, higher resolution images with more details have been acknowledged.

During a full demonstration of the ACTION-TV set-up, a second user test with nine participants was conducted to evaluate the users' impression regarding the different aspects of the demonstration scenario: Active collaborating users, peers' get-together, and passive TV watching. Three groups, consisting of three acquainted persons each, took part in the test. Each test session for a group comprised three so called 'test trials', i.e. test conditions:

a) The full ACTION-TV quiz scenario (active collaborating users and two of his/her social peers): A participant is performing a quiz against a pre-recorded studio candidate. Two members of the social group watch in separate rooms and communicate with each other via 2nd screen.
b) Just-chatting condition (social peers' get-together while watching non-interactive TV): All participants watch a pre-recorded non-interactive show in separate rooms. Two participants can communicate with each other via 2nd screen.
c) Couch condition (active collaborating user and passive TV-watching): A participant is performing the quiz against a pre-recorded candidate. Both members of the social group watch together in the same room and talk to each other.

Before the test session, each group of participants received instructions explaining the procedure and their task. They filled in a questionnaire on demographic data and we asked about their affinity for technology as well as their regular social media usage. After each test trial, participants received a questionnaire containing items to assess how they felt during this group exercise – the so called 'shared user experience'. At the end of the whole test session, participants had to fill in several additional questionnaires and were interviewed.

Four out of six participants who took the role of the active candidate performing the quiz round at some point during a test session named the respective round as their favourite for that very reason: Being active candidate was appraised to be experienced as 'something new'. There seems to be a certain appeal to participate as an active candidate. The fully integrated ACTION-TV quiz scenario (condition a) was chosen as favourite by four participants. However, as a means of communication, participants still preferred a common conversation on a couch (condition c), which was chosen as favourite by six participants. Overall, participants described their shared experience with, e.g. 'funny group experience'. Sitting at home, while participating in a TV game show with others was stated to be a simply good idea, novel and cool.

CONCLUSION

In this paper, the results of the European research project ACTION-TV have been presented. A novel concept for future interactive broadcast has been discussed and the system architecture and the production workflow have been detailed. Novel 3D capturing and rendering techniques have been investigated for this type of application in order to support low delay and high quality photo-realistic rendering instead of using unpleasant looking artificial characters. In order to allow for immediate interactivity by actively participating users in a broadcasted TV show, a new concept for distributed processing and transmission of various types of data has been developed exploiting cloud based architecture and P2P links. The overall end-to-end system has already been demonstrated to the public and user experience studies have been performed. As one out of many results, the designed end-to-end content delivery system meets the challenges for real-time, low delay processing and transmission to become an interactive, scalable, innovative media cloud system.

REFERENCES

1. http://www.action-tv.net/
2. N.M. Thalmann, and D. Thalmann, “Digital actors forinteractive television,” Proc. IEEE, vol. 83, no. 7, pp. 1022-1031, Jul. 1995.
3. “Rev-TV: Multimodal Avatars,” https://team.inria.fr/panama/projects/revtv/, accessed 05 July 2016.
4. “REVERIE,” http://www.reveriefp7.eu/, accessed 05 July 2016.
5. M. Satyanarayanan, (2015), “A Brief History of Cloud Offload: A Personal Journey from Odyssey Through Cyber Foraging to Cloudlets”. GetMobile: Mobile Comp. and Comm. 18, 4 (January 2015), 19-23. DOI=http://dx.doi.org/10.1145/2721914.2721921
6. G. Kulupana, D. S. Talagala, H. Kodikara Arachchi and A. Fernando, "Optimized resource distribution for interactive TV applications," in IEEE Transactions on Consumer Electronics, vol. 61, no. 3, pp. 344-352, Aug. 2015. doi: 10.1109/TCE.2015.7298294.
7. X. Chen, M. Koskela, “Skeleton-based action recognition with extreme learning machines”, Neurocomputing, Volume 149, Part A, 3 February 2015, Pages 387-396, ISSN 0925-2312, http://dx.doi.org/10.1016/j.neucom.2013.10.046.
8. G.J. Sullivan, J.R. Ohm, W.J. Han, T Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649-1668, 2012.
9. S. Ebel, W. Waizenegger, M. Reinhardt, O. Schreer, I. Feldmann, “Visibility-driven Patch Group Generation”, *IEEE Int. Conf. on 3D Imaging (IC3D), Liege, Belgium*, Dec. 2014

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the ACTION-TV project for their contributions to the work reported here. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7-ICT-2013-10) under grant agreement no. 611761.