Synchronization in hybrid multi-view entertainment distribution systems

A brief of the approaches taken in the DIOMEDES and ROMEO projects

Erhan Ekmekcioglu, Safak Dogan, and Ahmet M. Kondoz
I-Lab Multimedia Communications Research,
Centre for Vision, Speech and Signal Processing,
University of Surrey
Guildford, Surrey, GU2 7XH, UK
{E.Ekmekcioglu, S.Dogan, A.Kondoz}@surrey.ac.uk

***Abstract*— Synchronization is an essential part of any application that involves multimedia. In the most basic sense, synchronization is needed in any played back media (or likewise, streamed media) for comprehensiveness and a comfortable and natural perception. This refers to displaying the video frames in the correct time order and also playing the associated audio in a time-aligned fashion with respect to its video component. Hence, synchronization is an aspect that is tightly coupled with time. Ongoing technological advancements in content capturing, coding and transmission, as well as display technologies paved the way towards realising more interactive and immersive multimedia applications, including broadcast services. Accordingly, the dimensions of those services that need the incorporation of synchronisation aspects at different levels have increased. This paper is devoted to describe those various dimensions in a number of multi-view entertainment media distribution frameworks that have been targeted in two recent European Commission (EC) funded projects, called DIOMEDES and ROMEO. This paper also outlines the methods how the various synchronization issues are tackled in those projects.**

Keywords– multimedia synchronization; 3D multimedia; hybrid delivery; heterogeneous networks

I. INTRODUCTION

Traditionally, synchronization refers to the temporal relations existing between the different objects in a multimedia representation, as well as inside a multimedia object itself. These multimedia objects are time-dependent, unlike other kind of information sources, such as still images and textual information. For instance, television signals that we receive everyday are a good example, as they comprise synchronized visual information as a sequence of video frames with the corresponding acoustical information, both of which are aligned with each other. However, the synchronization concept does not only refer to the timely aligned play back of separate media objects in harmony. In order to facilitate timely aligned play back of media, synchronization needs to be addressed at various stages within an end-to-end media delivery and consumption chain. This need is basically a result of the constraints brought up by the underlying network infrastructure, real-time requirements of the application, as well as the hardware system utilized in the process (e.g., memory, processing speed, etc.). In a survey conducted on media synchronization [1], the notion of synchronization in multimedia systems is considered as a combination of various criteria, which involve the type of media (i.e., whether it is time-dependent or not), number of concurrent media objects, and the level of integration of various media objects. Accordingly, a multimedia system has been defined as an application that supports the integrated processing of several media objects [1].

Multi-view based immersive 3D experiences combine depth perception in a scene with an ability of freely changing the desired viewpoint by means of either eye/head position tracking or a personal control unit (e.g., a joystick). 3D tele-immersive visual systems have also been recently researched [2]. The reconstructed visual environment is accompanied by the rendered spatial audio, which aims to localize the sound sources in a spatial alignment with the visual objects in the scene. This refers to another dimension of synchronization, i.e., spatial synchronization of audio and video objects in the reconstructed audio-visual scene (audio-video spatial congruency), which needs to be ensured. In two recent collaborative European Commission (EC) funded projects, namely DIOMEDES (Distribution of Multi-View Environment using Content-Aware Delivery Systems) and ROMEO (Remote Collaborative Real-Time Multimedia Experience over the Future Internet), the distribution of multi-view 3D entertainment media is addressed, where all stages of the underlying end-to-end chain are addressed. A brief overview of both projects' system architectures is provided in the subsequent section. One of the distinct features that both projects have is that both of them aim at utilizing two separate content delivery networks in a harmonious hybrid fashion, such that part of the compressed streaming media is delivered through one of the networks and the rest through the other. These networks are DVB-T2 and the Internet, which are independent from each other in terms of interconnectivity. Thus, it is essential that time-wise synchronicity between the multimedia sub-streams delivered over two independent distribution media is maintained. Besides, in a remote collaborative 3D media enjoyment scenario, it needs to be ensured that media are played back at the same or with negligible (unperceivable) time difference, in all involved

The work presented in this paper was developed partly by the DIOMEDES (grant number: 247996) and ROMEO (grant number: 287896) projects, which were funded under the EC FP7 ICT collaborative research programme.

ISBN - 978-90-5986-410-8

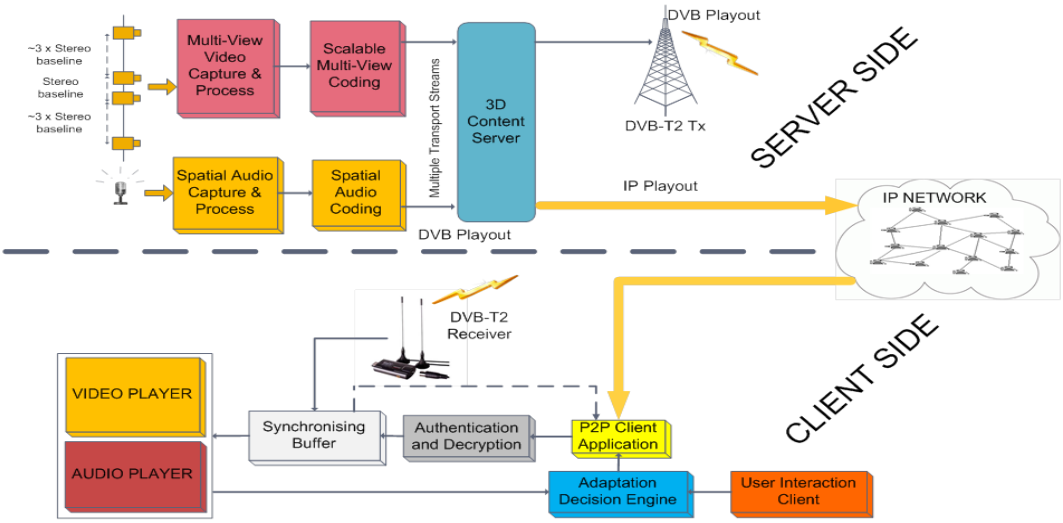

Figure 1. System building blocks for the DIOMEDES server and client architectures

parties' display sets. This requires the synchronization of multiple collaborators in addition to the synchronization of the multimedia complex within itself.

The rest of this paper is structured as follows: In Section II, a brief overview of both projects is provided, with an emphasis on the synchronization requirements arisen by the deployed system architectures. In Section III, the novel methodologies applied in those projects to tackle the synchronization at various stages of the end-to-end delivery chain are described. Finally, Section IV summarizes the paper.

II. Overview of DIOMEDES and ROMEO

This section briefly outlines the system architectures of the DIOMEDES and ROMEO projects.

A. DIOMEDES system architecture

An overview of the server and client architectures developed in the DIOMEDES project is shown in Figure 1 [3]. The projects aims is to capture multi-view content and multi-channel audio, compress them in a scalable manner, transmit them to clients through both the DVB-T2 aerial link (only part of the compressed bit-stream) and a Peer-to-Peer (P2P) overlay network over the Internet, and allow the users to render the 3D audio-visual scene in real-time with high quality and in a synchronous way. In order to facilitate efficient bandwidth adaptation, multiple camera views are encoded using the Scalable Video Coding (SVC) extension of the MPEG-4/H.264 Advanced Video Coding (AVC) standard. DVB standard mandates the use of MPEG-2 Transport Stream (TS) encapsulation of encoded 3D multimedia. MPEG-2 TS specifies a container format that encapsulates multiple packetized elementary streams (i.e., scalable view layers, depth maps, and multi-channel audio) with stream synchronization features for maintaining transmission integrity. To facilitate inter-stream association by overcoming the burden of parsing the headers of different types of container formats, MPEG-2 TS is also fed into the P2P overlay network. On the other hand, previous research has shown that using MPEG-2 TS over the Internet based packet video communications is not optimum in terms of bandwidth utilization, compared to native Real-time Transport Protocol (RTP) streaming [4]. Nevertheless, it should be noted that in a purely Internet based video delivery architecture, there is not such a concern as associating the delivered video packets time-wise with the video packets delivered through another independent network. The 3D content server stores the compressed and MPEG-2 TS multiplexed multi-view plus depth and multi-channel audio content. A security server is deployed to prevent unauthorized streaming of the stored content. The DVB-T2 play-out server is fed with the AVC-compatible base layer stream of the stereoscopic camera pair (i.e., the stereo baseline, as seen in the centre of the 4-camera set shown in Figure 1). The transport streams of the remaining cameras and depth maps are packetized into encrypted, variable size, and fixed duration units, called chunks, which hold extra data in their corresponding headers including time reference (inferred from the underlying TS packets) and an authentication key. They are fetched by peers to get disseminated over a P2P overlay as necessary.

The P2P scheme utilized in the project is pull-based, where peers are organized in a mesh structure. Peers have the ability to track the activities of other peers in the swarm, and a chunk selection and scheduling mechanism runs in each peer. The chunk scheduling mechanism is guided by a content and context driven adaptation decision engine. Context is defined by the users' current desired viewpoint, whereas the stream content is appended with Quality of Experience (QoE) related metadata. At times of bandwidth scarcity, the P2P streaming engine discards some of the sub-multimedia streams (e.g., viewpoints or quality enhancement layers) based on the stream priority list generated by the adaptation decision engine. The P2P streaming engine sends the received chunks to the authentication and decryption block, which then forwards the payload of the decrypted chunks (i.e., TS packets) to a synchronizing buffer. The synchronizing buffer has two inputs,

ISBN - 978-90-5986-410-8

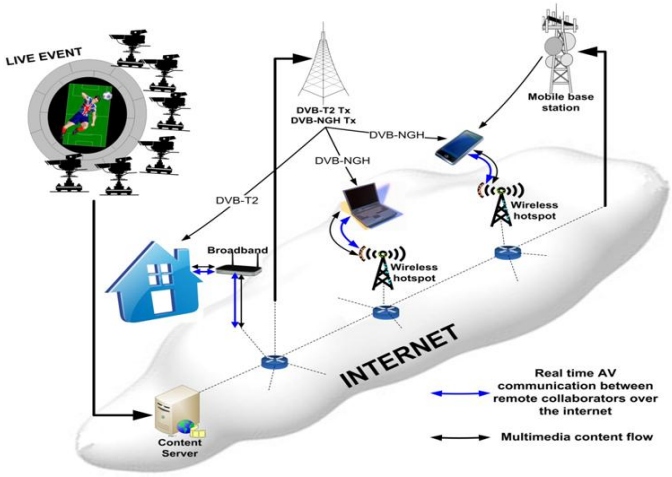

Figure 2. System overview diagram of ROMEO

one for the incoming TS packets from the DVB-T2 receiver and the other for the forwarded TS packets from the P2P gateway. The synchronizing buffer maintains a separate buffer for each received sub-multimedia bit-stream in order to compensate for inter-arrival time differences and feeds them to the media player units accordingly.

B. ROMEO system architecture

The end-to-end system that is developed in ROMEO project [5] has conceptual similarities with that of the DIOMEDES project. An overview of the ROMEO system is illustrated in Figure 2. The main similarity is the way the multi-view captured video accompanied by the multi-channel audio are delivered to users over a hybrid distribution path, which comprises both the DVB-T2 link and the Internet. The Internet video delivery is realized on a P2P overlay network. However, there are also some significant differences between the target application scenarios and building blocks of both projects. The first one of them is that mobile peers addressed in ROMEO are also a part of the P2P swarm unlike in DIOMEDES, where all users are fixed with wired broadband connection. Multiple mobile access networks, such as Wi-Fi (as denoted with wireless hotspot in Figure 2) and 3G/ LTE, are also considered to serve mobile users. Mobile users with their limited processing capability terminals, varying channel conditions, and multiple network access interfaces have a big impact on the design of the end-to-end content delivery system. The system should be able to offer these users a smooth stream either by offering the best network available or using different networks simultaneously. Second, in DIOMEDES, users enjoy the broadcast multi-view content individually. On the other hand, in ROMEO it is aimed at allowing multiple users to enjoy the broadcast multi-view content "collaboratively", where users watch the content simultaneously while communicating with each other over a separate audio-visual communication overlay. Thus, synchronization of users in addition to the synchronization of all time-dependent multimedia objects delivered to a particular user is essential, as the users are expected to communicate with each other commenting on an ongoing broadcast event simultaneously.

Different from the P2P video packet dissemination strategy applied in DIOMEDES, the P2P strategy in ROMEO is based on a multicast multi-tree architecture, where the special transmission units (i.e., chunks) of the quality layers of multiple views are distributed over separate pre-computed trees. The reason is that the peer collaboration facility of ROMEO imposes strict peer synchronization requirements to the P2P structure. Keeping this point in mind, tree-based structure is evaluated as the best candidate for the ROMEO architecture due to the more deterministic behavior of the tree systems compared to the mesh-based ones. A centralized network topology building mechanism is responsible for creating and regularly updating the multi-tree structure, based on the network activity information it collects from the participating collaborator peers. Accordingly, malfunctioning inter-peer links can be inherently inferred and updated, so that the multi-tree structure avoids utilizing the corresponding link until it is healed. Additionally, a scalable multiple-description coding scheme is adopted to further increase the delivery system's robustness against peer churns or mobile access networks' packet failures. Thus, it is the description packets effectively distributed through the multiple trees.

III. Applied Synchronization Mechanisms

In DIOMEDES, there are three major steps, in which synchronization is essential: (i) synchronization of the captured multi-view video and object based audio, (ii) synchronization of the hybrid delivery path (i.e., streams delivered through DVB-T2 and P2P/IP), and (iii) synchronization of audio and video players. Capturing and recording units for audio and video were independent from each other. The synchronization with simultaneous video and audio recordings has been achieved during recording using timestamps with common time code formats (e.g. SMPTE 12M Timecode - LTC). In order to facilitate spatially aligned rendering, the scene geometry is also recorded, where the approximate distances of individually recorded audio objects as well as the multi-camera geometry are noted and reflected in the content metadata. The content metadata arrives at the client terminals' media players ahead of the compressed streams to correctly configure the rendering units. The clock reference and time stamps of all streams (inc. audio and video) originate from the same system time clock. The timing information (PCR) is contained within the DVB stream, and is used as a reference time to the User Terminal clock. The PCR provides a clock with a resolution of 27MHz and an accuracy of 500ns. The synchronization block reconstructs the clock using the PCR from the DVB transport stream. The reconstructed clock is streamed to the audio and video players. DVB streams are delayed in the buffer, to allow the parallel P2P delivered streams have enough time to arrive. The reconstructed PCR clock should be delayed more than the buffer delay time. This method allows the video and audio players to receive the compressed streams before they should be played back according to the presentation time stamp. The P2P engine receives periodic PCR updates from the synchronization block. In this way, the P2P streaming engine can find the adequate chunk to start from for scheduling and

ISBN - 978-90-5986-410-8

downloading. As aforementioned, each chunk carries the reference time information in its header inherited from the payload it contains, i.e., the transport stream. Once the PCR update is received, the P2P engine first checks if it is falling behind the schedule or not, with respect to the ongoing DVB broadcast. If the IP delivery is ahead of the DVB delivery, then no further steps are taken. But in the opposite case, the P2P engine tries to estimate the next chunk it should download "on time" using two parameters: the bit-rate of the content and instantaneous monitored download rate. In the conducted tests, the user tunes in to the on-going DVB-T broadcast and P2P streaming engine tries to catch it by synchronizing its chunk scheduling accordingly. The time that elapses until the P2P streaming engine starts delivering all permitted quality layers of all views in timely alignment with DVB-T delivery is measured. Four geographically distributed peers and a common content server have been used in the tests (server and two peers in Germany, one peer in the UK and one peer in Turkey). After several repetitions, the tests results have shown that the synchronization performance is directly linked to the network bandwidth capacity. In the tests, the download capacity of the traced peer with DVB-T connection is artificially constrained to x2, x1.5 and x1.2 times of that of the total multi-layer multi-view content bit-rate. When the average download capacity is sufficiently high (e.g., two times more than the average content bit-rate), the average time it takes to catch the on-going DVB-T stream is found to be ~5.5 seconds. On the other hand, it reaches ~10 seconds if the average download capacity is close to the content bit-rate (e.g., 1.2 times the average content bit-rate). On the other hand, according to the deployed adaptation rules, the P2P streaming engine attempts to start downloading the base quality layers' chunks first. In the conducted tests, the average time elapsed until the full retrieval of base layer chunks of all views (excluding the stereoscopic view delivered through DVB-T) was found to be under one second (i.e., almost instantaneously). Therefore, the user can start experiencing the content in multi-view in a very short time after tuning in to the P2P broadcast content.

A dedicated transport stream with a pre-specified Program ID (PID), containing only the current PCR is used as the synchronization mechanism for the audio and video players. The synchronization block sends this stream to both players. PCR is interpreted by both as the correct time, at which the audio and video renderers play back the frames with the corresponding presentation time stamp. Thus, the synchronization block has to send the compressed multimedia streams early enough to allow sufficient decoding time in the media players.

Very similar to DIOMEDES, in the ROMEO platform, the stream synchronization module synchronizes the incoming content, received through the DVB-T2 and P2P networks, effectively referring to their embedded time-stamps and metadata. To adapt to the varying P2P network conditions, ROMEO develops an adaptive stream buffering, which is highly dependent on the buffering capabilities of the end-users, applied P2P overlay topology, and media encoding and packetizing techniques. The buffer adaptation mechanism is initiated by an audio-visual adaptation block in cases of buffer under-run and over-run. The audio-visual adaptation block is the only module that is responsible for making the adaptation decisions. In the ROMEO platform, the adaptation decision block is in charge of deciding to discard some of the quality enhancement layers of various camera views, or even some of the camera views as a whole, depending on the prevailing bandwidth conditions. In addition, with the help of redundant paths in multiple multicast tree overlay, this block may request to change the path from the centralized topology builder. In this sense, the adaptation of the end users' stream synchronization blocks against changing network conditions is highly dependent on the adaptation decision engine.

The ROMEO project started in the last quarter of 2011 with duration of three years. A majority of the intended objectives are currently under active research, including the inter-peer synchronization (i.e., synchronizing the collaborating peers for simultaneous enjoyment of the broadcast media) and its interrelations with the concurrently running audio-visual communication overlay. More detailed technical solutions to tackle this additional dimension of the synchronization problem in the proposed immersive 3D media application platform will be provided in the upcoming public deliverables that will be made available through the project website.

IV. Summary

This paper has provided an overview of the system architectures of two recently conducted EC funded collaborative research projects, namely DIOMEDES and ROMEO, with an emphasis on the underlying synchronization requirements at different stages. Accordingly, a summary of the mechanisms to maintain the synchronicity at various stages in both projects is provided. Since ROMEO is still an active project, more elaborate research results regarding the synchronization aspects described in this paper will be further published in later phases of the project.

Acknowledgements

Authors would like to express their thanks to all of their partners involved in both collaborative projects for their efforts in developing the intended systems and contributions towards the achieved objectives.

References

[1] G. Blakowski and R. Steinmetz, "A media synchronization survey: reference model, specification, and case studies", IEEE Journal on Selected Areas in Communications, vol.14, no.1, pp.5-35, Jan. 1996

[2] Tele-immersive Environment for EVErybody – TEEVE Project (http://cairo.cs.uiuc.edu/projects/teleimmersion/index.html)

[3] Distribution of Multi-View Environment using Content-Aware Delivery Systems (http://diomedes-project.eu)

[4] A. MacAulay, B. Felts, and Y. Fisher, "IP streaming of MPEG-4: Native RTP vs MPEG-2 transport stream", white paper, Oct. 2005 (accessible from: http://www.envivio.com/files/white-papers/RTPvsTS-v4.pdf)

[5] Remote Collaborative Real-Time Multimedia Experience over the Future Internet (http://ict-romeo.eu)