QoE-aware Traffic Management for Mobile Video Delivery

Bo Fu, Gerald Kunzmann
DOCOMO Euro-Labs, Munich, Germany
{fu, kunzmann}@docomolab-euro.com

Michelle Wetterwald
EURECOM, Sophia Antipolis, France
michelle.wetterwald@eurecom.fr

Daniel Corujo
Universidade de Aveiro, Aveiro, Portugal
dcorujo@av.it.pt

Rui Costa
Alcatel-Lucent Bell Labs, Paris, France
rui_pedro.ferreira_da_costa@alcatel-lucent.com

***Abstract*— Video delivery has become a major challenge for mobile networks. The increasing spread of smartphones, along with the expanding availability of LTE coverage, has contributed to an extensively growing mobile video demand, bringing large volumes of video traffic and far exceeding the capacity of mobile networks. As a result, wireless access congestion is becoming more frequent, degrading the Quality of Experience (QoE) for mobile video consumers. In this paper, we present a novel QoE-aware traffic management scheme and architecture for scalable video delivery. Thereby, Scalable Video Coding (SVC) is able to organize video data into layers of different importance, thus facilitating the rate adaptation of video streams. The proposed cross-layer architecture is implemented and the optimization is validated in a real-time streaming prototype and testbed. Results demonstrate the benefit of applying QoE-awareness and cross-layer optimization in congestion scenarios. The perceived video quality is improved considerably for all users in the cell, under both light and heavy congestion.**

Keywords—QoE; traffic management; mobile video delivery; SVC; MIH; cross-layer optimization.

I. INTRODUCTION

Operators and service providers world-wide observe a significant increase of mobile video applications and, thus, a growing demand for high data rates over the wireless interface, which is even surpassing LTE bandwidth. Compared to the mobile core network, congestions occur more frequent in the wireless interface. Video quality is significantly impaired due to congestions and varying channel conditions in the cells, degrading the Quality of Experience (QoE) of users.

The EU project MultimEDia transport for mobIlE Video AppLications (MEDIEVAL) aims at evolving the current mobile Internet architecture focusing on the optimization for video services. By combining mobility aspects with in-network CDN and by designing proper traffic engineering techniques, it tries to meet the Quality-of-Experience (QoE) constraints of video flows [1]. The MEDIEVAL architecture applies a cross-layer framework to efficiently handle video traffic in the mobile network. Its traffic management strategies focus on QoE instead of traditional QoS parameters. Video-specific enhancements are introduced at different layers of the protocol stack. Moreover, cross-layer approaches provide better video support at a lower exploitation cost.

QoE-based traffic management has been developed for single layer video applications [2]. Keeping up with a new trend, Scalable Video Coding (SVC) [3] is emphasized in MEDIEVAL. SVC has attracted great attention in recent years. Scalability is achieved by organizing video data into layers of different importance. Various use cases have been identified to take advantage of SVC [4], including flexible adaptation in multicast and unequal protection. However, due to complexity of the codec and lack of software support, SVC video streaming remains a big challenge. Recently, some companies have started to partially deploy SVC in their commercial products, e.g., Radvision Ltd.[14]. The author of [5] have developed a testbed for real-time SVC streaming over WLAN and implemented a scheduler that considers the importance of SVC layers.

This paper presents results of our real-time streaming prototype and testbed. It demonstrates the QoE-aware traffic management for scalable mobile video delivery within the MEDIEVAL architecture. To the best of our knowledge, the testbed is the first QoE-aware SVC streaming prototype integrated with a real LTE platform. The novel QoE-aware transport optimization algorithm leverages the utility functions of the streamed videos to optimize the overall perceived quality for all users in a cell. Thus, it provides a significantly higher overall QoE for users facing a limited capacity due to congestion in the network or the wireless access. The prototype utilizes the SVC layers, to facilitate rate adaptation of the video streams. Amongst others, it consists of a SVC streaming server, a QoE-based transport optimization, a traffic shaper, and extended functionalities in the wireless access.

Sections II and III introduce the MEDIEVAL architecture and describe our QoE-aware cross-layer transport optimization scheme. Section IV presents our testbed deployment and the evaluation results. Section V concludes this paper.

II. MEDIEVAL ARCHITECTURE

The MEDIEVAL global architecture consists of four main entities, namely Video Services (VS), Transport Optimization (TO), Mobility Management, and Wireless Access (WA). Its reference network architecture is the Third Generation Partnership Project (3GPP) Evolved Packet System, but including also non-3GPP networks such as WLAN hotspots for e.g. traffic offloading. Fig. 1 shows the functional building

blocks. The key components of the proposed model relevant to this work are introduced in the following.

The VS controls session management, video control, and content adaptation. It also marks video stream packets with priorities based on their importance to video quality. The priority marking is carried in the IP headers and can easily be accessed by the entities adapting the video stream in the core network and the wireless access. Thus, deep packet inspection in the network or at the access can be avoided. Mobility Management is responsible for connection and flow management, which do not have a major role in the experiments described in this work.

The WA supports connectivity through heterogeneous access technologies, based on the IEEE 802.21 standard [6]. It deploys a Media Independent Handover Function (MIHF) acting as an abstract interface between different wireless technologies and the video-aware upper layers, allowing media-independent, QoE-based optimization schemes. It interfaces with link layers through Service Access Points (SAPs), implementing primitives and parameters for the abstracted information to flow in a cross-layer way. The interface is also referred to as Layer 2.5 and was extended to optimize mobility control in video service environments [7]. Regarding the LTE link layer, the Video Frames Selection and Interface Configuration (VFSIC) component selects and prioritizes videos frames. Through information received via the MIHF, the VFSIC filters video frames in the user plane of the radio interface, only allowing the ones adapted to the wireless link load. The VFSIC works jointly with a monitoring module, the Measurements and Medium Access Strategies (MMAS), providing link quality information between the user equipment and the eNodeB. The MMAS monitors the occupancy of the Layer 2 queues in the eNodeB, thus being able to detect congestion in the wireless cell. For WLAN, these functionalities are provided by the WLANMM (Monitoring Module) and the WMDC (Dynamic Configuration).

The TO is a central component of the architecture, improving the transport from the VS to the user, using in-network CDN functionalities and optimizing the video delivery chain in terms of resources and perceived video quality. Thereby, the cross-layer traffic optimization (XLO) coordinates video forwarding along the network path and how traffic shaping handles congestions. In the latter case, the objective function is not limited to QoS parameters, but achieves an optimal overall QoE for all users affected by the congestion.

Video traffic can be adapted at different levels of the video path: Content adaptation at the source (e.g. by selecting a different encoding of the video) or in the network (e.g. by dropping packets at the Traffic Engineering (TE) module), or by decreasing resources used in the wireless access and in the core network. However, information about the congestion must be first transmitted to the TE. Thus, although it is perfectly suited to address persistent congestions in the network or the wireless access, it is not able to fast react to dynamic variations of the channel conditions. Such fine-granular traffic adaptation is done in the wireless access by the VFSIC/WMDC which are aware of the current channel conditions. This operation is

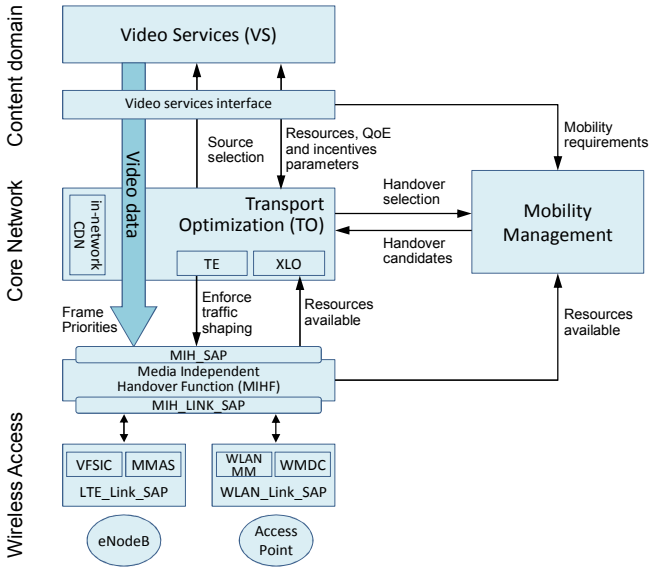

Fig. 1. MEDIEVAL functional architecture

accomplished after the packets have been decapsulated from the GTP-U tunnel and before they are encapsulated in the PDCP protocol [8].

The different layers of the architecture coordinate with each other to react to different traffic conditions. The TO obtains event notifications about the current link conditions via MIHF signaling. In a first phase, a link signal level threshold is configured where, upon detection of bad link conditions, the link starts dropping packets, based on their priority marking. In a second phase, a secondary threshold generates a link event notification to the TO about the aggravation of link conditions and triggers QoE-based traffic management in the network. More discussion on function interfacing can be found in [9].

III. QoE-based Traffic Management

Traffic management in the core network is designed to achieve optimal QoE for all users given network resource constraints. The core network has an overview of the current network and traffic conditions. The traffic is managed within the available network resources before arriving at the wireless access. Traffic management is divided in two coupled functionalities: the QoE-based optimization (in the XLO) and traffic shaping (in the TE). The QoE-based optimization is the intelligence behind rate adaptation actions, enforced by traffic shaping. For SVC video stream adaptation, the optimization indicates the target layers to be dropped by the traffic shaping.

The optimization aims to maximize the overall QoE of multiple users by allocating optimal transmission bitrates to users. It considers bandwidth conditions to estimate the maximum achievable bitrate of each user. To understand the effect of traffic management on QoE, utility functions are used to describe the relation between transmission bitrates and perceptual quality in terms of MOS values (Fig. 2).

The utility functions are derived offline for each video (or video category). For each bitrate, an adaptation of the video is performed and the corresponding video quality (QoE) is

estimated by an objective quality metric. In this work, videos are adapted by dropping SVC layers. The corresponding QoE is estimated by the objective Video Quality Metric (VQM) [10]. VQM takes into account parameters of distortions in spatial and temporal dimensions and is adopted in ITU Recommendations [15]. VQM gives an estimation of the perceptual quality in terms of Differential Mean Opinion Score (DMOS) ranging from 1 (no correlation) to 0 (same quality) which describes the perceptual quality difference between the original videos and the processed videos. It is mapped to Mean Opinion Score (MOS) ranging from 1 (low) to 5 (high) describing the absolute perceptual level: MOS = 5 - 4·DMOS.

Fig. 2 shows the utility functions of two videos used in our tests, namely News and Soccer. The videos are encoded into 3 Coarse Grained Scalability (CGS) layers and 5 temporal layers. Each point on the curves contains a set of information: the combination of layers, the bitrate of the combination, and the resulting MOS value. On the three dashed curves from left to right, the number of CGS layers increase from 1 to 3 layers. On the five points from bottom to top on each dashed curve the number of temporal layers increases from 1 to 5 layers. The utility function is the envelope displayed with the dotted curve. At some bitrates, multiple layer combinations exist, but MOS values differ largely. The layer combinations on the envelope give the optimal QoE.

The utility functions are content-dependent. In Fig. 2 the two videos show different combinations on their envelopes and different requirements of bitrates to achieve the same level of QoE. Based on the utility functions, the QoE-based optimization allocates the transmission bitrates of different videos to achieve the optimal average QoE within the available total bandwidth. In congestion mode, the optimization is able to avoid high demanding videos to occupy too much of the constraint resources and maintain an optimal average QoE in order to improve the satisfaction for the majority of users.

IV. EVALUATION

To validate our developed work, we implemented a simplified derivation of a MEDIEVAL demonstration focusing on congestion and adaptation.

A. Prototype

Fig. 3 depicts the testbed used for the experimentation and supporting some of the MEDIEVAL components. The testbed is fully operated under Linux (Ubuntu 10.04). The core network is represented by a MAR (Mobile Access Router) serving a point of access, e.g. an LTE eNodeB. Two mobile nodes (MN) are attached to the mobile network. Two SVC video servers provide two different videos, namely News and Soccer, with SVC encoding settings as described in Section III. The videos are streamed as unicast traffic to the requesting MN. The Live555 RTP Streamer [12] and the Open SVC Decoder [13] are used for real-time SVC streaming and rendering. More details of the scenario can be found in [16].

The MAR contains the traffic management elements (XLO and TE), shaping the traffic passing through it. Both MAR and MNs implement the MIHF functions used for the cross-layer

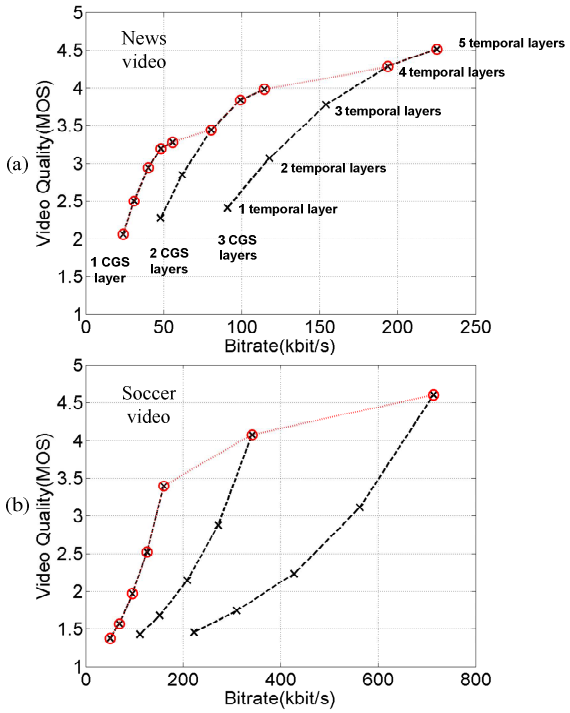

Fig. 2. Utility functions of SVC encoded videos

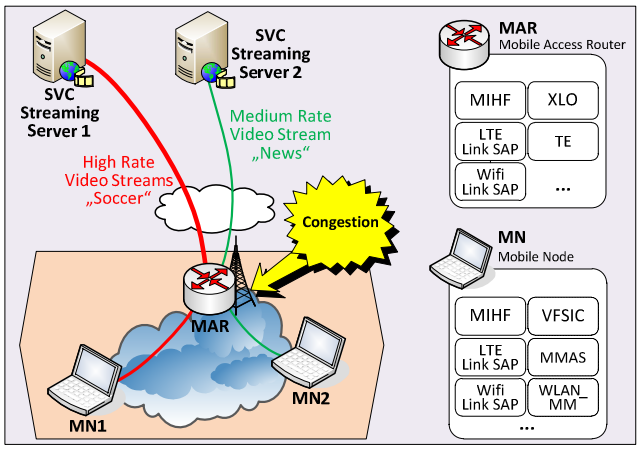

Fig. 3. Testbed setup and deployment of modules

interactions. In the current version of the testbed we were using standard WiFi, 3G and LTE access technologies and the cross-layer interactions were simulated using manual commands. The final version of the prototype will include the LTE and WiFi Link SAPs (Service Access Points), enabling the configuration of different wireless interfaces respectively. Therefore, the OpenAirInterface [11] will be used as LTE access.

B. Experiment results

With the SVC streaming prototype, we are able to show the benefits of the QoE-aware traffic management through two different scenarios. The first scenario shows the QoE-optimal SVC layer dropping for a single user. A second scenario shows the QoE-aware traffic management among multiple users. We use VQM to evaluate the received video quality.

In the single-user scenario, the video stream was adapted to maintain an optimal QoE under different levels of congestion. In the experiment, the Soccer video was streamed from Server1 to MN1 through the MAR. During the streaming session, a light level and a heavy level of congestion were simulated sequentially by limiting the bandwidth of the link using the Unix traffic control 'tc'. For demonstration purpose, the congestions were left ongoing for a while. Then, the XLO and TE were triggered by congestion notifications to handle the congestions. They performed the QoE-optimal layer dropping based on the utility function of the Soccer video.

Fig. 4 shows the significant variations of the throughput (solid curve) and the perceived QoE level represented in MOS values (dashed curve) of the video received on MN1 during the session. The experiment had five phases denoted with I to V. In phase I the bandwidth was sufficient for the video to be streamed with all layers. In phases II and IV, two levels of congestion were introduced, but not handled. In phases III and V, QoE-optimal layer dropping was performed to reduce the impact of the congestions.

In phase I, the throughput and MOS value are in their maximum, as shown in Fig. 5a. In phase II, the available bandwidth was reduced to 500Kbit/s, which was beyond the requirement of the video stream with all layers. The video was thus impaired severely by random packet dropping at the link and significant distortions resulted in low QoE (Fig. 5b). In phase III, the XLO was triggered to handle the congestion. From the utility function it determined the optimal point below the given bandwidth constraint, which was to drop the highest CGS layer. When performing this dropping, the two remaining CGS layers were successfully transmitted within the available bandwidth, resulting in a good QoE (Fig. 5c). In phase IV, the available bandwidth was further reduced below 200Kbit/s, which is even beyond the requirement of the video stream with two CGS layers. The video was again impaired severely, similar to Fig. 5b. In phased V, based on the utility function, the XLO decided to drop two CGS layers and let only the base layer pass through. After performing this dropping, the base layer was successfully transmitted, resulting in an acceptable quality, even under this heavy congestion (Fig. 5d).

In the single-user scenario, the layer dropping is performed based on the utility function representing the envelope which provides the QoE-optimal set of layers given different available bandwidths. The XLO adapts the video stream with the optimal selection of layers to be dropped and the perceived quality is significantly improved under congestion.

In the multi-user scenario, the video streams of two users is managed to efficiently utilize different amounts of available bandwidth. During the experiment, the News and Soccer video were streamed simultaneously to MN1 and MN2, respectively. The available total bandwidth of the link was changed from 0kbit/s (complete congestion) to 1100kbit/s (no congestion) in 50kbit/s intervals using the 'tc' command. Given the available bandwidth, the XLO performs the optimization to allocate transmission bitrates to the two streams. Two different metrics for allocation of layers were evaluated: The *proportional* allocation, allocates the bandwidth of the two streams according to the ratio of their bandwidth requirements. In contrast to that, the *MaxSumMOS* maximizes the sum of the MOS value of all streams.

The QoE of the two received videos depends on which allocation metric is used. Fig. 6a shows the influence of the allocation metric on the average perceived quality for different amounts of total available bandwidth. In general, the average QoE achieved by *MaxSumMOS* is higher than for the *proportional* allocation. The individual QoE of the two videos is shown in Fig. 6c for *proportional* allocation, and in Fig. 6d for *MaxSumMOS* allocation.

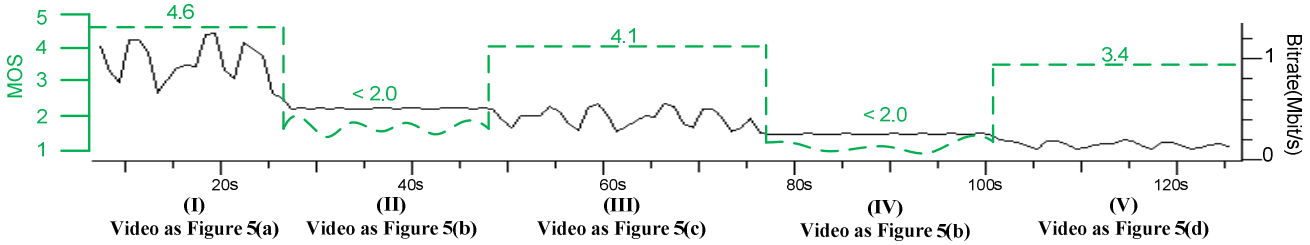

Fig. 4. The throughput and QoE levels in the single-user scenario

(a) (b) (c) (d)

Fig. 5. Perceived video in the single-user experiment: (a) No congestion. Perceptual level: Excellent (Imperceptible). (b) Suffering from congestion. Perceptual level: Bad (Very annoying). (c) Congestion handling by dropping one enhancement layer. Perceptual level: Good (Perceptible but not annoying). (d) Congestion handling by dropping two enhancement layers. Perceptual level: Fair (Slightly annoying).

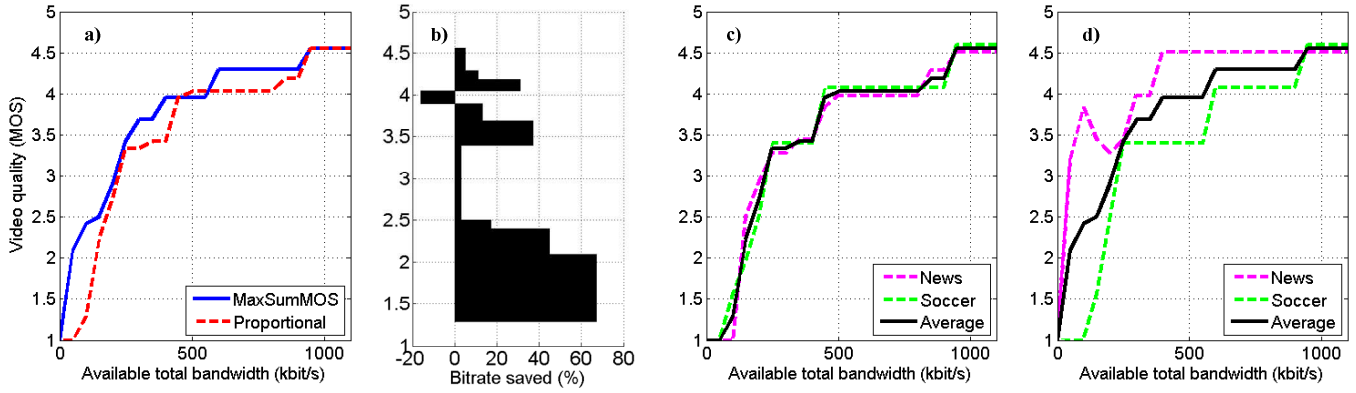

Fig. 6. Video quality under different bandwidth conditions in the multi-user scenario:
(a) Average MOS of two allocation metrics. (b) Bitrate saving from Max. alloc. vs. Prop. alloc.. (c) Proportional allocation. (d) MaxSumMOS allocation.

In the case of *proportional* allocation, the two streams fairly receive their proportional share of bandwidths. With the increase of total bandwidth, their shares grow at the same rate. In the case of *MaxSumMOS* allocation, their shares are no longer proportional, but based on their utilities. Compared to the Soccer video, the News video requires much less bandwidth to achieve the same QoE. Therefore, comparing Fig. 6c and Fig. 6d, when applying the *MaxSumMOS* allocation the News video was given more share of the total bandwidth to achieve a higher average QoE for all users. When the total bandwidth is extremely low, most of the resources are allocated to the News video until the total bandwidth is sufficient for the Soccer video to improve the average QoE. This also explains the spike on the curve of the News video in Fig. 6d.

With *proportional* allocation, users are given fairness, but videos with high demand occupy large amounts of the available bandwidth and influence the QoE of all the other users. With *MaxSumMOS* allocation, a better average MOS is achieved by sacrificing some of the resources (and thus quality) of videos with higher demand. The benefits of improving the average QoE increase with a larger number of users. Then, a higher satisfaction for the majority of users can be achieved, which is in the interest of mobile operators. *MaxSumMOS* allocation also saves bitrate to achieve the same average QoE, as shown in Fig. 6b. The bitrate saving compared to *proportional* allocation is measured at different MOS. It demonstrates the efficient resource utilization of *MaxSumMOS* allocation.

V. Conclusion & Future Work

This paper has described the MEDIEVAL cross-layer architecture as it targets traffic management for mobile video delivery in the core network as well as in the wireless access, taking into account QoE criteria in addition to more common ones and relying on a wireless abstract interface. We have presented the prototype implementation and the evaluation results obtained on the systems performance when handling network congestion. They clearly demonstrate the QoE improvement and efficient resource utilization obtained with QoE-aware traffic management. As next step we plan to assess the performance of the described algorithm and its impact on the wireless access after the LTE Link SAP has been integrated in the testbed.

Acknowledgment

The work is supported by the FP7 EU project MEDIEVAL (MultimEDia transport for mobIlE Video AppLications), grant agreement no. 258053.

References

[1] MEDIEVAL Project, Deliverable D1.3: "Final architecture design," December 2012.

[2] S. Thakolsri, W. Kellerer, E Steinbach, "QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation", Int. Conference on Communications, 2011.

[3] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding extension of the H.264/AVC standard," IEEE Trans. Circuits Syst. Video Technol., vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[4] T. Schierl, C. Hellge, S. Mirta, K. Bruneberg, T. Wiegand, "Using H.264/AVC-based Scalable Video Coding (SVC) for Real Time Streaming in Wireless IP Networks," IEEE International Symposium on Circuits and Systems, 2007.

[5] A. Detti, G. Bianchi, C. Pisa, F.S. Proto, P. Loreti, W. Kellerer, S. Thakolsri, J. Widmer, "SVEF: an open-source experimental evaluation framework for H.264 scalable video streaming," IEEE Symposium on Computers and Communications, 2009.

[6] IEEE Standard for Local and metropolitan area networks- Part 21: Media Independent Handover Services, IEEE Std. 802.21, 2008.

[7] D. Corujo, M. Wetterwald, A. De La Oliva, L. Badia, M. Mezzavilla, "Wireless Access Mechanisms and Architecture Definition in the MEDIEVAL Project," MediaWiN, June 2011, Corfu, Greece, Jun 2011.

[8] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2".

[9] D. Corujo, C. J. Bernardos, T. Melia, M. Wetterwald, L. Badia, and R. L. Aguiar, "Key Function Interfacing for the MEDIEVAL Project Video-Enhancing Architecture," MONAMI, Sep. 2011.

[10] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality", IEEE Trans. Broadcast., vol. 50, no. 3, pp. 312–322, Sep. 2004.

[11] OpenAirInterface, http://www.openairinterface.org

[12] LIVE555 Streaming Media, http://www.live555.com/liveMedia

[13] Open SVC Decoder, http://sourceforge.net/projects/opensvcdecoder

[14] Radvision Ltd, http://www.radvision.com

[15] ITU-T J.144, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference". 2001.

[16] MEDIEVAL Project, Deliverable D5.4: "Resource efficient mobile transport: final operational architecture," December 2012.