TOSCANE: TOwards SCalable Audiovisual Communication Networks

O. SALEM
Univ. Paris Descartes
F-75006 Paris, France

L. CARMINATI
Thomson Grass Valley
F-3553 Rennes, France

A. SENOUSSAOUI
NetQost
F-93000 Bagnolet, France

E. RENAN
Thales Communications
F-92700 Colombes, France

A-S. BACQUET, P. CORLAY, F-X. COUDOUX,
C. DEKNUDT, M. GAZALET, M. GHARBI,
M. ZWINGELSTEIN-COLIN ,
Univ. Valenciennes, IEMN-DOAE
F-59313 Valenciennes, France

M. GUARISCO , H. RABAH, Y. BERVILLER ,
S. WEBER
Univ. de Nancy HP, LIEN
F-54506 vandoeuvre-les-nancy, France

Abstract— **To optimally cope with heterogeneous end device capabilities and access network link dynamics, a novel integrated video streaming system is proposed. The TOSCANE system provides scalable video encoding, transmission, and quality monitoring and adaptation over wired (xDSL) and wireless (WiFi) access and residential networks. The various components of the system are designed for supporting both "Live" multicast and "on-demand" unicast video services**.

Keywords: H.264 Scalable video coding, xDSL, video quality monitoring and adaptation

I. Introduction

This paper is related to the work performed under the framework of the TOSCANE French collaborative project. The TOSCANE project aims at developing efficient solutions based on combined source and channel coding approach in order to optimize end-to-end video quality for the end user.

As described in [1][2], the most common network distribution modes pertaining to the scalable video are: 'multicast with a Media Aware Network Element' (MANE) to aggregate different sessions, 'multicast to terminals' with heterogeneous connectivity, and 'unicast' where the server aggregates in one RTP session possibly more than one layer.

In the context of TOSCANE project we rely on the two last distribution scenarios we respectively named "Live" and "VOD" scenarios. In both scenarios a streaming server (SVC content provider) has a repository of stored SVC coded streams to broadcast over an IP network up to a Digital Subscriber Line Access Multiplexer (DSLAM). The DSLAM next forwards RTP/UDP/IP packets up to a Modem/Gateway node that separates the ISP and the user domain.

In the 'Live' scenario, also called layered multicast, each layer is transmitted in its own IP multicast group. The modem/gateway subscribes to layers via IP multicast mechanisms (IGMP) depending on user-selected layouts as depicted in Figure 1. As this scenario is designed to optimize the network traffic in the core network, the modem/gateway node may convert a RTP multicast bitstream into a RTP unicast for a mobile application or a HD TV for instance. More details about so-called MST to SST re-paquetization are given below.

On the contrary in the 'VOD' scenario (Figure 2) the content provider aggregates multiple SVC layers into one single RTP session. Thus as this scenario supports personalized layout, the composing process is performed by the content server for each of the receiving endpoints: mobile, STB, HD TV, etc.

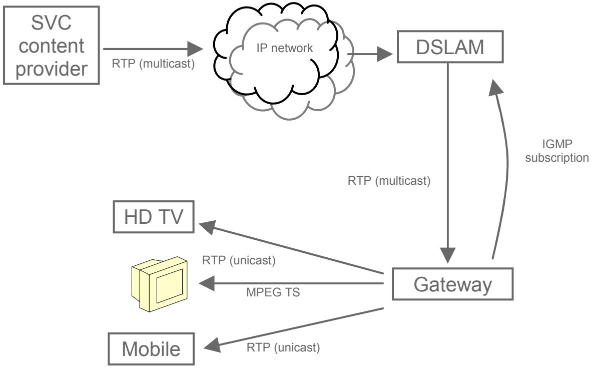

Figure 1. 'Live' Scenario: layered multicast

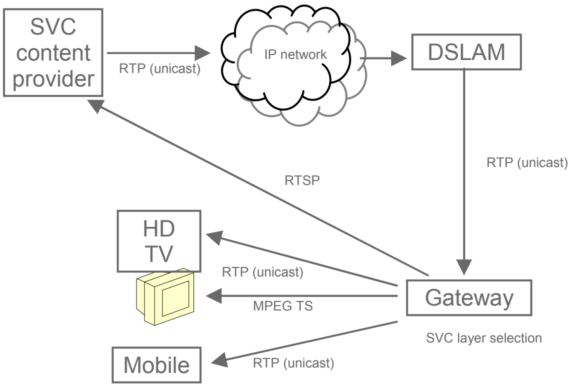

Figure 2. 'VOD' scenario: unicast with single RTP session for carrying multiple SVC layers

978-1-4244-4624-7/09/$25.00 ©2009 IEEE

The rest of this paper is organized as follows. Section II describes the transmission of SVC over RTP. In section III, we present the different methods used for SVC quality measurement. In section IV, we present the SVC adaptation model. Section V describes the real time adaptive architecture. Finally, Section VI presents concluding remarks.

II. SVC OVER RTP

A. Packetization mode

The TOSCANE architecture relies on RTP paquetization as transport protocol in IP packet-based networks. This implies that video may be transmitted in a layered multicast mode where each scalable layer is transported in its own multicast group or in a unicast mode where a single RTP session carries multiple layers. Thus the two basic modes introduced in [3] for the transmission of SVC data are addressed: Single Session-Transmission (SST) and Multi Session Transmission (MST) modes.

According to the current draft of the RTP payload format for SVC video [3], the following packetization modes for SST transmission have been retained:

- **Single NAL unit mode** considers the transmission order of single NAL unit packets comply with the NAL unit decoding order.
- **Non-interleaved** and **Interleaved** (resp.) modes differ in terms of whether NAL units are transmitted in decoding order for each session or not (resp.) and mechanisms for recovering NAL units in decoding order are quite different.

In the case of the MST mode all the non-interleaved modes are addressed here:

- **Non-interleaved timestamp based** (NI-T) mode employs a sampling time instance in RTP sessions to recover NAL units decoding order. The problem of synchronization between RTP and NTP timestamps is alleviated by RTCP Sender Report (SR) messages as discussed in [6].
- **Non-interleaved cross-session decoding order based** (CS-DON) mode implies to set a derived variable indicating the decoding order number. This variable is set in all RTP packets within all the session-multiplexed RTP sessions from the SVC bitstream.
- **Non-interleaved combined timestamp and CS-DON based** (NI-TC) mode offers both alternatives described before. Receivers are allowed to use their preferred mode.

MST interleaved mode (I-C) is not retained because it requires relatively high end-to-end latency and the decoding order recovery process is not as straightforward as non-interleaved modes.

B. Signaling of RTP streams

The signaling of SVC streams is based on the Session Description Protocol (SDP) [4]. The purpose of the SDP is to convey information about media streams contained in multimedia sessions. Due to the introduction of scalability by SVC, SDP defines a set of rules on signaling media decoding dependencies. As stated in [7] two types of media dependencies can be distinguished:

- **Layered/hierarchical decoding dependencies** considers that each layers (in case of SVC) may be decoded only when the layers it depends on are also decoded.
- **Multiple description decoding dependencies** assume that each layer forms an independent representation of the media but there is no hierarchy between layers.

In both cases SDP provides information about the potential dependencies between layers and media formats in that it allows for signaling a range of transport addresses in a certain media description. In our study the SDP is conveyed by the Real Time Streaming Protocol (RTSP) [5]. Thus the Modem/Gateway node acts as an RTSP proxy relaying RTSP messages from terminals to the streaming server.

III. SVC QUALITY MEASUREMENTS

TOSCANE project uses compression and channel coding to satisfy quality of service, and thus left us with a lot of questions about the impact of encoding and transmission on the perceptual quality of the video stream. The quality of video stream is quite complex and depends heavily on the codec type and its configuration, as well as networks performance parameters (latency, jitter, loss, errors).

In order to provide more comfort for user in the next generation IPTV (Internet Protocol TeleVision), the metric used for verifying the perceptual quality of received video must be defined. The main objective of this metric is to continuously verify the perceived quality of video by the end user, and not only at connection establishment. It must verify a short delay for channel switching and for stream reception, and must have a low CPU consumption, especially when used in embedded devices with low processing power. Used metric must have the same value for videos with the same quality, and must be independent of human intervention. Therefore, all subjective methods must not be used. Also, metric is largely depend of many environmental parameters, i.e. when reading a video conceived for mobile phone using high definition TV, the quality is very bad comparing to its quality when we play it in mobile phone.

This metric is used to measure the quality of the video at the end user. We want to extend the functionality of this metric to allow video adaption (quality enhancement) in the network, through sending back this metric to the network. Thus free the streaming server form quality adaption tasks. Adapting the channel with the quality of the video was one of the main objectives of TOSCANE project.

A. Video Quality Measurement Models

There are two types of method to measure the quality: passive and active. Some methods are used by service provider and others are used by end users. In active measure methods, agents are deployed like clients (or servers) to open new connections, get new flow, etc. Active methods are used for testing especially under heavy-loaded network conditions, but these methods can not monitor the whole network. Usually, active methods are used for network testing and during verification.

In passive measure methods, agents only monitor the existing video stream in the network, without any interaction with it. They provide information about the quality received by each user, without using a lot of bandwidth like in active method. The bandwidth used is just for transferring reports, and consequently requires less bandwidth than active methods. Sending reports can be done using SNMP, and a monitoring server can query the whole agents like in nagios or any other SNMP monitoring system.

Several methods are available for quality measurement, like the subjective Mean Opinion Score (MOS) and DSIS (Double Stinulus Impariement Scale) which require human intervention, and other objectives mathematical methods for comparing input and output video stream. Quality measurement in objectives or subjective methods may depend or not on a reference media. If the method doesn't use a reference (called 0-reference), it tests the intrinsic quality of the video, and not the quality with respect to another video. Partial reference methods (also known as reduced reference) use only a part of video as a reference, and have a good reliability and less complexity than full reference methods in which we compare the input and output video stream.

The subjective quality estimation methods (i.e. MOS), in which human gives a ratio for the perceptual quality of media, is too complex and not accurate. These methods are not real time and are costly, because we pay people that will give ratio to the quality. Therefore, several objective methods have been proposed for the evaluations of perceptual quality for video stream. Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR [11]) are objective methods, which compare the input and output video stream mathematically.

Despite the complexity of these methods (per-pixel processing and spatial alignment of input/output video), these methods don't correlate with the Human Visual System (HSV). The HSV is not superior to the complex algorithms of pixel-based measures used in these methods. Unfortunately, these computational intensive algorithms can only be used in lab for testing, pre-deployment test or troubleshooting.

There is another type of objective methods (PQR, DVQ, VQM, etc.), which are based on modeling human vision through the Discrete Cosine Transform (DCT), and their result are correlated with subjective measure.

A wide range of full reference algorithms which compares the input and output video stream, have been developed, like MPQM (Moving Pictures Quality Metric [12]), VQM (Video Quality Metric [10][14]) and CVQE (Continuous Video Quality Evaluation [12]).

The Video Quality Experts Group (VQEG) has conducted two phases of testing, and results to warrant recommendation for use, resulting in ITU-T Recommendation J.144 [12]. It contains only VQM from the previously listed algorithms. Full reference algorithms are not adequate for in-service assessment, because one must have the original video stream and the encoded one for comparison, e.g. make useless the transmission of the video over the network, as we already have it.

Zero reference [15] algorithms are generally more suitable for in-service monitoring of video services as they can analyze live streams [13]. This type of algorithm can consider fewer factors than a full reference algorithm and can be deployed in a much wider variety of scenarios (fixed and mobile video services).

A new subclass of zero reference algorithms, called Media stream based algorithms, has recently gained a high interest. These algorithms, such as Telchemy's VQmon/HD [16], Psytechnics's PVI [17], Symmetricom's V-factor [18], and Mehaoua's ServMon [19] analyze the IP stream and video transport protocols such as RTP/RTCP (time distribution of lost and discarded packets, jitter, etc.), to build up an assessment of video quality and expressing this as a perceptual video quality score.

Telchemy's algorithm is differentiated in its ability to analyze the time distribution of lost and discarded packets and to model the impact of transient IP problems on perceptual quality, based on their VQmon technology.

Our proposed algorithm in [19] is using active and continuous IP network performance monitoring (nQoS) along video delivery paths to derivate subjective perceived video quality (pQoS) of MPEG-encoded video applications. As far as no intensive video signal processing is required, the newly proposed video quality assessment algorithm is suitable for both mobile/handheld devices and residential setup-boxes.

There is standardization activity within the industry related to the definition and reporting of zero reference performance metrics. These metrics are:

- VSTQ (Video Service Transmission Quality) : a codec independent measure of the network's ability to transport video
- VSPQ (Video Service Picture Quality): a codec dependent estimate of the viewing quality of the video
- VSAQ (Video Service Audio Quality): an estimate of the quality of the audio stream
- VSMQ (Video Service Multimedia Quality): an overall QoE (Quality of Experience) metrics that encompasses picture quality, audio quality and audio-video synchronization
- VSCQ (Video Service Control Quality): a metrics that estimates the quality of the video control plane (e.g. response times)

Therefore, the simple approach of reporting only packet loss and jitter on the basis that if packet loss and jitter are low then quality must be good, is no longer enough. This approach may perform well in a scenario where quality is either perfect or terrible, but is likely to be less dependable when packet loss rates are "noticeable" and it becomes important to understand the impact of loss on the specific codec type and configuration.

TOSCANE is adopting a metric based on VQM (Video Quality Model) and calculated MOS by ServMon agent in DSLAM as shown in Figure 3. In fact, VQM has a correlation ratio of 95% with respect to MOS subjective method, for a video of 525 lines. It is based on DCT and recommended by ITU J.144 and ANSI T1.801.03-2003. Adpation algorithm in the DSLAM will take into account the measured network performance parameters, as well as the MoS received from the client, either to adapt the video coding accordingly if it is responsible of quality degradation, or it will instructs the Gateway about the new coding that it must use to enhance the MoS of user. In the second case, where the DSLAM tells the user gateway to change the media coding, it plays the role of Policy Decision Point (PDP) and the gateway is the Policy Enforcement Point (PEP). The architectural framework is shown in Figure 4.

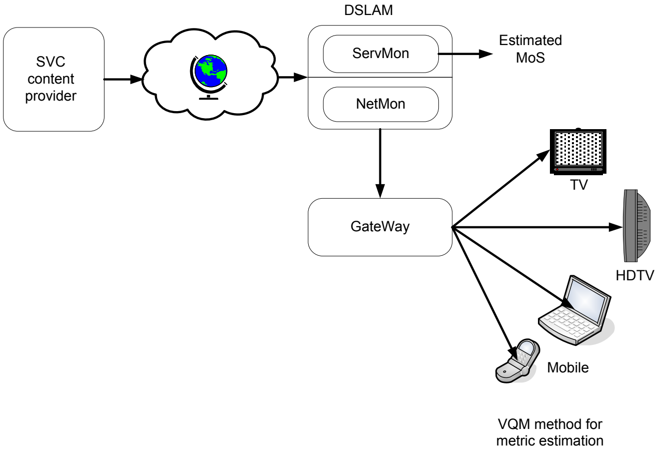

Figure 3. Adopted method: VQM at terminal & MOS estimation in DSLAM

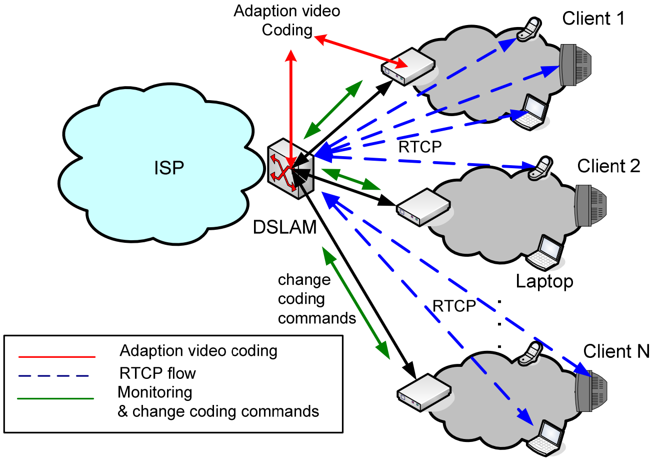

Figure 4. Policy Based Management Architecture

B. Signaling of Video Quality Measurements

Many protocols for sending back the value of performance metric, i.e. MPEG-21 Event Reporting (ER) feedbacks the DRM (Digital Right Management), the DIA (Digital Items Adaption). DIA contains the capacity of the network and host, NPI (Network Provider Interface), as well as adaptation policy (reducing resolution, etc.) in respond to existing constraint. RTCP protocol provides also information about the quality of the real time media flows through its feedback reports.

TOSCANE is adopting the use of RTCP protocol (Figure 5) for transferring the loss ratio and jitter from receiver to servMon agent in DSLAM. Usually, these information are transferred through the use of Receiver Report message in RTCP to synchronization source SSRC (DSLAM). In fact, ServMon in DSLAM exploits and translates these reports into perceptual quality (MOS) before sending them to adaption agent. This last may be located in DSLAM (IPTV) or in a remote server (VOD) by using the protocol SOAP (Simple Object Access Protocol).

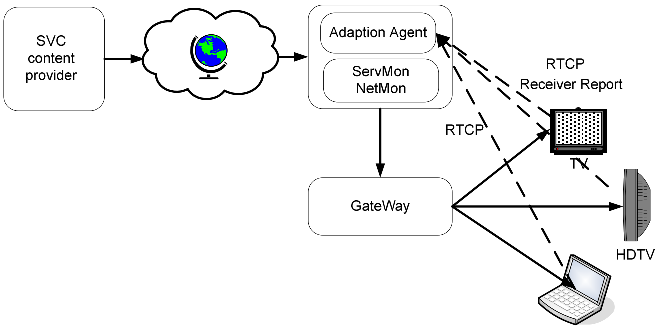

Figure 5. RTCP for transporting metric

IV. SVC ADAPTATION MODELS

In order to optimize end-to-end video quality for the end user, we developed efficient algorithms based on a combined source-channel coding approach. Two options have been considered in our study, depending whether the original compressed video stream has been encoded into a single layer or multiple layers.

A. H.264/AVC input bitstream

In this first part of the project, we have considered that the input video bitstream has been encoded into a single layer by means of the H.264/AVC video compression standard. Compressed video is transmitted from the DSLAM to the end user's residence, and then it is delivered on the addressed terminal. Different adaptation processes could be necessary depending on the spatial/temporal characteristics of the terminal and/or the available bandwidth. In all cases, video adaptation is based here on efficient transcoding /transrating schemes. Several original H.264/AVC transcoding and transrating algorithms have been developed during the TOCANE project. In particular, a transrating algorithm based on selective frequency filtering has been validated and implemented on FPGA architecture [8]. The algorithm operates in the pixel domain in order to avoid error propagation over the entire GoP. Moreover, it re-uses the coding mode decisions from the incoming bitstream in order to reduce time-consuming. Experimental results show that the proposed transrating method leads to pictures with very satisfying visual quality (Figure 6).

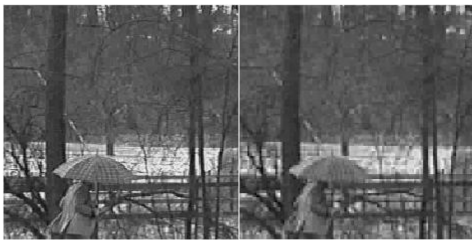

Figure 6. Visual comparison of original (left) and transrated (right) sequence.

Based on this architecture, we finally proposed a solution to estimate the bit rate at which the transrating operation has to be done in order to provide optimal video quality to the end user, under DSL transmission constraints.

B. H.264/SVC input bitstream

In the second part of the project, we have considered that the input video bitstream has been directly encoded by means of the scalable extension of the H.264/AVC video compression standard, called SVC. Hence, the input video bitstream is already encoded by means of SVC into several layers corresponding to different video size, frame rate or quality. These layers are combined into a single compressed bitstream. Typically, each layer corresponds to a NAL unit into the SVC bitstream. Then, it is possible to access to a given spatial/temporal/quality level of the compressed video sequence by selection of appropriate SVC layers. Practically, the analysis of each NALU header allows determining which NALUs should be extracted from the SVC bitstream in order to reconstruct the desired spatiotemporal resolution and bit rate. This selection process called SVC layer dropping is applied here to perform video adaptation. In the TOSCANE project, several study cases have been considered for digital video delivery over DSL + residential wireless networks, as illustrated in Section II. In particular, SVC layer dropping should allow to:

- Extend the coverage area for DSL video distribution, by adapting the bit rate to the quality of each subscriber line;
- Adapt to channel perturbations to ensure the continuity of video services to the end user;
- Address a wide variety of terminals (HDTV, SDTV, and PDA).

Figure 7 gives a synthetic overview of the scenarios studied in the TOSCANE project including SVC layer dropping.

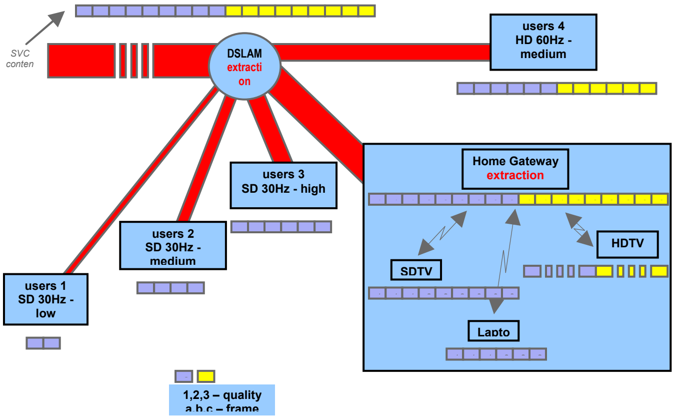

Figure 7. HD/SD distribution over DSL to residential network (from [9])

Moreover, it is well known that layered video coding can be combined efficiently with channel coding and modulation in order to increase robustness to channel errors. Hence, Unequal Error Protection (UEP) consists in protecting the video layers in a different way according to their relevance on reconstructed quality. This research topic has been also addressed by TOSCANE partners, who have studied different solutions in order to apply UEP accounting for channel properties and based on:

- Forward Error Correction using RS codes;
- Multicarrier modulation using Hierarchical QAM;
- Combined use of the two previous solutions.

Finally, compression and channel coding (including modulation) should be combined optimally in global transmission architecture to deliver video contents to the user with the best end-to-end visual quality.

V. Real time adaptive architecture

The process of adaptation for optimized scalable transmission of digital video depends on the originally coded data and the required format, which depends on the channel bandwidth, and addressed terminals. The adaptation can be performed by using various techniques with different complexities. A first technique consists of dropping packets; this supposes that the different resolutions have been defined and transmitted over tagged packets to assist the dropping procedure. The second technique consists of performing a decoding followed by encoding after a scaling operation. The complexity of this transcoding will depends on the depth of decoding which can be full or partial. To be efficient, the adaptation process must integrate an adaptive channel coding by allocating efficiently and optimally data to different carriers. All this operations have to be executed in real time and then require efficient hardware/software architecture capable of adapting its throughput to transmission requirements. Figure 8 represents the block diagram of this adaptive processing architecture.

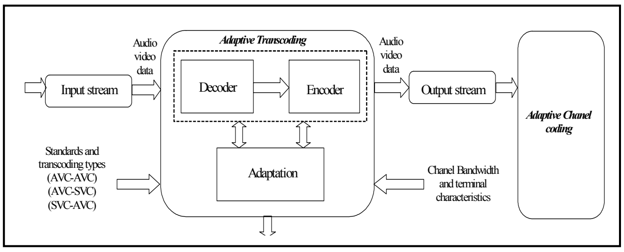

Figure 8. block diagram of adaptive processing architecture

This architecture is capable of performing the following operations:

- Conversion of a non-scalable AVC stream to scalable stream SVC, comprising at least one layer
- Conversion of an SVC stream to an AVC one comprising at least information contained in the SVC base layer
- Conversion of an AVC stream to another AVC stream consisting of a down scaled version (reduction of number of image per second, the size, or bit rates)
- Conversion of an SVC layer to another SVC layer with lower representation (rate reduction)

Most of the system's tasks manipulate data blocks and pass the results to next task in a stream processing way. Some of these tasks can be implemented in software, and other in dedicated hardware with reconfiguration capabilities, depending on the required throughput and their complexities. The complexity of tasks and their stream processing nature allowing parallel processing, led us to the choice of a reconfigurable heterogeneous multiprocessing architecture. One of the advantages of this architecture is its scalability that let us foresee a joint transcoding and channel coding by simple extension.

VI. Conclusion

The paper is presenting a new framework for the transmission of scalable video coding streams towards heterogeneous terminal over wired (xDSL) and wireless (WiFi) access networks. This work is performed under the framework of the TOSCANE French collaborative project. The TOSCANE project aims at developing efficient solutions based on combined source and channel coding approach in order to optimize end-to-end video quality for the end user. The framework is composed of three sub-systems to encode, transmit, monitor the quality, and finally adjust source and channel coding parameters to cope with variable terminal and links characteristics.

REFERENCES

[1] Y.K. Wang, “System and Transport Interface of the Emerging SVC standard”, Joint Video Team, Doc. JVT U145, Hangzhou, October 2006

[2] S. Wenger, Y.K. Wang and T.Schierl, “Transport and Signaling of SVC in IP Networks”, IEEE Transactions on Circuits and Systems for Video Technology, vol 17 no 9, pp 1174-1185, September 2007

[3] S.Wenger, Y.K. Wang, T.Schierl and A.Eleftheriadis “RTP Payload Format for SVC Video”, IETF Internet Draft Draft-IETF-Avt-Rtp-Svc-18.txt, March 2009”

[4] M.Handley, V. Jacobson and C. Perkins, “SDP: Session Description Protocol” IETF RFC 4566, July 2006

[5] H. Schulzrinne, A. Rao and R. Lanphier, “Real Time Streaming Protocol (RTSP)”, IETF RFC 2326, April 1998

[6] C. Perkins, “Rapid synchronisation of RTP Flows”, Draft Draft-perkins-avt-rapid-rtp-sync-01 (work in progress)

[7] T. Schierl and S. Wengeret, “Signaling media decoding dependency in Session Description Protocol (SDP)”, Draft draft-ietf-mmusic-decoding-dependency-06 (work in progress), February 2009

[8] C. Deknudt et al, “Transrating by frequencies selectivity for H264/AVC Intra pictures”, accepted to *IEEE BMSB 09*, Bilbao, 13-15 May 2009.

[9] E. François et al., “Scalable video coding: scalable extension of H.264/MPEG-4 AVC”, Internal Report, Apr. 2006.

[10] S.Grgic, M.Mrak, B.Z-Cihlar, “Performance Analysis of Image Compression Using Wavelets”, IEEE Transactions on Industrial Electronics, vol. 48, NO. 3, June 2001.

[11] Stefan Winkler, Digital Video Quality, Wiley 2005

[12] C J. van den Lambrecht, Color Moving Pictures Quality Metric, , PhD Thesis, EPFL 1996

[13] Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system. Wolf, S., Pinson, M. Proc SPRI Multimedia SYstems, 1999

[14] ITU-T Recommendation J.144-2004, Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the presence of a Full Reference.

[15] R. Venkatesh Babu, A.S. Bopardikar, A. Perkis, and O.I. Hillestad, “No-reference Metrics for Video Streaming Applications,” in Proceedings of International Packet Video Workshop, December 2004

[16] Telchemy, “VQMON HD”, http://telchemy.com/vqmonhd.html.

[17] Psytechnics, “Psytechnics Video IP monitor” www.psytechnics.com.

[18] Symmetricom, “V-factor probes”,
www.symmetricom.com/products/qoe-assurance/v-factor/

[19] A. Mehaoua, et al., "A Novel Cross Layer Monitoring Architecture for Media Services Adaption Based on Network QoS to Perceived QoS Mapping", in Journal of Signal, Image and Video Processing, Springer Verlag Editor, London, nov. 2008.