DIOMEDES

SEVENTH FRAMEWORK PROGRAMME

DIstribution Of Multi-view Entertainment using content aware DElivery Systems

DIOMEDES

Grant Agreement Number: 247996

D4.4

Report on the developed audio and video codecs

Document description	
Name of document	Report on the developed audio and video codecs
Abstract	This report describes the final audio and video codecs, including performance evaluation results. This report is an update to the interim report, D4.3. The codecs are assessed in terms of performance (bit-rate vs. quality for audio and video encoders, speed for decoders), with respect to the state of the art. Objective evaluation results as well as subjective evaluation results are reported in this deliverable.
Document identifier	D4.4
Document class	Deliverable
Version	1.0
Author(s)	Erhan Ekmekcioglu, Safak Dogan, Stewart Worrall (UNIS), Thomas Korn (IDMT), Goktug Gurler (KU)
QAT team	Tamir Adari (OPTEC), Haluk Gokmen (ARC)
Date of creation	06/09/2011
Date of last modification	29/10/2011
Status	Final
Destination	European Commission
WP number	WP4

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Purpose of the document

The purpose of this document is to outline the final status of the developed audio and video codec architectures. It is aimed to explain the updated design features of the developed audio and video codecs from the interim design that was reported in D4.4. This document presents the performance results of the developed final prototype audio and video codecs using objective scores as well as subjective scores based on subjective test results.

1.2 Scope of the work

The scope of the work presented in this document is to report the final design features of the audio and video encoder-decoder pairs within the DIOMEDES architecture. Audio and video encoders supply the compressed content to be encrypted and stored in the 3D content server, whereas the decoders are facilitated within the audio and the video cluster units of user terminal devices. The scenarios considered in this document are referred from the use-case scenarios defined previously in deliverable D2.1 and updated in D2.2. The implemented functionalities are based on the specifications previously reported in D4.1 and updated in D4.3. Media transportation related design features were previously reported in D4.2.

1.3 Achievements

The functional and working final prototypes for the audio and video codecs are built. For the audio encoder, subjective tests are performed to assess the final design. It is reported in the upcoming section (Section 2) that the audio codec design is not changed significantly from the interim design reported in D4.3. Similarly, the final visual attention adaptive scalable multi-view encoder design is compared to the state-of-the-art simulcast SVC in terms of perceived quality degradation under equal bit-rate adaptation conditions. Subjective tests are performed to find out the impact of applying visual attention based encoding on the perceived stereoscopic video quality during bit-rate adaptation process (i.e. truncation of quality enhancement layers). Video decoder is tested to find out its performance in terms of decoded frames per second for real-time playback. It should be noted that the joint performance of the final video codec and the designed final P2P system will be reported in D4.5.

1.4 Structure of the document

Section 2 gives the details of the final audio codec. The subjective performance results of the audio codec are also presented. Section 3 discusses the final design of the video codec, by mainly noting the updates from the interim design previously reported in D4.3. More specifically, this comprises the implementation details of visual attention based scalable video coding. Similarly, the design features of the final real-time scalable video decoder are explained in detail, along with the presented performance results. Finally, Section 4 concludes this deliverable by giving final remarks.

2 FINAL AUDIO CODEC

2.1 Overview

This section reports the functionality, structure and specification of the final spatial audio codec. The performance of the audio codec is assessed with respect to the criteria set in the technical annex. Since project task T4.2 “Spatial Audio Compression” was already finished before deliverable D4.3 and no major changes in the audio codec structures had to be implemented after this interim status, the following sections contain mainly the audio codec descriptions of D4.3. Several refinements of specifications have been added to complete the final descriptions.

2.2 The object based audio scene format and the encoder-decoder pair

Within the DIOMEDES architecture, a dedicated spatial audio codec is needed to transmit the object based audio scene in a compact representation suitable for the use within broadcast contexts (e.g. DVB) or application with P2P networks, which is specified as main audio rendering mode for the DIOMEDES architecture. Data compression of the audio object’s signals to a stream of selectable bit-rate is a main feature of the codec. Furthermore, the codec shall allow the transmission of audio object description data that accompanies the audio signals.

Deliverable D4.1 listed the functional requirements for the spatial audio codec with the following aspects:

- Object based audio scene transmission via P2P.
- Independence of audio format and exact reproduction loudspeaker setup.
- Offer data representation for streaming and storage use cases.
- Audio objects shall have descriptive properties describing their reproduction with respect to defined listener and screen positions in addition to adaptation to loudspeaker setup. Perceived congruency (spatial alignment) of audio objects and visible objects shall be achieved.

The Object Based Audio Scene format is represented by an MPEG2 Transport Stream (MPEG2-TS) container carrying a Packetized Elementary Stream (PES) that contains a sequence of coded audio scene blocks.

The object based audio scene transmission is possible by transmitting audio data and accompanying object description data that controls the rendering of each transferred audio object.

The independence of audio format and reproduction loudspeaker setup is a consequence of the accompanying description data: no assignment to an output loudspeaker channel of the reproduction system is made within the object description, but object properties like source position allow a rendering of the audio object using a wide range of nearly arbitrary reproduction system configurations (varying mainly in speaker number and position).

Streaming is the main use case of the coded audio scenes: transmission from the broadcast site to the receiver via DVB or P2P and streaming within the receiver terminal to the audio rendering cluster. For this reason the object based audio format is based on data blocks of the temporally segmented audio scene. The data blocks are formed by data atoms containing audio data and object description data. Each block is dedicated to be contained within one elementary stream packet of an MPEG2 Transport Stream. The MPEG2-TS container directly allows a DVB transmission and provides data structures for synchronisation of audio/video rendering that is needed especially within the distributed DIOMEDES rendering architecture. Furthermore, the MPEG2-TS encapsulation allows storage of audio scenes as files. As the MPEG2-TS format is predestined to carry sets of different media streams, the transmission and storage of object based audio scenes along with established audio formats (AC-3, MPEG-1 Layer 2) and video formats in one stream or file is possible. This option is also used in the

DIOMEDES production workflow.

Figure 1 shows the structure of the Object Based Audio Scene format. Audio and description data are transmitted separately in individual data atoms. The atomic structure was chosen using the structure of the MP4 file format [1].

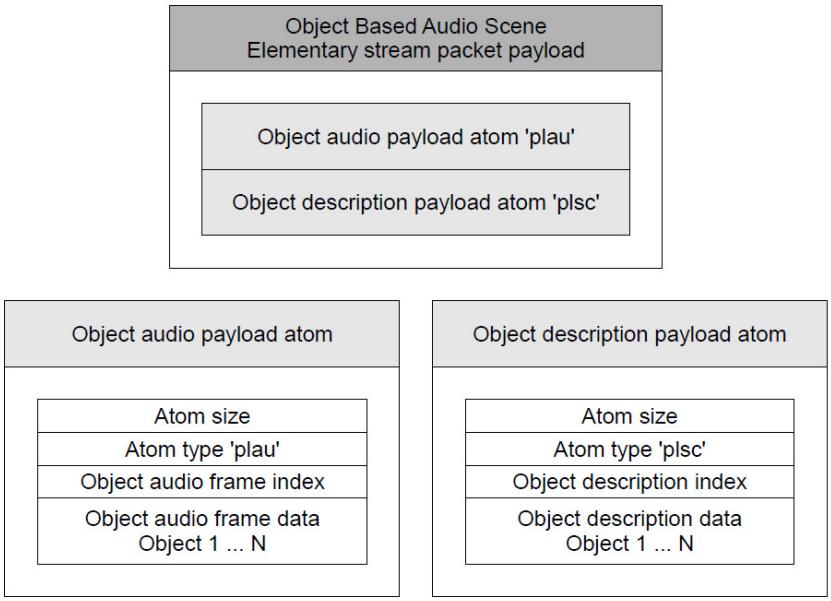

Figure 1: Object Based Audio Scene Format - Components of an Elementary Stream packet

2.3 Final audio encoder specifications and functionality

The spatial audio encoder generates the described format from the audio signals of all simultaneously existing audio objects and from the object descriptions. The first encoder for this format was implemented in DIOMEDES as a real-time encoder. Figure 2 shows the structure of this encoder implementation.

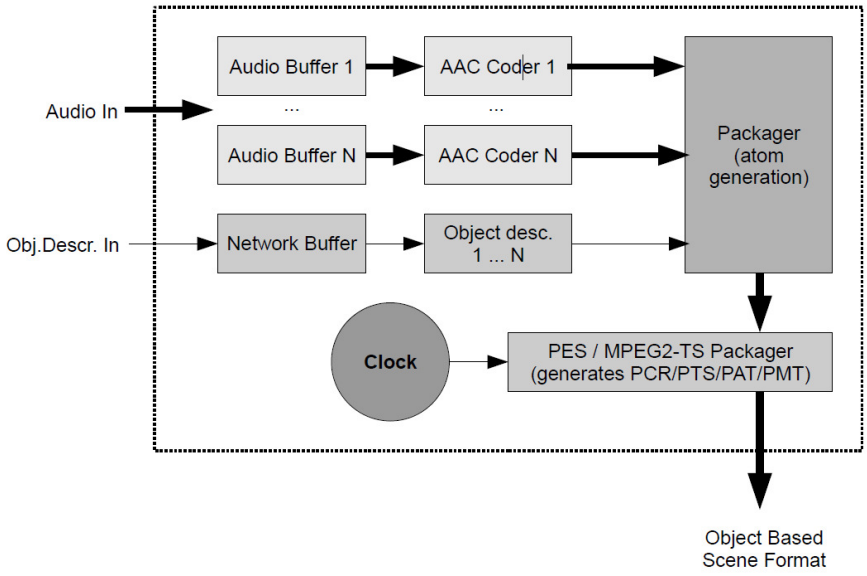

Figure 2: Object based audio scene encoder structure

A multichannel audio input feeds N input buffers. The maximum number of simultaneous audio objects in a scene determines N, where N remains constant during runtime of the encoder. Each buffer holds the 1-channel signal of one audio object. The input audio buffers feed an array of N audio coders (within the DIOMEDES architecture, AAC-LC was chosen as the signal encoder).

The object description input interface reads UDP description packets from the network and updates the audio object description data structure for N audio objects.

The audio scene format is generated by packaging the audio encoder data frames and the object description data into the atomic data structure that forms the elementary stream packet payload. The elementary stream and the embracing transport stream are encapsulated within this structure.

To generate PTS and PCR timestamps, an internal clock is derived from the audio sample clock. Finally, program information data is added to the stream in periodic intervals. These data include the Program Association Table (PAT) and the Program Map Table (PMT).

2.3.1 Audio packetisation

The audio scene PES consists of a series of ES packets carrying the audio scene segments.

- Each temporal audio scene segment is transmitted as the payload of one elementary stream packet.
- Each elementary stream packet contains the audio signal and object description data of all simultaneous objects of the audio scene.
- The audio signal and object description data transmitted in one elementary stream packet have the same presentation time reference to allow correctly timed play-out of the audio scene
- Presentation timestamps (PTS) are transmitted within the PES packet header
- The coded audio signal frames of all audio objects on one elementary stream represent the same audio segment duration.
- The coded audio signal frames and object descriptions are transmitted in individual “atom” structures: the object audio payload atom contains the coded audio frames, the object description payload atom contains the object descriptions
- The signal or description content of each audio object is individually addressed using an index table preceding the signal or description data.
- Additional atom types can be introduced to transmit additional audio scene data, e.g. abstract content descriptions.

Within the structure of MPEG2-TS, a series of elementary stream packets can be transmitted for streaming or stored in files.

The structure of the Elementary stream payload has a byte structure according to the following field sequence:

Field	Length (Bytes) / Type	Description
Audio atom size	4 (unsigned integer)	Number of bytes contained in audio atom
Audio atom type	4 (char array)	Char string “plau” (Payload Audio)
Audio frame index – scene size	2 (unsigned integer)	Number of audio objects in current scene (determines the number of the following index entries and audio frames)
Audio frame index sequence – series of index entry pairs: object channel index fame size	 2 (unsigned integer) 2 (unsigned integer)	Sequence of data field pairs assigning the following audio frames to the object/channel indices of the scene. Number of entries is determined by scene size. Object/channel index of audio frame Size of audio frame (number of bytes)
Audio frame data sequence	-	Sequence of coded audio frame data according to order and sizes indicated in index sequence; overall size results from cumulating all frame sizes in index. Number of data frames is determined by scene size.
Scene atom size	4 (unsigned integer)	Number of bytes contained in object description (scene) atom
Scene atom type	4 (char array)	Char string “plsc” (Payload Scene)
Object description index – scene size	2 (unsigned integer)	Number of objects in current scene (determines the number of the following index entries)
Object description index sequence – series of index entry pairs: object channel index description size	 2 (unsigned integer) 2 (unsigned integer)	Sequence of data field pairs assigning the following object description structures to the object/channel indices of the scene. Number of entries is determined by scene size. Object/channel index of object description Size of object description (number of bytes)
Object description data sequence	-	Sequence of object description data according to order and sizes indicated in index sequence; overall size results from cumulating all description sizes in index. Number of object descriptions is determined by scene size.

Within the DIOMEDES architecture, the Elementary Stream containing the object based audio scene format is signalled using the Program Map Table (PMT) in the MPEG-2 Transport stream [2].

The following fields of the Program Map Table entries are used for identifying the object based audio stream. (In the DIOMEDES demonstrator the stream_type value is defined as sufficient to identify object based audio streams to allow for a simple generation of PMT in the preceding modules of the audio cluster.

Field	Value	Comment
stream_type	0x87	Value indicates a user private stream
Registration descriptor: format_identifier	“WFS0”	Char string (4 bytes) indicating the object based audio format.

In DIOMEDES architecture, one PES payload unit represents one coded segment of an audio scene. Due to the optimisation of bit allocation in the utilised AAC-LC audio compression, the

payload unit data size of subsequent packets is variable. If constant bitrate (CBR) audio coding is performed, only a near-constant size can be achieved. If a constant bitrate coded TS is necessary, stuffing with Null-TS-packets has to be applied.

2.3.2 Specifications and operating range

In the following, Table 1 summarises the specifications and the intended operating range of the multi-channel spatial audio encoder.

Table 1: Specifications of the DIOMEDES audio encoder

Audio sampling rate [Hz]	48000
Audio frame size (uncoded) [samples]	1024
Audio coding of individual object signals	MPEG- 4 AAC-LC (1 channel)
Avg. audio coding bitrate range (CBR)* [kbit/s]	56 … 288
Max. number of simultaneous objects within a scene	Configurable, typ. 32**
Total TS bitrate	Depending on max. object number and AAC bitrate setting

* CBR (constant bitrate) and VBR (variable bitrate) modes are applicable. VBR mode is useful for efficient object based audio scene coding with higher numbers of simultaneous objects.
** The maximum possible number of simultaneous audio objects of a scene is limited by computation capacity during encoding/decoding and audio rendering and thus varying with the used hardware.

2.4 Final audio decoder specifications and functionality

The structure of the object based audio scene decoder of the DIOMEDES system was presented in deliverables D2.2 and D4.1. Figure 3 shows a refined diagram of the decoder module structure, as it is implemented in the current version of the audio cluster. Besides the ability for object based audio scene decoding, the implemented module is also incorporating decoder libraries for channel based audio formats (MPEG-1 Layer 2, AC-3) that are also supported in the DIOMEDES architecture. The diagram shows only the elements that are related to the novel object based decoding.

The decoder delivers a set of audio channels that are assigned to the transmitted audio objects and a stream of object description messages to the spatial audio rendering which is described in the D3.5/D3.6.

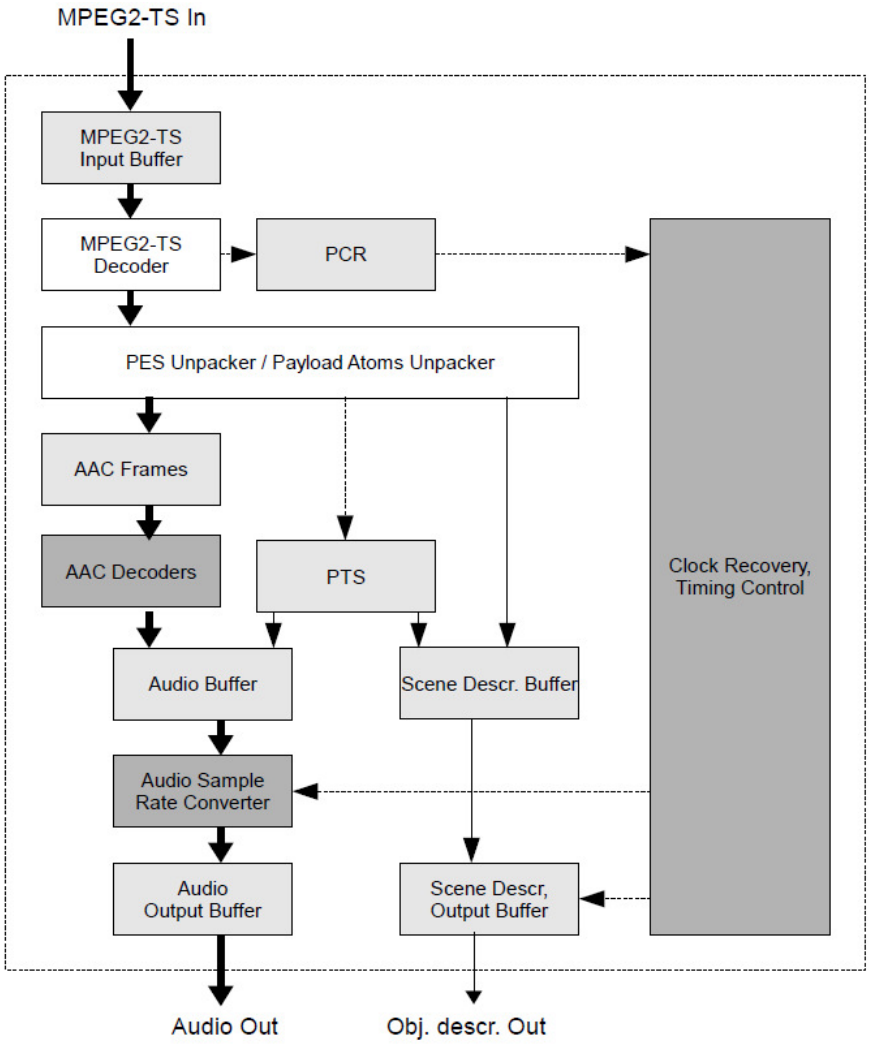

Figure 3: Object based audio scene decoder structure

Due to the use of this decoder structure in a streaming context, clock recovery from the network stream input is highlighted here. The incoming MPEG2-TS packets carry PCR timestamps that are fed into clock recovery. The TS payload is accumulated to extract Elementary Stream packets. If available, their PTS timestamps are extracted along with the audio scene segment payload.

After decoding the audio frames of all audio objects, the audio signals are fed into an audio buffer with an indication of their presentation timestamp (PTS). The object descriptions are fed into a scene description buffer with an indication of the same presentation timestamp (PTS). Both buffers are used to read audio and description data at the correct time.

The audio output signals are retrieved by applying advanced sample rate conversion / interpolated audio delay line methods. The recovered system clock controls the momentary sample rate and compensates for the PCR drifts and the TS jitter. From the scene description buffer, the scene with the closest distance of PTS to the current recovered clock is selected for output. To handle TS program changes or occurrences of discontinuous PCR values, the decoder module has methods for timing control.

The object based audio scene decoder retrieves a set of multiple audio channels and a sequence of accompanying object description data at the output.

2.5 Performance results

2.5.1 Implementation performance

Real-time encoding and decoding of audio scenes was tested during the implementation of the first DIOMEDES demonstrator. Both audio scene encoding and decoding have worked in real-time without errors or interruptions for a tested number of maximum 32 simultaneous audio objects, using a local network connection. Testing was successful for both CBR and VBR bitrate modes (tested bitrates: approx. 500 kbps up to 4000 kbps). Within the local network, clock recovery worked reliably.

2.5.2 Subjective quality tests

To judge the effect of AAC-coding on the perceived quality of the object based audio scenes, a series of subjective tests was conducted, which is described in more detail in Deliverable D3.4. The quality test should give an indication of which audio bitrate must be used for audio compression of the objects signals.

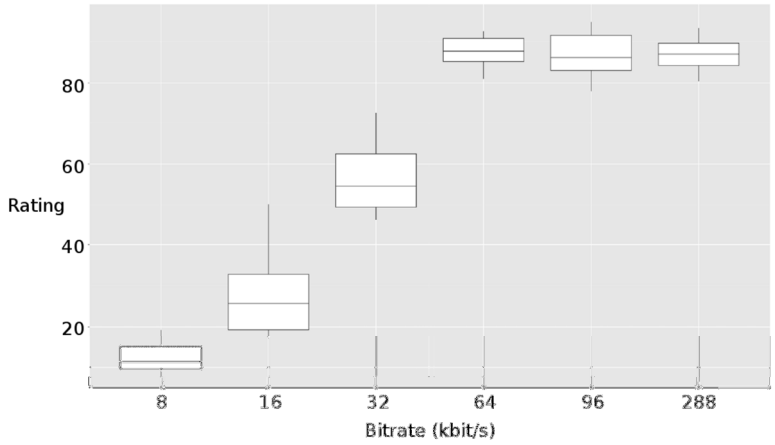

Figure 4: Box-plots for ratings of all stimuli of the respective bit rate

An Absolute Category Rating (ACR) experiment was conducted with 28 participants. A Wave Field Synthesis (WFS) reproduction system was used to reproduce different object based audio scenes containing audio material of different genres of up to 32 simultaneous audio objects. All stimuli were presented multiple times using varying bitrates for audio compression ranging from 8 kbps to 288 kbps (CBR AAC-LC - Fraunhofer IIS Encoder Library) per audio object (the highest bitrate used allows a near-transparent coding).

The participants rated the stimuli with a value from an 11-point quality scale (0…100 = Bad…Excellent). Mean opinion scores (MOS) were derived from these ratings. Figure 4 shows a box-plot of the experimental results. “Excellent” results were observed, when audio objects were coded at 64 kbps or higher bitrates. Bitrates of 32 kbps per object and lower lead to reduced average perceived quality. To provide excellent reproduction quality in CBR bitrate mode, the audio coding bitrate should be set to a value that is higher than a minimal bitrate between 32 and 64 kbps. A more detailed description of the results and discussion is documented in Deliverable D3.4.

3 FINAL VIDEO CODEC

3.1 Overview

This section reports the status of the final developed 3D scalable, visual attention adaptive

video codec. Updates in terms of functionality on top of the interim developed video encoder and decoder pair are explained in this section. The rate-distortion performance of the video codec is assessed objectively and subjectively. Please note that the terms visual saliency map and visual attention map are used interchangeably in the rest of this section.

3.2 Overview of the visual attention based quality scalability and the interim encoder-decoder pair

As depicted previously in D4.3, the strategy in utilising the visual attention model in DIOMEDES is to distribute the available bit-rate between the base layer and the enhancement layer(s) of the SVC coded camera viewpoints, such that the perceived quality drop between two successive layers at network congestion times is minimised. The overall objectives of the visual attention based scalable video encoder can be found in D4.3.

To group the visually salient frame regions in to distinct groups, Flexible Macroblock Ordering (FMO) tool in AVC standard and in its extension SVC, was decided to be utilised in the interim encoder design. FMO is primarily an error resilience tool, but is also a useful feature to manage Region-Of-Interest (ROI) coding. Because, according to its design, the macroblocks in a frame can be reorganised explicitly (with a formerly decided macroblock-to-slice group assignment map). This tool was handy and required least amount of modifications on the source code. But, later it was found that there are implications of forming arbitrary shaped slice groups based on the visual attention map on the standards conformance. Contrary to AVC, the slice group map type syntax element associated with the FMO tool is indeed restricted to take some values only in the SVC context. More specifically, the only allowed slice group formation method for FMO is to select rectangular macroblock groups with a left-over slice group covering all the macroblocks not taking place in one of the rectangular shaped macroblock groups. This restriction however prevents the use of the visual attention maps that do not naturally provide strictly rectangular saliency regions. Since the initial visual attention maps for the test video sequences were not ready to use at the time the first encoding tests were carried out with the interim encoder design, rectangular shaped manually sketched visual attention models were used. Hence, the appropriate slice group map type could be selected. However, in general to comply with the standard while exploiting the visual attention maps, the use of FMO was given up. The encoding results associated with the interim design can be found in D4.3.

3.3 Final video encoder specifications and functionality

It was explained in the interim report on the developed audio and video codecs (D4.3) that the quality scalability feature of the proposed multi-view video encoder is coupled with the utilisation of the content based visual attention information extracted prior to encoding. It was aimed that the bit-rate distribution among the frame sections in different layers results in the minimised loss of perceived quality of the decoded video at times of network adaptation. It was proposed to unequally allocate the spared bit-rate for a particular quality enhancement layer (or base quality layer), such that the pixels of visually salient frame regions are reconstructed with higher precision in that particular layer. Figure 5 depicts this procedure.

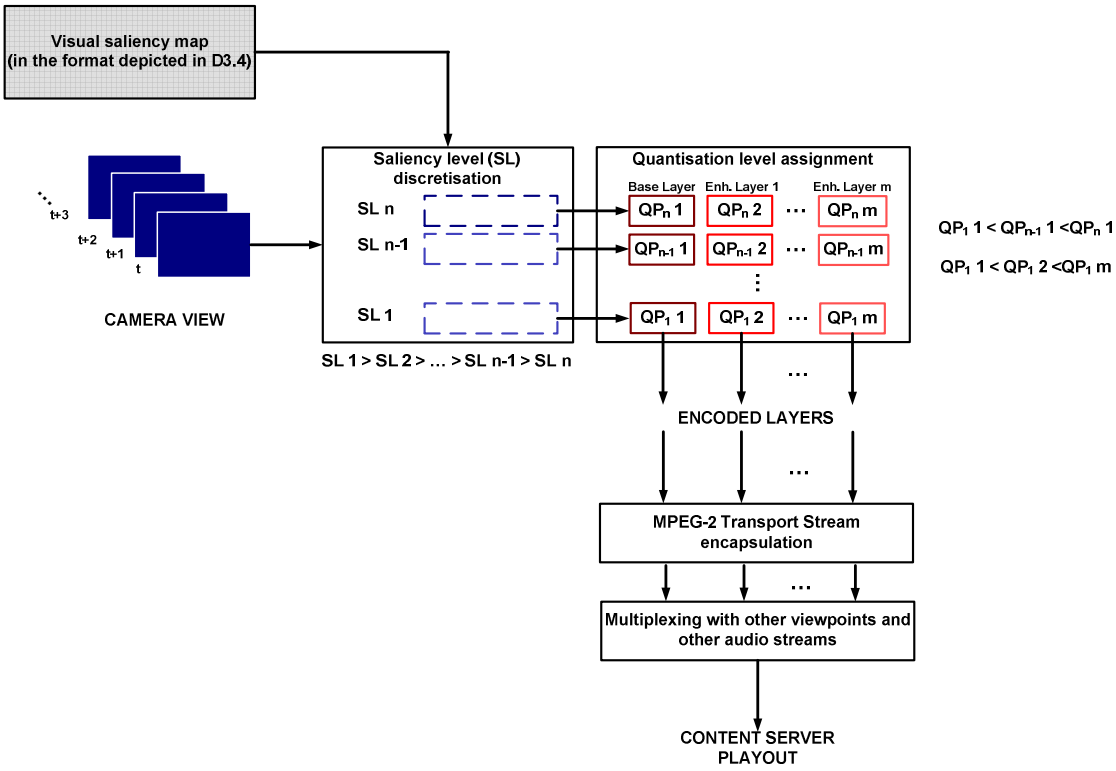

Figure 5: DIOMEDES encoding block for camera views using visual attention information

It is seen from Figure 5 that based on the content and consequently the extracted saliency map (for more information on how the visual attention map is extracted based on the developed visual attention model, please refer to Deliverables 3.2 and 3.4), each video frame is categorised into sub frame regions that have different visual importance. SL 1 is denoted as the most visually salient collection of pixels in the video frame. Please note however that since the target video encoder (SVC) is a block based encoder, the boundaries between discrete saliency regions are aligned with macroblock boundaries. Then, in the modified SVC encoder, for the base layer and for coarse grain scalability (CGS) layers, more than one Quantisation Parameter (QP) values are employed based on the visual saliency value of each processed macroblock (MB). The amount of the generated quality enhancement layers, as well as the step size between the QP values of two adjacent quality enhancement layers is primarily a matter of how the network adaptation will work. The selection of these two factors affects the bandwidth adaptation range and the granularity of adaptation to dynamic network state. The modification affects the assignment of the actual QP value used for the coding mode decision of an MB in slice encoding stage. Unless the specifically indicated QP value is assigned to a particular macroblock, the reference encoder enforces the base QP value in the encoded slice header to be assigned as the MB QP value. The implicit referral of the visual attention map value in assigning different quantisaion parameters does not necessitate encoding extra control information for correct decoding in the user terminal side. Because, the SVC (originally AVC) syntax element called *mb_qp_delta*, which depicts the difference between the QP value of a particular macroblock and the other syntax element *slice_qp_delta*, is coded to a non-zero value and can be decoded in the user terminal side to apply inverse quantisation process correctly. Please note that the utilisation method of the visual attention map in the final encoder design is pretty different than the method previously applied in the interim design. Because, at that time only rectangular shaped regions of interest could be formed (based on restricted standards allowance) and a layer was inherently fragmented into more than one NAL units. This was enforced by the FMO concept. However in the final design, arbitrary shaped visual attention maps can be employed in assigning various QP values to macroblocks

within a single slice group, and therefore a layer of an access unit can be packed into a single NAL unit packet. In the following section, several test results are presented to show the rate distortion performance of the proposed final encoder design with respect to native SVC encoder that has the same amount of layers at similar bit-rates.

3.3.1 Rate-distortion performance and subjective test results

Based on previously set encoding requirements depicted in D4.1 and D4.2, the total number of layers per camera viewpoint (only colour component) was selected as 2, i.e. one base layer and one quality enhancement layer. Hence, *m* is set to 1 in Figure 5. Similarly, we have chosen to discretise the visual attention values into two bins, such that a macroblock is either a salient macroblock or is a member of the left-over region. Hence, there are two saliency levels used in the experiments.

In Figure 6, we first show a set of encoding results for *Music* scene, where than the base quality layer is homogeneously encoded (i.e. all macroblock QP's within the base layer are equal), whereas only the quality of the salient macroblocks are enhanced in the quality enhancement layer. Seven different combinations of QP values are shown in Figure 6, where the QP combinations depicted with numbers are shown in Table 2. The results are for a combination of two cameras setting up the stereoscopic video.

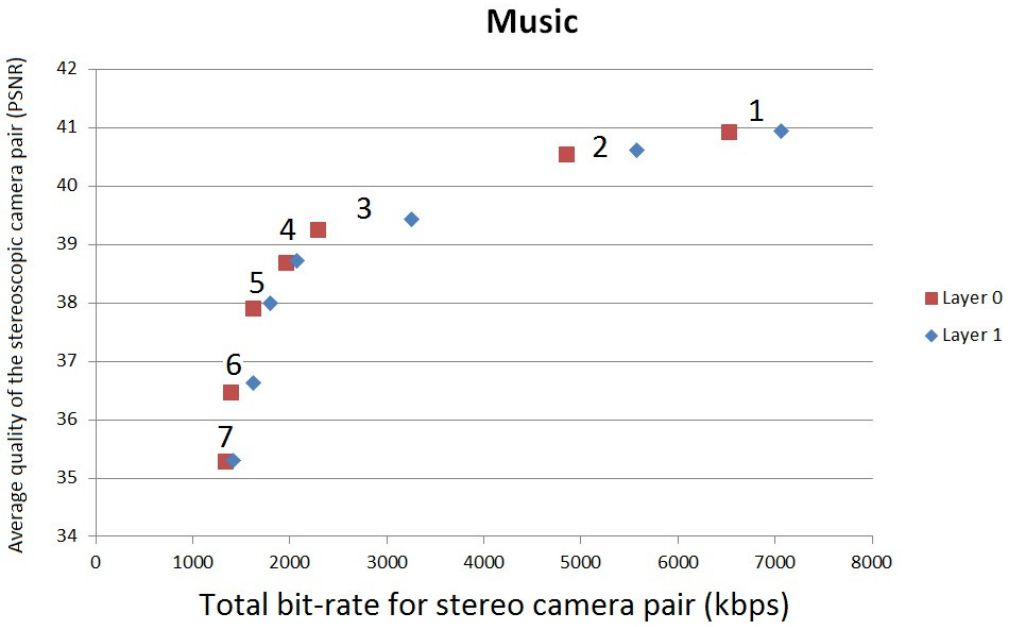

Figure 6: R-D curve for stereoscopic *Music* scene coded with two layers

Table 2: QP combinations used for encoding attention area and leftover region

Reference	QP over attention area	QP over leftover region
1	20	22
2	20	24
3	20	30
4	30	32
5	30	35
6	30	40
7	40	44

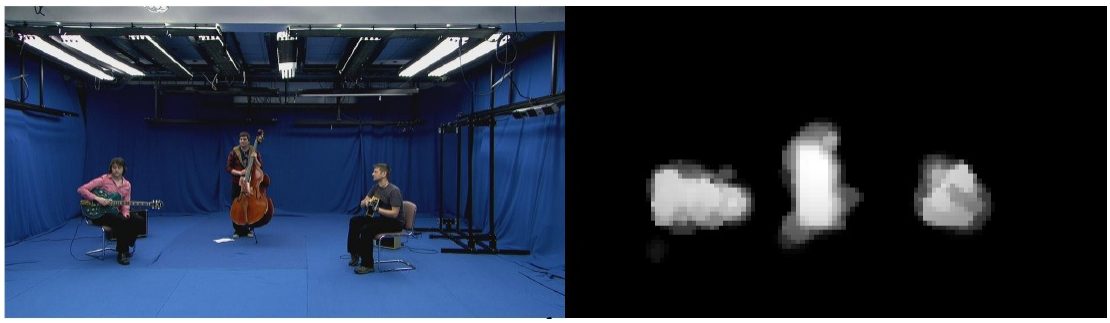

Figure 7: Sample frame from *Music* scene with the corresponding visual attention map

Figure 7 shows a sample frame from the test scene and its corresponding macroblock based visual attention map, where all black blocks represent left-over regions and other blocks are counted as salient regions. The same results shown here are also used in D3.4, where they are assessed subjectively. From the objective R-D performance results shown in Figure 5, it can be seen that the PSNR difference between two layers in almost all test points is marginal. In other words, purely based on PSNR judgement, it is unlikely to say that the additional bit-rate necessary for the quality enhancement layer is worthwhile. Such results are expected, since objective evaluation based on PSNR is not taking into consideration perceptual attributes. Looking at Figure 7, it is easy to see that the proportion of the visually salient macroblocks to the overall image area is much smaller than that of the left-over macroblocks. Therefore, it is a comparatively smaller region whose reconstruction fidelity is increased in the enhancement layer. The rest is not affected. Hence, the quality increase in the overall image is bounded by the quality increase in the proportionally smaller region. However, based on the analysis and subjective tests done and reported in D3.4, it is rather easy to see that the increase in the perceptual quality (mean opinion score) between layers is significant as opposed to the marginal change in PSNR. This is to draw the conclusion that objective scores based on the widely used PSNR metric cannot be decisive in assessing the performance of the visual attention adaptive scalable video encoder. However, as the QoE model for L-R stereoscopic video (as part of Task 3.2) and the quality metric that is based on the individual qualities of left and right videos are being developed while preparing this document, the objective performance analysis of the codec, along with the P2P streaming overlay, based on the actual QoE model will be further elaborated in D4.5 (Report on performance of integrated MD-SMVD and P2P system).

In the other set of experiments, we have used 5 different stereoscopic HD test sequences to compare the performance of the proposed DIOMEDES codec under the intended objective (e.g. quantising the perceptually significant macroblocks with higher precision in the base quality layer, whereas enhancing the quality of only the left-over macroblocks in the

subsequent enhancement layers) against the anchor scheme, where no visual attention information is taken into account (e.g. all macroblocks within a layer have equal QP values). The test stereoscopic video sequences we used are DIOMEDES test sequences *Music* and *Lecture*, and other test sequences such as *Café*, *Street* and MUSCADE *Band*. The test is done in a single operating point, where the base layer bit-rate is set between 3 Mbps – 4.5 Mbps and the total bit-rate of the stream is set around 6 Mbps – 8.3 Mbps (depending on the content). The bit-rate values are depicted in Table 3 below. The first bit-rate specified in Table 3 is for the base layer only, whereas the second bit-rate value is the total bit-rate of the stream.

We encoded 20 second long clips from each stereoscopic test video. In order to simulate a realistic network adaptation scenario during P2P streaming, we have manually truncated the enhancement layer packets of a block of frames (i.e. chunks, because they are the smallest transmission unit in the DIOMEDES P2P overlay) from the downloaded content.

Table 3: Operting point bit-rates for the test stereoscopic video sequences

	Anchor SVC	Visual Attention adaptive SVC
Band	4.4 Mbps – 8.3 Mbps	4.5 Mbps – 8.3 Mbps
Café	3.8 Mbps – 6 Mbps	3.9 Mbps – 6.2 Mbps
Street	3.4 Mbps – 7.1 Mbps	3.4 Mbps – 6.9 Mbps
Music	3.6 Mbps – 7.2 Mbps	3.7 Mbps – 7.3 Mbps
Lecture	3.4 Mbps – 6.5 Mbps	3.3 Mbps – 6.4 Mbps

We truncated equal amount of packets at arbitrary places in the streams, for both the anchor and proposed encoding scenarios. In order to find out the effectiveness of the proposed visual attention based scalable multi-view video encoding under the need for network adaptation, and in order to compare it with the anchor scheme, we have done a subjective test. Stimulus Comparison Adjectival Categorical Judgement (SCACJ) method specified in ITU-R BT.500-11 standard is used to assess the performance. 7 expert subjects are asked to visually compare the test videos each time, where the formation of the decoded test videos is as depicted in Figure 8. In each test, subjects compare Video A against Video B.

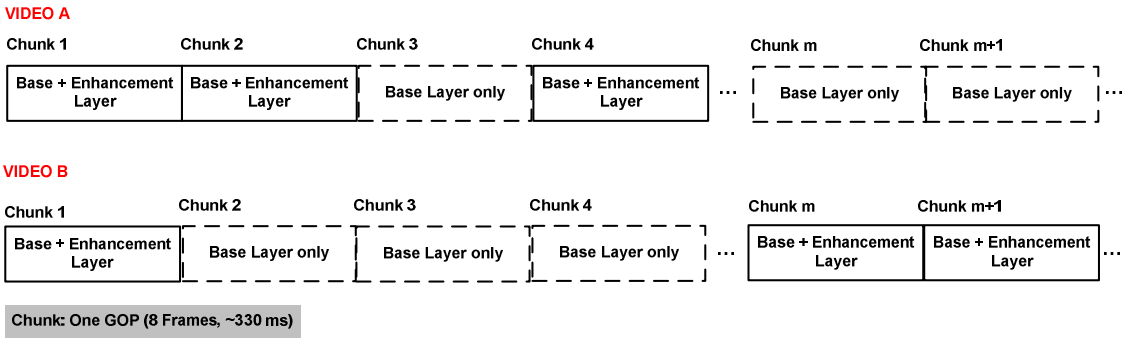

Figure 8: Test video formation for subjective experiments

Figure 9 shows the corresponding Mean Opinion Score (MOS) results, where the performance of both encoding schemes is plotted for each test sequence separately, including the 95% confidence interval. The opinion score axis is normalised to the range of [-1, 1].

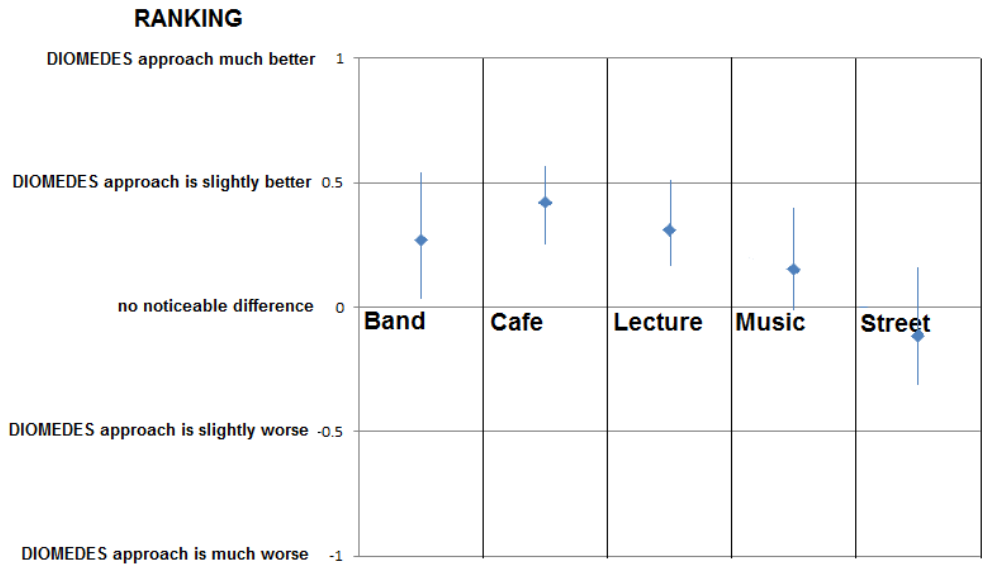

Figure 9: Subjective test results

It is seen from the graph that the majority of the presented results are in the positive range, meaning that the application of the visual attention based quality scalable video coding improves the quality of experience with respect to not applying it, especially at the times of bandwidth adaptation. It should be noted that at the times, where the bandwidth state is fine and the 3D view pair can be streamed with both layers, the perceptual quality is equivalent for both the anchor coding scheme and the proposed visual attention based scalable video coding.

3.4 Final video decoder specifications and functionality

The decoder adopted in DIOMEDES project is a modified version of an open source library named as "openSVCdecoder". This was described before in D4.1 and D4.3. It is built upon the famous FFmpeg decoder and several modifications had been implemented on it to support SVC decoding. In this sense, the decoder is composed of optimized codes that take advantage of MMX instructions to increase the performance.

Unfortunately, it was not possible to use the library as it is, because it was originally implemented in MPlayer. Therefore, we could not obtain reconstructed raw images and forward them to the DIOMEDES 3D player. Yet another problem was that it did not originally have the networking functionality that is required in DIOMEDES.

In order to build a decoding platform that is compatible with the other modules in the user terminal device, we have pinpointed the parts that are specific to decode SVC elementary bit-streams. On top of that, we have implemented the networking functionality, which enables the decoder to receive NAL units sent by the elementary stream (ES) demultiplexer module. That input stream is already in decodable format, meaning that the base layer chunks and the enhancement layer chunks are merged in the appropriate way.

The decoding operation of a frame can only start when the NAL unit of the next frame is available in the buffer because the decoder determines the size of the current NAL unit using the NALU start indicator (0x00 0x00 0x01) of the next unit. Moreover, the decoding should be performed on access unit basis, meaning that all the NAL units regarding a frame should be available. This means that even if the base layer NAL unit received, it cannot be directly decoded because the enhancement layer NAL unit should be present while decoding the base layer NAL unit. Nevertheless, the decoder can start decoding the base layer NAL unit, even if the enhancement layer NAL unit has not arrived yet, if the prefix NAL unit for the next access unit arrives at its buffer. The delay between the base layer and enhancement layer NAL units

is negligible. Therefore, the only critical issue is to implement an algorithm that performs decoding operation only when appropriate.

In order to deal with these requirements, we have adopted the following algorithm for initiating the decoding process. The figure presents the bit-stream of SVC encoding video. It starts with non-video coding layer (non-VCL) NAL units followed by the VCL NAL units. The SVC is backwards compatible. Therefore, the base layer NAL units do not have any information regarding SVC. Instead, SVC related information for that particular base layer NAL unit is provided in the prefix NAL unit. In DIOMEDES we use these prefix NAL units as markers. The decoder waits until the prefix of the next frame is received. Only then it starts the decoding process for that particular access unit. Figure 10 depicts the mentioned SVC bit-stream format.

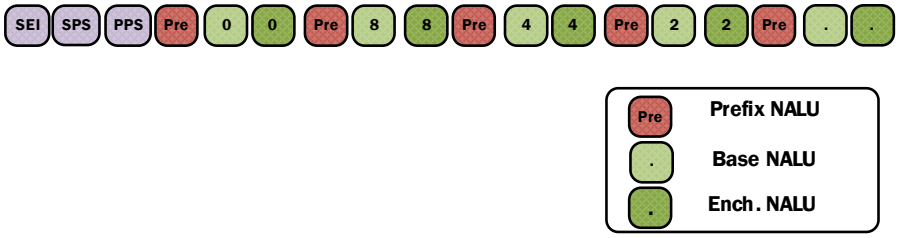

Figure 10: SVC bit-stream format

The final decoder also has a special task for the adaptation purposes. There are Key Performance Index (KPI) NAL units embedded into the SVC bit-stream, which provide information necessary for video adaptation decision taking. The details of video adaptation are described in detail in D3.4. It is the task of the decoder to extract these SEI NAL units and forward them to the Adaptation Decision Engine within the user terminal device. Extraction is performed by analysing the NAL_UNIT_TYPE field in the NAL unit header.

3.4.1 Real-time compatibility results

The current focus in simultaneous SVC based multi-view video decoder module is its decoding performance. This section provides the current state we have achieved in SVC video decoding.

Up to now, we have performed decoding operation using Intel-i5 processor based PCs, which provided adequate performance for both HD-Ready and Standard Definition (SD) resolutions considering base layer video only. With either resolution, the decoder exceeds the minimum needed decoding rate (e.g. video playback rate), even if more than 2 viewpoints are decoded simultaneously (e.g. single view decoding rate is over 70 fps, for 4 views in a single PC the decoding rate is around 27-28 fps). However, when quality enhancement layer NAL units are decoded, the performance is significantly degraded.

In order to increase the decoding speed and obtain an acceptable performance when the enhancement layer NAL units are also decoded, more advanced hardware is purchased. The following shows the specific PC configuration used to test the decoder performance:

- Intel Core i7 990X (Extreme) 3.4 GHz 6 Cores (12 processes with Hyper Threading)
- Main Board: Asus Rampage Extreme (2200 MHz with O.C.)
- RAM: 8Gb Kingston (4 x 2Gb)

Using this configuration, the processing rate of the CPU could be overclocked up to 4.1 GHz. At this computing rate, the decoder could simultaneously decode 4 HD-Ready coded views with quality enhancement layers at around ~23 fps. Nevertheless, in order to avoid inflicting damage to the hardware as a result of overclocking the native computing rate of the CPU,

specific caution needs to be taken. Therefore, hardware specialists are consulted in the meantime of the preparation of this report and improved and more extensive decoding results will be reported as part of D4.5 tests.

4 CONCLUDING REMARKS

This deliverable has outlined the details of the design and the implementation of the final prototype audio and video codecs. Namely, the additional features on top of the ones defined in the interim codec architectures, reported previously in D4.3 are implemented in the final prototype codecs to analyse their performances under particular DIOMEDES use-cases. Multi-channel audio object coder and decoder is implemented and demonstrated to function in real-time. Subjective experiments are carried out to verify the design features. Final prototype visual attention based scalable video encoder is implemented and subjective tests are carried out to assess its performance against the simultaneous SVC based multi-view video coding. Final prototype real-time scalable multi-view video decoder for 4 concurrent camera views with quality enhancement layers is implemented using special hardware design for improved decoding speed. Decoding tests are done on an overclocked quad-core architecture to show the real time decoding performance. For the video encoder and decoder, further performance results will be reported in the upcoming deliverable D4.5.

REFERENCES

[1] ISO/IEC 14496-14 “MPEG-4, Information technology — Coding of Audio, Picture, Multimedia and Hypermedia Information — Part 14: MP4 file format”.

[2] ISO/IEC 13818-1 “Information technology — Generic coding of moving pictures and associated audio information: Systems”, Second edition 2000-12-01.

APPENDIX A: GLOSSARY OF ABBREVIATIONS

AAC	Advanced Audio Coding
ACR	Absolute Category Rating
AVC	Advanced Video Coding
CABAC	Context-Adaptive Binary Arithmetic Coding
CBR	Constant Bit-Rate
CGS	Coarse Grain Scalability
CPU	Central Processing Unit
DVB	Digital Video Broadcasting
ES	Elementary Stream
FMO	Flexible Macroblock Ordering
FPS	Frames per Second
GOP	Group of Pictures
HD	High Definition
ITU-R	International Telecommunications Union-Recommendation
KPI	Key Performance Index
MB	Macroblock
MD-SMVD	Multiple Description – Scalable Multi-View Depth Coding
MMX	Single instruction, Multiple Data instruction set (by Intel)
MOS	Mean Opinion Score
MP4	MPEG-4 Part 14 (formally ISO/IEC 14496-14:2003)
MPEG	Motion Picture Experts Group
MPlayer	A free and open source media player
MVC	Multi-View Coding
NAL	Network Abstraction Layer
NALU	NAL Unit
P2P	Peer-to-Peer
PAT	Program Association Table
PCR	Program Clock Reference
PES	Packetised Elementary Stream
PMT	Program Mapping Table
PSNR	Peak Signal-to-Noise Ratio
PTS	Presentation Time Stamp
QoE	Quality of Experience
QP	Quantisation Parameter
RAM	Random Access Memory
R-D	Rate-Distortion
ROI	Region of Interest
SCACJ	Stimulus Comparison Adjectival Categorical Judgement
SD	Standard Definition
SNR	Signal-to-Noise Ratio
SVC	Scalable Video Coding
TS	Transport Stream
UDP	User Datagram Protocol

VAM	Visual Attention Model
VBR	Variable Bit-Rate
VCL	Video Coding Layer
WFS	Wave Field Synthesis