ETSI TR 126 905 V13.0.0 (2016-01)

Digital cellular telecommunications system (Phase 2+);
Universal Mobile Telecommunications System (UMTS);
LTE;
Mobile stereoscopic 3D video
(3GPP TR 26.905 version 13.0.0 Release 13)

Reference
RTR/TSGS-0426905vd00

Keywords
GSM,LTE,UMTS

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Intellectual Property Rights

IPRs essential or potentially essential to the present document may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Foreword

This Technical Report (TR) has been produced by ETSI 3rd Generation Partnership Project (3GPP).

The present document may refer to technical specifications or reports using their 3GPP identities, UMTS identities or GSM identities. These should be interpreted as being references to the corresponding ETSI deliverables.

The cross reference between GSM, UMTS, 3GPP and ETSI identities can be found under http://webapp.etsi.org/key/queryform.asp.

Modal verbs terminology

In the present document "**shall**", "**shall not**", "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Contents

Foreword

This Technical Report has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

- x the first digit:
  - 1 presented to TSG for information;
  - 2 presented to TSG for approval;
  - 3 or greater indicates TSG approved document under change control.
- y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.
- z the third digit is incremented when editorial only changes have been incorporated in the document.

1 Scope

The present document provides a study of stereoscopic 3D video services over 3GPP networks and terminals. Technical definitions, use case descriptions, working assumptions, subjective tests results and technical studies are presented.

This document identifies the gaps within the Release 10 3GPP specifications in order to enable the implementation of the mobile 3D video use cases.

2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non-specific.
- For a specific reference, subsequent revisions do not apply.
- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TS 26.114: "IP multimedia subsystem (IMS); Multimedia telephony, Media handling and interaction".

[3] 3GPP TS 26.234: "Transparent end-to-end packet switched streaming service (PSS); Protocols and codecs".

[4] 3GPP TS 26.346: "Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs".

[5] 3GPP TS 26.247: "Transparent end-to-end Packet-switched Streaming Service (PSS); Progressive Download and Dynamic Adaptive Streaming over HTTP (3GP-DASH)".

[6] 3GPP TS 26.140: "Multimedia Messaging Service (MMS); Media formats and codecs".

[7] 3GPP TS 26.245: "Transparent end-to-end Packet-switched Streaming Service (PSS); Timed Text Format".

[8] 3GPP TS 26.430: "Timed Graphics".

[9] 3GPP TR 26.904: "Improved Video Coding Support".

[10] IETF RFC 3261: "SIP: Session Initiation Protocol".

[11] IETF RFC 3264: "An Offer/Answer Model with the Session Description Protocol (SDP)".

[12] IETF draft draft-ietf-payload-rtp-mvc-01: "RTP Payload Format for MVC Video".

[13] IETF personal draft draft-greevenbosch-mmusic-sdp-3d-format-002: "Signal 3D format".

[14] ITU-R Recommendation BT 1788: "Methodology for the subjective assessment of video quality in multimedia applications".

[15] 3GPP TS 26.244: "Transparent end-to-end packet switched streaming service (PSS); 3GPP file format".

[16] ISO/IEC 14496-15: 2010: "Information technology – Coding of audio-visual objects – Part 15: Advanced Video Coding (AVC) file format".

[17] ISO/IEC 14496-12:2008/Amd2 | 15444-12:2008/Amd2: " Part 12: ISO base media file format AMENDMENT 2: Support for sub-track selection & switching, post-decoder requirements, and color information ISO base media file format".

[18] ISO/IEC 14496-12:2008 | 15444-12:2008: "Information technology – Coding of audio-visual objects – Part 12: ISO base media file format" | "Information technology – JPEG 2000 image coding system – Part 12: ISO base media file format".

[19] JM H.264/AVC Reference Software, http://iphome.hhi.de/suehring/tml/download/.

[20] JMVC H.264/MVC Reference Software, Version 8.5, March 2011.

[21] Proceedings of the IEEE: "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard", A. Vetro, T. Wiegand, and G. Sullivan, ,, vol. 99, no. 4, p. 626642, 2011.

[22] Doc. JVT-AE022:"Coding performance of stereo high profile for movie sequences, London, U.K., Joint Video Team (JVT)", T. Chen, Y. Kashiwagi, C. S. Lim, and T. Nishi , Jul. 2009.

[23] Broadcasting, IEEE Transactions on: 'Studies on the bit rate requirements for a HDTV format with 1920x1080 pixel resolution, progressive scanning at 50 Hz frame rate targeting large flat panel displays,' H. Hoffmann, T. Itagaki, D. Wood, and A. Bock, , vol. 52, no. 4, pp. 420–434, 2006.

[24] 3GPP TS 26.237: "IP Multimedia Subsystem (IMS) based Packet Switch Streaming (PSS) and Multimedia Broadcast/Multicast Service (MBMS) User Service; Protocols ".

[25] IETF RFC 6381: "The "Codecs" and "Profiles" Parameters for ``Bucket´´ Media Types", Gellens R., Singer D. and Frojdh P., August 2011.

[26] ISO/IEC 23009-1: 'Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats.

[27] Open Mobile Alliance: "User Agent Profile Version 2.0", February 2006.

[28] 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON): ,"Adaptive parallax for 3D television", Ide, K. and Sikora, T., 2010, vol., no., pp.1-4, 7-9 June 2010.

3 Definitions and abbreviations

3.1 Definitions

For the purposes of the present document, the terms and definitions given in TR 21.905 [1] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in TR 21.905 [1].

3.2 Abbreviations

For the purposes of the present document, the abbreviations given in TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in TR 21.905 [1].

AVC	Advanced Video Coding
DASH	Dynamic Adaptive Streaming over HTTP
HDMI	High-Definition Multimedia Interface
HTTP	Hypertext Transfer Protocol
IMS	IP Multimedia Subsystem
IP	Internet Protocol
IPD	Interpupillary Distance
IR	Infrared
LTE	Long Term Evolution
MBMS	Multimedia Broadcast/Multicast Services
MPD	Media Presentation Description
MTSI	Multimedia Telephony Services for IMS
MVC	Multiview Video Coding
PSS	Packet Switched Streaming Service
RTP	Real Time Protocol
SDP	Session Description Protocol

4 General

4.1 Introduction

This Technical Report provides a study on mobile 3D stereoscopic video in 3GPP. Use cases and technical solutions are investigated regarding a variety of setups using 3GPP's streaming, multicast/broadcast, download and progressive download as well as conversational services. Clause 5 provides a definition of the stereoscopic 3D video technologies and terminology as well as a video codecs performance comparison. Clauses 6 and 7 focus on use cases for which the working assumptions and the operation points are defined before providing a technical analysis, whereas clause 8 provides a set of use cases in which further study is required so as to identify the gaps. Clause 9 introduces subjective tests conducted on a 3D capable mobile terminal and clause 10 presents a generic approach for 3D content adaptation depending on the client terminal charasteristics. The conclusion summarizes the recommended way forward for the introduction of 3D stereoscopic video support in 3GPP specifications.

5 Technology description

5.1 Mobile 3D rendering technologies

5.1.1 Introduction

Stereoscopy is the method of combining two plane pictures in order to produce a depth perception by the human brain. Each eye seeing a different angle of a scene, the human visual system - with subjective assessments - is able to interpret the depth information.

In the scope of the present document, this section provides some information on how the rendering technologies provide the depth perception. These technologies are split into two categories; the glasses based systems and the glasses free systems.

5.1.2 Glasses-free 3D video rendering technologies

5.1.1.1 Parallax barrier

The parallax barrier consists in a grid placed over the screen. When electrically activated, this barrier prevents the eyes of the user from viewing all the pixels of the display such as depicted in the figure 1. The resulting quality of experience is half the resolution per view compared with the 2D mode (i.e. when the barrier is switched off).

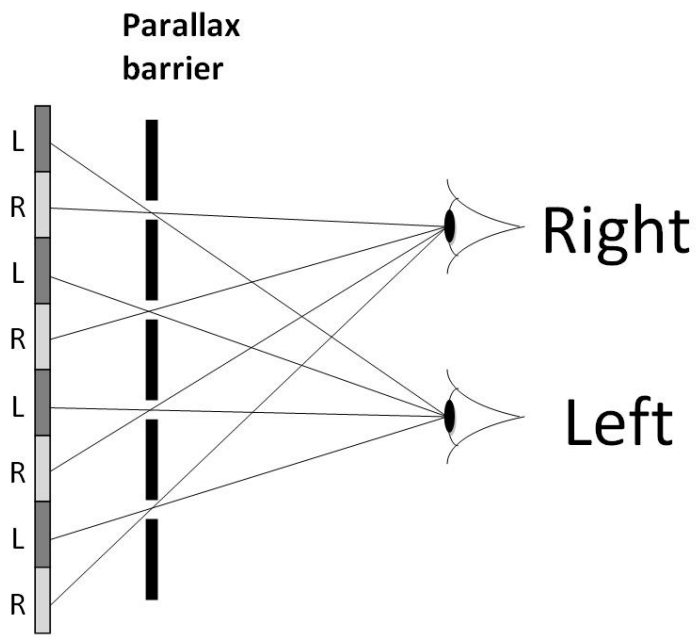

Figure 1: Parallax barrier

5.1.1.2 Lenticular lens sheet

This rendering technology is based on a lens sheet. It consists in a series of vertical hemi-cylindrical lenses placed so as to direct light in different viewing angles. When correctly placed, each eye can receive a different view from the other, as shown on the figure 2.

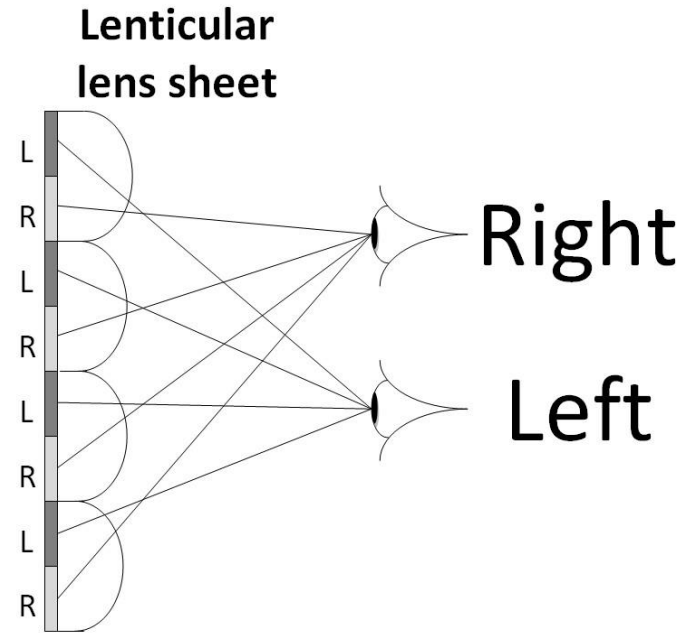

Figure 2: Lenticular lens sheet

5.1.3 Glasses-based 3D video rendering technologies

5.1.3.1 Active-shutter glasses

The active-shutter glasses are synchronized with the 3D display (potentially with IR signal transmitted from the glasses to the terminal) which displays alternatively the left and right views of a video. The figure 3 below illustrates such a case.

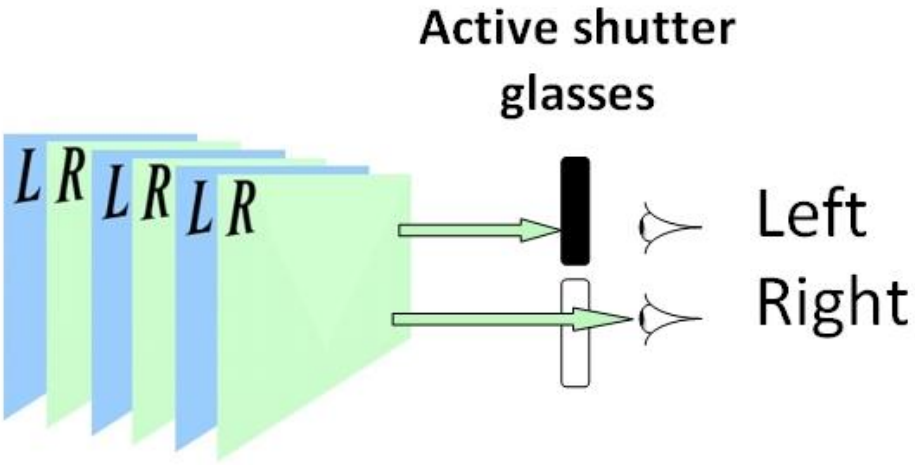

Figure 3: Active shutter glasses

5.1.3.2 Passive glasses

Passive glasses use a polarized filter placed on both the screen and the glasses. For example, the current 3D displays can interlace the left and right views in a single image on the screen whereas the filters on the glasses only allow the left eye to see the odd lines (in red on figure 4) and the right eye to see the even lines of the screen (in green on figure 4). In this case, image resolution is halved if compared to active systems but new systems such as active retarder will attempt to solve this problem.

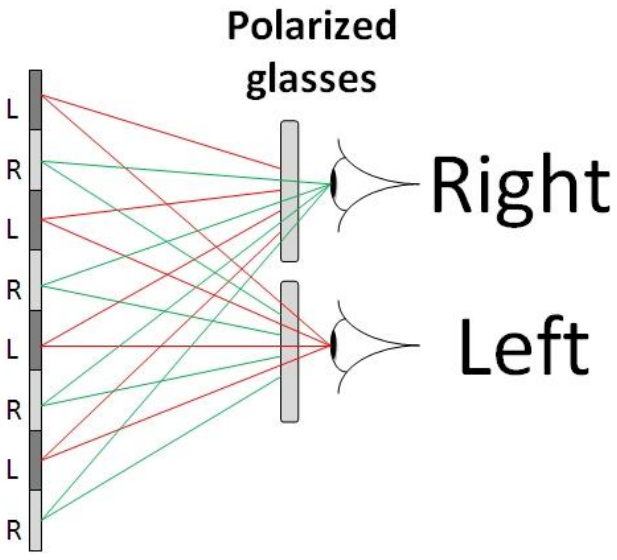

Figure 4: Passive polarized glasses

5.1.4 Potential impacts on a 3D service implementation

Given the fact that the rendering technologies offer different levels of quality of experience such as the resolution per view, the viewing angles… a service may benefit from adapting the provided 3D video format to the rendering technology in use. In this case appropriate signalling is necessary to either describe the different formats such that the client can select/request the format or the appropriate signalling of the rendering technology is important such that the server can select or annotate the appropriate format.

Depending on the service, these formats may have to be mapped to the different signalling frameworks in which the 3D video is offered, e.g. MPD in 3GP-DASH, SDP for MTSI and PSS, etc.

5.2 Stereoscopic 3D frame packing formats

5.2.1 Frame-compatible packing formats

The frame-compatible packing format consists in sub-sampling the two views which compose a stereoscopic 3D video and pack them together in order to produce a video signal compatible with a 2D frame infrastructure.

In a typical operation mode, the spatial resolution of the original frames of each view and the packaged single frame, have the same resolution. The spatial packing arrangement may use a side-by-side, top-bottom, interleaved, or checkerboard format as illustrated in figure 5 and the down-sampling process should be performed accordingly.

In most commercial deployments only side-by-side or top-bottom frame packing arrangements are applied.

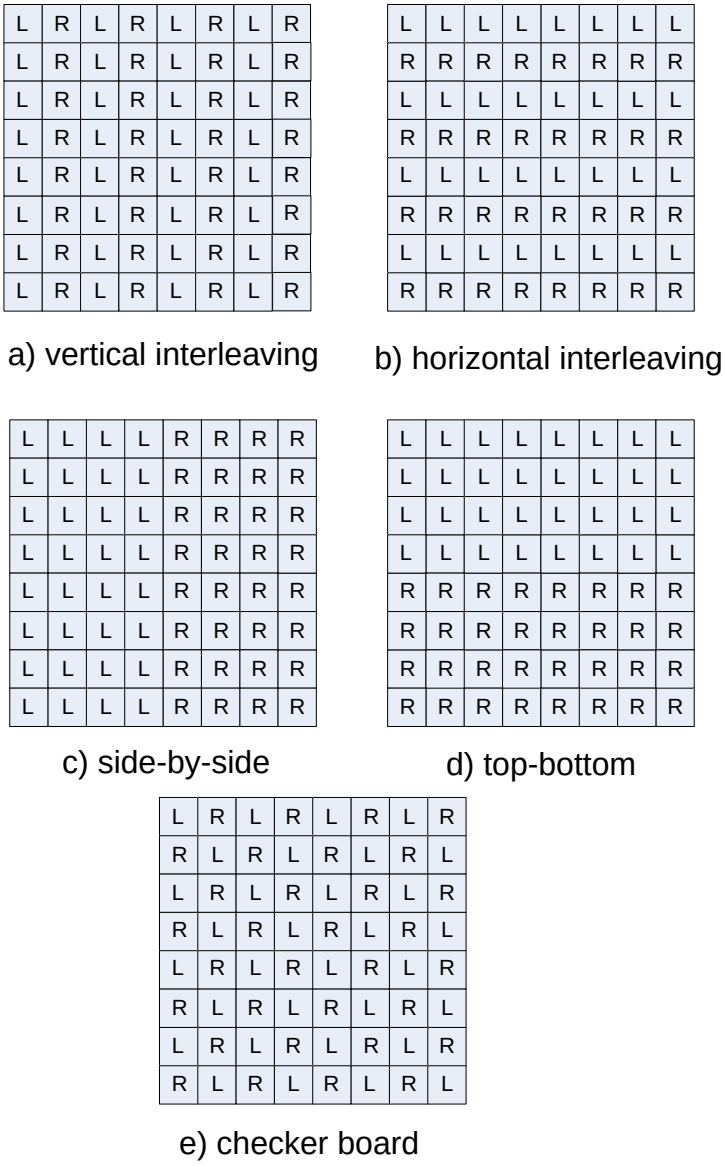

Figure 5: Spatial frame packing formats

5.2.2 Full resolution per view packing formats

In order to avoid the lack of definition introduced by the frame-compatible packing formats, it is possible to transmit both views at full resolution. In this case, the amount of data is twice as much as the frame compatible packing formats. Although the spatial packing format can be used in order to generate a twice bigger image, the most common format is the frame packing for which the left and right views are temporally interleaved such as shown on the figure 6 below.

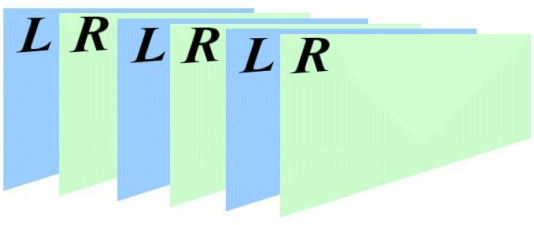

Figure 6: Temporal interleave packing format

5.3 Video codecs for stereoscopic 3D

5.3.1 H.264/AVC for frame compatible packing formats

In frame-compatible stereoscopic video, at the encoder side a spatial packing of a stereo pair into a single frame is performed and the single frames are encoded. The output frames produced by the decoder contain constituent frames of a stereo pair. The encoder side indicates the used frame packing format by including one or more frame packing arrangement supplemental enhancement information (SEI) messages as specified in the H.264/AVC standard into the bitstream. The decoder side should decode the frame conventionally, unpack the two constituent frames from the output frames of the decoder, do up-sampling to revert the encoder side down-sampling process and render the constituent frames on the 3D display.

5.3.2 H.264/AVC for temporal interleaving packing format

In temporal interleaving, the video is encoded at double the frame rate of the original video as illustrated in figure 7. Each pair of subsequent pictures constitutes a stereo pair (left and right view). The rendering of the time interleaved stereoscopic video is typically performed at the high frame rate, where active (shutter) glasses are used to blend the incorrect view at each eye. This requires accurate synchronization between the glasses and the screen.

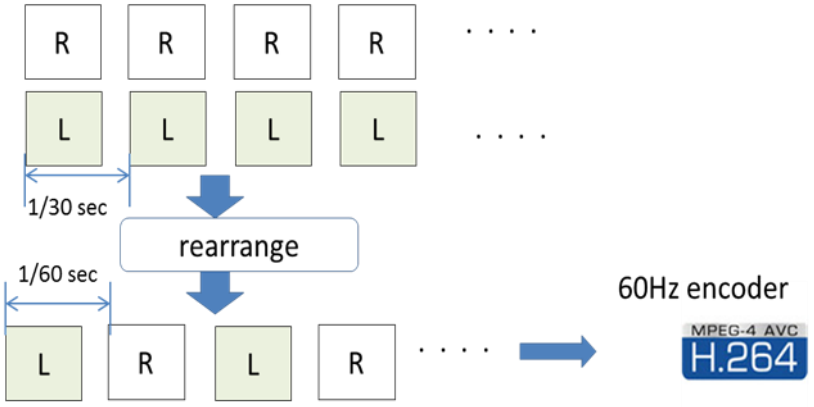

Figure 7: H.264/AVC with temporal Interleaving

5.3.3 MVC (Multiview Video Coding)

Multiview Video Coding was standardized as an extension (annex) to the H.264/AVC standard. In MVC, the views from different cameras are encoded into a single bitstream that is backwards compatible with single-view H.264/AVC. One of the views is encoded as "base view". A single-view (e.g. constrained baseline or progressive high profile) H.264/AVC decoder can decode and output the base view of an MVC bitstream. MVC introduces inter-view prediction between views exemplarily as illustrated in figure 8. MVC is able to compress stereoscopic video in a backwards compatible manner and without compromising the view resolutions. If the server is aware of the UE capabilities, it can omit sending the view components of the non-base view to a device that does not support 3D or does not have enough bitrate to deliver both views.

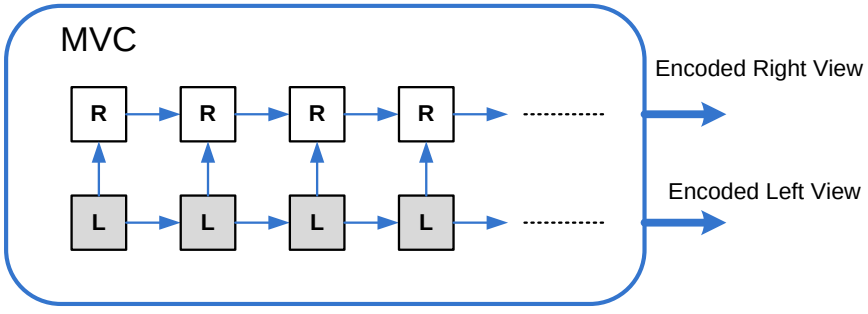

Figure 8: MVC encoding with exemplary inter-view prediction, the left view being the base view

5.3.4 Performance evaluation of the compression efficiency

5.3.4.1 Simulation setup

The following formats for stereoscopic 3D video are compared:

- Side-by-Side frame packing (SBS)
- Top-Bottom frame packing (TB)
- Side-by-Side full resolution frame packing (SBSF)
- Temporal Interleaving (TMP)
- Multiview Video Coding (MVC)

For side-by-side and top-bottom frame packing formats, the left and right views are sub-sampled to yield a packed frame that has the same resolution as the original view resolution.

For side-by-side full resolution frame packing the left and right views are not sub-sampled. The resulting packed frame has the same height and double width.

In temporal interleaving, the video is encoded at double the frame rate of the original video. Each pair of subsequent pictures constitutes a stereo pair (left and right view).

For the H.264/AVC encoding of the packed formats and time interleaving JM [19] has been used. The JMVC [20] encoder has been used to encode the MVC sequences.

NOTE: JMVC and JM reference software differ in maturity of the optimization level, and enhancements to both software implementations are possible. Therefore, results generated by these software implementations should not be taken as the definite.

The following encoding parameters have been used:

- Fixed QP for I, P, B pictures:

  20-34 for Side-by-Side and Top-Bottom frame packing

  24-38 for Side-by-Side full resolution frame packing, Temporal Interleaving, and MVC
- Reference Frames: 2
- Frame Rate: 30
- GOP period: 16 frames
- High profile conformance
- Motion estimation search range: 16

The test sequences that have been used are:

- Flower3: 640x360, 111 frames
- Car: 640x360, 234 frames
- Horse: 640x360, 139 frames
- Caterpillar: 640x360, 100 frames

For side-by-side and top-bottom frame packing formats, the left and right views were unpacked and up-sampled before PSNR calculation. For the down-sampling and the up-sampling procedure the code from JMVC reference software [20] has been used.

NOTE: The type of chosen up-sampling algorithm has a significant impact on results achieved by the Side-by-Side (SBS) and Top-Bottom (TB) frame packing formats. There may exist an up-sampling algorithm that would provide better performance than the up-sampling algorithm used in the simulation.

5.3.4.2 Simulation results

The following figures depict the Rate-Distortion curves for different formats of stereoscopic 3D video, compression configurations, and different video sequences. Figures 9 to 12 present results for all QPs mentioned in section 5.3.4.1 while Figures 13 to 16 present results for a smaller range of QPs for better readability.

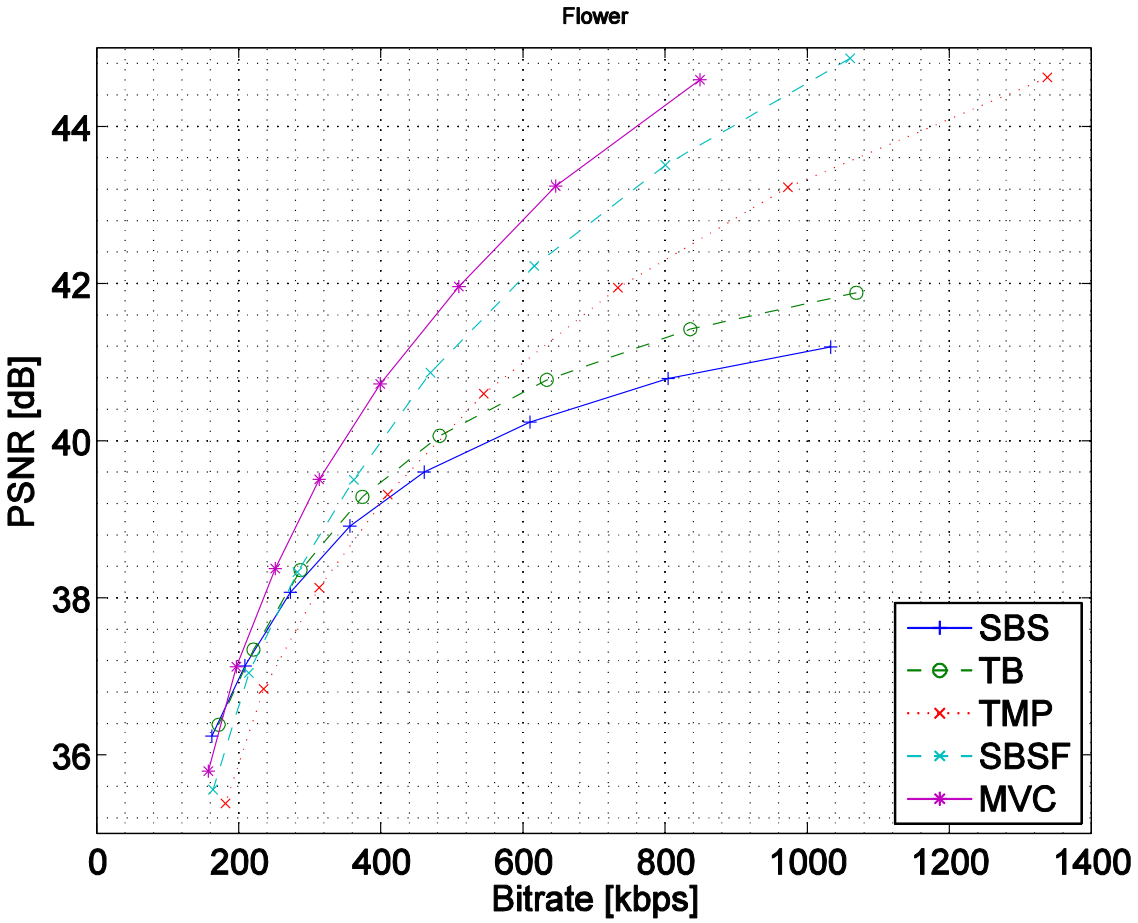

Figure 9: Flower video sequence

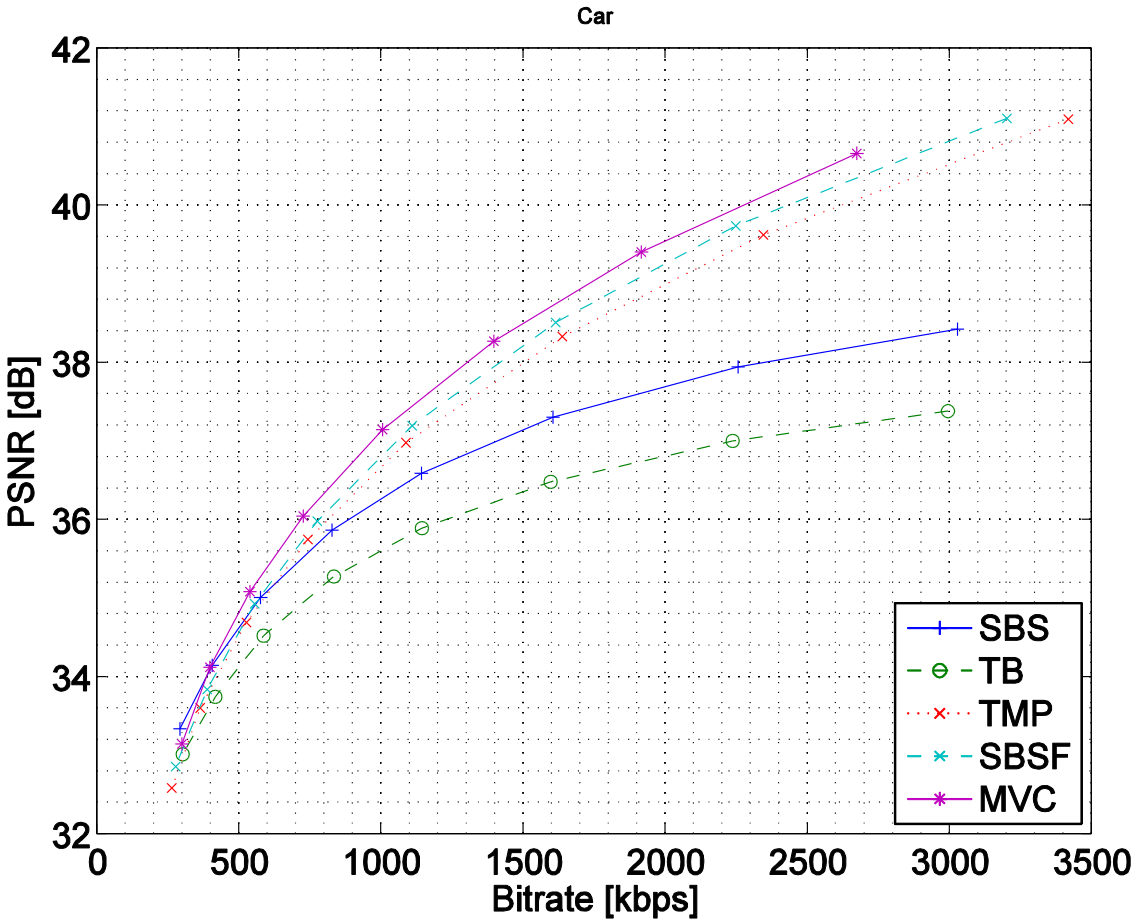

Figure 10: Car video sequence

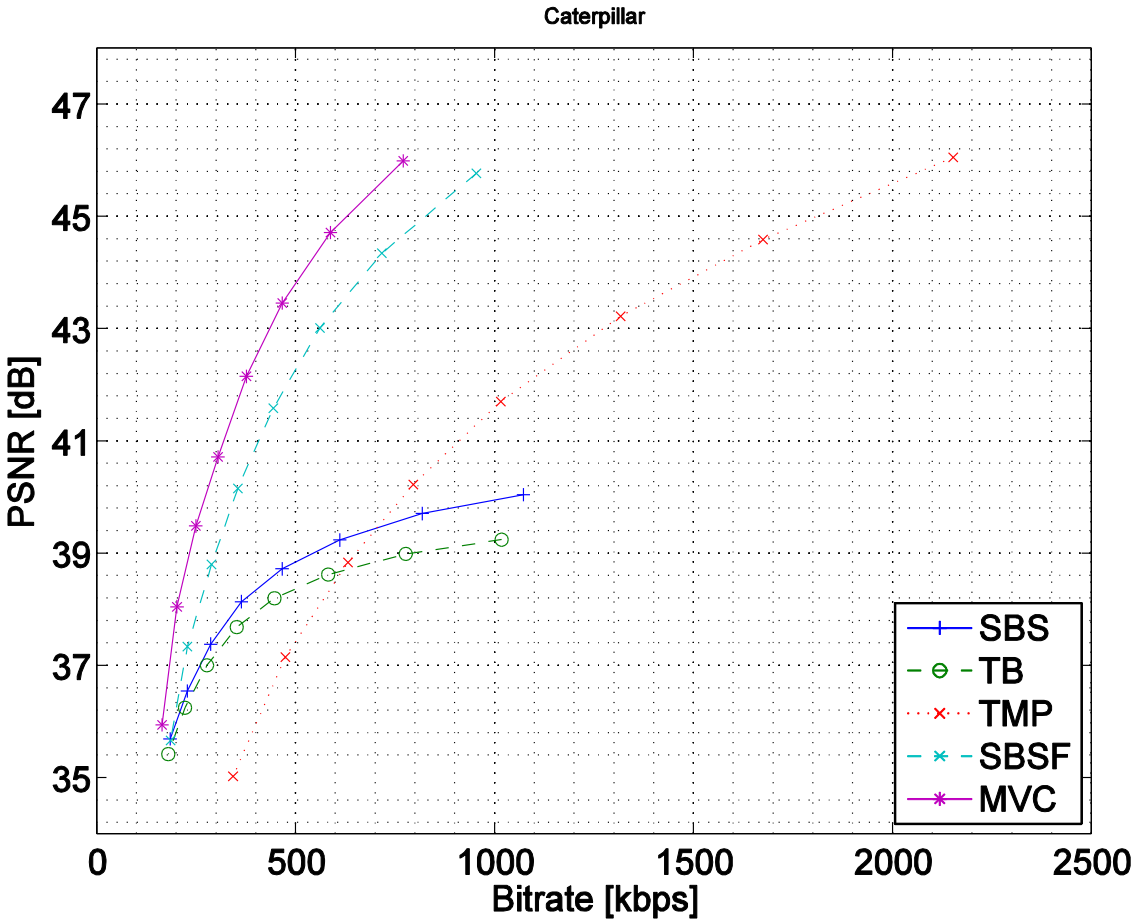

Figure 11: Caterpillar video sequence

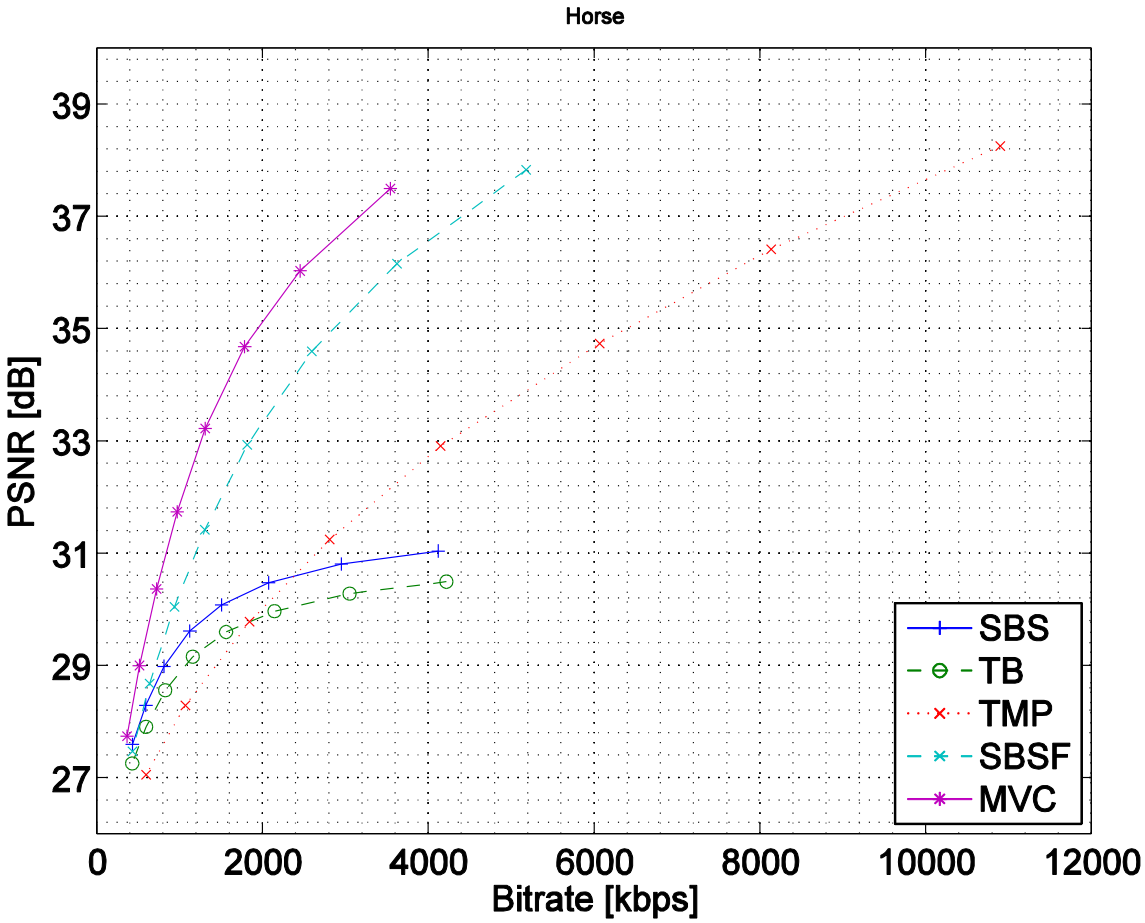

Figure 12: Horse video sequence

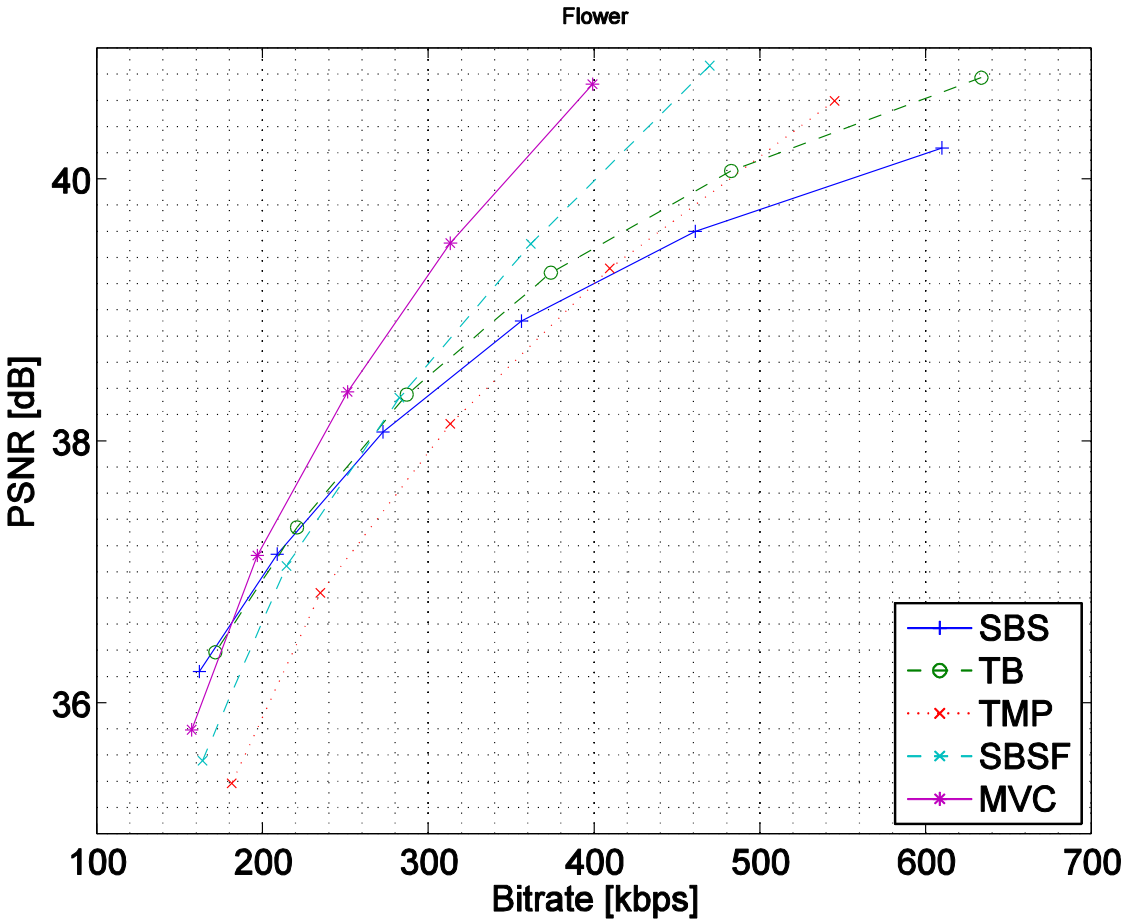

Figure 13: Flower video sequence

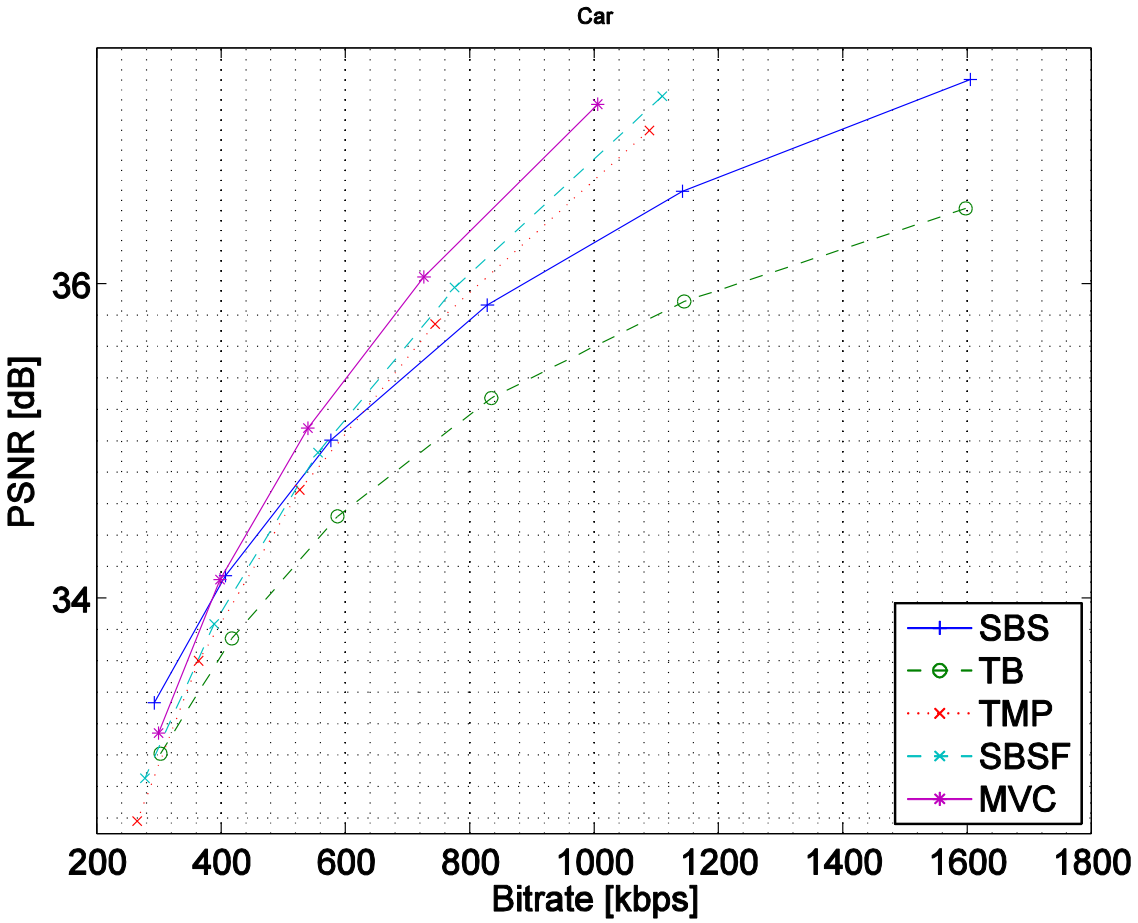

Figure 14: Car video sequence

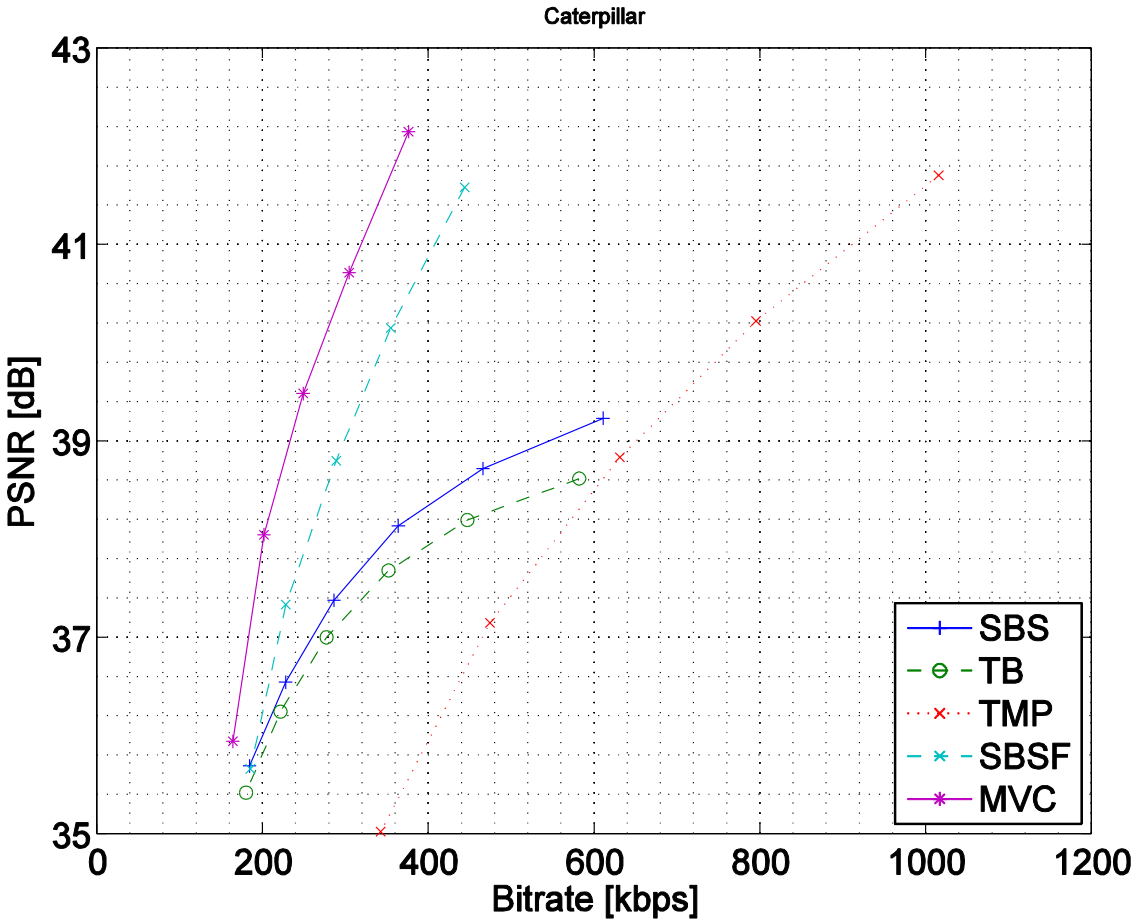

Figure 15: Caterpillar video sequence

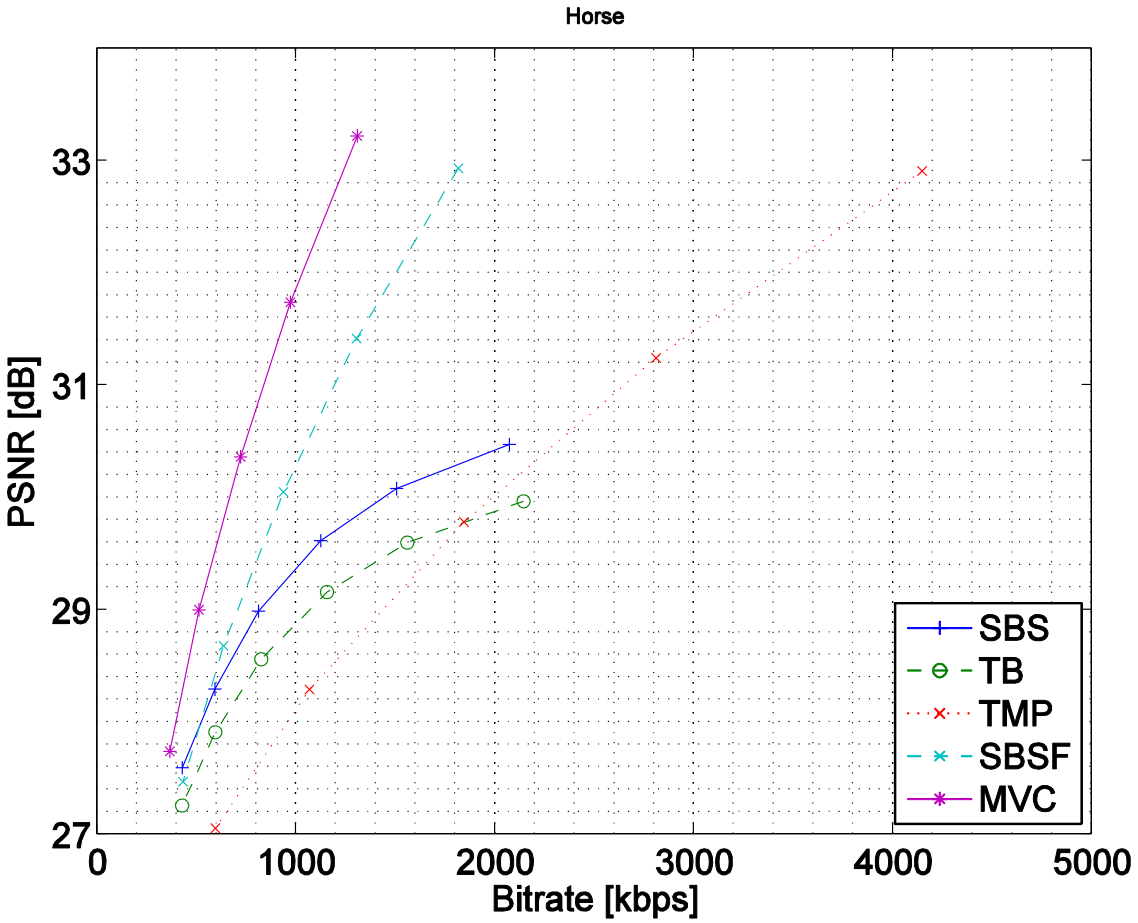

Figure 16: Horse video sequence

5.4 3D signalling

5.4.1 SIP/SDP codec and format signalling

To extend MTSI/IMS for 3D video, during session setup signalling is needed to negotiate whether 3D video is supported, and if so, which codec and which format to use. This signalling should be done during SIP negotiation [10], which is based on the SDP offer/answer model [11].

Currently, the IETF is working on several drafts to extend SDP to support 3D codec negotiation. For example, the document [12] draft aims at delivering MVC using RTP, and extends SDP to signal it. At the time of this contribution, the draft is in an early stage and does not provide signalling for frame-compatible formats that can be used with regular 2D codecs and frame-packing format negotiation. For this purpose, [13] specifies SDP signalling for several frame packing formats, simulcast of left and right view, and 2D + depth. The drafts are still work in development, but potentially provide the needed hooks for SIP in MTSI/IMS.

In [13], the SDP signalling of 3D video is done as follows:

a=3dFormat:<Format Type> <Component Type>

Where <FormatType> can take the values "FP" (frame packed), "SC" (simulcast) and "2DA" (2D + auxiliary). The <Component Type> provides additional information. For example, for frame packing "SbS" indicates Side-by-Side framepacking, whereas "Seq" indicates frame sequential frame packing.

The following SDP example illustrates the usage of the attribute to describe a video stream transmitted using frame packing:

v=0

```
o=Alice 2890844526 2890842807 IN IP4 131.163.72.4
s=The technology of 3D-TV
c=IN IP4 131.164.74.2
t=0 0
m=video 49170 RTP/AVP 99
a=rtpmap:99 H264/90000
a=3dFormat:FP SbS
m=audio 52890 RTP/AVP 10
a=rtpmap:10 L16/16000/2
```

Based on the use cases in this TR additional requirements may be collected and communicated to facilitate the work in IETF.

5.4.2 File format signalling

5.4.2.1 Introduction

3D video coded in any of the coding formats described in clause 5.3 can be encapsulated and delivered in 3GP files [15] (or 3GP segments in the case of DASH). Frame compatible H.264/AVC and temporally interleaved H.264/AVC use the traditional AVC file format [16] where information about the stereo arrangement is carried in an SEI message "frame packing arrangement SEI". Multiview Video Coding MVC on the other hand uses extensions of the AVC file format [16] which specify separate signalling for MVC streams.

NOTE: It is recommended to use the "frame packing arrangement SEI" rather than the "stereo video SEI".

Storing frame compatible or temporally interleaved 3D video in a 3GP file as described above ensures that a UE can decode the bitstreams correctly (if it has the corresponding decoding capability), but it does not ensure that a UE renders the 3D video correctly. For instance, a UE that is not aware of the SEI message indicating that a bitstream represents frame compatible 3D or temporally interleaved 3D will simply render the video frames as consecutive 2D frames. The output will most likely look like garbage or with disturbing artifacts to the viewer.

The above problem can be avoided by enforcing post-decoder requirements with the restricted video mechanism specified in an amendment [17] to the ISO base media file format [18]. The mechanism is similar to the content protection transformation where sample entries are hidden behind generic sample entries, `"encv"`, `"enca"`, etc., indicating encrypted or encapsulated media. The analogous mechanism for restricted video uses a transformation with the generic sample entry `"resv"`. The method should be applied when the content should only be decoded by clients that present it correctly. For the above cases with frame compatible and temporally interleaved 3D video, the scheme type for stereoscopic video `"stvi"` [17] should be used.

In addition, UEs consuming content provided in the 3G file format expect to identify the content based on the MIME Type of the 3GP file in order to accept or reject content. RFC 6381 [25] provides an ability to signal profile and codec parameters and may be considered to be used in this context as well. For more details refer to clause 5.4.2.6.

The following subclauses describe 3D file format signalling in more detail including the case of mixed services.

5.4.2.2 Frame compatible H.264/AVC

Frame compatible H.264/AVC is stored in a 3GP file as defined for H.264/AVC in the AVC file format [16] where the AVC sample entry has been transformed according to the restricted video mechanism using the sample entry `"resv"` and the stereo video scheme type `"stvi"` [17]. The stereo scheme of the stereo video box is 1, e.g., the stereo indication type identifies the frame packing arrangement type by using the values defined by the frame packing arrangement SEI (Table D-8 of ISO/IEC 14496-10).

5.4.2.3 Temporally interleaved H.264/AVC

Temporally interleaved H.264/AVC is a special case of frame compatible H.264/AVC. It is stored in the same way as frame compatible H.264/AVC in the above subclause with stereo indication type value 5 signalling temporally interleaved H.264/AVC.

5.4.2.4 Multiview Video Coding MVC

Multiview Video Coding MVC is stored and signalled in a 3GP file as defined for MVC in the AVC file format [16]. When at least one track uses sample entry type "avc1", compatibility with H.264 (AVC) file readers and decoders is ensured.

5.4.2.5 Mixed 2D/3D video

Decoding and rendering requirements are signalled in the sample entry descriptions as detailed above. In fact, each video sample in a 3GP file is associated with a sample entry description, which in turn specifies if the video data is 2D H,264/AVC or any of the above types of 3D H.264/AVC. If a file contains both 2D and 3D video, two sample entry descriptions are used where 2D parts of the file are associated with the 2D sample entry description and 3D parts of the file with the appropriate 3D sample entry description.

5.4.2.6 MIME type signalling for 3D stereoscopic video files

UEs consuming content provided in the 3G file format expect to identify the content based on the MIME Type of the 3GP file in order to accept or reject content. RFC 6381 [25] provides an ability to signal profile and codec parameters and may be considered to be used in this context as well.

To signal content provided in MVC, the codecs parameter as defined in RFC6381 [25] may be used. The details on how to signal MVC content are provided in RFC 6381 [25], clause 3.3. The section addresses also the use case when MVC content is coded in an H.264/AVC-compatible fashion. In this case it is recommended that the two configuration records both be reported as they may contain different H.264/AVC profile, level, and compatibility indicator values. Thus, the codecs reported would include the sample description code (e.g., 'avc1') twice, with the values from one of the configuration records forming the 'avcoti' information in each.

To signal 3D stereoscopic content provided in a frame packing arrangements as introduced in section 5.4.2.2 and 5.4.2.3 and encoded in H.264/AVC, there exists no MIME type support until to signal frame packing arrangement.

It is desirable to provide such signalling capabilities when introducing 3D stereoscopic content provided in a frame packing arrangements into 3GPP services.

A possible solution to this problem may be the definition of a new compatibility brand that signals the necessity of a specific frame packing arrangement support to process the content. The corresponding brand may then be added to the profiles parameter in the MIME type. Other solutions may be considered if the above use case is considered relevant.

In case of mixed content, all required capabilities may be signalled in the MIME type parameters.

5.4.3 Device capability exchange signalling of supported 3D video codecs and formats

The device capability exchange signalling in 3GPP"s PSS [3] specification enables servers to provide a wide range of devices with content suitable for the particular device in question. In order to optimize delivery of stereoscopic 3D video content to the client terminal, a new set of attributes need to be included in the PSS vocabulary for device capability exchange signalling (PSS vocabulary in [3] is common to other mentioned services). These proposed attributes should describe the 3D video decoding and rendering capabilities of the client terminal, including which 3D video codecs and frame packing formats the client supports. This may for example allow the server and network to provide an optimized RTSP SDP or DASH MPD to the client terminal, as well as to perform the appropriate transcoding and 3D format conversions in order to match the transmitted 3D video content to the capabilities of the client device.

As an example, a solution may be considered by:

- defining a new attribute carrying a list of 3D formats.
- incorporating this attribute into the ThreeGPFileFormat component and/or Streaming component of the PSS vocabulary.

- defining the list of possibly supported 3D video format based on a list of existing code points, e.g. the ones available in '3dFormat" syntax introduced in clause 5.4.1, or the ones defined in the H.264/AVC specification or in another existing repository of code points for 3D video formats and codecs.

5.4.4 Inclusion of 3D video information in the DASH MPD

Signalling of the stereoscopic 3D video content information to the DASH client as part of the MPD is essential in order to provide the client sufficient information about the related 3D video codecs and formats for DASH representations and help the client determine whether/how it can decode and render the content.

MPEG DASH already supports these use cases and signalling:

- by using the Frame-Packing Descriptor as defined in ISO/IEC 23009-1 [26], section 5.8.4.6.
- by using the DASH multiple views scheme as defined in ISO/IEC 23009-1 [26], section 5.8.5.6.
- by using codec information including H.264/AVC and MVC-based encoding, e.g., "avc1", "avc2", "mvc1", "mvc2" in IETF RFC 6381 [25].

This enables a DASH client to accept or reject Representations without accessing the actual media streams.

6 Streaming use cases

6.1 PSS and MBMS-based 3D video services

6.1.1 Use case description

This use case describes the delivery of 3D video content to 3D-enabled UE devices over PSS [3] and MBMS [4] services. Figure 17 illustrates the use case.

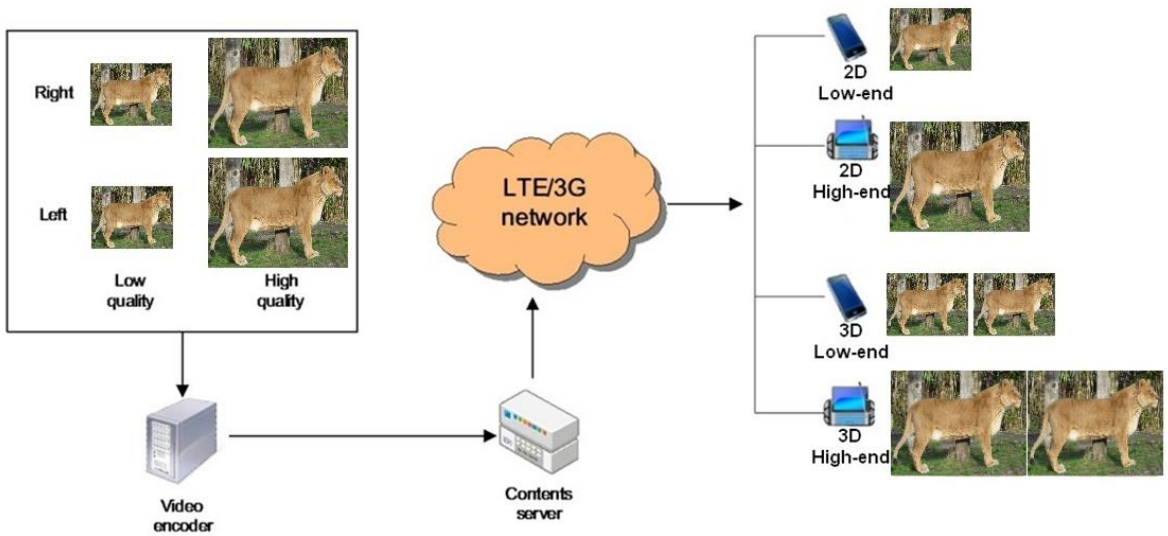

Figure 17: Use case for PSS/MBMS-based 3D video delivery

6.1.2 Working assumptions and operation points

PSS/MBMS-based 3D video services require 3D-enabled UEs to decode and render 3D video at their UE devices. Most of these devices have 3D screens either 3D glasses-free or with glasses. Furthermore, these devices are typically equipped with 3D video hardware and high-performance software drivers for watching high-quality 3D videos and playing 3D video games.

Services such as PSS and MBMS provide the means for distributing the content to 3D capable mobile devices. The specified delivery options include multicast/broadcast, RTP streaming, adaptive HTTP streaming and progressive download [3], [4] and [5].

6.1.3 Technical analysis

While 3D capable devices will enjoy the 3D video, it should also be possible to author so that legacy devices can consume the same content in 2D.

For PSS [3] and IMS-based PSS and MBMS services [24], device capability change signalling of supported 3D video codecs and formats can be handled as described in clause 5.4.3.

6.2 DASH-based streaming of 3D content

6.2.1 Use case description

In a variant of the use case 6.1, the 3D video is made available in multiple bitrates at the server and offered as DASH content to be consumed as streaming services.

6.2.2 Working assumptions and operation points

An appropriate MPD signalling for the 3D content as well as appropriate segment formats are required for the 3D content. The signalling needs to be specified. See clause 5.4.2 for encapsulation of 3D video in 3GP files and clause 5.4.4 for inclusion of 3D video information in the DASH MPD. Additionally, when one representation carries both the base view and the non-base view of an MVC bitstream, SubRepresentation element @level and @dependencyLevel attributes are used for the representation to indicate the base view and the non-base view in separate sub-representations, and the Level Assignment box and a Subsegment Index box per each Segment Index box that indexes only leaf subsegments are used in the segment format to indicate the presence and location of base view and non-base view data within segments. Moreover, when the base view and non-base view of an MVC bitstream are carried in separate representations, the @dependencyId attribute is used to indicate their dependency relationship.

6.2.3 Evaluation of DASH-based streaming with HTTP-caching

6.2.3.1 Introduction

Adaptive streaming using HTTP, typical for a VoD service, takes advantage of widely deployed network caches to relieve video servers from sending the same content to a high number of users in the same core network multiple times. Streaming adaptivity allows for elegant provisioning of multiple representations that match the UEs capabilities and its currently available bandwidth. One possibility to provide stereoscopic video via DASH is to encode each representation with H.264/AVC at the server and offer them side-by-side. Another is to offering several representations embedded in a single file using Multiview Video Coding (MVC) [21]. The presented simulations compare the impact of multiple content representations on the caching efficiency using H.264/AVC or MVC.

In the given use case, high and low quality representation of content are offered. Furthermore, the content is offered as 2D and as stereo representation, thus a total of four representations can be chosen. Figure 18 schematically shows the representation setup of the video library, the network infrastructure and several UEs with varying capabilities.

The operator of the access network (i.e., the cloud in the figure 18), offers connectivity to its customers via access links and connects to the Internet (where the content library is offered on an origin server by a third party) over a 'transit' link. In that way the customers of the access network operator can access video content, in particular the movies on the origin server. The network operator deploys a proxy and a cache in its network to minimize the amount of transmitted data through the 'transit' link relieving the server of having to send an extremely high amount of video data. Since the cache is usually too small to host the complete video library and the content library on the origin video server often changes, the video files that are stored in the cache at every moment need to be carefully selected.

This is accomplished by an appropriate caching algorithm and many different cache replacement algorithms that have been proposed over the last years that optimize the caching performance based on certain criteria. Most algorithms make decisions based either on how recently an object has been requested or on how frequently an object has been requested over a time period or a combination thereof.

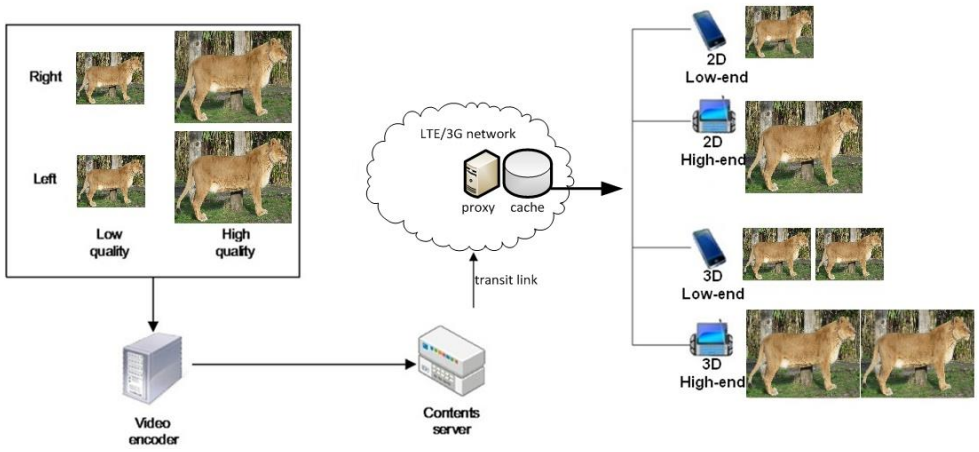

Figure 18: DASH-based streaming with caching infrastructure

6.2.3.2 Coding of VoD content items

Two schemes of video coding for the VoD contents of the given representation setup are compared. For the setup shown in figure 18, the evaluation considers the case where all representations are encoded using the High Profile of H.264/AVC as specified in 3GPP TS 26.237 [24] and the case where 2D and stereo representation at a certain (low or high) quality are encoded as a single MVC bit stream using the Stereo High Profile.

In the H.264/AVC case, the right (LQ_R) and left view (LQ_L) of the low-quality representation as well as the right (HQ_R) and left view (HQ_L) of the high-quality representations are assumed to require the same bitrate for equal visual quality of the views as the content is very similar.

For the MVC-based coding scheme, the non-dependent base (left) view of each representation quality is assumed to require the same bitrate as the respective H.264/AVC coding with the same quality, since both bit streams are H.264/AVC compatible and benefit from the same High Profile coding tools. The average coding gains of the dependent (right) view using the Stereo High Profile of MVC compared to individual encodings of the second view using the High Profile of H.264/AVC is around 30% on average with a maximum of around 40% on 1080p sequences, as shown in [22].

Assuming a resolution of 720p for the low quality representations and 1080p for the high quality representations, it was shown in [23] that the bitrate ratio between the two qualities is roughly 1:2 using H.264/AVC, i.e. the low quality representation requires half of the bitrate of the high-quality representation.

The distribution of the bitrates, concluded from the above findings and normalized to the lowest bitrate, is given in table 1. Since the gain of coding each second dependent view with MVC compared to H.264/AVC dominates the performance in the overall evaluation, MVC gains of 20%, 30% and 40% with respect to an H.264/AVC coded independent second view are simulated, denoted as MVC-20, MVC-30 and MVC-40 in the following.

Table 1: Normalized bitrates of representations for investigated codings.

Representation Coding	Low Quality – Left (LQ_L)	Low Quality – Right (LQ_R)	High Quality – Left (HQ_L)	High Quality – Right (HQ_R)
H.264/AVC	1	1	2	2
MVC_20	1	0.8	2	1.6
MVC_30	1	0.7	2	1.4
MVC_40	1	0.6	2	1.2

The coding schemes as well as their effect on the caching performance are illustrated in figure 19. It can be seen that by using MVC, more content items can be stored using a given cache size compared to H.264/AVC due to the bitrate gains with MVC when encoding the dependent right view at low and high quality.

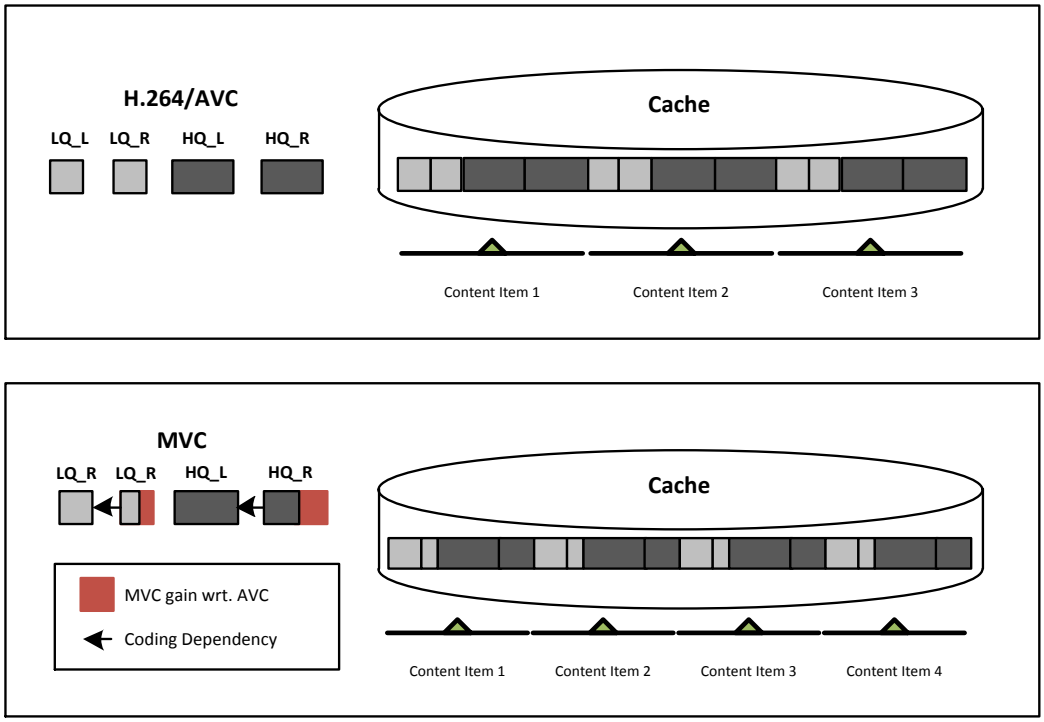

Figure 19: H.264/AVC and MVC coding schemes and caching performance

6.2.3.3 Simulation model

The simulations are based on real data statistics as introduced in Annex C of [9]. The requests have been extracted from the observation of a deployed VoD service. The statistics have been measured within the time period of one month. The provided VoD service offers a wide variety of movies of more than 5000 files for the users to choose from. In these statistics an average of about 3400 requests per day is reported. Since the used VoD service does not offer multi-representation video, it is assumed that half of the user request 3D representations and half of the user can decode high quality representations, which is realised using a uniform distribution of requests for the available representations.

Figure 20 shows the pattern of the requests during 31 days in more detail, specifically the number of request aggregated for each hour of the period of 31 days. The requests extracted from the real data are distributed among the users connected to the service. The performance of the cache is analysed using the Last Recently Used (LRU) caching algorithm that favours recently request content items. The cache capacity is measured in media units, which are equivalent to the file size of a content item of 90 minutes with a single view at the lowest quality. A cache capacity range from 50 up to 9800 media units is simulated in order to cover a wide operational area and reasonable cache performance.

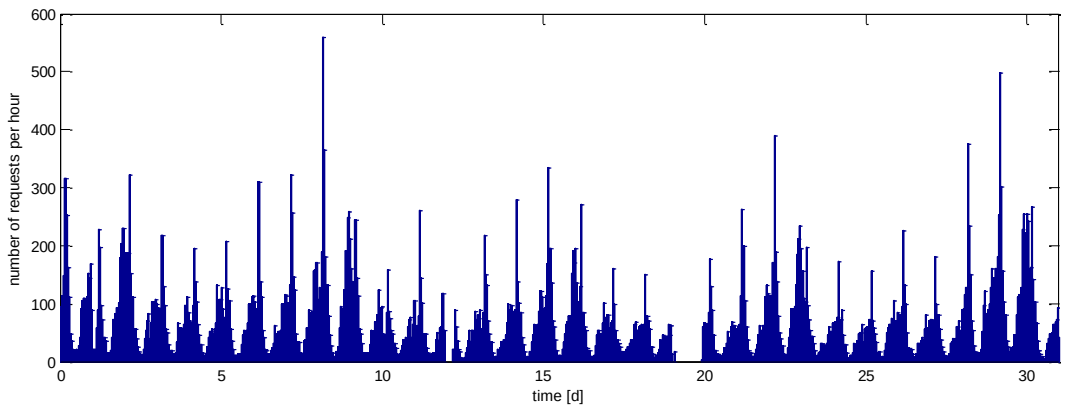

Figure 20: Requests characteristics for the period of 31 days.

6.2.3.4 Simulation results

The results shown in figure 21 depict the difference between the use of H.264/AVC and the three MVC variants (MVC-20, MVC-30 and MVC-40) in terms of caching performance. It can be observed that the use of MVC leads to an increase of cache hit ratio with respect to H.264/AVC, because more content items can be stored in a cache of the same size. This is due to the fact that all second view representations are encoded with significantly smaller bitrates on average when using MVC compared to H.264/AVC. Figure 22 shows the absolute increase in cache hit ratio shown when using MVC with respect to H.264/AVC for the three MVC variants. The increase grows linearly with the MVC gain of encoding the dependent second view.

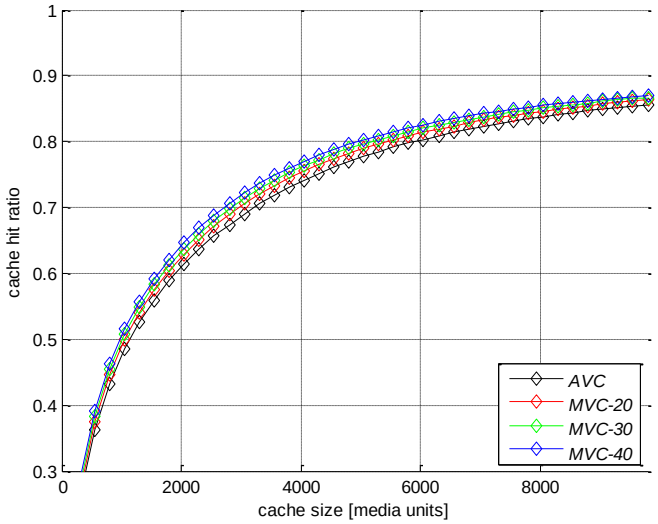

Figure 21: Cache hit ratio of the MVC encodings with respect to H.264/AVC

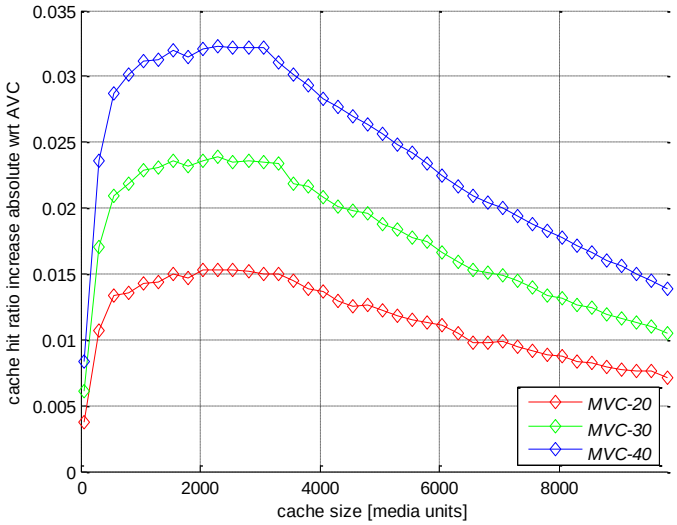

Figure 22: Absolute cache hit ratio increase of MVC encodings with respect to H.264/AVC

The higher cache hit ratio when using MVC makes it more likely for UEs to be served from the cache rather than from the content server. This is reflected in a reduction of traffic on the transit link that connects the content server with the 3GPP network. The average transit link traffic is shown in figure 23. Figure 24 shows the relative reduction of transit link traffic with respect to H.264/AVC for the three MVC variants. As can be seen from the figure, the MVC-30 coding leads to a reduction of transit link traffic of up to 20% and the other variants range from 14% to 27% maximum transit link traffic reduction.

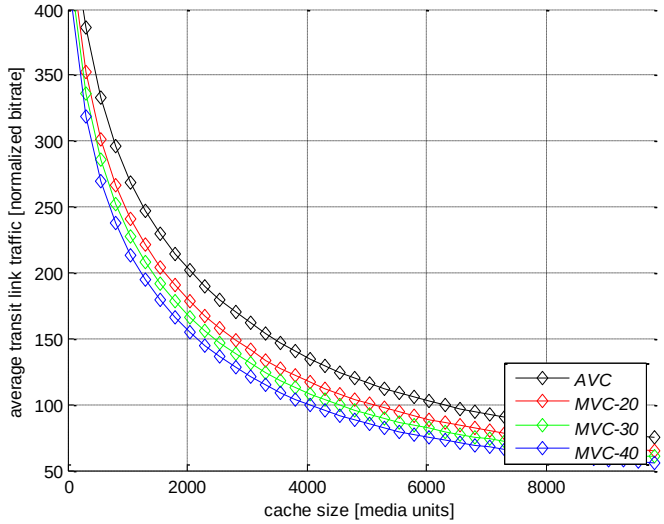

Figure 23: Average traffic through the 'transit' link.

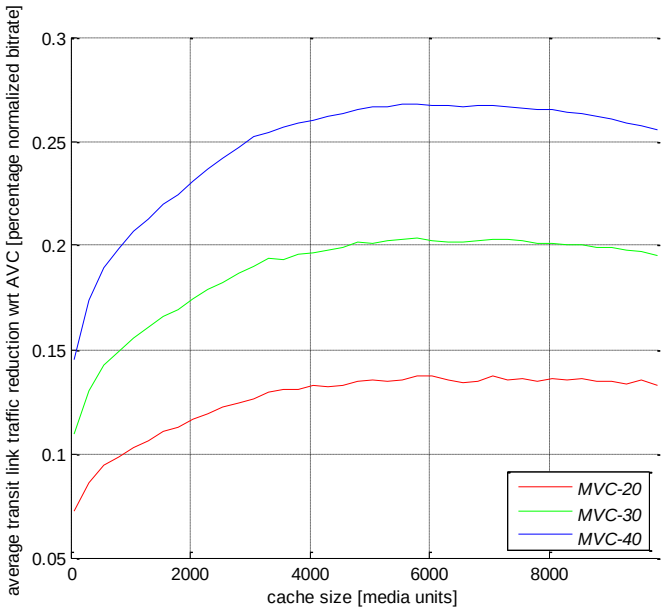

Figure 24: Relative traffic reduction on the transit link of MVC encodings with respect to H.264/AVC

6.3 Common provisioning of 2D and 3D content for download and streaming

6.3.1 Use case description

In the case where there is a coexistence of a variety of device capabilities (e.g. 3D devices and 2D devices) within a 3GPP system. The support for heterogeneous devices is particularly important and service providers generally have two optional approaches to encode and store the contents. Therefore, in an extension to the above use cases, not only the 3D version is available, but also a 2D version of the same content is made available as shown on figure 25.

In one way, the same content of different source files (2D and 3D) are encoded into separated content items and they are made available as separated content. In this case for the delivery of the 3D content is covered by the use cases of sections 6.1 and 6.2 from above.

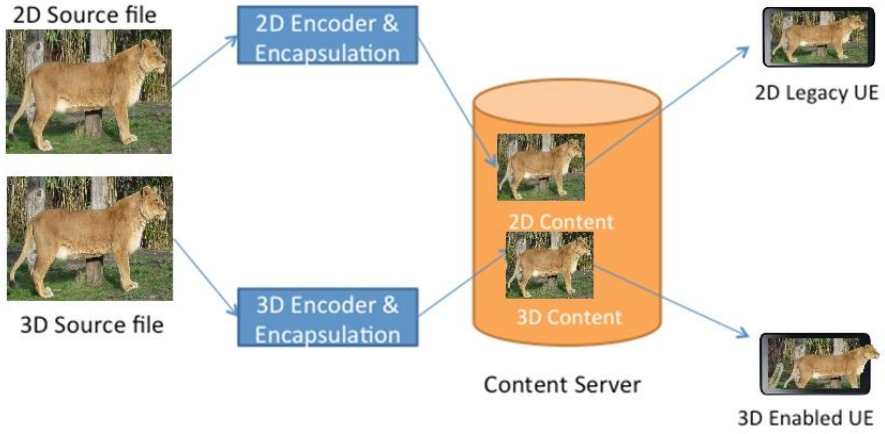

Figure 25: Video encoded into two different files

While encoding different content items into separate videos (2D and 3D) may keep the solution simple, there may exist optimizations in storage, caching and delivery for treating the 2D and the 3D content jointly.

The other approach is to encode the source files into one common provisioning format and offer 2D and 3D content to the UEs in this content as figure 26 depicts:

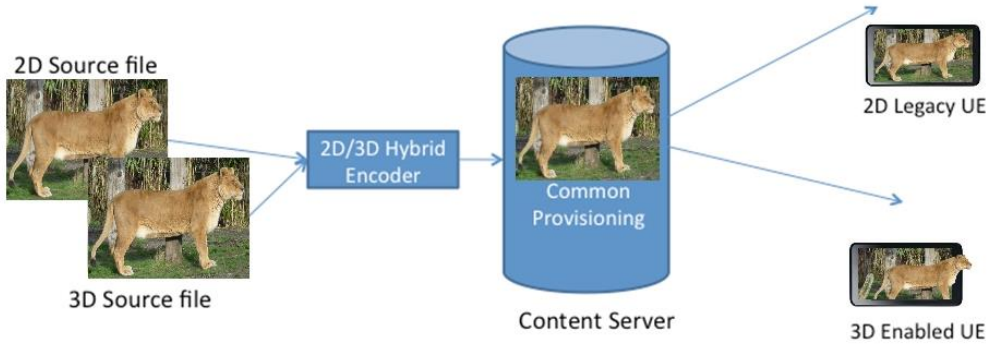

Figure 26: Video encoded into one generic file

The 2D legacy UE then "extracts" the 2D content from the common provisioning format whereas the 3D enabled device extracts the 3D content. The exact processing for the extraction depends on the provisioning format. Extraction may be done by the UE (for example selecting a Representation or doing byte range access) or by the server.

A common provisioning approach may have some advantages that should be explored:

a) With proper codec design, one 2D/3D content item may take up less storage than two separately 2D and 3D content items combined.

b) With proper codec design, one 2D/3D content item may be delivered more efficiently in certain environments such as MBMS and DASH.

c) It may simplify the video management and thus helps to achieve lower operation cost.

However, common provisioning may also result in more complexity on UEs and encoding and the benefits and drawbacks need to be carefully accessed.

6.3.2 Working assumptions and operation points

Common Provisioning may include the following options:

- separate encoding of the two views and common provisioning of the same content;
- joint encoding of the two views and common provisioning of the same content.

In any case, the provisioning formats need to be defined such that 2D content and 3D content can be appropriately rendered. The same formats as considered in use cases of section 6 may be used, but may need adaptation to provide this joint provisioning. The extraction process should be as simple as possible and backward compatibility issues to legacy UEs should be taken into account.

6.3.3 Technical analysis

The exact processing for the extraction depends on the provisioning format.

When left and right views are packed together in frame-compatible format, the 2D extraction process includes decoding of the frame, and rendering (UE extraction) or transmission (server extraction) of only the left view of the pair.

When left and right views are encoded in separate frames, only the master view is decoded, and the secondary view is skipped. The master view is the one specified in provisioning format (e.g. view 0 in H.264/MVC) or the left view by default if unspecified (e.g. time-interleaved H.264/SEI).

6.4 3D Timed Text and Graphics

6.4.1 Use case description

3GPP SA4 has worked on timed text and graphics for 3GPP services which resulted in TS 26.245 [7] for timed text and TS 26.430 [8] for timed graphics. Both formats enable the placement of text and graphics in a multimedia scene relative to a video element. 3GPP Timed Text and Timed Graphics are composited on top of the displayed video and relative to the upper left corner of the video. A track positioning is defined by giving the coordinates of the upper left corner (t_x, t_y) and the box values are defined as the relative values from the top and left positions of the track, as illustrated in figure 27 below.

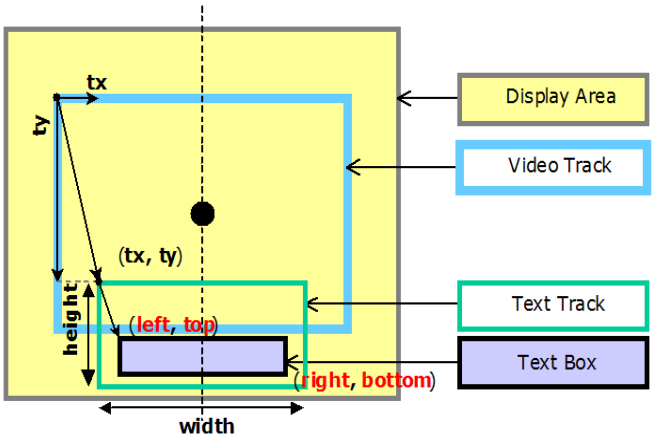

Figure 27: Example of text rendering position and composition defined by 3GPP Timed Text in 2D coordination system

With the introduction of stereoscopic video support, the placement of timed text and graphics will be more challenging. Simply overlaying the text or graphics element on top of the video will not result in satisfactory results, as it may confuse the viewer by communicating contradicting depth clues. As an example, A timed text box which is placed at the image plane with disparity 0, would over-paint objects in the scene with negative disparity.

In addition, text and graphics elements may be placed with an additional degree of freedom in the scene and as such would benefit from higher flexibility in laying out the element in the scene.

6.4.2 Working assumptions and operation points

The timed element should appear correctly on the stereoscopic display, showing correct depth cues. The 3GPP Timed Text and Timed Graphics formats need to be extended appropriately to support correct and flexible layout.

For terminals without 3D support the layout of the text or graphics element should fallback to 2D placement that is backwards compatible with 3GPP Timed Text and Timed Graphics, whenever possible.

6.4.3 Possible solution

The above problem can be resolved by introducing a new dimension when describing the position of the timed text and timed graphic tracks. The new dimension would be perpendicular to the image plane (display surface).Due to the new dimension, text and graphics elements may be placed with an additional degree of freedom in the scene allowing to benefit from higher flexibility in laying out the track in the 3D scene, as presented on figure 28.

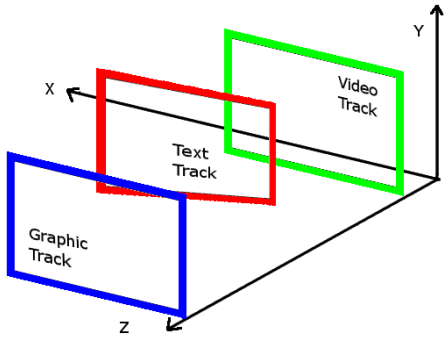

Figure 28: Example of flexible plane tracks overlay

The flexible positioning of the timed text and graphics tracks can be based on signalling the position of one corner of the box (t_x,t_y,t_z) in the 3D space, the width and height of the box (width, height). Furthermore, rotation ($alpha_x$, $alpha_y$, $alpha_z$) and translation ($trans_x$, $trans_y$) values may also be signalled. These information would allow a terminal to calculate the position of all corners of the box in the 3D space. For example, the bottom right corner would be calculated as follows:

$$\begin{bmatrix} tx\_br \\ ty\_br \\ tz\_br \end{bmatrix} = \begin{bmatrix} tx+width \\ ty+height \\ tz \end{bmatrix} * R + T \quad \textbf{(1)}$$

where, the rotation matrix $R=R_x*R_y*R_z$, where

$$R_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha_x) & \sin(\alpha_x) \\ 0 & -\sin(\alpha_x) & \cos(\alpha_x) \end{bmatrix} \quad (2)$$

$$R_y = \begin{bmatrix} \cos(\alpha_y) & 0 & -\sin(\alpha_y) \\ 0 & 1 & 0 \\ \sin(\alpha_y) & 0 & \cos(\alpha_y) \end{bmatrix} \quad (3)$$

$$R_z = \begin{bmatrix} \cos(\alpha_z) & \sin(\alpha_z) & 0 \\ -\sin(\alpha_z) & \cos(\alpha_z) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

, and the translation vector:

$$T = \begin{bmatrix} trans_x \\ trans_y \\ trans_z \end{bmatrix}. \quad (5)$$

3GP file format can be then extended by introducing new boxes that describe the position of the timed text and graphics tracks and boxes in the 3D space. The new boxes if present would override the information of the timed text and graphic tracks positioning in 2D space. Consequently, a natural fallback to 2D placement for the terminals without 3D support would be ensured.

A terminal which supports the new 3D related signalling information would be able to project the box onto the target views (i.e. the left and right view) and create the 3D timed text and graphic reflecting correct depth placement with relation to the objects in the 3D video. This projective transform can be performed on the terminal side for each of the left and right views and based on the following equation (or any of its variants, including coordinate system adjustments):

$$s'(x, y) = s\left(cx + (x - cx)\frac{Vx}{Vx - z}, cy + (y - cy)\frac{Vy}{Vy - z} \right) \quad (6)$$

where Vx and Vy represent the pixel sizes in horizontal and vertical directions multiplied by the viewing distance, cx and cy represent the coordinates of the centre of projection for each of the left and right views, and (x,y,z) are the coordinates of the 3D point (in pixels) to be projected. Figure 29 depicts the perspective projection procedure, which results in the required perceived depth.

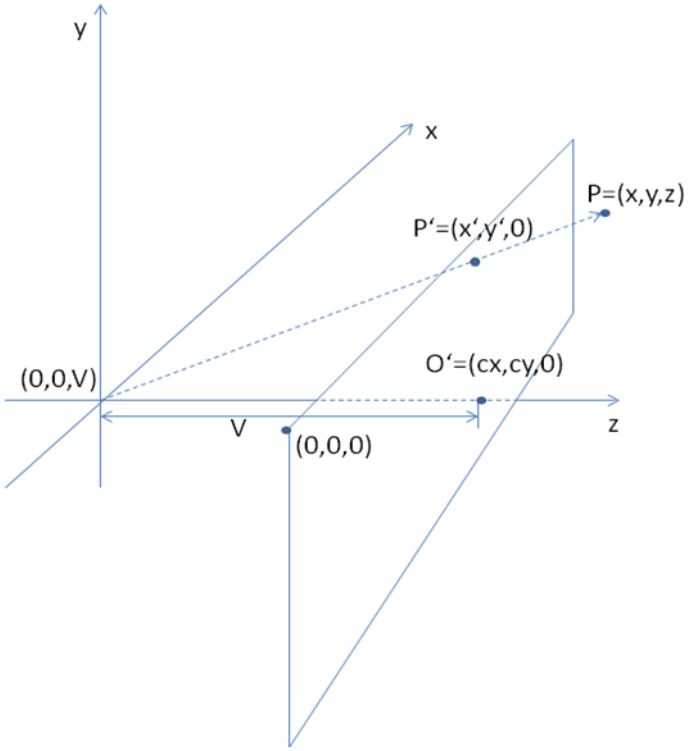

Figure 29: Perpective projection

6.5 2D/3D mixed contents service

6.5.1 Use case description

This use case describes a 2D and 3D mixed contents service. For example, music and drama contents are delivered in 3D and advertisement content is delivered in 2D and those contents might contain 2D-3D switches which might happen at the scene level.

Figure 30 illustrates the use case.

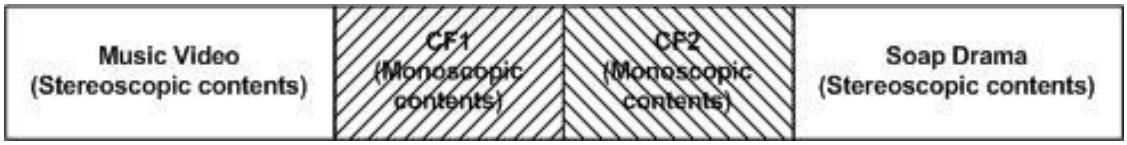

Figure 30: Use case of 2D/3D mixed contents service

6.5.2 Working assumptions and operation points

The mobile device is able to render the 2D/3D combined contents. The identification for indicating pure stereoscopic contents and 2D/3D mixed content is available to the rendering device. Sufficient boundary information for identifying 2D and 3D content segments is signaled to the rendering device.

6.5.3 Technical analysis

See clause 5.4.2 for encapsulation of mixed 2D/3D video in 3GP files. Switches between 2D and 3D modes (there could be several 3D modes depending on 3D formats) could require re-configuration of the codec dataflow as well as potentially the UE display sub-system (memory allocations, parameters, etc.). The optimization of such switches is for further study.

6.6 Service provisioning based on depth range of the 3D content

6.6.1 Use case description

A service provider offers 3D content that is available as a DASH streaming services, MBMS service, or PSS service to a 3GPP 3D service capable UE. The 3D content is available in multiple representations differed in bitrates, resolutions, etc. similarly as is the case with the 2D content. However, the 3D content is also available with different depth ranges. Each depth range is targeted for specific display parameters, viewing distance, and a user"s preferences to ensure high quality 3D experience.

6.6.2 Working assumptions and operation points

Appropriate MBMS, DASH, and PSS signalling indicating the depth range or the target screen size of stereoscopic 3D video should be specified. MBMS, DASH, PSS clients should be able to conclude based on such signalling the applicability/quality of the stereoscopic 3D video based on the prevailing contextual information. MBMS, DASH, PSS client should be able also to conclude if MBMS/DASH service content is suitable for a given hardware, i.e. if the depth range of the content is in the range of the comfort zone for a given hardware.

6.6.3 Possible solution

One possible solution would be to indicate the depth range in form of maximum and minimum disparity values of the stereoscopic 3D video provided over DASH, PSS, or MBMS services. The depth range of the stereoscopic 3D video could be then calculated based on the following equation:

$$D = \frac{V}{\frac{1}{s_D * d} - 1} \quad (7)$$

where D is perceived 3D depth, V is viewing distance, I is inter-pupil distance of the viewer, s_D is display pixel pitch of the screen (in horizontal dimension), and d is disparity.

Another possible solution would be to indicate the screen width (or diagonal) or range of screen widths (or range of diagonals) for which the stereoscopic 3D video, provided over DASH, PSS, or MBMS services, is targeted for.

The calculated depth range based on maximum and minimum disparity or provided screen width (or diagonal) or range of screen widths (or range of diagonals) would allow DASH, PSS, or MBMS client to choose version of stereoscopic 3D video which would ensure the highest quality of the 3D experience on a given hardware.

In case of DASH service the maximum and minimum disparity, screen width (or diagonal) or range of screen widths (or range of diagonals) could be signalled as part of MPD describing DASH service. In case of PSS and MBMS services the required signalling information could be as part of SDP describing MBMS or PSS service. In case of PSS service yet another possible solution, where the stereoscopic 3D video for a given PSS client is chosen by a PSS server based on information in UAProf [27] and the target depth range, screen size, or range of screen sizes of the stereoscopic 3D video.

7 Download use cases

7.1 Download of 3D video

7.1.1 Use case description

A service provider offers 3D content hosted at an HTTP-Server that is available for HTTP-based download to a 3GPP-service capable UE. The same content may also be distributed over eMBMS using MBMS download delivery method as illustrated in the figure 31 below.

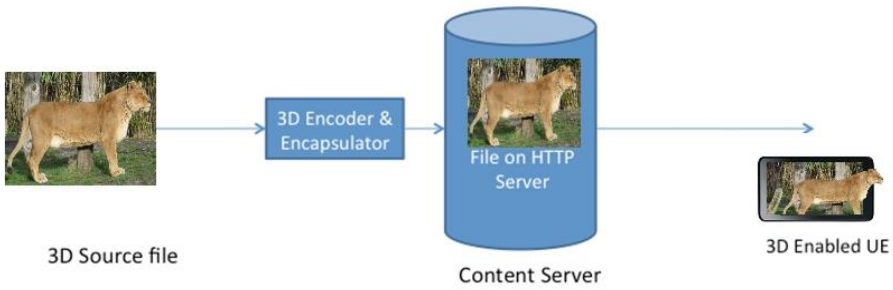

Figure 31: Download of 3D video

7.1.2 Working assumptions and operation points

Download of 3D video requires the encapsulation of the video in a file format that is supported by the 3GPP-service capable UE. It also requires that the file is advertised appropriately in the Content-Type to signal that it is a 3D-content and what are the required codecs and rendering capabilities to support this. It is expected that a 3GP file format based encapsulation format is used.

7.1.3 Technical analysis

The details of the file format as well as the appropriate signalling need to be defined to support this use case. See clause 5.4.2 for encapsulation and MIME type signalling of 3D video in 3GP files.

7.2 Progressive download of 3D video

7.2.1 Use case description

In a variant of the use case 7.1, the 3D video is progressively downloaded to start decoding and rendering of the 3D video before the download is completed.

7.2.2 Working assumptions and operation points

In addition to the requirements for the use case 7.1, the file format is expected to support an arrangement of the video content such that playout before completely downloading the file is supported.

7.2.3 Technical analysis

The details of the file format as well as the appropriate signalling need to be defined to support this use case. See clause 5.4.2 for encapsulation and MIME type signalling of 3D video in 3GP files.

7.3 Correct rendering of downloaded 3D video

7.3.1 Use case description

In variants of the use cases 7.1 and 7.2 (and also 6.2), the 3D video is provided with rendering requirements in the 3GP files carrying the 3D bitstreams. The client downloading the 3D video is either a legacy client or a Release 11 client.

Although a legacy client may be capable of decoding the bitstream, it is not supposed to decode and render the video unless it can detect that the video is in 3D and identify that it is capable of performing required post-decoder transformations. The same procedure applies to a conforming Release 11 client.

7.3.2 Working assumptions and operation points

The file format needs to support signalling of post-decoder requirements (for frame compatible and/or temporally interleaved 3D video) in order to:

stop legacy clients from decoding and improperly displaying 3D video contents.

allow Release11 conforming clients to inspect files for post-decoder requirements before rendering.

7.3.3 Technical analysis

See clause 5.4.2 for encapsulation of 3D video in 3GP files and using post-decoder requirements.

8 Use cases for further study

8.1 Introduction

The following use cases are introduced as extension of the already existing 3GPP services to support 3D. The identifications of the deltas with the current 3GPP specifications and the technical analysis on the codecs support and signalling need requires more investigation which could be part of a future study. These use cases address the 2D to 3D conversion, conversational services and specific mobile 3D video adaptations to the bitrate variations.

8.2 3D video delivering based on 2D video warehouse

8.2.1 Use case description

Apart from creating new 3D video content, existing 2D video resources may also be delivered and displayed as 3D content. This is especially important to cover for the still low amount of 3D content. Users may decide to watch an originally 2D content in 3D. Figure 32 illustrates this scenario.

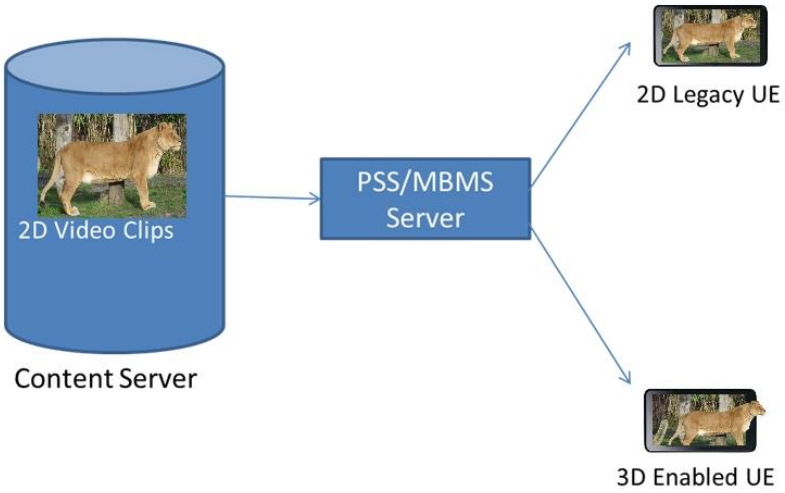

Figure 32: 3D video delivery from a 2D video database

8.2.2 Working assumptions and operation points

Firstly the 2D to 3D conversion is required at certain point of the system. It could be either done by an offline conversion operation (which is not covered by the present document), or by converting in real time.

In one way, the PSS/MBMS servers decode the 2D video, and then re-encode it into 3D video, and finally distribute it to the UEs. The 3D-enabled UEs are expected to be able to decode and render 3D video at their UE devices.

Another way is to let the PSS/MBMS servers simply deliver the 2D video to the UEs, and the UEs are expected to conduct 2D to 3D conversion and 3D rendering.

In the first approach, the conversion procedure is done completely at the server and so the complexity at the terminal is reduced. In the latter approach, the terminal needs to provide sufficient processing power to implement the 2D to 3D conversion in real-time. Hybrid approaches, where the depth map extraction is performed at the server and the 3D rendering is done based on a 2D view plus the depth map at the client may be possible.

The trade-off between complexity and bandwidth usage is to be studied in these different approaches. The feasibility of real-time 2D to 3D conversion either fully or partly at the UE is to be studied as well.

8.3 3D video conversational services

8.3.1 Use case description

Apart from streaming of 3D video content to users from network servers, 3D conversational services have also been supported by the newest 3D-enabled products and these applications have been gaining significant popularity among the consumers recently. For instance, the latest 3D-enabled laptops have 3D web cameras that capture and record 3D video, and transfer it to others over the internet, enabling services such as 3D video chat, 3D video conferencing, 3D gaming and 3D video phone using web applications.

Figure 33 illustrates the use case for 3D conversational video based on MTSI [2] over a video conferencing application.

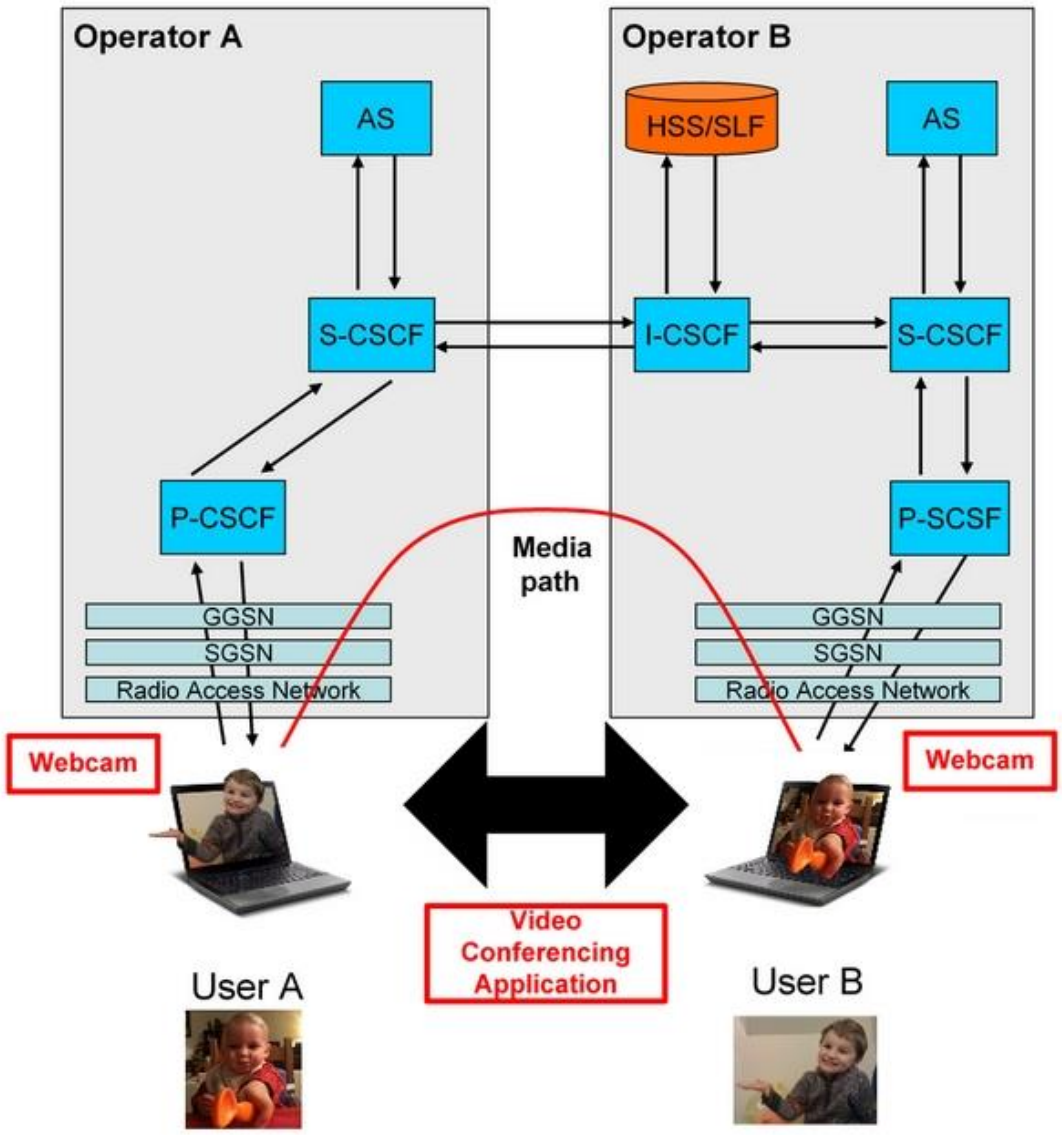

Figure 33: Use case for 3D conversational service over a video conferencing application

8.3.2 Working assumptions and operation points

3D conversational video services require 3D-enabled UE devices that capture and record 3D video via the availability of 3D web cameras, and transfer it to other devices over the 3GPP network, e.g., through the use of MTSI services. In addition, 3D-enabled UEs at both ends of the link are required to decode and render 3D video at their UE devices. Most of these devices have 3D screens either 3D glasses-free or with glasses. Furthermore, these devices are typically equipped with 3D video hardware and high-performance software drivers for watching high-quality 3D videos and playing 3D video games.

8.4 Multiple-party 3D video conference

8.4.1 Use case description

This use case describes a multiple-party 3D video conference service, a meeting host can setup the 3D video conference call, and other 3D phone users can join the call. The 3D video information is exchanged among the multiple 3D-enabled phones users over MTSI system, the users can join or drop off the 3D video group call at different locations. The figure 34 illustrates the use case.

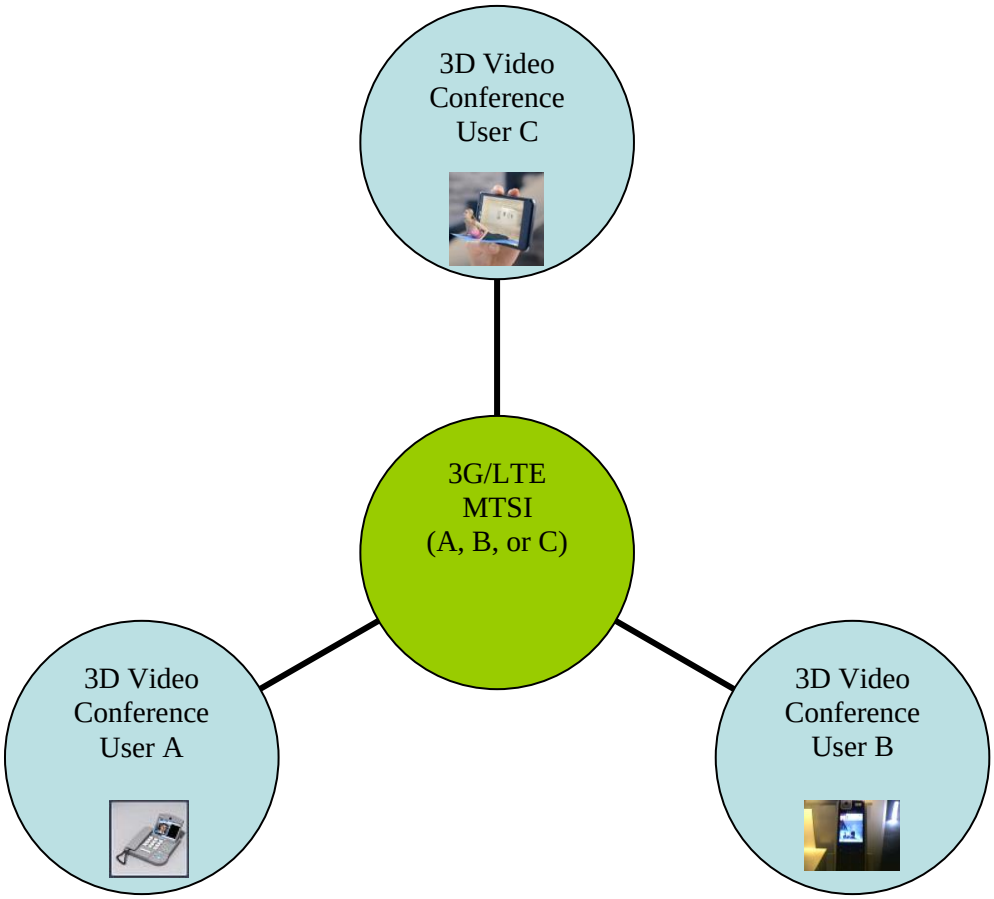

Figure 34: Multiple-party 3D video conference

8.4.2 Working assumptions and operation points

The multiple-party 3D video conference requires 3D-enable mobile terminals to capture, record, store and transmit the 3D video sequence and audio via the MTSI system over the 3GPP operators" mobile network, which also require 3D-enabled UEs to decode and render the 3D video display to the conference users. The device should have 3D screens either with 3D glasses or glass-free display.

8.5 3D video call fall back to legacy phone

8.5.1 Use case description

A 2D video mobile or a legacy voice only phone user wants to join a 3D video conference group; however, the 3D video call or multiple-party conferencing scenario requires that the other ends of the participating parties be 3D video capable devices. To guarantee backward compatibility, a legacy phone needs to be able to join the conference as a 2D video call or voice only service, while keeping at the same time the other 3D video phone users in active 3D conferencing service. The figure 35 depicts the scenario for 3D video call fallback. A related use case is a regular call between a legacy phone and a 3D video phone with 3D video call capability, for which the negotiated call will end up as a 2D video or a voice call.

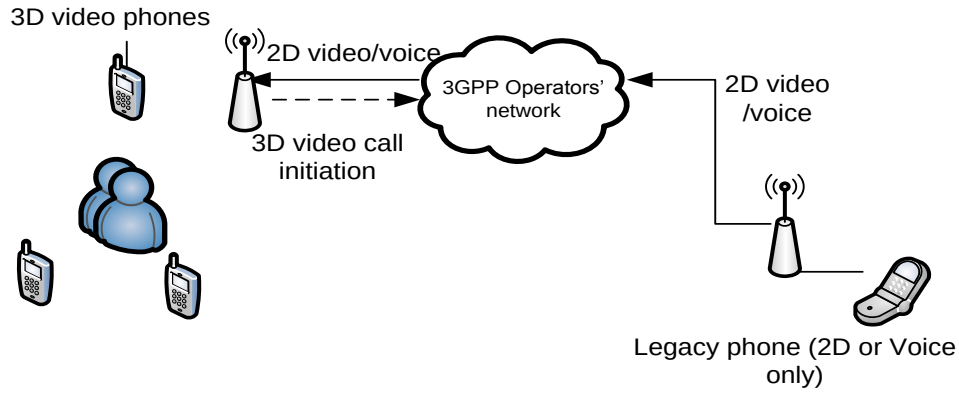

Figure 35: The 3D video call fall back to a legacy phone

8.5.2 Working assumptions and operation points

The 3D video phone has the functions to capture 3D video sequences as well as having 3D display screen, in which the powerful software and hardware process power is equipped. The 3D video phones have the capability to transfer the 3D video to other 3D devices over the 3G/LTE via MTSI system. 3D UEs are expected to perform session negotiation in such a way that during the call setup, a proposed 3D session can be downgraded to 2D or even voice in a manner compatible with existing IMS procedures.

8.5.3 Gap analysis on supporting 3D video call fallback between 3D video phones

The fallback procedures are already handled by MTSI specifications. The only gap identified for MTSI with 3D video support is SDP signalling of the 3D video media as described in section 5.4.1.

8.6 3D video call fall back between 3D capable phones

8.6.1 Use case description

For the wireless environment, the wireless fading, the cell-edge performance and high latency may prevent the network from providing the full bandwidth, capacity and QoS to enable a 3D video call service for users. For such scenario, the 3D video call system needs to fall back to 2D video call, for the worst case, only the voice call can be conducted, the system may prevent the frequent back and forth transition for better user experience.

8.6.2 Working assumptions and operation points

3D video phones are normally also capable of 2D video calls. During an ongoing 3D call the phones have a mechanism to fall back from 3D video to 2D video or even to voice only.

8.6.3 Gap analysis on supporting 3D video call fallback between 3D video phones

The 3D video call fall back to the 2D video call can be determined based on the SDP-based capability negotiation procedures described in clause 5.4.1.

8.7 3D content in messaging

8.7.1 Use case description

The Multimedia Messaging Service (MMS) is defined by 3GPP in TS 26.140 [6]. MMS allows users to embed multimedia content into messages and send them to other users. With the advance of 3D technology and the steadily growing adoption by manufacturers, users with devices equipped with a stereo camera are able to shoot video and pictures in 3D.

Figure 36 illustrates the use case, showing a user sharing a captured picture over MMS with another user.

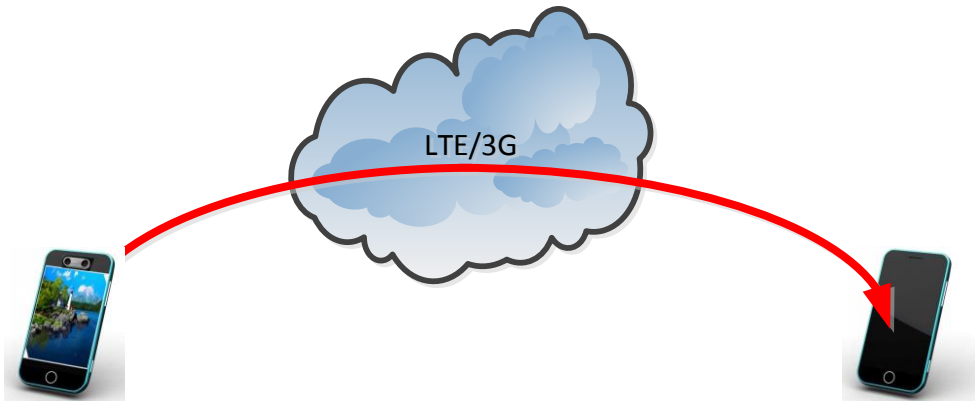

Figure 36: 3D pictures and videos over MMS

8.7.2 Working assumptions and operation points

In order to enable this use case, the UE of the first user is required to be equipped with a stereoscopic camera and the necessary processing power to capture stereoscopic images and video. The MMS service would need to be extended to support the 3D content formats and appropriate signalling.

The used formats are required to be backwards compatible, to enable users with non-3D capable UEs to still view the content, albeit in single view.

8.8 3D service in the converged environment

8.8.1 Use case description

In this scenario the user can use mobile devices to capture 3D video contents which can be shared with other devices, e.g. 3D TVs. The 3D contents can be delivered directly to the other devices.

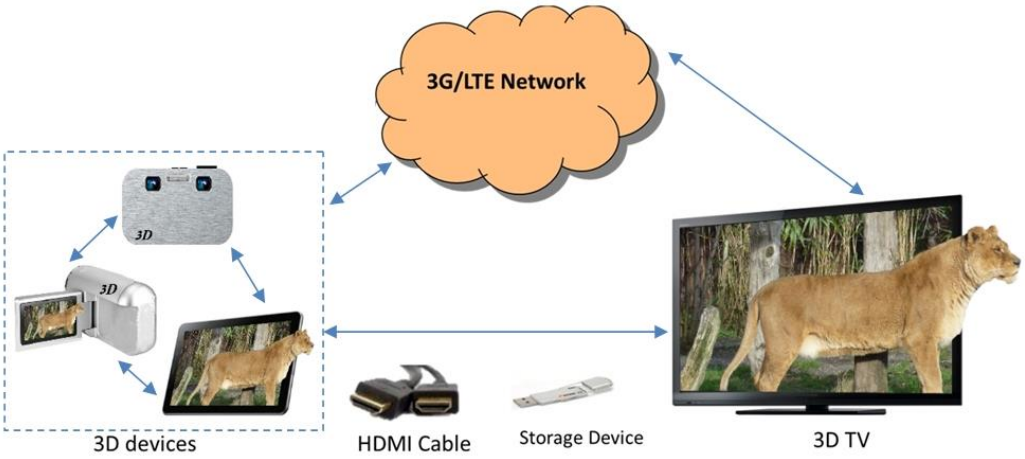

Figure 37: Use case of 3D service in the converged environment

As Figure 37 illustrates, 3D devices are able to share 3D videos among one another via cable (e.g. via HDMI). In addition, they can share videos via 3GPP networks, where a video server could be required for NAT traversal and video storing/distributing/converting.

8.8.2 Working assumptions and operation points

The mobile device such as a smart phone supports the function to capture 3D contents and send to other devices such as TV, for rendering or storage. The captured contents may be sent using either an uncompressed form (e.g. via HDMI) or a compressed form (using a wireless interface, e.g. 3G/LTE). The latter assumption implies that a mobile device can encode 3D contents. For storage of 3D contents, it is assumed that standard container formats (e.g. those derived from the ISO base file format) are used.

8.9 Bitrate adaptation

8.9.1 Introduction

Mobile networks offer different types of radio access networks with different access bitrates. In addition, depending on reception and load conditions, available bitrates in mobile networks may be more or less restricted.

Furthermore, mobile network conditions may change over time which may require that the application adapts to the changing bitrates by changing the quality of the delivered media.

Some use cases explaining such circumstances are provided below.

8.9.2 Restricted access bandwidth

A service provider may want to offer a 3D streaming service to 3D capable phones. To ensure a sufficient quality for the service, the service operator requires choosing a minimum quality for which the streaming service can be offered also for 3D capable phones that have restricted access bandwidth. However, it is expected that there is a certain lower boundary where a 3D service can still be considered as being sufficient quality and below this quality the service may be unwatchable due to artefacts that are 3D specific.

8.9.3 Rate adaptation in PSS and DASH

A service provider may want to offer a 3D streaming service to 3D capable phones over PSS and/or DASH. To ensure service continuity also in temporarily low bitrates, streams (in RTP-based PSS) or Representations (in DASH) need to be provided to ensure this service continuity, but still to have acceptable quality. In 2D-video, low quality video for service continuity may be achieved by reducing the spatial and temporal resolution as well as to increase the quantization noise. As a last resort, the omission of video transmission entirely and relying on audio only may be applied. In 3D-video, in principle the same mechanisms are available, but the effects of applying these schemes in terms

of quality may be more severe and there may be preferences in either dispending earlier or operate along other axis to avoid annoying perceptual artifacts. Another option is to apply a 2D-fallback.

8.9.4 Rate adaptation in MTSI

In MTSI, the feedback from the far-end decoder could be used to facilitate real-time 3D video encoding according the prevailing network conditions. In the event of severely reduced throughput, the encoder has several options, such as coarser quantization of transform coefficients, reduced spatial resolution, or downgrading from stereoscopic 3D video to single-view video.

8.9.5 Rate adaptation due to shared radio resources

Another situation where rate adaptation may be needed is where multiple service users converge in a cell and available bandwidth capacity per user therefore depletes quickly. In such case, service to lately incoming UEs may be refused, or all UEs in the cell may suffer severe 3D quality degradation. The situation can be improved when bandwidth of the streams can be reduced adaptively. The service quality is recovered as congestion state of the cell is relieved.

8.10 View scalability for graceful degradation

8.10.1 Introduction

Mobile network conditions may change as time and location changes. As experienced by the application in the UE, the changing network conditions may result into varying residual error rate, varying throughput, and/or varying throughput delay that may distort media data delivery. As 3D media data can be sensitive to delivery disturbances such as burst errors in wireless channels, 3D user experience might be severely degraded. View scalability may offer possibilities for graceful degradation when 3D data delivery is affected by network conditions.

In the following sub-clauses specific use cases where graceful degradation may help are introduced.

8.10.2 Graceful degradation in MBMS when entering bad reception conditions

Unlike a PSS service, an MBMS service cannot adapt to individual receivers needs. That is, users entering difficult reception conditions may experience sudden service interruption instead of soft degradation of e.g. 3D video quality. To keep users satisfied with the mobile 3D experience, graceful degradation of the broadcast service is a desired feature. Such a feature can be applied to a broadcast service by allowing differentiation transmission robustness for different parts of the video stream. The service should allow minimum acceptable quality to the user perception at the service coverage configured by the operator.

8.10.3 Graceful degradation in MTSI

In MTSI, the feedback from the far-end decoder could be used to facilitate real-time 3D video encoding according the prevailing network conditions. It may be possible to use different robustness for different parts of the video stream to improve the subjective quality of the reconstructed video under severe network conditions.

8.10.4 Combined support of heterogeneous devices and graceful degradation

It is expected, that there will be a coexistence of a variety of device capabilities (e.g. 3D devices and 2D devices) within a 3GPP system and each of these devices may be in different reception conditions. Therefore to cope with both of these challenges in an efficient way, a service should be able to support the heterogeneous devices and to provide Graceful Degradation behaviour at the same time. The support for heterogeneous devices is particularly important to be coped in MBMS where there are limited possibilities to adapt the transmitted content to the present receivers and the bitrate resources in the radio link may be scarce.

9 Mobile 3D subjective tests

9.1 Introduction

This section presents the results of subjective tests conducted on a 3D capable mobile terminal. The objective is to evaluate the impact of the frame rate and the resolution of video on the perceived quality of experience of the user.

9.2 Test description

9.2.1 Video sources

The subjective tests contained 3 different sequences with different characteristics in terms of movement and textures.

Football sequence: with significant movements.

Interview sequence: with almost static content.

Cartoon sequence: an animated movie with the main characteristic of containing simplified textures.

9.2.2 Content preparation

9.2.2.1 Frame rate evaluation

In this test the resolution of the sequences was fixed to the one of the device, i.e. qHD (960x540).

The source contents in 1080i25 were converted to 1080p50 using a deinterlace filter.

Then different frame-rates were produced by dividing the initial frame rate (50 fps) by 2, 3, 4, 5 and 6 in order to get respectively 25, 16,6, 12,5, 10 and 8,33 fps.

Frame rates which are not an integer divider of 50 fps were not produced in order to avoid the frame interpolations which might introduce additional visual artefacts.

9.2.2.2 Resolution evaluation

In this test the frame rate of the sequences was fixed to 25 fps.

Initial fullHD per view contents were then converted to different resolutions in side-by-side 3D format.

The produced resolutions were: WQVGA (320x176), WHVGA (480x272), WVGA (848x480), qHD (960x540) and 720p (1280x720).

NOTE: If different interpretations are made between the name of the resolution and the number of pixel per rows and per columns, the last mentioned should be taken into account.

9.2.3 Encoding profiles

Since the hardware of the 3D mobile terminal is not able to decode uncompressed video sequences at these formats, encoding step was needed.

Video sequences were compressed using a high enough bitrate to avoid any encoding visual artefact.

The encoding profile was the same for all the sequences, as follow:

Encoder common features:

Constant Bit Rate (CBR)

MP4 File format

Source encoding profile:

MPEG4 AVC/H.264 at 5Mb/s CBR

Depending on the resolution of the content, the Profile/Level change to Baseline@L3.0 to High@L4.0

9.2.4 Subjective test conditions

9.2.4.1 Methodology

The SAMVIQ (ITU-R BT 1788) [14] method was used. It is based on a continuous scale (0-100) graded with following five quality items: Bad, Poor, Fair, Good and Excellent.

It is an adapted method discriminating quality levels and providing accurate quality scores.

9.2.4.2 Implementation

A mobile application was created to allow the playback of the different sequences and also to rate their perceived quality directly on the mobile terminal.

The 3D mobile terminal has the following technical features:

qHD (960x540) screen resolution

4.3" screen size

Cell matrix parallax barrier

9.2.4.3 Observers

28 participants took pat of the subjective tests. In the worst case 9 observers were rejected, 3 observers in the best case.

This means that the participants whose results were too much different from the average results were rejected. In order to define the number of participants to reject a correlation factor is computed and a threshold is set. In our case the correlation factor was 82%.

9.3 Test results

9.3.1 Frame rate evaluation

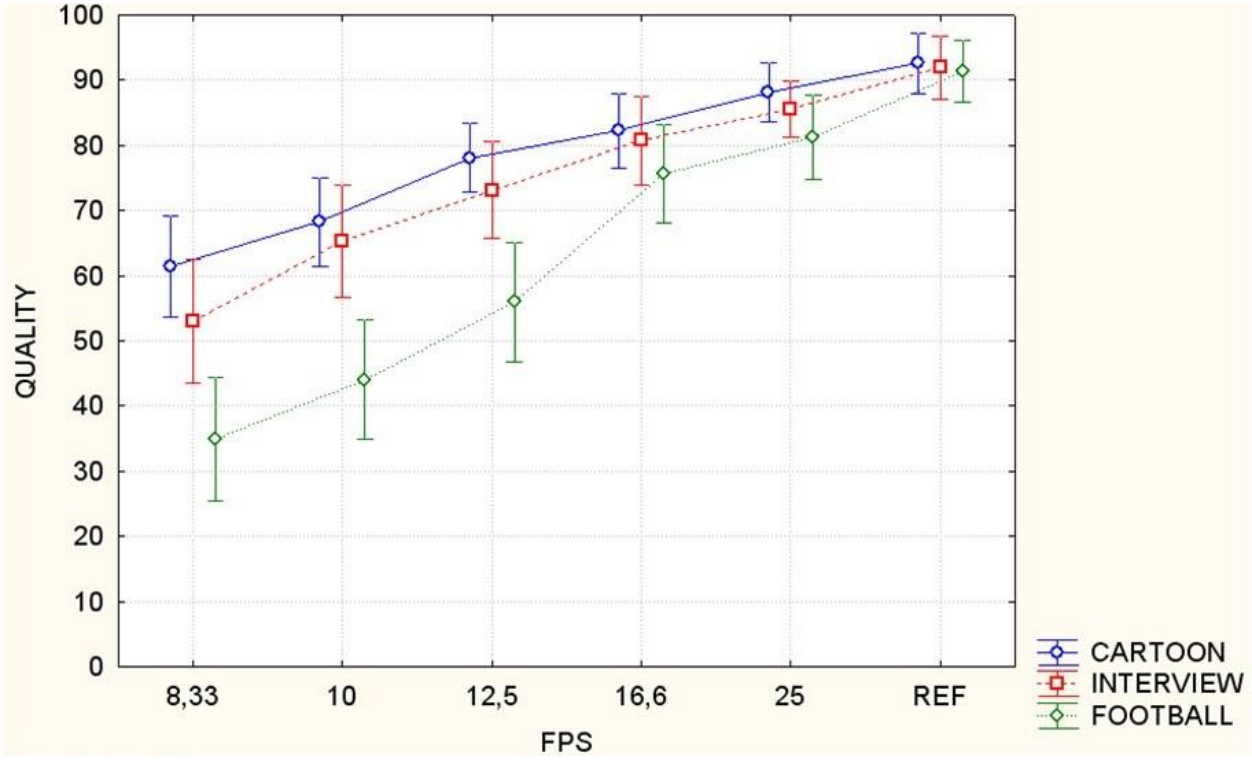

Figure 38: Subjective tests results on the impact of frame rate

On the figure 38 it can be noted that a good quality is achieved for Cartoon and Interview from 10 fps. The same level of quality needs at least 16,6 fps for Football.

Only 19/28 user"s ratings were taken into account. This high number of rejected testers is due to the fact that the difference between the low frame rates is not significant enough.

9.3.2 Resolution evaluation

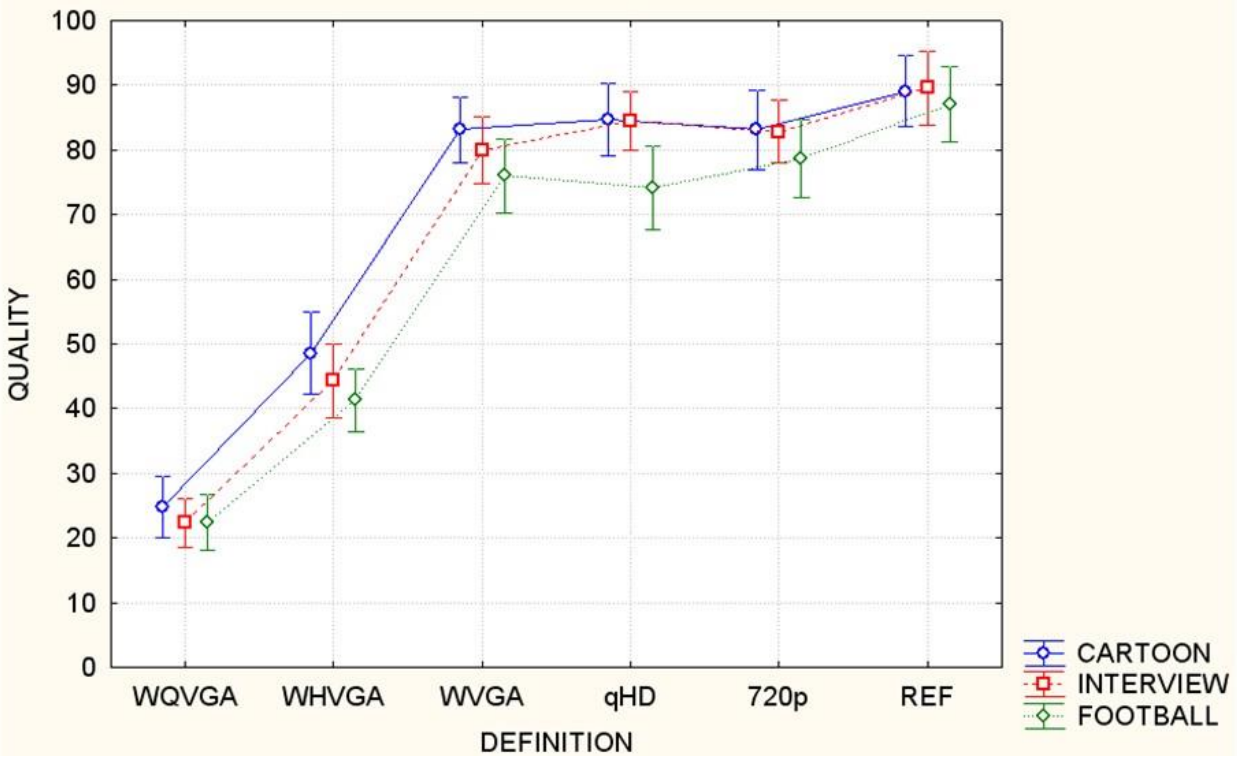

Figure 39: Subjective tests results on the impact of definition

In this test, which results are shown on the figure 39, the excellent quality is reached from WVGA resolution with Cartoon and Interview sequences whereas it is only reached with the hidden reference.

Only 3 users were rejected over 28 participants.

9.4 Conclusion of the test

This section provides the results of subjective tests conducted on a 3D capable mobile terminal with the following statements within the test conditions described above:

- The minimum frame rate required for video contents with not a lot of movements is 10 fps whereas sequences with a higher level of moving objects needs at least 16.6 fps.
- Poor to fair quality with WQVGA and WHVGA profiles are achieved whatever the contents are used.
- The minimum resolution for reaching a good to excellent quality is WVGA.

10 Content re-targeting

10.1 Introduction

3D Content is rarely created for exclusive consumption on the mobile device. It is expected that most of the content will be re-purposed from content that was originally created for large TV consumption. The task is not simply done by downscaling the left and right view of a stereoscopic video to fit the target screen. As was shown in TR 26.904 [9], the change of display resolution will affect the perceived depth. Furthermore, the change of the pixel density of the target screen will also have an impact on the perceived depth.

The perceived depth (D) is governed by the disparity according to the following formula for:

$$D = \alpha V \frac{d}{b - \alpha d} \quad (8)$$

where α is the pixel pitch, b is the interpupillary distance (IPD), and d is the measured disparity in pixels. By changing the target screen to a mobile screen, the parameters V and α will change. If the resolution of the target device differs from the content resolution also the parameter d changes due to down-sampling or up-sampling operation. Consequently, the perceived depth will be distorted unless the disparity is adjusted.

In addition, other limitations may apply, for instance the comfort zone implied by the accommodation-vergence problem, where the eyes are focusing on the center of the display, while converging on objects at a different depth. This problem gives a range of perceived depth that is comfortable and acceptable to the viewer. Objects of perceived depth outside the comfort zone may result in discomfort and shadow images due to the impossibility of fusing the left and right views.

In order to preserve the same depth perception also on mobile devices and by consequence a reasonable user experience, the content re-targeting has to be performed appropriately. The disparity map (i.e. the depth information of the stereoscopic 3D video content) needs to be adjusted correctly to correspond to the new viewing distance, screen resolution, and pixel density.

10.2 Down-sampling/Up-sampling

The content down-sampling (up-sampling) is essential in the operation of content re-targeting. The main options available for the down-sampling(up-sampling) are whether to perform the down-sampling (up-sampling) operation on both stereo views first and then perform the depth adjustment, or whether to first extract the depth map and then perform the down-sampling (up-sampling) operation on one view and the depth map as illustrated in figure 40.

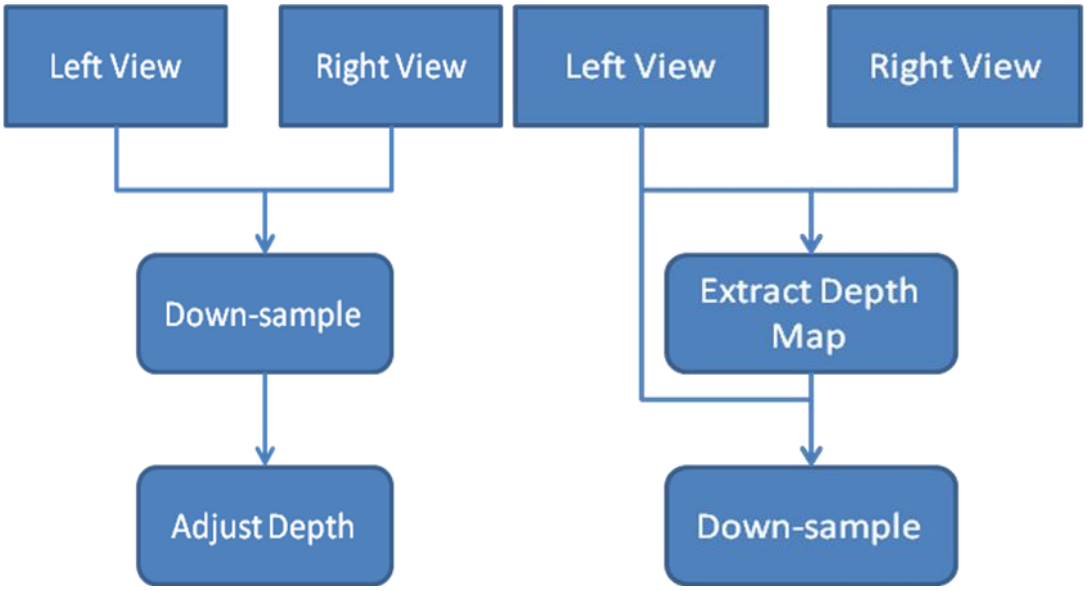

Figure 40: Down-sampling and then depth adjustment (left); Down-sampling of a left view and a depth map (right)

10.3 Extraction of depth map

The depth map is an essential component of the content re-targeting operation. The depth map is typically not available as part of the original stereo content. It is thus necessary to extract the depth information, typically by estimating the disparity map. For the estimation of the depth map, a dense stereo matching operation is performed. The operation is relatively computing intensive, however it is simplified given the fact that the left and right cameras are calibrated and that the views are rectified.

Finally, the disparity map may either be processed directly or it may be converted into a depth map. The conversion to depth map is performed using equation (8) and by setting the parameters according to the target viewing setup of the original content. For instance, if the content is intended for home viewing (e.g. in form of a 3D BluRay disc), then the viewing distance is set as a typical viewing distance in a living room, the pixel pitch is set to a typical pixel pitch value in commercial TVs, and the IPD is set to an average IPD value.

10.4 Occlusion handling

The disparity map estimation will usually leave some areas that cannot be mapped in both views due to occlusion problem, i.e. an object can be visible in one view but not in another one. Those areas will remain as areas of uncertainty and it will not be possible to assign a depth value to those pixels.

Occluded areas should be assigned a depth value that is inferior to the main objects depth value, thus causing them to be hidden in the right view.

Due to the depth adjustment the dis-occluded areas in the new right view may visible. These areas may be filled through inpainting taking the original right view as a source of information.

Occlusion handling tries to fill the dis-occluded areas and remove the occluded areas reliably in the right view.

10.5 Depth adjustment

As mentioned in the previous subclauses, the depth adjustment should take into account several aspects such as the display characteristics, the new viewing distance, and the comfort zone. The depth map is first normalized, taking values in the range [0-1] and extracting the Znear and Zfar values. The depth range may then be adjusted by adjusting the Znear and Zfar values, thus allowing shifting and streteching/shrinking of the depth range. The figure 41 illustrates such a case.

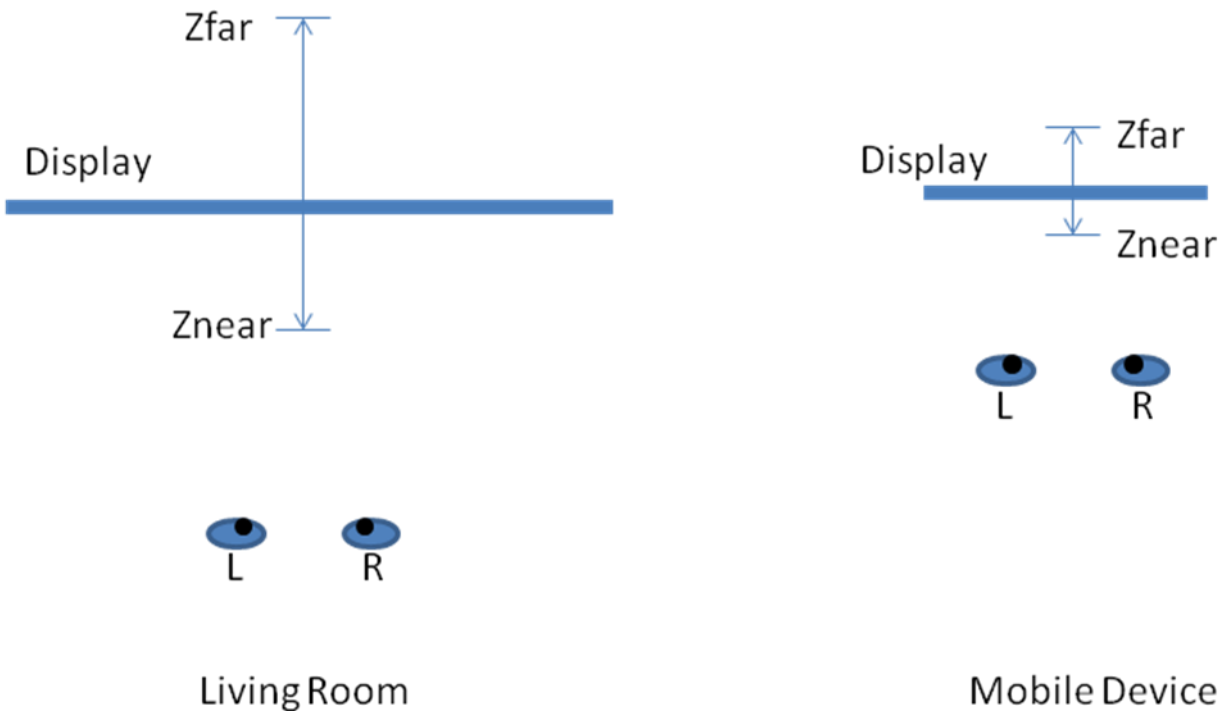

Figure 41: Example of perceived depth range on a TV-set and a mobile device.

Note also that it may be possible to perform the adjustment operation directly on the disparity map, avoiding the depth-to-disparity and disparity-to-depth conversion steps. However, due to the non-linear relationship between depth and disparity, this might result in significant distortions for the viewer [28].

10.6 Creation of the second view

After adjusting the depth map and before filling the dis-occluded areas, the second view (typically the right view) needs to be generated. This operation is performed pixel by pixel, taking the disparity value for each pixel (when available) and locating the new pixels position in the target view. The operation is performed with sub-pixel accuracy and takes into account the depth value of each pixel to resolve overlapping conflicts and to fill occluded areas. The dis-occluded areas are marked appropriately to be filled later as described in 6.3.

11 Conclusions

11.1 Introduction

This Technical Report provides analysis on different aspects relating to the introduction of stereoscopic 3D video support for 3GPP services. The most relevant use cases for mobile stereoscopic 3D video have been identified and the relevant technical solutions that can be used to enable the use cases (namely usage of the H.264 Frame Packing formats and the Multi-view Video Coding extension of H.264) have been studied. It is concluded that both the technical solutions have distinct benefits and depending on the target devices, service environment and the content to be served, one of these solutions may be more appropriate to provide the content. For instance, content that is available at half resolution side-by-side format may be more appropriately provided using the corresponding frame packing format, whereas content targeted for high end devices with H.264 High profile support may instead be better suited for MVC encoding.

It is further concluded that it is desirable to provide restrictions to the full feature sets of both frame packing and MVC based solutions in order to provide interoperability points, that address the important use cases; for example the most appropriate frame packing formats and packetization approaches are to be identified and specified.

The following sections give more details about the suitability of each of the solutions and identify the gaps that need to be filled to have support for the solution.

11.2 Frame Compatible Format for Stereoscopic Video Coding

The Frame Packing formats, namely Side-by-Side (SbS) and Top-and-Bottom (TaB), are recommended for the following services. These services are listed in the priority order:

3GPP DASH

3GPP Progressive Download and MBMS Download

3GPP PSS and MBMS RTP streaming services

3GPP MMS

3GPP MTSI

This recommendation is based on the following reasons:

A significant amount of content is available in frame packing format with half resolution and are suitable for the auto-stereoscopic displays of mobile devices

A large amount of 3D capable devices is available in the market and can be supported immediately with the SbS and TaB formats as they are inherently supported by those devices

compression efficiency is sufficient and suitable for the aforementioned 3D services

available 2D and 3D devices support the H.264/AVC Baseline profile and are thus able to support the 3D enhanced 3GPP services without hardware updates and with no or minor software updates

frame packing formats are widely supported by consumer electronics such as 3D TVs and as such content can be played out directly on external devices even by 2D devices

frame packing formats may use the baseline profile of H.264/AVC and as such show low complexity and are suitable for services such as MMS and MTSI

in many cases, 2D content and 3D content are authored separately so that different content is offered to 2D and to 3D devices

For the introduction of frame packing formats in the above listed 3GPP services, the following enablers still need to be specified:

Appropriate signalling for DASH, Progressive download (File Format), PSS (e.g. PSS base vocabulary update), MBMS, MMS, and MTSI to ensure appropriate (2D or 3D) content is offered to the different devices

Support for the frame packing arrangement SEI message for the 3D capable devices

Extensions to the services to signal the format of the content and to provide fallback solution/alternative content to 2D devices

It is recommended to use the following restrictions to the frame compatible formats:

The video is encoded progressively and only the frame packing arrangements Side-by-Side with id 3 and Top-and-Bottom with id 4 are used when frame packing is supported

For Side-by-Side the left view is packed on the left side and the right view on the right side

For Top-and-Bottom the left view is packed on top and the right view is packed on the bottom

The quincunx sampling is not used

Flipping of the views is not allowed

UE supporting frame compatible formats are able to render content formats as listed above.

11.3 Stereoscopic Multi-view Video Coding

The analysis carried out in this Technical Report indicates that the H.264/MVC Stereo High profile:

Enables full resolution stereoscopic 3D services with no inherent signal degradation

Outperforms frame packing based approaches when it comes to compression efficiency at medium to high bitrates

Provides full backwards compatibility for 2D clients supporting H.264 High profile, i.e. clients implementing H.264 High profile, but not MVC, will see a 2D version (the base view) of the stream

Has a coding toolset identical with that of H.264 High profile and thus can be implemented with small changes to typical H.264 High profile codecs (main burden seen in the system testing and verification, similarly to the case of frame packing based solutions)

Can help improve the cache hit ratios of DASH services

Can help keep server storage space for common provisioning of 2D and 3D for download and streaming as small as possible

Can help avoid transcoding in MCU and MMS relays/servers when H.264 Stereo High profile bitstreams are served to 2D clients with H.264 High profile support

Enables compression of the content to the desired bit-rates without risk of "cross-eye" compression artifacts

Has MVC file format support for storing the second eye as a separate stream, which in turn can be made available separately in DASH, saving bit-rate for clients that do not implement MVC or wish to display 2D

Has no risk of clients seeing "corrupted" video (frame-packed video displayed without unpacking)

Has mature file format and signaling available for the DASH, Download and MMS services

Based on the considerations above, H.264 Stereo High profile is recommended for the following services. These services are listed in the priority order:

3GPP DASH

3GPP Progressive Download and MBMS Download

3GPP MMS

For the introduction of MVC in the above listed 3GPP services, the following enablers still need to be specified:

Appropriate signalling for DASH, Progressive download (File Format) and MMS to ensure appropriate (2D or 3D) content is offered to the different devices

Extensions to the services to signal the format of the content and to provide fallback solution/alternative content to 2D devices

Annex A:
Change history

Change history							
Date	**TSG #**	**TSG Doc.**	**CR**	**Rev**	**Subject/Comment**	**Old**	**New**
2012-06	56	SP-120227			Presented at TSG SA#56 (for approval)		2.0.0
2012-06	56	SP-120330			Presented at TSG SA#56 (for approval, figures updated)	2.0.0	2.0.1
2012-06	56				Approved at TSG SA#56	2.0.1	11.0.0
2014-09	65				Version for Release 12	11.0.0	12.0.0
2015-12	70				Version for Release 13	12.0.0	13.0.0

History

Document history		
V13.0.0	January 2016	Publication