A QoS Solution for Three-Dimensional Full-HD H.264/MVC Video Transmission over IP Networks

Diogo M. Bento
Departamento de Eng. Eletrotécnica
Instituto Superior de Engenharia, UAlg
Faro, Portugal
db_eee@yahoo.com

Jânio M. Monteiro
INESC Inovação, Lisbon, Portugal &
Instituto Superior de Engenharia, UAlg
Faro, Portugal

Abstract— Tridimensional video streaming has recently drawn significant attention from users and content providers. This has led to an implementation of 3D transmission over IP networks that extend the legacy 2D solution to support multiple views within each image. More recently, a Multiview Video Coding amendment of the H.264 standard has been approved which, among other applications, is being used on the encoding of 3D content in Blu-ray discs. This latter solution has shown to be capable of improving the compression ratio when compared with the Side-by-Side encoding, by exploring inter-view redundancies. In this paper, we evaluate the challenges of encoding and transmitting 3D content in MVC, for Full HD content distribution over IP networks. We also design and evaluate a Weighted RED queuing mechanism for QoS capable networks based on the MVC structure is capable of improving the quality of the received video.

Keywords: H.264 Multi-View Video Coding (MVC), 3D video transmission, High-Definition

I. INTRODUCTION

Users' interest in stereoscopic or tridimensional (3D) video using 3D glasses has grown tremendously in recent years, driven not only by the recent surge in 3D display terminals and Set-top-Boxes [1], but also by content availability, including 3D movies and 3DTV broadcasts over Digital Video Broadcast (DVB) or Internet Protocol Television (IPTV). Together with 3D, there has also been an improvement in video definition and Signal-to-Noise Ratio (SNR) qualities, which are desirable in order to be able of exploiting the full potential of such solutions.

In terms of optical media, Multiview Video Coding (MVC) has become the de facto solution for encoding tridimensional video, which includes 3D Blu-ray discs and High Definition (HD) cinema. These solutions use the H.264/MVC extension [2] (Annex H) of the H.264 standard. The basic principle of MVC coding is to go beyond the reduction in intra-frame spatial and temporal redundancies of consecutive frames, exploring the similarities between frames of different views. Applying this process, there is a reduction of the data transmission rate, when compared with the transmission of two separate views, without compromising the final quality of the reconstructed video.

The encoding and transmission of Multi-View is still an open research topic, as many issues are still being explored, including the subjective quality or Quality of Experience (QoE) evaluation [3]. Until now, results about the perception of stereoscopic 3D video have shown that the human visual system (HVS) is able to asymmetrically tolerate a reduction of high frequency components in one of the views. These results suggest that stereoscopic video streaming can exploit asymmetric quality distribution among views [4].

There is also a need for Quality of Service (QoS) and Unequal Error Protection strategies, requiring the understanding of the MVC interdependency structure and focusing on the maximization of the decoded QoE.

In terms of H.264 transmission, while the RTP payload format for H.264/AVC [5] is now fully implemented in many IPTV networks, the Internet Engineering Task Force (IETF) is currently working on the definition of an equivalent standard for H.264/MVC [6]. One of the potential problems faced by the transmission of 3D HD content is that the average size of Network Adaptation Layer (NAL) units tends to significantly grow much beyond the fixed size of IP packet payloads. Due to fragmentation at the transport layer, each of these NAL units is typically fragmented in multiple RTP/UDP/IP packets, making the encoded structure very sensitive to packet losses. In H.264/MVC this problem is expected to be combined with its complex and interdependent structure.

While the WRED algorithm can be used to allocate differential protection to distinct video packets accordingly to their importance (as shown in [7]), as far as we know none of these studies were applied to the transmission of 3D and Full-HD MVC sequences.

Given these considerations, the aim of this paper is to evaluate and improve a Full HD 3D transmission using H.264/MVC. It starts by evaluating the loss resilience of the structure to Binary Erasure Channels (BEC) [8], like wired and wireless IP networks; it continues by evaluating the improvements introduced by an unequal loss protection mechanism; and given these results it specifies and evaluates a selective marking and discard solution, based in the Weighted Random Early Detection (WRED) congestion avoidance mechanism [9].

The rest of this paper is organized as follows. Section II presents the MVC bit stream structure. Section III describes the 3D Full HD video sequences used in the tests and the results of the H.264/MVC encoding. Section IV measures the effect of

Packet Losses on the encoded sequence and analyses the advantages of an unequal loss protection mechanism in terms of quality assurance. Section V designs and evaluates a queuing mechanism for erasure protection and finally, section VI concludes the paper.

II. H.264/MVC STRUCTURE

The MVC bit stream is composed of multiple-views combined with temporal scalability. Temporal scalability is a technique that allows support of multiple frame rates. In MVC, temporal scalability is implemented by using hierarchical prediction structures, in a process similar to the solution used in Scalable Video Coding (H.264/SVC) [1] (Annex G).

MVC views are very interdependent. This means that the loss of a NAL unit of a certain view or temporal layer may cause a severe reduction of quality (see Figure 1) or even prevent the proper decoding of other layers.

Figure 1. Example of visual artifacts in a partially discarded H.264/MVC sequence

For instance, if a 3D video sequence is encoded in H.264/MVC, with a Group-of-Pictures (GOP) size of 4 and an Intra period of 8, the interdependency structure between NAL units will be the one represented in Figure 2. Each of these NAL units may be identified within the encoded bit stream, using the following identifiers: View Identifier (VID), the NAL Reference Index (NRI) and the temporal layer identifier, as described in [1][6].

The NRI is a two-digit binary field. A value of '00' means the content of a certain NAL unit is not used to rebuild reference pictures for future predictions. These NAL units can be discarded without losing the integrity of the reference pictures in a certain view. A value higher than '00' means that the decoding of a NAL unit is needed to maintain the integrity of reference pictures on a certain view, or that the NAL unit contains a parameter set.

Regarding the structure illustrated in Figure 2, we can observe that the most important information that should be kept away from packet losses are the frames with higher NRI levels (i.e., with NRIs 2 and 3).

Inside a certain GOP, if a NAL unit is lost, all NAL units that depend on it will not be used in the decoding process, even if they arrive properly. Based on the example illustrated in Figure 2, for instance, if a BRef NAL unit of the base layer (i.e. VID=0) is discarded, all lower NAL units of the same view will not be used for decoding. The same also happens with other interdependent NAL units belonging to the other view.

This problem is also amplified by the fragmentation of each NAL unit in several packets, which is the typical behavior on IP-based networks. In fact, if a retransmission based transport protocol or Application Layer - Forward Error Correcting (AL-FEC) mechanism is not used, the loss of only one fragment in a NAL unit implies the complete discard of all the other fragments that might be received.

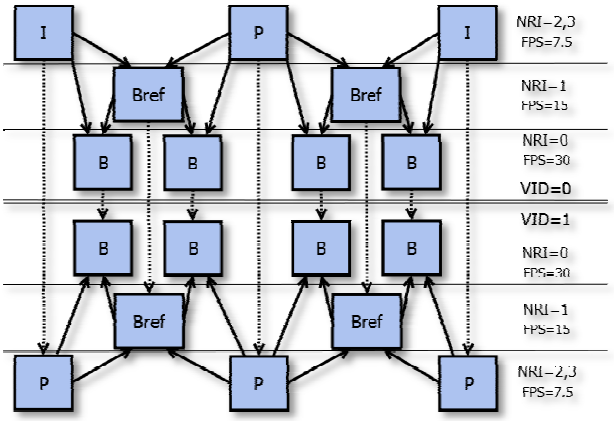


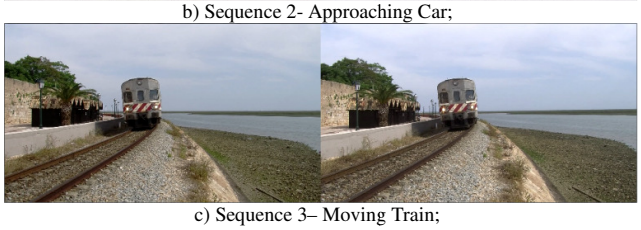
Figure 2. Example of an interdependency structure in an MVC encoded video with two distinct views, where arrows represent NAL unit dependencies

III. ENCODED SEQUENCES

In this section, we will describe the results of encoding several video sequences using the standard H.264/MVC encoder.

a) Sequence 1- Football Match;

b) Sequence 2- Approaching Car;



c) Sequence 3– Moving Train;

Figure 3. Encoded MVC Video Sequences used in the tests

Figure 3 shows a Side-by-Side representation of the sequences used in this paper. These video sequences were

recorded using two High-Definition cameras (each view with 1920×1080 pixels, in progressive scan), at a frame rate of 30 fps and with a spacing between camera lens of 12 cm. All video sequences had a length of 10 seconds.

Sequence 1 presents a football training session. This sequence is characterized by a low level of spatial information and by medium levels of movement. Sequence 2 is characterized for presenting an image of a car approaching a fixed camera, with medium levels of both spatial and temporal information. Sequence 3 shows a train passing close to the video camera, which follows the train movement. It has medium levels of spatial information and high levels of temporal information.

These videos were afterwards encoded as described in the following sub-sections, using the H.264/MVC encoding structure. The encoded video sequences can be found in [10].

A. *H.264/MVC Encoding*

Each video sequence was encoded in H.264/MVC, using the MVC reference software JMVC (Joint Multiview Video Coding), version 8.5. A GOP size of 4 and an Intra period of 8 were used for all the sequences. The length of each sequence was adjusted to 300 frames, for a corresponding duration of 10 seconds.

Table I shows the bit rates of the encoded video sequences. The different columns of transmission rate represent the extraction points supported by the temporal scalability, using the distinct NRI values. There are no visual artifacts when the streams are stripped according to the NRI levels. When discarding all layers having NRI=0, the resulting video sequences present a frame rate of 15 fps, since all lower-level B-frames have been discarded; when discarding all NAL units with NRIs 0 and 1, all video streams ran at 7.5fps, as all B_{Ref} and B frames have been discarded. The selective discard of temporal layers yields an average reduction on the transmission rate of 14.7% and 30.0%, respectively.

TABLE I.
SCALABILITY OPTIONS OF THE ENCODED H.264/MVC SEQUENCES

	Y-PSNR (dB)	Transmission Rate (kbps)		
		NRI		
		All	***1, 2 and 3***	***2 and 3***
Sequence 1	40.49	20209.41 (100%)	18972.31 (93.9%)	14939.55 (73.9%)
Sequence 2	38.95	34184.41 (100%)	28419.61 (83.1%)	24337.22 (71.2%)
Sequence 3	39.00	39980.42 (100%)	31475.25 (78.7%)	25906.25 (64.8%)
Decoded Frame Rate		30 fps	15 fps	7.5 fps

IV. RANDOM VERSUS SELECTIVE DISCARD OF H.264/MVC SEQUENCES

In this section, we analyze the impact of random packet losses in a 3D Full-HD video stream encoded in H.264/MVC, comparing it with a simple differential protection mechanism.

A. *Description of the implemented discard process*

The results of random packet losses were quantified using the percentage of data that reaches the receiver but cannot be used on the decoding process due to NAL unit dependencies (as explained in section II) and the Y-PSNR metric.

In order to approximate the discard process of what occurs on a real IP-based video transmission system, NAL units longer than 1450 bytes were fragmented into several Real-Time Transport Protocol (RTP) packets [5][6].

RTP packets were randomly discarded with increasing values of packet loss probability (0.1%, 0.2%, 0.3%, 0.5%, 0.7% and 1%).

To simulate what happens in the receiver side all NAL units were reassembled and decoded according to the following rules:

1) Since NAL units cannot be truncated or partially received, if one of its fragments is lost, therefore, it should be completely discarded. Hence, long NAL units are expected to be very sensitive to packet losses.

2) If one NAL unit was completely received but a previous one from which it depends on wasn't received or decoded, then it cannot be decoded. For instance, if a NAL unit with (VID,T)=(1,1) depends on a NAL unit with (VID,T)=(0,1), and considering that the latter was not received or decoded, this implies that the prior one cannot also be decoded. This process is hereafter referred as interlayer dependency.

If the interdependencies between NAL units are not carefully considered, it may lead to streams which cannot be decoded by the JMVC decoder. In order to prevent this from happening, we have implemented an interdependency analysis function that is used before the decoding process.

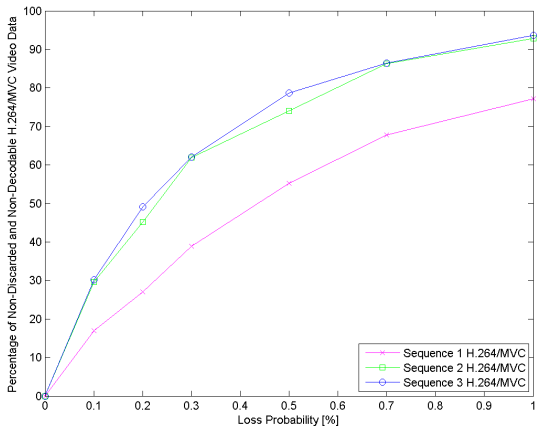


Figure 4. Impact of random packet losses in data dependency of H.264/MVC Video Data

B. *Random Losses of the H.264/MVC Sequences*

Figure 4 illustrates the interdependency results of random packet losses when the sequences 1, 2 and 3 are encoded in H.264/MVC. They show that the H.264/MVC encoding is very sensitive to packet losses. As can be seen, a small increment in the loss probability of packets causes a significant increment of the amount of data that is received but cannot be decoded, as a consequence of fragment losses and interlayer dependency. In this figure, the sum of all discarded data at the receiver

includes the discarded data caused by fragment loss and the discarded data caused by the absence of lower layer NALs from which it depends on.

Table II shows the average number of packets needed to carry Full-HD H.264/MVC units over the RTP protocol. As can be seen, the average number of packets needed to carry a NAL unit increases as we increase the NRI level (and hence, the accordingly frame type). This is an important factor because if the receiver buffer is small, there is a big chance that there will be congestion at the receiver side, and lower-level frames will need to be discarded.

TABLE II.
AVERAGE NUMBER OF PACKETS REQUIRED FOR CARRYING H.264/MVC NAL UNITS OVER RTP

	Average Number of Packets per NAL Unit			
	NRI			
	All	***0***	***1***	***2 and 3***
Sequence 1	19.27	8.94	12.66	46.07
Sequence 2	32.60	11.04	15.64	91.62
Sequence 3	38.13	16.29	21.34	97.52

C. Selective Discard of the H.264/MVC Sequences

Given the several ways to differentially protect such encoding from packet losses or errors, on the following we analyze how to increase the resilience of MVC transmission, supporting a gracefully quality degradation.

We have implemented a selective discard mechanism that allocates higher levels of protection to higher NRI values. In this solution, under certain Packet Loss Ratio (PLR) values only NAL units having a $NRI \leq \lambda$ are randomly discarded, where λ is dynamically adjusted to meet the required PLR.

Each of the sequences was submitted to this selective discard mechanism. Figure 5 illustrates the results in terms of reduction of the Y-PSNR, for each of the 300 frames of the Sequence 1, when it is exposed to a random PLR of 1%, comparing it with the selective discard solution. As can be verified, the decay in the Y-PSNR values is very pronounced in the random discard case, while the selective discard solution is able to maintain the quality of key images, which leads to an average quality increment. The selective discarding solution was able to assure a quality increase in all H.264/MVC decoded sequences, as the PLR increases. In fact, although the quality of some frames belonging to certain views decreases (as shown in Figure 5), the overall quality of all frames is maintained above the random discard counterpart.

Figures 6 and 7 respectively illustrate the average Y-PSNR and mean frame rate results for all MVC sequences, comparing the random losses with the selective discard mechanism. It can be verified not only the Y-PSNR of the decoded frames is maintained close to the maximum, but that there is also a significant increment in the number of decodable frames, when compared with the random packet losses.

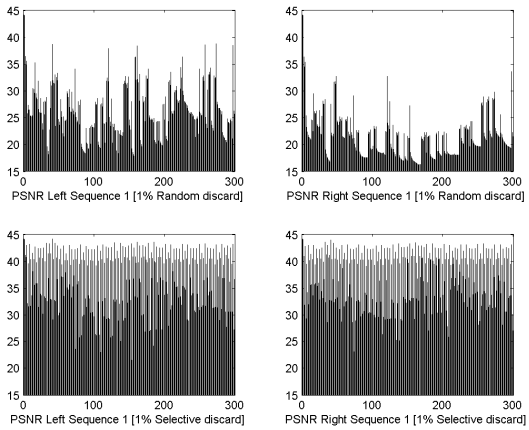


Figure 5. Y-PSNR of the decoded frames of Sequence 1 encoded in H.264/MVC for a random (upper part) versus selective (bottom part) Packet Loss Ratio of 1%.

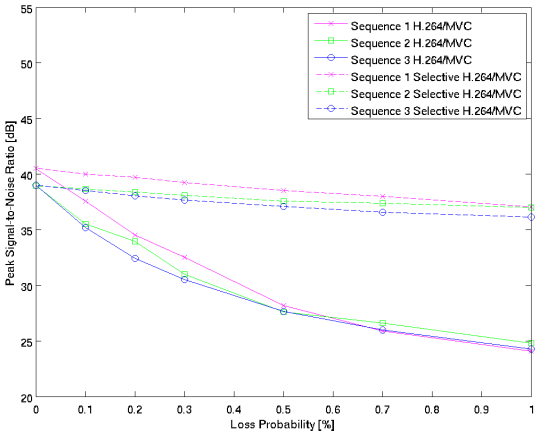


Figure 6. Impact of the selective discard mechanism in the Y-PSNR of video sequences encoded in H.264/MVC, when compared with the random packet losses

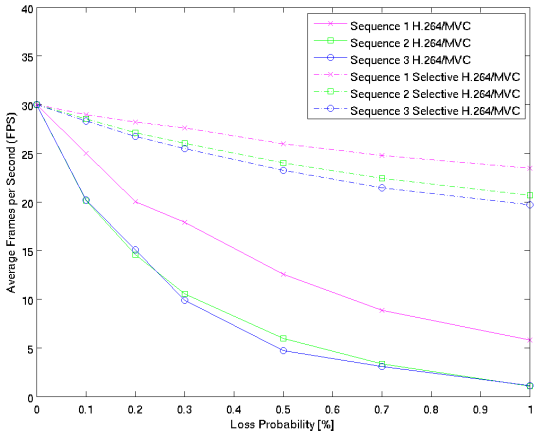


Figure 7. Impact of the selective discard mechanism in the average frame rate of H.264/MVC video sequences when compared with random packet losses

V. Definition of a QoS queuing mechanism for Full-HD 3D Video transmission

Given the results of previous section, it becomes important to define and evaluate a QoS mechanism capable of supporting such differential protection of Full-HD 3D H.264/MVC sequences. Since the results of previous section have only considered PLRs up to 1%, a solution should be defined that supports much higher values of congestion. To do it, in this section we consider a Weighted RED (WRED) queuing mechanism with four levels of priority (from P_0 to P_3). The proposed method can be used as a Queuing mechanism in QoS capable routers or wired/wireless bridges. Given the four priority levels, in the following we will evaluate how this queuing mechanism can be used to protect H.264/MVC sequences.

A. *QoS algorithm for Full-HD 3D video streams:*

Figure 8 illustrates the four WRED priority levels considered in the following tests, which was implemented using Algorithm 1.

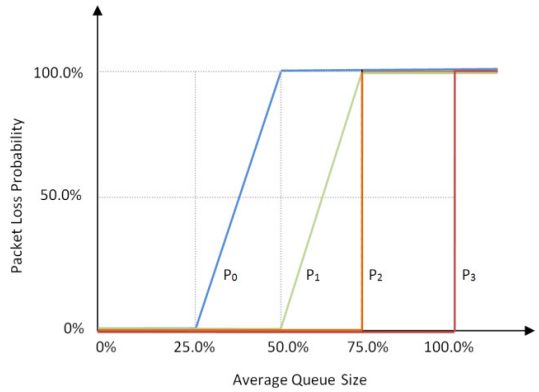


Figure 8. WRED-based algorithm used for the Full-HD 3D MVC sequences

Algorithm 1. Discard Algorithm for encoded H.264/MVC Full-HD Sequences

```
WRED(Priority, Packet)
λ := GetQueueFillRatio()
if (λ < 0.25)
        Enqueue(Packet)
else if (λ < 0.50)
        if (Priority==0 && rand()<= 4.0 × λ − 1.0)
                Discard(Packet)
        else
                Enqueue(Packet)
else if (λ < 0.75)
        if (Priority==0 || (Priority==1 && rand()<= 4.0 × λ − 2.0))
                Discard(Packet)
        else
                Enqueue(Packet)
else if (λ < 1.0)
        if (Priority <= 2)
                Discard(Packet)
        else
                Enqueue(Packet)
else
        Discard(Packet)
End
```

Table III presents the QoS marking solution which was implemented together with the WRED solution. As we can see, video packets belonging to NAL units with NRI=0 and 1 were respectively associated with priorities P_0 and P_1, enabling a reduction of 50% and 75% of the frame rate of both views as the buffer length increases to nearly 75%. Above 75% all packets that belong to NRI>=2 and VID=1 were marked as P_2 and therefore were discarded. When this happens, only those packets marked as P_3 are forwarded; as such, packets belonging to VID=1 are completely discarded and the receiver should get a single view with a frame rate of ¼ of the original value. In this case, the receiver application should replicate the same view in both eyes.

TABLE III.
QoS Marking for the encoded H.264/MVC video sequences

	NRI Levels		
VID Level	***0***	***1***	***2 and 3***
0	P_0	P_1	P_3
1	P_0	P_1	P_2

In order to test the previous algorithm, a simulator in MATLAB was implemented, comparing the WRED solution proposed here against the standard Drop-Tail (or First-In-First-Out, FIFO) queue management. A maximum queue length (i.e., a buffer) of 1000 packets was used and the output bit rate was adjusted to meet a certain congestion level. After that, we have measured the average Y-PSNR values obtained for different levels of congestion and the queue length, for congestion levels varying from 0% to 100%.

B. *Results*

Figure 9 illustrates the average queue length of both FIFO and WRED solutions, according to the congestion level. As can be verified, the FIFO/Drop-Tail solution tends to quickly fill the buffer, which leads to uncontrolled losses of packets when the buffer fills. On the contrary, the proposed queue management algorithm starts discarding parts of the video as the queue length increases, which contributes to a higher stability of discards.

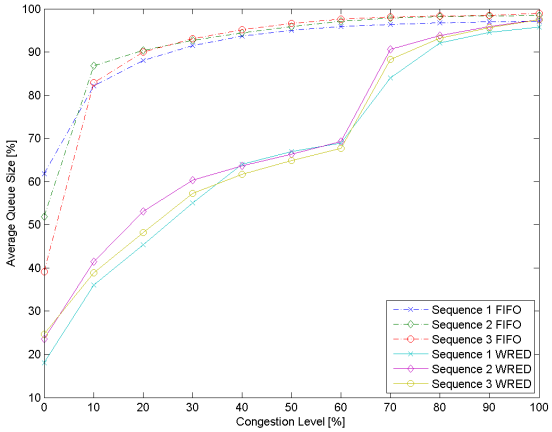


Figure 9. Average Queue Size for different Congestion Levels (FIFO vs WRED)

The stability effect of the proposed WRED solution is illustrated in Fig 10, which shows the Y-PSNR of each decoded frame, comparing the FIFO *versus* WRED solutions for a congestion level of 20%. The WRED solution is capable of achieving a higher Y-PSNR for both views, with a higher stability in intra and inter left and right views, which is also important to assure a higher quality of perception of users.

Finally, Figure 11 illustrates the Y-PSNR decay of both solutions as a function of the congestion level. As can be verified, the proposed WRED solution obtains a significant improvement in terms of the decoded Y-PSNR when compared with the FIFO method. Particularly, it can be verified that for congestion levels as high as 20% and 30%, the gain introduced by this method surpasses 6 dB of Y-PSNR, when comparing it with the Drop-Tail solution.

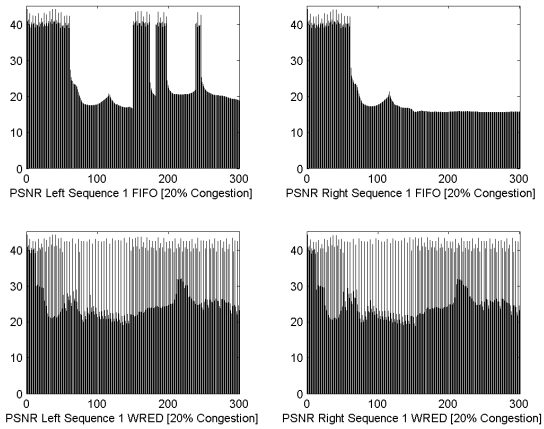


Figure 10. Measured Y-PSNR for each frame and view of each H.264/MVC video stream comparing the Drop-Tail (upper images) with the WRED (bottom images) for a congestion level of 20%

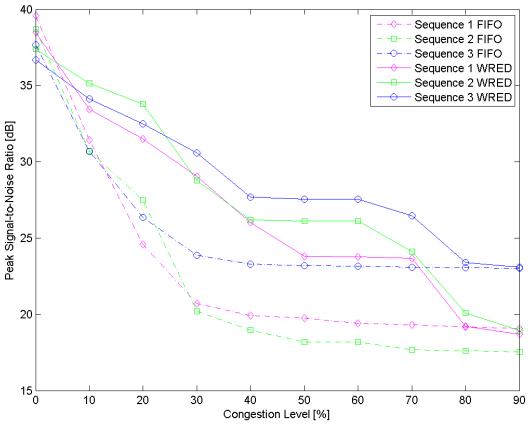


Figure 11. Measured Y-PSNR for each frame and view of each H.264/MVC video stream (50% congestion)

The decay in the Y-PSNR values of the WRED was mainly due to two effects. First, as the queue fills up, part of the frames are discarded due to the selective marking and discard solutions chosen. This leads to a reduction in the represented frame rate, which in turn affects the measured Y-PSNR values, since the last decoded frame is used as reference for its computation. Additionally, when the number of packets in the queue surpasses 75% of the maximum length, all frames of the second view are discarded. In this case, the Y-PSNR of the right view decays significantly.

VI. CONCLUSIONS

The results in this paper have demonstrated that the transmission of Full-HD 3D video over IP networks using H.264/MVC requires the specification of a proper erasure protection mechanism. In fact, the results of the random packet losses made in Section IV have shown that the H.264/MVC encoding structure is very sensitive to PLR values lower than 1%. Such random and sparse packet losses cause a quality degradation of nearly 15 dB due to the combined effect of large NAL unit sizes resulting from the Full-HD definition, with an increased complexity introduced by the H.264/MVC 3D structure. Using a proper selective discard mechanism we have achieved Y-PSNR decays of nearly 4 dB and an average frame rate reduction lower than 10 fps, when imposing a 1% PLR.

In order to implement such a differential protection mechanism we have considered a WRED queuing solution that was specially adapted for MVC transmission. The proposed solution has been capable of introducing considerable quality resilience to packet losses, when compared with the tests made in section IV and with a Drop-Tail queuing mechanism.

We have also verified that the random discard tests performed in Section IV caused higher Y-PSNR decays than the ones obtained with the Drop-Tail queue. This was due to the fact that the Drop-Tail solution tends to discard several consecutive video packets when the buffer is full, while in the random discard packets are sparsely discarded, which, for the same PLR, tends to affect more NALs units.

REFERENCES

[1] Broadcom, “BCM7425 - Full-Resolution HD 3DTV Set-Top Box SoC Solution”, Available: http://www.broadcom.com/products/Cable/Cable-Set-Top-Box-Solutions/ BCM7425, accessed: Feb. 7, 2012.

[2] ITU-T and ISO/IEC JTC 1, ”Advanced video coding for generic audiovisual services, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (AVC),” Version 8 (including MVC extension), Jul. 2007.

[3] G. Saygili, G. Gurler, A. M. Tekalp, “Quality assessment of asymmetric stereo video coding,” in Proc. IEEE Int. Conf. on Image Processing (ICIP), Hong Kong, Sep. 2010.

[4] W. J. Tam, “Image and depth quality of asymmetrically coded stereoscopic video for 3D-TV,” Joint Video Team document JVT-W094, Apr. 2007.

[5] S. Wenger, M. Hannuksela, T. Stockhammer, M. Westerlund, D. Singer, "RTP Payload Format for H.264 Video", RFC 3984, Feb. 2005.

[6] Y.-K.Wang, T. Schierl, and R. Skupin, “RTP Payload Format for MVC Video,” IETF Internet Draft, Work in Progress, draft-ietf-payload-rtp-mvc-01.txt, Sept. 2011.

[7] W.-T. Lee, F.-H. Liu, H.-F. Lo, “Improving the performance of MPEG-4 transmission in IEEE 802.15.3 WPAN”, IEEE CIT 2008, Jul. 2008.

[8] P. Elias, “Coding for two noisy channels,” in Proc. 3rd London Symp. on Information Theory, London, U.K., 1955, pp. 61–76.

[9] Cisco, “Distributed Weighted Random Early Detection”, Cisco Press White Paper. Available: http://www.cisco.com/en/US/docs/ios/11_1/feature/guide/WRED.html. Accessed Feb. 14, 2012.

[10] D. Bento, J. Monteiro, A. Milene, “3D Full HD Video Sequences”, Available: http://w3.ualg.pt/~jmmontei/Res.htm, accessed: Apr. 9, 2012.