Turk J Elec Eng & Comp Sci, Vol.20, No.4, 2012, © TÜBİTAK
doi:10.3906/elk-1010-880

Trellis-based optimization of layer extraction for rate adaptation in real-time scalable stereo video coding

Nükhet ÖZBEK
Department of Computer Engineering, Yaşar University, Bornova 35100, İzmir-TURKEY
e-mail: ozbek.nukhet@gmail.com

Received: 23.10.2010

Abstract

The concept of quality layers (QLs) has been adopted in the scalable video coding standard to enable optimal rate adaptation of precoded video in the rate-distortion sense. QLs were previously extended to stereo and multiple-view scalable video for efficient transport of 3DTV over the Internet. However, it is not possible to use the QL method in applications that require real-time encoding since the priority determination process assumes the availability of the whole video sequence. In this work, a trellis-based online rate adaptation is proposed for real-time scalable stereo video coding, with a delay of 1 group of pictures (GoP). The delay can be controlled by selection of the GoP size according to the application, such as 16 frames for live broadcast or 8 or 4 frames for videoconferencing. In addition, the joint optimization of layer extraction for scalable multiview coded stereo video is also proposed. It is assumed that the encoder/extractor is aware of the available dynamic network bandwidth in order to perform rate-distortion optimized medium-grain fidelity scalability layer selection for each GoP. Experimental results show that the performance of the proposed online method is very close to that of QLs that would require the whole video sequence.

Key Words: *3DTV, dynamic rate allocation, quality layers, scalable video coding, scalable multiview video coding, trellis-based optimization*

1. Introduction

Recently, scalable video coding (SVC) and stereo and multiview video have gained wide interest. The new SVC and multiview video coding (MVC) standards, which are extensions of H.264/AVC [1], MPEG-4 part 10 (ISO/IEC 14496-10:2005/AMD3), were developed by the Joint Video Team to respond to market needs for Internet video and 3D video, respectively. The SVC amendment was finalized in 2007 and MVC was added to the standard in 2009 [2,3]. The joint scalable video model (JSVM) [4] and the joint multiview video model (JMVM) [5] were developed as reference codecs to provide software implementation and to demonstrate nonnormative encoding tools. Although the JMVM was implemented based on the JSVM, in order to take advantage of some of the interfaces and transport mechanisms introduced for SVC, the JMVM currently does not support scalable coding.

For adaptive streaming applications, packet-based fidelity scalability and optimized rate adaptation are highly desirable. A low-complexity but high-performance method for packet-based fidelity scalability, also referred to as medium-grain fidelity scalability (MGS), was adopted as a normative element of SVC [2]. MGS operates in the transform domain and allows fragmentation of a given fidelity enhancement, which means frequency-selective grouping of the transform coefficients [6]. This splitting of coefficients among fragments enables graceful degradation if fragments are dropped during adaptation.

The quality layers (QLs) concept, which was designed for transmitting a priority value for each network abstraction layer unit (NALU), was adopted in the SVC standard in order to enable an optimal adaptation in a rate-distortion (R-D) sense [7]. However, the method of deriving suitable priority_id values is not part of the standard. The example method in [7] also presents a way to extract NALUs according to priority_ids for a given target bitrate. The QL method is used as a ground truth and the goodness criterion is how close the proposed method is to the QL method in terms of R-D performance.

Latest advances in 3DTV technology have led to new approaches for efficient coding and transport of multiview video (MVV). There are several approaches for the encoding of MVV, which provide a trade-off between random access, ease of rate adaptation, and compression efficiency, allowing simulcast coding, scalable simulcast coding, MVC, and scalable multiview video coding (SMVC). MVC based on hierarchical B-pictures in temporal and interview dimensions has proven to have the best performance in exhaustive experiments conducted in the context of MPEG standardization. The effectiveness of this approach was demonstrated by an experimental analysis of temporal versus interview prediction in terms of a Lagrange cost function in [8].

In order to combine the advantages of scalable coding and MVC, SMVC was recently introduced, which is an extension of SVC [2] for MVC and presented coding results that were superior to simulcast SVC of stereo and multiple views with effectiveness in view and/or layer switching [9,10]. SMVC uses hierarchical B-pictures in both temporal and interview prediction. QLs were also extended to stereo and multiview scalable video for adaptive optimized 3DTV streaming over the Internet [11,12]. However, it is not possible to use QLs in applications that require real-time encoding since the priority determination process assumes availability of the whole stereo video sequence and the R-D data are computed for each NALU in which computing the distortion is the most time consuming.

Trellis-based approaches are widely used in bit allocation by optimal quantization [13-15], video summarization [16], and optimal fidelity scaling for spatial layers [17]. In this paper, a trellis-based online rate adaptation is proposed for real-time scalable coding of stereo videos. The algorithm assumes that the encoder/extractor is aware of the available network bandwidth and thus R-D optimized (RDO) MGS layer selection can be performed dynamically for each group of pictures (GoP). The delay can be controlled by selection of a suitable GoP size according to the application, such as 16 frames for live broadcast or 8 or 4 frames for videoconferencing. The paper is organized as follows: Section 2 reviews scalable stereoscopic video coding, which was developed earlier, along with the QL concept and previous work on stereo extension of QLs. Section 3 introduces the proposed algorithm and discusses implementation and complexity issues. Section 4 provides experimental results with monocular and stereoscopic test sequences. Conclusions are drawn in Section 5.

2. Background

2.1. Scalable stereo video coding

The SMVC design [9] exploits the temporal scalability feature of the JSVM reference software [4] by sequential interleaving of the first (right) and second (left) views in each GoP. The prediction structure, which is given in detail in [9], supports adaptive temporal or disparity-compensated prediction such that every frame in the left view uses past and future frames from its own view and the collocated frame from the right view for prediction. In every view, except the first view, the first frame of each GoP in the coding order, which is generally called the key frame, uses only interview prediction so that extraction of any view at the desired temporal resolution is possible. For the first view, key frames are encoded by using temporal prediction.

Since 2 views are interleaved, the effective GoP size is half of the original GoP size of the JSVM, where even- and odd-numbered frames in the display order correspond to the right and left views, respectively. Thus, the decoder and bit stream extractor modules of the JSVM are modified to recover the last temporal level of the interleaved bitstream as the left view. The spatial and signal-to-noise ratio scalability functionalities of the JSVM remain unchanged; however, the decoder is modified in order to decode the right and left views at different spatial resolutions

2.2. QLs and QL-based bitstream extraction

The QL principle assigns a prioritization order to various elements constituting the SVC bitstream. This prioritization exhibits a virtual layered organization of the stream to be used for adaptation such that stream elements are transmitted according to their priority.

The proposed QL method is designed as a postprocessor to evaluate the SVC bitstream and signal a preferred extraction order for the rate and distortion of the various enhancement information pieces. QL-based extraction significantly outperforms the basic extractor that was initially implemented in SVC [7]. For each NALU, a priority identifier is calculated and embedded in the NALU header. Thanks to the signaling of a preferred extraction order in the header of the NALUs, through the syntax element priority_id, or in a supplemental enhancement information (SEI) message, the adaptation can be performed with a simple parser.

The computation of the QL information is performed in 4 steps. First, rate and distortion values are calculated for each picture that is encoded using a base representation and quality refinement levels. Second, the R-D values for all pictures are used to establish the R-D curve. Third, the R-D points lying on the convex hull of the curve are sorted according to their R-D slopes. Finally, the priority_id value is calculated from this slope.

2.3. QLs for scalable stereo video coding

The distortion of a picture depends on the distortion of the pictures from which it has been predicted. The JSVM employs a hierarchical temporal prediction, so dependency constraints can be represented using a hierarchical structure.

The QL principle was previously extended to the case of SMVC for 2 and 8 views [11,12]. According to temporal and interview prediction structure in SMVC, the list of dependants must change for each frame. The first odd-numbered frames (the last temporal level) have no dependants in the single-view case, but this is not true for the stereoscopic case, since they represent the second view. For example, in the GoP size = 16 scenario that is shown in Figure 1, the list of dependants of picture 17 in display order includes pictures 9, 11, 13, 15, 19, 21, and 23. The list of dependants of a first-view frame must also have its neighbor frame, which has an incremented picture number due to the dependency between the 2 views, such that the list of dependants of picture 16 includes picture 17, as well.

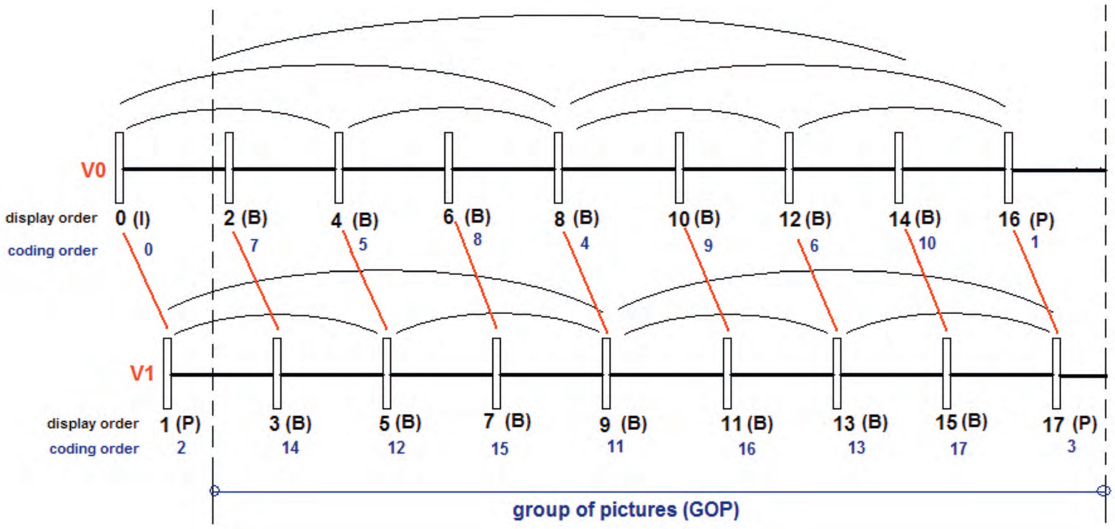

Figure 1. SMVC prediction structure for 2 views and GoP size of 16. Note that there is also dependency between neighbor frames (i.e. 2 and 0/4, 3 and 1/5).

3. Trellis-based RDO layer extraction

This section formulates the RDO layer extraction problem, explains the concept of the proposed algorithm, and discusses implementation and complexity issues.

3.1. The concept

We assume an adaptive streaming scenario with real-time SMVC encoding, where the extractor is aware of the available dynamic network bandwidth in order to perform RDO MGS layer selection for each GoP. The SMVC bitstream is encoded with 1 base and 1 enhancement layer at a single spatial resolution, and the enhancement layer is encoded using MGS with key pictures. The proposed trellis-based MGS layer extractor module in the sender periodically receives the sending rate R_{TFRC} from the transmission control protocol-friendly rate control (TFRC) module [18].

Determination of how many MGS layers should be allocated for each frame in each view is a resource allocation problem in which there are dependencies between temporal levels as well as the 2 views. Thus, a trellis-based optimization can be used to solve the problem, as follows.

Maximize the quality of each GoP of a stereo bitstream, defined by:

$$Q_{GoP} = \frac{PSNR_{GoP}^{left} + PSNR_{GoP}^{right}}{2}, \tag{1}$$

subject to:

$$\sum_{TL=0}^{N} R_{TL} \leq R_{TFRC}, \tag{2}$$

where Q_{GoP} indicates the average peak signal-to-noise ratio (PSNR) over 1 GoP for the left and right channels and R_{TL} is the total rate, base, and enhancements for temporal level TL.

In order to solve this problem, a trellis is built, which links the base quality (initial) node to the nodes in the final stage showing all possible dependencies within a GoP. The base quality node is the initial node (denoted by a_{-1}) that corresponds to the base quality (with no MGS) layer video at full temporal resolution. In SMVC, the higher temporal levels are dependent on lower temporal levels, so the trellis stages are arranged in the increasing order of temporal levels. The number of trellis stages is equal to the number of temporal levels. Each stage represents a temporal level for which the number of MGS fragments shall be determined for each view. Each node is represented by a pair, with accumulated stereo bits of all quality increments up to that stage (denoted by a_i) and the achieved stereo PSNR for the GoP (Q_{GoP}).

Figure 2 shows the example trellis diagram with a GoP size equal to 16; the number of temporal levels is 4 and the MGS layer has 3 fragments, denoted by MGS = 1, MGS = 2, and MGS = 3, respectively. Therefore, there are 4 stages in the trellis and 9 nodes for each stage.

The rate of possible nodes in a stage is given by Eq. (3).

$$\begin{gathered}
a_{-1}\left[0\right]\left[0\right] = R_{Base}^{left} + R_{Base}^{right} \\
a_i\left[j\right]\left[k\right] = a_{i-1}\left[0\right]\left[0\right] + R_{Enh}^{left}(TL=i, MGS=j) + R_{Enh}^{right}(TL=i, MGS=k), \\
0 \leq i < N \\
1 \leq j, k \leq M
\end{gathered} \tag{3}$$

Here, N and M represent the number of temporal levels and MGS layers, respectively. R_{Base} is the rate of the base quality node (MGS = 0) in the trellis diagram and R_{Enh} is the rate of enhancement quality nodes on Viterbi stages representing MGS quality increments for each temporal level. All rate values are calculated over 1 GoP.

Although the optimal Viterbi solution efficiently evaluates all possible paths at each stage of the trellis [19], a Viterbi-like suboptimum algorithm is proposed here to find a high-quality, feasible solution; it is summarized in Figure 3. The proposed algorithm selects the node that has the highest PSNR among all feasible paths that satisfy the rate constraint (2) at each stage. The Viterbi algorithm minimizes cost, while the proposed algorithm aims to maximize PSNR while matching the available network rate. The recursive cost computation of the Viterbi algorithm is, instead, performed for PSNR calculations. The proposed algorithm is suboptimal, because the number of surviving paths is limited to one by pruning the others at each stage by assuming MGS fragments of temporal levels as dependent quantizers and the monotonicity in the predictor's quantization level [20].

3.2. Implementation details, complexity, and delay

The proposed Viterbi-like algorithm determines the number of MGS fragments that will be extracted for each temporal level in each view. In 1 GoP, every temporal level may have different number of MGS fragments, but every frame in the same temporal level has the same number of MGS fragments within a view. If the rate constraint is not satisfied, the MGS value of the temporal level remains zero, i.e. a_{-1}, the initial node.

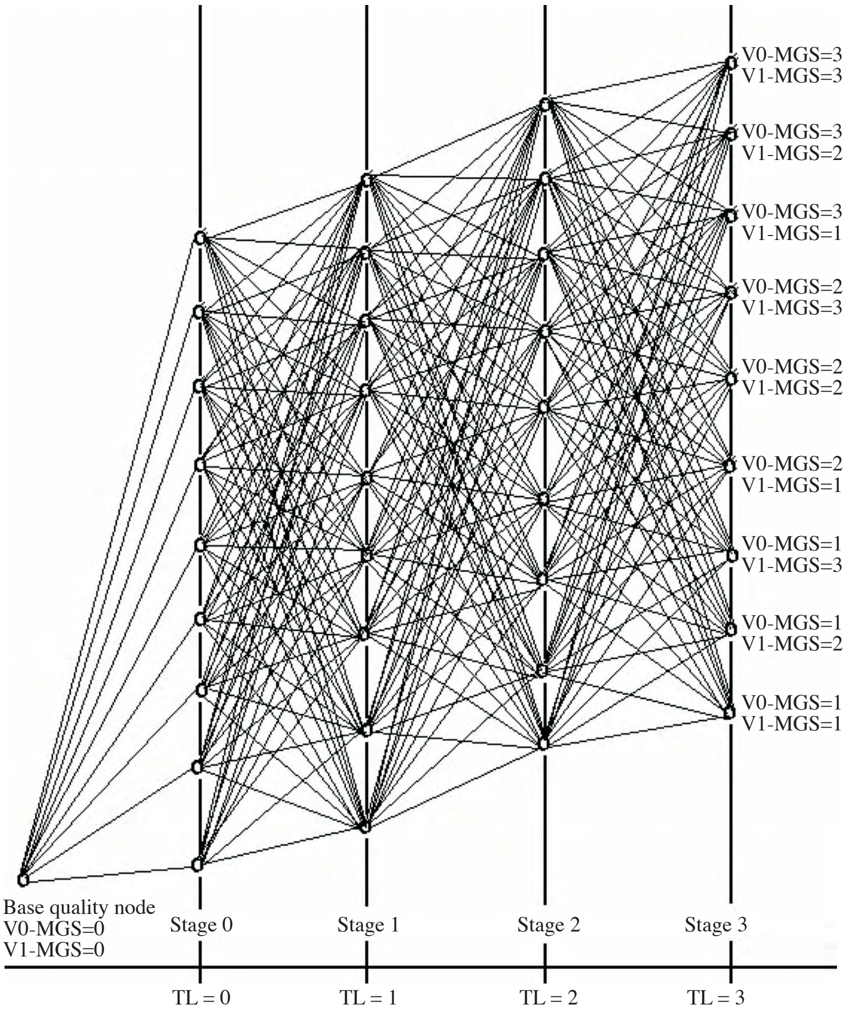

Figure 2. Trellis diagram for Viterbi-like rate adaptation for stereo GoP size = 16, number of MGS fragments = 3.

```
Starting with the base quality node
for all Viterbi stages i
  for all MGS nodes j
    evaluate the rate $a_i$ for all branches j satisfying condition (2)
    select the path giving the highest quality (1)
    move to the next stage i
```

Figure 3. The proposed Viterbi-like RDO layer extraction algorithm.

The QL concept for deriving the priority values is not suitable for real-time applications since optimization is done as postprocessing after the whole stereo video is encoded, and every frame is decoded as many times as there are quality increments. Instead, the proposed method performs a GoP-by-GoP optimization for rate adaptation. Similar to all dynamic programming algorithms with greedy search or node pruning, the computational complexity can be $O(NM)$, where N and M are related to the number of temporal levels and MGS layers and thus may be 6 and 16 at most, respectively.

In order to build the Viterbi trellis for each GoP, after 1 GoP is coded, the rate and PSNR values are saved to the memory for further calculation. Rate values are obtained for each NALU, but PSNR values are only available for base and full quality layers. For other intermediate MGS points, piecewise-linear interpolation is employed using the lowest and highest PSNR values and number of fragments.

Acceptable processing delay can be up to 400 ms for 2-way video communication [21]. If the GoP size is 8 frames, then 1 GoP delay will be acceptable since it corresponds to 260 ms for a video of 30 fps, which is less than ITU-T Rec. G.1010 [21]. However, if the GoP size is reduced to 4 frames, then the delay would be less than the more strict requirement of 150 ms.

4. Experimental results

Three stereo videos, Soccer2, Balloons, and Flowerpot, were used; they are of quarter source input format (QSIF) resolution and 30 fps. All videos are 240 frames long. Characteristics of the Soccer2 sequence can be summarized as outdoors, sports (football), complex object motion, complex camera motion, high detail, complex depth structure, and natural stadium light. Balloons: Outdoors, children running with balloons, complex object motion, partially simple camera motion, partially no camera motion, scene cut, high detail, complex depth structure, natural light. Flowerpot: Outdoors, street/square scenery, simple object motion, simple camera motion, high detail, complex depth structure, natural light.

The encoding configuration was selected as the base layer and 1 MGS layer. All extraction points were selected at 30 Hz since no temporal scaling was allowed for any of the views. The R-D plots in Figure 4 give R-D performance comparisons for the stereo test videos where the MGS layer consists of 3 fragments with 3, 3, and 10 scan positions. The x-axis corresponds to the bitrate of the extracted SMVC bitstream, namely the total bitrates of the first view (V0) and the second view (V1). The y-axis corresponds to visual quality, namely the stereo PSNR, which is calculated as the average of the V0 and V1 PSNR values. In Figures 4 and 5, the basic and QL curves were captured from a previous work [11]. Note that the average PSNR is a conventional metric, but it does not represent perceptual stereo quality well. Although there is no generic visual quality metric for stereo video, current research shows that the image quality of stereo images with a different degree of blur in the left and right views seems to be dominated by the high-quality component; if both components are degraded by blockiness, the perceived quality seems to be an average of the image quality of the left and right views [22,23]. Furthermore, blockiness and blur measures are very close to each other when V0 and V1 and QL and the Viterbi-like extractor (VT) are compared. Hence, we conclude that the average PSNR is a good approach for this study.

The Figures clearly show that the shape of the R-D curve is still concave when Viterbi-like extraction is used; the VT curves are very similar to the QL curves for the stereoscopic case, as well. For all tests, the maximum difference between the R-D curves may reach 0.1 dB in the worst case. The Viterbi-like method provides a fine rate granularity and improved quality of the reconstructed points thanks to the RDO extraction process.

The way in which stereo bitrates are allocated among the 2 views with the 3 extraction methods was further investigated. The R-D plots in Figure 5 show individual bitrate and PSNR values of V0 and V1 obtained with the corresponding extraction rate points in the stereo R-D curves. As with QLs, the introduction of the Viterbi-like method also results in the concave behavior of the R-D curves of V0 and V1.

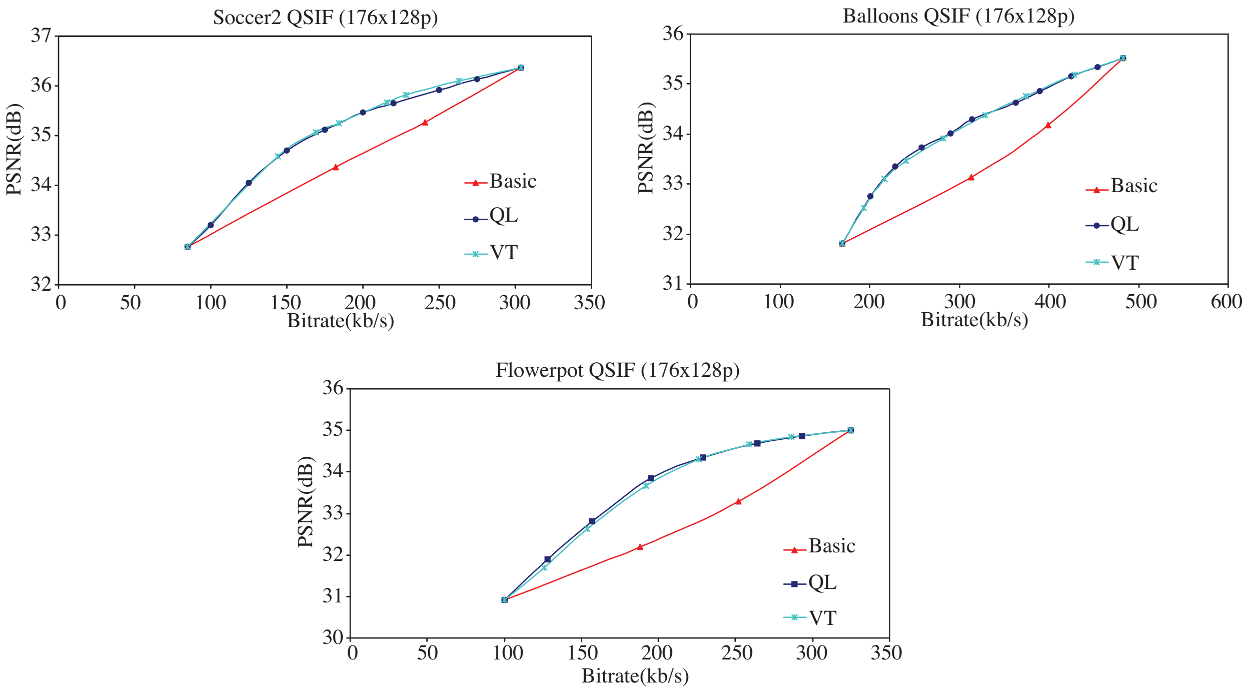

Figure 4. Stereo R-D performance comparison for QSIF scenario among JSVM-basic extractor, quality layers-based extractor (QL), and Viterbi-like extractor (VT) in MGS 3310 fragmentation.

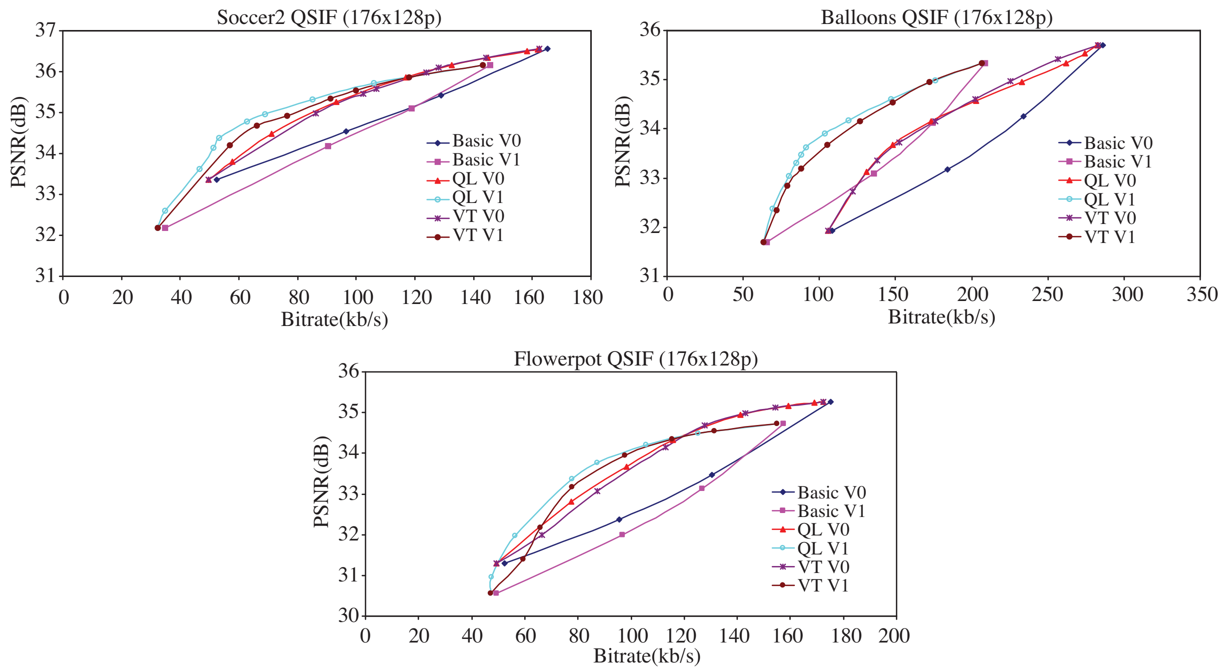

Figure 5. Individual (V0 and V1) R-D performance comparison for QSIF scenario among JSVM-basic extractor, quality layers-based extractor (QL), and Viterbi-like extractor (VT) in MGS 3310 fragmentation.

The way in which MGS packets are distributed among the temporal levels with the 2 extraction methods was also investigated. Figure 6 is depicted to analyze QL and VT extraction methods in terms of MGS layer selection for different temporal levels, where the y-axis represents the number of NALUs. Figures 6a and 6b show analysis results of the Balloons QSIF with MGS 3310 fragmentation (3 MGS fragments) for QL and VT, respectively. Within the Figures, there are a total of 7 stereo R-D points, located with odd-numbered points as V0 and even-numbered points as V1. More MGS layers are allocated for lower temporal levels with the QL method compared to the VT method. For instance, key frames get very small values of priority_id, which shows the importance of key frames. A higher MGS layer gets larger priority values for each frame. Since the dependant lists of the first-view frames also include the second-view frames, the first-view frames get higher priority. Thus, more MGS layers are extracted for the first view with the QL method compared to VT.

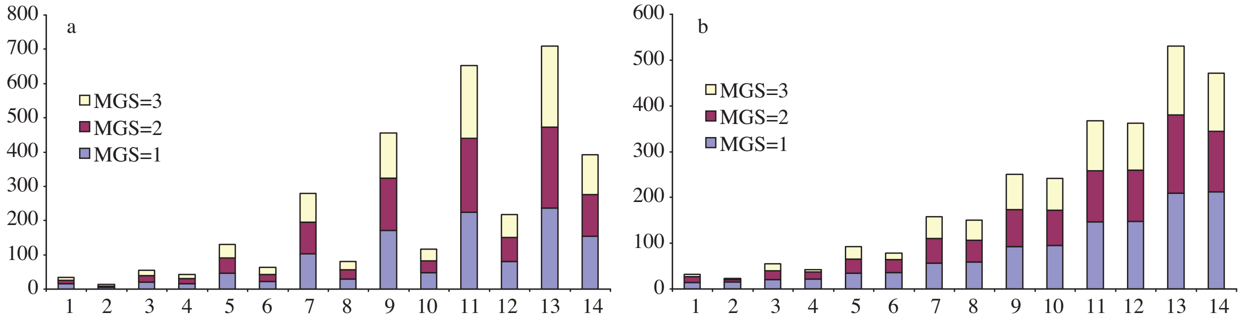

Figure 6. Analysis for MGS layer distribution in cases: a) Balloons for QL, b) Balloons for VT. Odd-numbered points on the x-axis refer to V0 R-D points while even ones refer to V1 R-D points.

Finally, a dynamic network load change scenario was simulated such that a linearly increasing and suddenly exponentially decreasing (due to possible congestion) network bandwidth was assumed. Figure 7 is given to demonstrate how well the Viterbi-like algorithm reacts to dynamic network load change for the Flowerpot sequence. When the available bandwidth (RateIn) linearly increases, the stereo PSNR also increases slowly, and after 20 GoPs, the available network bandwidth starts exponentially degrading so that the abrupt change in the stereo PSNR occurs by rate adaptation with trellis-based optimized layer extraction.

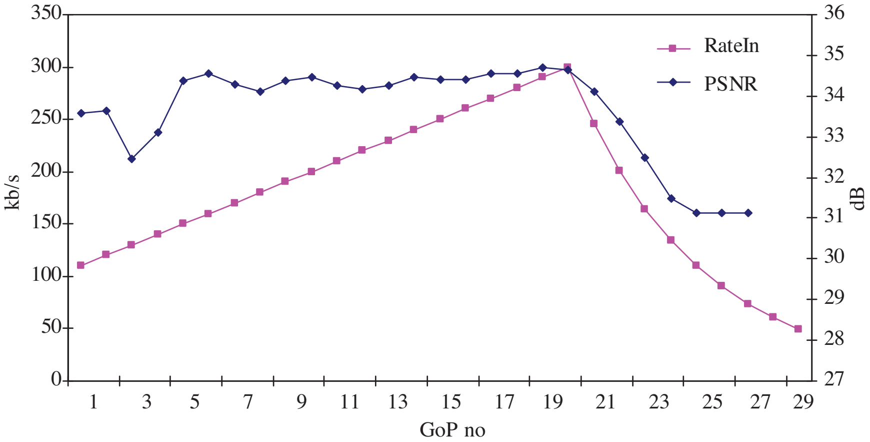

Figure 7. Flowerpot sequence under dynamic network adaptation by the proposed Viterbi-like algorithm.

5. Conclusion

A simple but high-performance method for online adaptive extraction of scalable stereo video coding was proposed. The trellis-based algorithm has low complexity, is efficient, and performs online optimization, while its performance is comparable to that of QLs, which require the whole stereo video sequence for priority determination. Thus, the proposed method can be an alternative approach for online applications. The new extraction method is suitable for use in applications that require real-time encoding where 1 GoP processing delay is acceptable (e.g., live broadcasts or videoconferencing with small GoPs), noting that reducing delay is related to GoP size versus encoding performance issue.

Acknowledgment

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK), Project Number 109E145.

References

[1] Joint Video Team of ISO-IEC MPEG & ITU-T VCEG, ISO/IEC ITU-T Rec. H.264: Advanced Video Coding for Generic Audiovisual Services, 2003.

[2] Joint Video Team of ISO-IEC MPEG & ITU-T VCEG, Text of ISO/IEC 14496-10:2005/FDAM 3 Scalable Video Coding, Lausanne, N9197, 2007.

[3] Joint Video Team of ISO-IEC MPEG & ITU-T VCEG, Text of ISO/IEC 14496-10:2008/FDAM 1 Multiview Video Coding, 2008.

[4] Joint Video Team of ISO-IEC MPEG & ITU-T VCEG, Text of ISO/IEC 14496-4:2001/PDAM 19 Reference Software for SVC, N9195, 2007.

[5] JSVM Software Manual, JSVM 9.17, 2009, available at cvs://garcon.ient.rwthi- aachen.de:/cvs./jvt.

[6] H. Kirchhoffer, D. Marpe, H. Schwarz, T. Wiegand, "A low-complexity approach for increasing the granularity of packet-based fidelity scalability in scalable video coding", Picture Coding Symposium, Lisbon, 2007.

[7] I. Amonou, N. Cammas, S. Kervadec, S. Pateux, "Optimized rate-distortion extraction with quality layers in the scalable extension of H.264/AVC", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, pp. 1186-1193, 2007.

[8] P. Merkle, A. Smolic, K. Müller, T. Wiegand, "Efficient prediction structures for multi-view video coding", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 17, pp. 1461-1472, 2007.

[9] N. Ozbek, A.M. Tekalp, "Scalable multi-view video coding for interactive 3DTV", IEEE International Conference on Multimedia & Expo, Toronto, pp. 213-216, 2006.

[10] N. Ozbek, A.M. Tekalp, "Unequal inter-view rate allocation using scalable stereo video coding and an objective stereo video quality measure", IEEE International Conference on Multimedia & Expo, Hannover, pp. 1113-1116, 2008.

[11] N. Ozbek, A.M. Tekalp, “Rate-visual-distortion optimized extraction with quality layers for scalable coding of stereo videos”, IEEE International Conference on Image Processing, San Diego, pp. 2104-2107, 2008.

[12] N. Ozbek, A.M. Tekalp, “Quality layers in scalable multi-view video coding”, IEEE International Conference on Multimedia & Expo, New York, pp. 185-188, 2009.

[13] A. Ortega, K. Ramchandran, M. Vetterli, “Optimal trellis-based buffered compression and fast approximations”, IEEE Transactions on Image Processing, Vol. 3, pp. 26-40, 1994.

[14] G.M. Schuster, A.K. Katsaggelos, “A video compression scheme with optimal bit allocation among segmentation, motion, and residual error”, IEEE Transactions on Image Processing, Vol. 6, pp. 1487-1502, 1997.

[15] M. Luttrell, J. Wen, J.D. Villasenor, “Trellis-based R-D optimal quantization in H.263+”, IEEE Transactions on Image Processing, Vol. 9, pp. 1431-1434, 2000.

[16] Z. Li, G.M. Schuster, A. Katsaggelos, B. Gandhi, “MINMAX optimal video summarization”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, pp. 1245-1256, 2005.

[17] T.C. Thang, J.W. Kang, J.J. Yoo, Y.M. Ro, “Optimal multi-layer adaptation of SVC video over heterogeneous environments”, Advances in Multimedia 2008. doi: 10.1145/1280940.1281006.

[18] M. Handley, S. Floyd, J. Padhye, J. Widmer, “TCP friendly rate control (TFRC)”, Internet Engineering Task Force RFC 3448, 2003.

[19] G.D. Forney, “The Viterbi algorithm”, Proceedings of the IEEE, Vol. 61, pp. 268-278, 1973.

[20] K. Ramchandran, A. Ortega, M. Vetterli, “Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders”, IEEE Transactions on Image Processing, Vol. 3, pp. 533-545, 1994.

[21] ITU-T, “End-user multimedia QoS categories”, Recommendation G.1010, 2001.

[22] D.V. Meegan, L.B. Stelmach, W.J. Tam, “Unequal weighting of monocular inputs in binocular combination: implications for the compression of stereoscopic imagery”, Journal of Experimental Psychology: Applied, Vol. 7, pp. 143-153, 2001.

[23] L.M.J. Meesters, W.A. IJsselstejn, P.J.H. Seuntiens, “A survey of perceptual evaluations and requirements of three-dimensional TV”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 14, pp. 381-391, 2004.