# SCALABLE MULTI-VIEW VIDEO CODING FOR INTERACTIVE 3DTV

*Nukhet Ozbek[1] and A. Murat Tekalp[2]*

[1]Ege University, International Computer Institute
Bornova, Izmir, Turkey, 35100
[2]Koc University, College of Engineering
Sariyer, Istanbul, Turkey, 34450

## ABSTRACT

A standard for scalable video coding (SVC) is currently being worked on by the ISO MPEG Group. Work on standardization of multiple-view video coding (MVC) has also recently started under the ISO MPEG. Although there are many approaches published on SVC and MVC, there is no current work reported on scalable multi-view video coding (SMVC). This paper presents new coding structures for scalable stereo and multi-view video coding. The proposed structures are implemented as extensions to the JSVM software and resulting bitrates and PSNR are demonstrated. SMVC can be used for transport of multiview video over IP for interactive 3DTV by dynamic adaptive combination of temporal, spatial, and SNR scalability according to network conditions.

## 1. INTRODUCTION

Multiple view video coding (MVC) serves emerging applications, such as interactive, free-viewpoint 3D video and TV, where we encode multiple views of the same scene with possibly high correlation between them. This inter-view redundancy can be exploited by performing disparity-compensated prediction across the views. MPEG Ad-Hoc Group for 3D Audio and Video (3DAV) is now working on the MVC standard [1], where new prediction structures as well as processing tools are being investigated for efficient multi-view video coding. Some of the proposed algorithms are reviewed in [2].

In [3], a stereoscopic video codec based on H.264 is introduced, in which the left view is predicted from other left frames, whereas the right view is predicted from all previous frames. In [4], a novel scheme is presented for coding multi-view video sequence based on global motion prediction between adjacent views, where the left-most view is compressed as a reference sequence using standard block-based motion compensated prediction coding, and the other views are compressed using global motion prediction from the reference left view. In [5], the relationship between the coding efficiency, frame rate and the camera distance is discussed. A multi-view codec based on MPEG-2 is proposed for view scalability in [6].

In [7], the concept of GoGOP (a group of GOP) is introduced for low-delay random access, where all GOPs are categorized into two kinds: base GOP and inter GOP. A picture in a base GOP may use decoded pictures only in the current GOP. A picture in an inter GOP, however, may use decoded pictures in other GOPs as well as in the current GOP. In [8], a Multi-View Video Codec based on H.264 has been proposed using disparity and motion estimation/ compensation. The buffering structure of H.264 is modified and several referencing modes are implemented. Results show that the new codec outperforms simulcast H.264 coding for closely located cameras.

The scalable extension of H.264/AVC is selected as the starting point of the SVC work [9]. It specifies temporal scalability by means of a lifting framework on motion-compensated temporal filtering (MCTF). For spatial scalability, a combination of motion-compensated prediction and over-sampled pyramid decomposition is proposed. SNR scalability is achieved by residual quantization with little modification to H.264/AVC. In [10], combined scalability support of the scalable extension of H.264/AVC is examined. For any spatio-temporal resolution, the corresponding spatial base layer representation must be transmitted at the minimum bitrate. Above this, any bitrate can be extracted by truncating the FGS NAL units of the corresponding spatio-temporal layer and lower resolution layers in a suitable way.

It is well-known that for appropriate 3D perception from stereo video, the right and left views need not be encoded with full temporal, spatial, and SNR resolutions. This can be used to benefit in effective transport of multiple view video, where one of the views is sent with full resolution, whereas the spatial, temporal and/or SNR resolution of other view(s) can be dynamically adapted according to video content and network conditions. With scalable coding of multi-view video, the encoding can be done once and off-line. In a point-to-point transmission scenario, bitstreams at various spatial, temporal and SNR resolutions can be extracted dynamically on demand. Alternatively, transport of interactive (free-view) 3DTV over IP can be achieved by receiver-driven multicast, where the receiver can subscribe to receive each view at some desired temporal, spatial and/or SNR resolution.

This paper presents novel coding structures for scalable stereo and multi-view video coding. The proposed structures are implemented as extensions to the JSVM software and resulting bitrates and PSNR are demonstrated. Section 2 describes the stereoscopic (N=2) SVC implementation. Section 3 explains the proposed new structure and extensions to the JSVM for multi-view (N>2) scalable coding. Section 4 provides experimental results. Conclusions are drawn in Section 5.

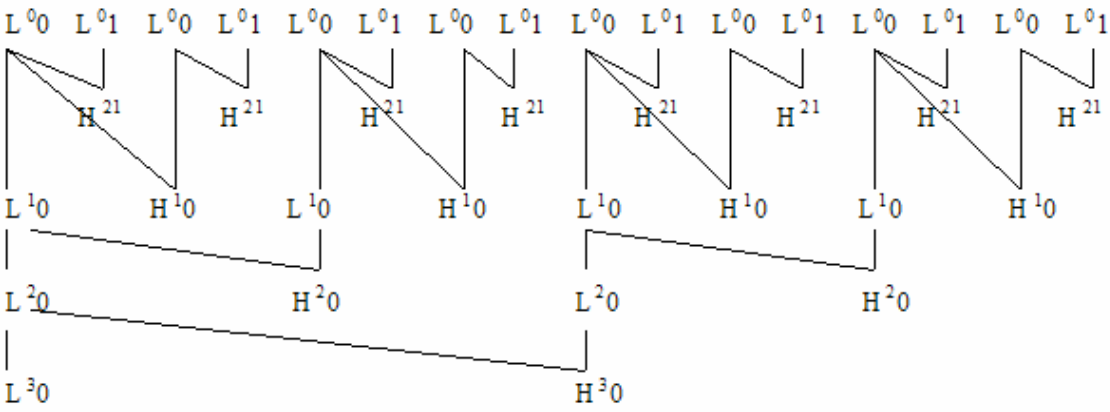


Fig. 1 Prediction structure for stereoscopic scalable video coding for N=2 and GOP=16.

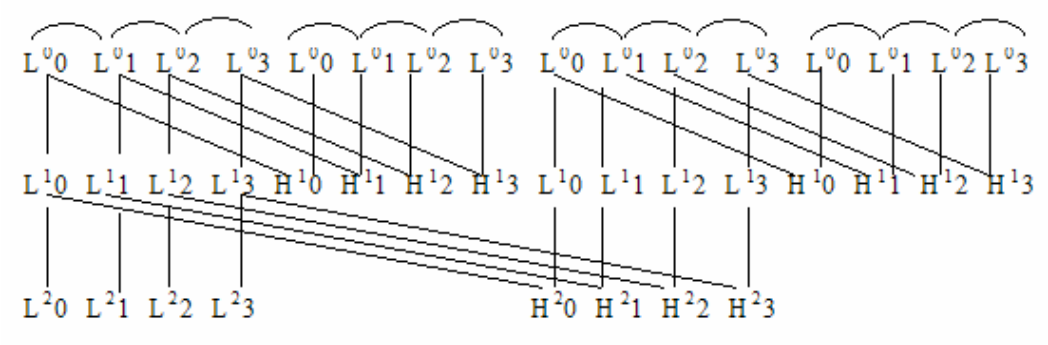


Fig. 2 Prediction structure for multi-view scalable video coding for N=4 and GOP=16.

## 2. SCALABLE STEREO VIDEO CODING

The JSVM reference software [11] naturally supports scalable coding of stereo video by sequential interleaving of right and left views using the present SVC MCTF structure without update steps. It has already been shown that the coding efficiency with hierarchical B pictures (MCTF without updates) and a closed-loop control is higher or similar to that of the MCTF-based coding [12].

The prediction structure for multi-view stereo coding is illustrated in Fig. 1 for GOP=16, where the right view is predicted temporally from itself and the left view is predicted from the right view. In this figure, the difference frames denoted by $H^{21}$ are temporally independent frames and correspond to the left view. Hence, the left view can be served at half temporal resolution by discarding odd or even numbered H frames.

The bit stream extractor and decoder modules need to be modified in order to recover the last temporal layer as the left view. Since we have two views, the effective GOP size reduces to half the original GOP size shown in Fig. 1, where odd numbered Level 0 frames at the decomposition stage correspond to the left view, and $L^2$ frames become $L^1$ frames of the right view in our notation. Further, $H^2$ frames in the implementation correspond to $H^1$ frames of the right view and so on.

Temporal scalability is fully supported for the right view with all possible layers in the corresponding GOP size whereas it is not supported by the current syntax for the left view. However, spatial and SNR scalability functionalities remain unchanged with the proposed structure. The only limitation of the proposed method could be that temporal prediction is not allowed within the left view. We note that this limitation is directly related to the current version of the reference codec software [11], and should be overcome in the future. This may be a disadvantage in terms of compression efficiency for multi-view video coding, when there is large disparity between different views.

## 3. SCALABLE MULTI-VIEW VIDEO CODING

When the number of views (*N*) is larger than 2, a new MCTF structure needs to be designed. To this effect, we propose a novel prediction structure, which uses more than two L/H frames as input to produce difference (H) frames. This is illustrated in Fig. 2 for N=4 and GOP=16, where only the first view is independently temporally predicted, and includes both H and L frames. However, other three views include only H frames, and they each depend on the previous view in addition to temporal prediction. Only temporal correlation is utilized for the first view whereas temporal plus inter-view correlations are employed for all other views. Multi-view scalability can be possible with this scheme for all views.

The implementation is performed by modifying the current JSVM software with the GOP size is 16. To that end, instead of just one key frame, N key frames, namely

16, 17, 18, 19 for N=4, are encoded to incorporate the other views but these frames are encoded as high pass and predicted from the previous view. Further, these key frames are copied to the first N frames of the next GOP structure after reconstruction. Every frame in V1, V2 and V3 uses past and future frames from its own view and the same frame from the previous view for prediction. In every view, only first frame of the GOP use just inter-view prediction so that subscribing to receive any view at some desired temporal resolution can be possible.

## 4. RESULTS

We use five sequences, xmas, race2, flamenco2, race1 and ballroom, in our experiments. All sequences are 320x240 in size and 30 fps. For the Xmas sequence, the distance between the cameras is 60 mm for stereoscopic video coding while it is 30 mm for multi-view video coding. For Race2 sequence, the camera distance is larger (20 cm). Other three sequences also have large disparities.

In Tables 1, 2, 3, and 4, results for N=2 and N=4 cases are stated for GOP=8. Bitrates and PSNR values for each temporal and FGS layers are given. We also report the ratio (R) of the sum of left and right bitrates to that of the reference (the right) view. This ratio, which is near 2 for simulcast, indicates the extra cost of the left view in reference to that of the right view. For GOP size 16, the results are shown in Tables 5 and 6. In those tables, the right view has several temporal resolutions which vary from 3.75 Hz to 30 Hz, whereas the left view has only 30 Hz resolution.

Among the results, the Xmas sequence gives the best performance with the smallest Ratio values for both stereo and multi-view coding cases. It is due to the smaller camera distance and thus lower disparity.

In Tables 7 to 11, the proposed multiview (N>2) MCTF results are presented for all test sequences. Each view has 3 temporal resolutions, but only 15 and 30 Hz values are given. In those tables SMVC results are compared to simulcast single view (SV) results. For a fair comparison, SV results are taken under the same conditions using the original JSVM software except GOP size is set to 4 in SV case. A fixed quantization parameter (QP), which is 28, is used and 48 frames per view are totally encoded during these tests.

For the flamenco2 sequence, the results show that bitrate decrease at V1, which is predicted from V0, is not as high as the one at V2, which is predicted from V1. Also, though the SV bitrates are the same for V2 and V3, MV bitrate of V3 is higher than the one of V2. These facts are closely related to similarity between the adjacent views. Especially for this sequence the illumination (spotlights, shadows, etc.) varies largely over the multi-view images due to the lighting conditions.

**Table 1: Xmas sequence N=2, GOP=8 results.**

| Substream | QL = 0 Bitrate | R | QL = 1 Bitrate | R | psnr Y (dB) |
|---|---|---|---|---|---|
| Right 7.5 Hz | 284 | 1.3 | 644 | 1.3 | 39.56 |
| Right 15 Hz | 378 | 1.2 | 746 | 1.3 | 38.34 |
| Right 30 Hz | 477 | 1.2 | 877 | 1.2 | 37.38 |
| Left 30 Hz | 78 | - | 216 | - | 35.23 |

**Table 2: Race2 sequence N=2, GOP=8 results.**

| Substream | QL = 0 Bitrate | R | QL = 1 Bitrate | R | psnr Y (dB) |
|---|---|---|---|---|---|
| Right 7.5 Hz | 168 | 2.3 | 476 | 2.0 | 39.06 |
| Right 15 Hz | 198 | 2.1 | 534 | 1.9 | 37.87 |
| Right 30 Hz | 229 | 1.9 | 602 | 1.8 | 37.00 |
| Left 30 Hz | 221 | - | 488 | - | 34.86 |

**Table 3: Xmas sequence N=4, GOP=8 results**.

| Substream | QL = 0 bitrate | R | QL = 1 bitrate | R | psnr Y (dB) |
|---|---|---|---|---|---|
| V0 7.5 Hz | 284 | 1.2 | 644 | 1.2 | 39.56 |
| V0 15 Hz | 377 | 1.1 | 745 | 1.1 | 38.34 |
| V0 30 Hz | 477 | 1.1 | 876 | 1.1 | 37.38 |
| V1 30 Hz | 43 | - | 101 | - | 35.78 |
| V2 7.5 Hz | 283 | 1.2 | 644 | 1.2 | 39.58 |
| V2 15 Hz | 375 | 1.1 | 744 | 1.2 | 38.35 |
| V2 30 Hz | 476 | 1.1 | 873 | 1.1 | 37.42 |
| V3 30 Hz | 43 | - | 110 | - | 35.81 |

**Table 4: Race2 sequence N=4, GOP=8 results.**

| Substream | QL = 0 bitrate | R | QL = 1 bitrate | R | psnr Y (dB) |
|---|---|---|---|---|---|
| V0 7.5 Hz | 168 | 2.3 | 476 | 2.0 | 39.06 |
| V0 15 Hz | 198 | 2.1 | 534 | 1.9 | 37.87 |
| V0 30 Hz | 229 | 1.9 | 602 | 1.8 | 37.00 |
| V1 30 Hz | 221 | - | 488 | - | 34.86 |
| V2 7.5 Hz | 164 | 2.3 | 454 | 2.0 | 39.02 |
| V2 15 Hz | 196 | 2.1 | 507 | 1.9 | 37.90 |
| V2 30 Hz | 232 | 1.9 | 576 | 1.8 | 37.09 |
| V3 30 Hz | 217 | - | 495 | - | 34.84 |

**Table 5: Xmas sequence N=2, GOP=16 results.**

| Substream | QL = 0 Bitrate | R | QL = 1 Bitrate | R | psnr Y (dB) |
|---|---|---|---|---|---|
| Right 3.75 | 203 | 1.4 | 416 | 1.5 | 42.71 |
| Right 7.5 Hz | 277 | 1.3 | 486 | 1.4 | 41.61 |
| Right 15 Hz | 374 | 1.2 | 590 | 1.4 | 40.84 |
| Right 30 Hz | 480 | 1.2 | 735 | 1.3 | 40.25 |
| Left 30 Hz | 77 | - | 215 | - | 39.74 |

**Table 6: Race2 sequence N=2, GOP=16 results.**

| Substream | QL = 0 Bitrate | R | QL = 1 Bitrate | R | psnr Y (dB) |
|---|---|---|---|---|---|
| Right 3.75 | 130 | 2.6 | 326 | 2.5 | 40.37 |
| Right 7.5 Hz | 162 | 2.3 | 376 | 2.3 | 38.78 |
| Right 15 Hz | 198 | 2.1 | 440 | 2.1 | 37.75 |
| Right 30 Hz | 230 | 1.9 | 516 | 1.9 | 36.97 |
| Left 30 Hz | 220 | - | 488 | - | 34.88 |

**Table 7: Xmas sequence N=4, GOP=16 results.**

| Substream | SV bitrate | SV psnr Y | MV bitrate | MV psnr Y |
|---|---|---|---|---|
| V0 15 Hz | - | - | 567.34 | 37.11 |
| V0 30 Hz | 869.65 | 36.95 | 869.59 | 36.95 |
| V1 15 Hz | - | - | 116.22 | 36.60 |
| V1 30 Hz | 863.23 | 36.98 | 224.58 | 36.60 |
| V2 15 Hz | - | - | 159.35 | 36.49 |
| V2 30 Hz | 861.34 | 37.00 | 286.64 | 36.52 |
| V3 15 Hz | - | - | 171.32 | 36.42 |
| V3 30 Hz | 860.05 | 37.00 | 306.91 | 36.47 |

**Table 8: Race2 sequence N=4, GOP=16 results.**

| Substream | SV bitrate | SV psnr Y | MV bitrate | MV psnr Y |
|---|---|---|---|---|
| V0 15 Hz | - | - | 346.75 | 36.82 |
| V0 30 Hz | 495.54 | 36.65 | 495.96 | 36.65 |
| V1 15 Hz | - | - | 324.31 | 36.73 |
| V1 30 Hz | 475.90 | 37.00 | 500.30 | 36.75 |
| V2 15 Hz | - | - | 249.21 | 36.44 |
| V2 30 Hz | 451.00 | 36.80 | 413.31 | 36.48 |
| V3 15 Hz | - | - | 290.70 | 36.30 |
| V3 30 Hz | 462.60 | 36.70 | 448.39 | 36.34 |

**Table 9: Flamenco2 sequence N=4, GOP=16 results.**

| Substream | SV bitrate | SV psnr Y | MV bitrate | MV psnr Y |
|---|---|---|---|---|
| V0 15 Hz | - | - | 542.21 | 38.33 |
| V0 30 Hz | 772.00 | 38.17 | 768.30 | 38.14 |
| V1 15 Hz | - | - | 487.80 | 37.87 |
| V1 30 Hz | 719.00 | 38.37 | 709.00 | 38.02 |
| V2 15 Hz | - | - | 338.13 | 37.12 |
| V2 30 Hz | 836.26 | 37.75 | 566.83 | 37.20 |
| V3 15 Hz | - | - | 393.75 | 37.43 |
| V3 30 Hz | 837.70 | 37.96 | 627.60 | 37.50 |

**Table 10: Race1 sequence N=4, GOP=16 results.**

| Substream | SV bitrate | SV psnr Y | MV bitrate | MV psnr Y |
|---|---|---|---|---|
| V0 15 Hz | - | - | 617.84 | 36.82 |
| V0 30 Hz | 870.30 | 36.71 | 865.29 | 36.68 |
| V1 15 Hz | - | - | 423.01 | 36.23 |
| V1 30 Hz | 867.42 | 36.71 | 682.78 | 36.30 |
| V2 15 Hz | - | - | 389.84 | 36.26 |
| V2 30 Hz | 859.32 | 36.76 | 641.58 | 36.31 |
| V3 15 Hz | - | - | 417.56 | 36.37 |
| V3 30 Hz | 832.33 | 36.87 | 656.89 | 36.47 |

**Table 11: Ballroom sequence N=4, GOP=16 results.**

| Substream | SV bitrate | SV psnr Y | MV bitrate | MV psnr Y |
|---|---|---|---|---|
| V0 15 Hz | - | - | 409.22 | 35.97 |
| V0 30 Hz | 587.38 | 35.91 | 582.18 | 35.88 |
| V1 15 Hz | - | - | 451.30 | 35.14 |
| V1 30 Hz | 614.78 | 35.72 | 648.50 | 35.24 |
| V2 15 Hz | - | - | 377.76 | 35.33 |
| V2 30 Hz | 612.61 | 35.95 | 571.76 | 35.45 |
| V3 15 Hz | - | - | 409.43 | 34.84 |
| V3 30 Hz | 656.55 | 35.72 | 610.64 | 34.94 |

## 5. CONCLUSIONS

In this study, we propose modifications to the SVC scalable codec for scalable stereo and multi-view video coding. We observe that the current SVC structure inherently supports scalable stereo coding as another temporal level, when update steps are removed, by interleaving left and right view frames at the input.

For scalable multi-view video coding, N>2, we propose a new prediction (MCTF) structure which takes all N views within a single GOP, and supports adaptive temporal or disparity compensated prediction. We report bitrates and PSNR that are superior to simulcast SVC coding of the multiple views.

For the future work, subjective evaluation of different views at different resolutions will be performed. QP selection criteria for the other (nonreference) views in SMVC is still a research issue.

## 6. REFERENCES


[1] Aljoscha Smolic and Peter Kauff, "Interactive 3-D Video Representation and Coding Technologies," *Proceedings of the IEEE*, Vol. 93, No. 1, Jan. 2005.

[2] A. Vetro, W. Matusik, H. Pfister, J. Xin, "Coding Approaches for End-to-End 3D TV Systems", Picture Coding Symposium (PCS), December 2004.

[3] B. Balasubramaniyam, E. Edirisinghe, H. Bez, "An Extended H.264 CODEC for Stereoscopic Video Coding", Proceedings of SPIE, 2004.

[4] X. Guo, Q. Huang, "Multiview Video Coding Based on Global Motion Model", PCM 2004, *LNCS 3333*, pp. 665-672, 2004.

[5] U. Fecker and A. Kaup, "H.264/AVC-Compatible Coding of Dynamic Light Fields Using Transposed Picture Ordering", *EUSIPCO 2005*, Antalya, Turkey, Sept. 2005.

[6] J.E. Lim, K.N. Ngan, W. Yang, and K. Sohn, "A Multiview Sequence CODEC with View Scalability", *Signal Processing: Image Communications*, Vol. 19, No. 3, pp. 239-365, 2004.

[7] H. Kimata, M. Kitahara, K. Kamikura, and Y. Yashima, "Free-viewpoint Video Communication Using Multi-view Video Coding", *NTT Technical Review*, Aug. 2004.

[8] C. Bilen, A. Aksay, G. Bozdagı Akar, "A Multi-View Codec Based on H.264", *IEEE International Conference on Image Processing (ICIP)*, 2006 (submitted).

[9] J. Reichel, H. Schwarz, M. Wien (eds.), "Scalable Video Coding – Working Draft 1," Joint Video Team (JVT), *Doc. JVT-N020*, Hong-Kong, China, Jan. 2005.

[10] H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, "Combined Scalability Support for the Scalable Extension of H.264/AVC", *IEEE International Conference on Multimedia & Expo (ICME)*, Amsterdam, The Netherlands, July 2005.

[11] J. Reichel, H. Schwarz, M. Wien, "Joint Scalable Video Model JSVM-4", *Doc. JVT-Q202*, Oct. 2005.

[12] H. Schwarz, D. Marpe, and T. Wiegand, "Comparison of MCTF and closed-loop hierarchical B pictures", *Doc. JVT-P059*, July 2005