Scalable Stereo Video Coding for Heterogeneous Environments[†]

Sehchan Oh, Youngho Lee, and Woontack Woo

GIST U-VR Lab.
Gwangju 500-712, South Korea
{soh, ylee, wwoo}@gist.ac.kr

Abstract. In this paper, we propose a new stereo video coding scheme for heterogeneous consumer devices by exploiting the concept of spatio-temporal scalability. We use MPEG standard for coding the main sequence and interpolative prediction scheme for predicting the P- and B-type pictures of the auxiliary sequence. The interpolative scheme predicts matching blocks by interpolating both motion predicted macro-block and disparity predicted macro-block and employs weighting factors to minimize the residual errors. To provide flexible stereo video service, we define both a temporally scalable layer and a spatially scalable layer for each eye's view. The experimental results show the efficiency of proposed scheme by comparison with already known methods and advantages of disparity estimation in the view of scalability overhead. According to the experimental results, we expect the proposed functionalities will play a key role in establishing highly flexible stereo video service for ubiquitous display environment where device and network connections are heterogeneous.

1 Introduction

Recent advancements in Internet and multimedia services have enabled immersive display, such as stereo and panoramic video display. Stereo video enables user feel more natural and immersed, but bandwidth requirement for stereoscopic transmission is twice that for conventional monocular transmission. The objective on a bandwidth-limited transmission system is to develop an efficient coding scheme that exploits the redundancies of the stereo image sequences.

A typical compression scenario of stereo video is exploiting both the effective motion compensation of individual image sequences and the reduction of disparity between the reference and target frames [1][2]. The reference (or left-view) frame is encoded by using conventional compression standard. However, the target (or right-view) frame is encoded by using disparity vector (DV) estimated from the reference frame and displacement compensated difference (DCD) instead of encoding the target frame itself. Therefore, estimating and representing DV and motion vector (MV) accurately are main objectives in a compression technique for stereo video.

[†] This work was supported by Korea Research Foundation Grant (KRF-2002-003-D00221).

V. Roca and F. Rousseau (Eds.): MIPS 2004, LNCS 3311, pp. 72–83, 2004.

Moreover, the availability of multimedia service, such as stereo video service, strongly depends on the network infrastructure and display environments of clients. For example, high quality video services require high-resolution display and broadband network. As shown in Fig. 1, there exist various types of consumer devices such as TV, LCD/CRT monitor, PDA, HMD, etc. Therefore, a new coding algorithm is needed to represent and deliver stereo video according to available network and display devices.

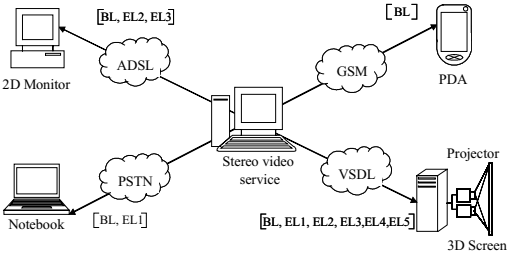

Fig. 1. Heterogeneous network and display systems with stereo video service

There are many research activities about stereo video compression using block matching algorithm (BMA) or its alternative implementations. The most general implementation method for motion/disparity estimation is BMA which is used in motion estimation of MPEG standard [3-5]. Recently, several stereoscopic video compression schemes have been developed by using multi-resolution based BMA in order to reduce a searching complexity of vectors and hierarchical disparity estimation which uses variable block size instead of using fixed block size, to raise accuracy of disparity estimation [6][7]. Most of the compression techniques of stereo video mainly focused on the developing the modified coding algorithm using the MPEG-2 multi-view profile [4][5]. These schemes modified various types of scalability to be suitable for stereo video and concentrated on improving a compression ratio using the similarity between the two views. Few research activities on coding scheme for providing flexible stereo video service have been reported [8]. However, they have a limitation of flexibility because of combining stereo coding scheme with individual scalability technique.

In this paper, we propose a highly scalable yet efficient stereo video coding method for various heterogeneous devices by exploiting the concept of spatio-temporal scalability. The proposed scheme uses MPEG-2 standard for encoding the left image and an interpolation based motion-disparity prediction scheme for predicting macro block (MB) of P- and B-type pictures of the right image sequence. Each MB of every target B-picture in auxiliary sequence is predicted by interpolating both bi-directional (forward and backward) motion predicted MB from I- and P-type pictures and disparity predicted MB from corresponding reference picture. Similarly, each MB of every target P-picture in auxiliary sequence is predicted by forward motion predicted MB and disparity predicted MB. In this scheme, we apply same weighting factor to each motion predicted MB and disparity predicted MB for estimating the best matching blocks.

To provide efficient stereo video service among heterogeneous clients, the proposed scheme uses the functionalities of spatio-temporal scalability [3][9][10]. The encoder produces one base layer (BL) bit-stream and several enhancement layer (EL) bit-streams. The BL bit-stream represents lower resolution of main sequences. The EL bit-streams provide frames of auxiliary-view as well as additional information for reproduction of the lower resolution frame with high spatial and temporal resolution. In general, spatial scalability offers flexibility of spatial resolution, but on the other hand, there is bit-rate overhead due to scalability. In stereo video coding, the spatial scalability overhead can be decreased by reducing redundancies between stereo pair.

The rest of this paper is organized as follows. In Chapter 2, we present system configuration and the proposed stereo video coder with spatio-temporal scalability. We evaluate the proposed coding scheme and analyze the experimental results in Chapter 3. Finally, some concluding remarks and possible extension of the proposed scheme are mentioned in Chapter 4.

2 Stereo Video Coding with Spatio-temporal Scalability

To deliver stereo video efficiently among heterogeneous display systems, we define one BL and several ELs as shown in Fig. 2. The BL and EL3 represent the lower resolution of left and right sequences respectively. The BL encoder performs motion compensation prediction (MCP) based encoding to remove unnecessary data, which is temporal redundancy, between current frame and previous decoded frame. The EL3 encoder performs disparity compensation prediction (DCP) based encoding as well as MCP based encoding. More specifically, it employs two types of prediction, one referencing a decoded left-view frame and the other referencing decoded right-view frames. The EL1 and EL4 generate additional data for providing full temporal resolution. The EL2 and EL5 encoder perform Intra coding to encode additional data, needed for providing full spatial resolution, without any prediction.

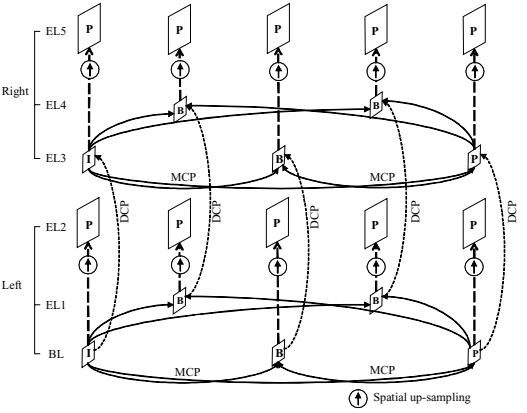

Fig. 2. Prediction for motion and disparity compensation

The proposed scalable stereo video coder is an extended version of compatible stereoscopic coding methods and can be divided into (1) stereo video coder, (2) spatial scalability coder, and (3) temporal scalability coder.

2.1 Stereo Video Coder

The structure of proposed stereo video coder is shown in Fig. 3(a). The reference frames are encoded by MPEG-2 standard. To encode the target frames, stereo encoder calculates DV with respect to temporally coincident left-view frame and estimates MV from the temporally closed right-view frames. The stereo encoder determines whether DV or MV provides higher compression efficiency, and finally, encodes both selected vector and residual information.

Disparity is the vectorial distance between the two points of a superposed stereo pair that correspond to the same point in the 3D scene. The estimation of disparity is indispensable for the prediction of the target image from the reference image. Then, the DCD is evaluated as follows:

$$DCD(x, y, DV) = I_R[x][y] - I_L[x + DV][y] \tag{1}$$

where I_R and I_L are pixel intensity value of right and left frames. The DV is defined as:

$$DV(x, y) = \arg\min_{dv \in S} \left| DCD(x, y, DV) \right| \tag{2}$$

where S means size of window searching area. As described in the expression (3), to find DV in real image, we calculate SAD for every macro block in the image and find a best matching-block which has minimum SAD. The DV is defined as coordinate of the point (x,y) of matching-block. However, y is around zero according to the general characteristic of DV.

Characteristics of the proposed coding scheme are as follows:

- *The reference or left video sequence are coded by non-scalable MPEG-2 video encoder.*
- *I-picture in the right-view sequence is coded by using a frame which results in disparity compensated prediction.*
- *P-picture of the right-view is coded by using the predicted frame which results in one of forward, disparity and bi-directionally (forward and backward) interpolated predictions. More specifically, the stereo encoder selects one prediction having minimum SSD to determine predicted frame.*
- *As shown in Fig. 4, B-picture is coded by using the predicted frame which results in one of forward, backward, disparity, and tri-directionally (forward, backward and disparity) interpolated predictions. Indeed, three reference frames used for prediction are the left-view frame coincidental with the right-view frame to be predicted, and the previous and next right-view frames in display order. The encoder selects one prediction according to its SSD.*

In case of P-picture, the interpolated prediction is generated by a weighted combination of a motion predicted frame and disparity predicted frame. The encoder uses interpolated prediction when a predicted macroblock with interpolated vector has minimum SSD. A macroblock predicted by the interpolated scheme is described as:

$$P_{pred}(v_f, v_d) = W_f R_{rec\ f}(v_f) + W_d R_{rec\ d}(v_d) \tag{3}$$

where P, v, W represents predicted macroblock of P-picture, vector, and weighting factor respectively, and R_{rec} is a reconstructed macroblock of I- or P-picture. Each indexes, f and d denote forward, disparity respectively, and each weighting factors, W_f, W_d are given as 0.5 and 0.5.

In case of B-picture, the predicted macroblock is specified as:

$$B_{pred}(v_f, v_b, v_d) = W_f R_{rec\,f}(v_f) + W_b R_{rec\,b}(v_b) + W_d R_{rec\,d}(v_d) \tag{4}$$

where B, v, W represents predicted macroblock of B-picture, vector, and weighting factor respectively. Each indexes, f, b, d denotes forward, backward, disparity respectively. In this case, the weighting factors, W_f, W_b, W_d are given as 0.25, 0.25, and 0.5.

Fig. 3(b) shows the stereo video decoder. The right-view frame is decoded with respect to the decoded left-view frame, coincidental with the right-view frame.

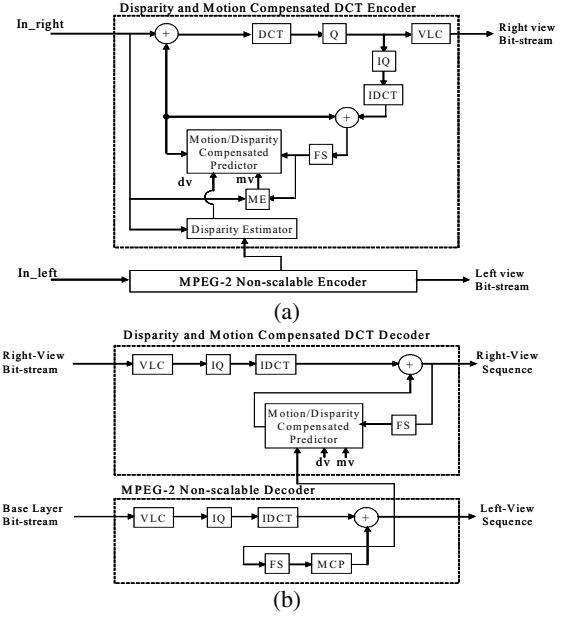

Fig. 3. Stereo video coder (a) Encoder (b) Decoder

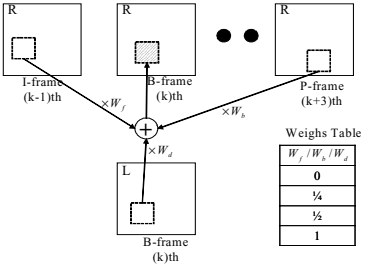

Fig. 4. Predictions in B-picture of right sequence

2.2 Spatial Scalability Coder

In addition to the bit-stream for lower spatial resolution image, spatial scalability EL encoder generates additional bit-stream for providing full resolution image as shown in Fig. 5(a). Stereo sequence is fed to the spatial scalability BL encoder after spatial downsampling. A locally decoded BL frame is spatially upsampled to the same sampling grid as the spatial scalability EL, and then subjected to EL encoder. The spatial scalability EL encoder encodes the difference between the decoded image from BL and original image at full resolution. Since the residual image loses the characteristics of natural image, it is Intra coded without MCP. In general, residual image has Laplacian distribution, which is highly peaked around zero, meaning that it can be compressed more easily than the original image. Therefore, a block having variance less than specific threshold value is skipped without encoding process in order to allot more bits to a block having large variance.

The decoding process of spatial scalability is the reverse of the encoding process as described in Fig. 5(b). To reconstruct the EL bit-stream, the bit-stream of temporally coincident frame in spatial scalability BL should be decoded first. The decoded frame in BL can be directly displayed in the client equipped with lower resolution display system. For example, when a client isn't equipped with 3D display and can provide only lower resolution display, it is efficient to provide directly lower resolution video decoded from spatial scalability BL decoder. However, to provide with full resolution video to a client equipped with high-resolution display system, it is resampled to full resolution and combines with the result from the spatial scalability EL.

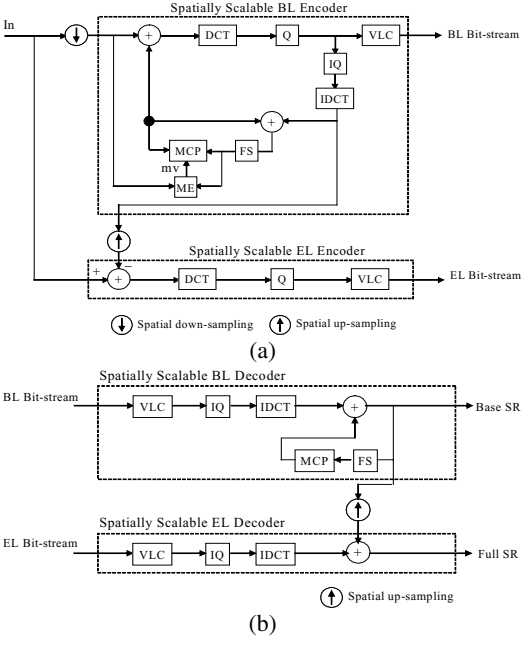

Fig. 5. Spatial scalability coder (a) Encoder (b) Decoder

2.3 Temporal Scalability Coder

Temporal scalability allows a video sequence to display as different temporal resolutions or frame rates according to the type of display and available channel capacity. In general, for temporal scalability providing a full and a half frame rate, an odd number of B-pictures is necessary [11]. In proposed temporal scalability encoder, the temporal demux (demultiplexer) splits up B-pictures, which are not used as reference frame, into temporal scalability BL and EL. For motion compensation, a BL frame is predicted only from the previous BL frames, whereas an EL frame can be predicted from both BL and EL frames. The clients can selectively receive the encoded frames of BL and ELs according to their display types and network connections. The decoded frames at the output of the temporal scalability BL can be shown by themselves at half frame rate of input video or can be temporally multiplexed in the temporal remux (remultiplexer) with the output of the EL decoder to provide full frame rate, the same as that of the input video.

2.4 Stereo Video Delivery

The proposed stereo video coding scheme aims at providing multicast-based stereo video streaming service. If a server provides stereo video service to N number of clients, unicast method requires each client to have its own video stream, separate from the others. Therefore, multiple end-to-end unicast cannot use network bandwidth efficiently because the network channel carries redundant portions of the same video stream. However, multicast-based method can remove these redundancies because the intermediate node copies the received bit-streams and sends selectively to its clients according to their capabilities.

Fig. 6 shows an example of stereo video streaming service. The video server captures stereo video from the stereo camera and then, encodes it. When an intermediate node requests one specific stereo video contents to server, the server sends relevant layered bit-streams. The intermediate node copies the received bit-streams and sends a portion or all the bit-streams to its clients. For example, client C in Figure 3.6 needs full resolution of mono video and accordingly the intermediate node provides BL, EL1 and EL2 bit-streams. The multicast clients can receive stereo video with various resolutions and can display stereo or mono video according to their types of display and bandwidth.

3 Experimental Results and Analysis

The experiments have been made with progressive scan 640×480, 24Hz refresh rate, and 4:2:0 chroma format test sequences. The picture structure is frame picture and the length of group of picture (GOP) is 15. A GOP structure includes 3 B-pictures between I- and P-pictures (M=3). Fig. 7 shows arbitrary left and right frames of the test sequence, *Laboratory* and *UbiHome*.

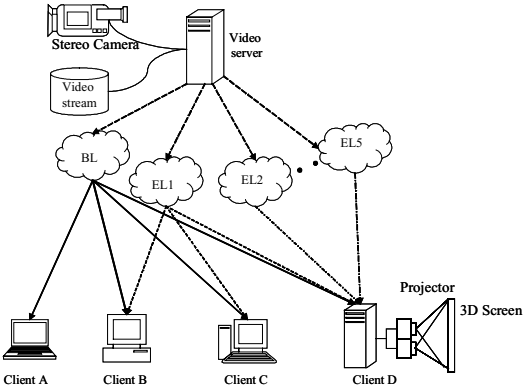

Fig. 6. An example of stereo video service using proposed stereo video coder

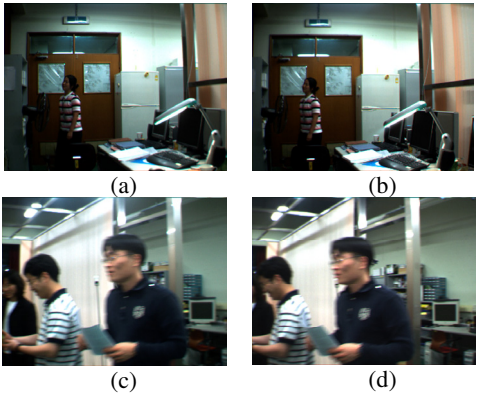

Fig. 7. Test Sequences (a) Left-view of *Laboratory* (b) Right-view of *Laboratory* (c) Left-view of *UbiHome* (d) Right-view of *UbiHome*

In Fig. 8, we compare proposed scheme with two already known methods, which are defined in MPGE-2 multi-view profile. We can easily observe that the proposed method provides a better overall performance compared with simulcast method and compatible method. In this experiment, the test sequences, *Laboratory* and *UbiHome*, are coded at 3 and 6 Mbps respectively. The bit-rates assigned to the left and right sequences are same. In this case, only spatial scalability BLs of right-view sequences are compared i.e. EL3 and EL4 layers.

In Table 1, the proposed coding scheme is compared with non-scalable MPEG-2 coder. We assigned same fixed bit-rate, 1Mpbs, to left and right sequence. For comparison of the efficiency of spatial scalability, we controlled bit-rate of the proposed scalability coder to yield the same quality. In general, the spatial scalability video coder needs more bits to provide same image quality as non-scalable coder. The additional data rate is defined as spatial scalability overhead. The proposed coding

scheme combines stereo coder with spatial scalability. Therefore, it can reduce spatial scalability overhead because the quality of right sequence is improved by DCP.

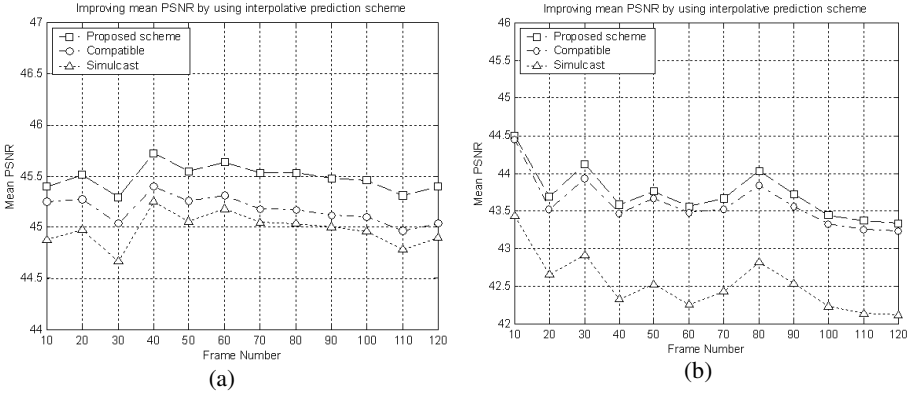

Fig. 8. Mean PSNR of right sequence with lower spatial resolution (a) *Laboratory* (b) *UbiHome*

The first results in the Table 1 show the spatial scalability overhead for the test sequence, *Laboratory,* whose motion is relatively slow. In this case, since most of macroblocks are coded with MV, the compression efficiency getting by the DCP is low. In case of *UbiHome* having fast motions and scene changes, however, most of macroblocks are affected by DCP instead of MCP. Therefore, the compression efficiency of sequence, *Laboratory*, is much improved when compared with that of simulcast coding. As a result, DCP in the proposed coding scheme can reduce the spatial scalability overhead.

Table 1. The spatial scalability overhead for test sequence *Laboratory*

Sequence		*Laboratory*		*UbiHome*	
		Left	Right	Left	Right
Single Layer Encoder	Bit-rate [Mbps]	1.0	1.0	1.0	1.0
	Average PSNR for luminance [dB]	41.03	40.25	36.50	36.33
	Mean PSNR [dB]	40.62		36.42	
Proposed Scalable Stereo Coder	Bit-rate [Mbps]	1.14	1.14	1.045	1.045
	Average PSNR for luminance [dB]	40.81	40.43	36.83	36.07
	Base layer bit-rate [Mb]	0.49	0.49	0.41	0.41
	Base layer bits-rate as percent of total bit-rate [%]	43	43	39.2	39.2
	Spatial scalability overhead [%d]	14		4.5	
	Mean PSNR [dB]	40.61		36.44	

Fig. 9 shows degree of disparity compensation and corresponding compression efficiency with PSNR of right sequence. Figure 9(a), shows the comparison of the results from the proposed scheme with simulcast and compatible coding, for the test sequences to correspond to a GOP. Since, the proposed stereo video scheme exploits both DCP and MCP, the compression efficiency is improved when it compared with existing schemes. The Fig. 9(b) and 9(c) show the percentage of encoded macroblock types of test sequences. In case of a slow motion sequence, Laboratory, most of macroblocks are coded with MCP instead of DCP, because of its small prediction errors,

whereas most macroblocks of sequence, *UbiHome*, are coded with DCP. As a result, in case of *Laboratory*, there is a little compression efficiency when compared with simulcast coding. However, in case of *UbiHome*, the compression efficiency is improved in proportion to the number of macroblock which referred to DCP.

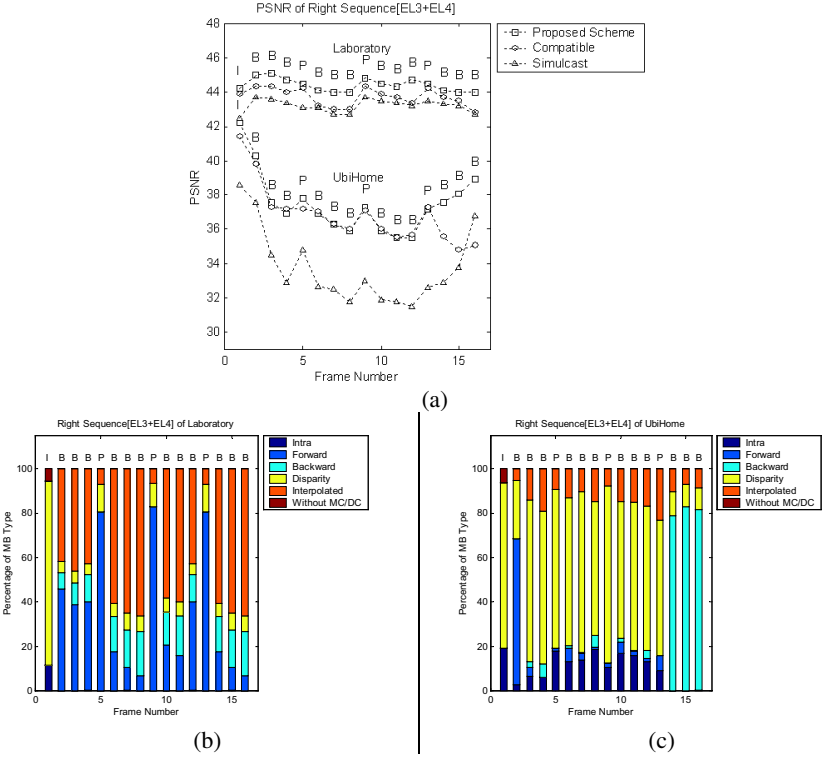

Fig. 9. The number of disparity estimated macroblock types according to the motion rate (a) PSNR comparison (b) Percentage of macroblock types for *Laboratory* (c) Percentage of macroblock types for *UbiHome*

Unlike non-scalable coding approach, the compression efficiency of scalable video coding varies according to the change of bit-rate to be allotted to each layer even if the total bit-rate is same. Therefore, it is need to find the best bit-rate partitioning point for each scalability layer. The mean PSNRs for left and right are differed according to the changes of bit allocations as presented in Table 2. In this case, the effects of the spatial scalability are not considered. Therefore, the layers for the left and right sequences are the spatial scalability BLs which are [BL+EL1] and [EL3+EL4]. A total bit-rate allotted to left and right sequences is 2Mbps. The compression efficiency is improved when we assign more bits to the layer for left sequence, instead of assigning same bits to left and right sequences. In the experiments, we can observe that the image quality is most high when the bit-rate for the left sequence is about 60% of the total bit-rate. This partition is well balanced for the overall test sequences.

Table 2. Mean PSNR according to bit partition in left and right sequences

Left[BL+EL1]/ Right[EL3+EL4]		1.4 / 0.6 [Mbps]	1.2 / 0.8 [Mbps]	1/1 [Mbps]	0.8 / 1.2 [Mbps]	0.6 / 1.4 [Mbps]
Sequence	*Laboratory*	43.02	43.85	43.83	43.61	42.74
	UbiHome	34.52	34.49	34.13	33.48	32.94

We can see in Table 3 and Table 4 that the mean PSNR according to the allotted bit-rate of spatial scalability BL and EL. When the total bit-rate is 1Mbps, the quality is increased as the spatial scalability BL bit-rate increases and the spatial scalability EL bit-rate decrease. However, when the total bit-rate is 6Mbps, the quality is increased as the spatial scalability BL bit-rate is reduced. If total bit-rate is low, the mean PSNR is affected by BL quality, but if total bit-rate relatively high, the mean PSNR is adversely affected by EL quality. The results suggest that to choose a partition point for the system, we need a measure which takes into account the qualities of both layers.

Table 3. Mean PSNR according to bit partition in spatial scalability base and enhancement layers (Total bit-rate = 1Mbps)

Base[(BL+EL1), (EL3+EL4)] / Enhancement[EL2+EL5]		0.7/0.3 [Mbps]	0.6/0.4 [Mbps]	0.5/0.5 [Mbps]	0.4/0.6 [Mbps]	0.3/0.7 [Mbps]
Sequence	*Laboratory*	35.36	35.06	34.69	33.91	33.23
	UbiHome	32.38	32.32	32.34	32.36	32.32

Table 4. Mean PSNR according to bit partition in spatial scalability base and enhancement layers (Total bit-rate = 6Mbps)

Base[(BL+EL1), (EL3+EL4)] / Enhancement[EL2+EL5]		4.2/1.8 [Mbps]	3.6/2.4 [Mbps]	3/3 [Mbps]	2.4/3.6 [Mbps]	1.8/4.2 [Mbps]
Sequence	*Laboratory*	45.54	45.95	46.10	46.32	46.58
	UbiHome	41.34	41.92	42.23	42.31	42.32

4 Summary and Future Work

We have proposed stereo video coding scheme for heterogeneous consumer devices by exploiting the concept of spatio-temporal scalability. The proposed scheme exploits interpolation based motion-disparity compensation and prediction to improve coding efficiency. The experimental results show the efficiency of proposed interpolative coding scheme by comparison with already known methods, compatible stereo and simulcast stereo, and the advantages of disparity estimation in terms of scalability overhead. To provide flexible stereo video service, we define both a temporally scalable layer and a spatially scalable layer for each eye-view. With this scheme, clients in the system are able to decode the stereo video based on their own display size, bandwidth availability and processing power by selectively receiving layered streams.

According to the experimental results, we expect the proposed functionalities will play a key role in establishing highly flexible stereo video service for ubiquitous display environment where devices and network connections are heterogeneous. Currently, we are pursuing a new quantizer for the residual images which are input of spatial scalability ELs.

References

[1] M.G.Perkins, "Data Compression of Stereopairs," Proc. of *IEEE Tr. on Communications*, vol. 40, no. 4, pp. 684-696, April 1992.
[2] A.Kopernik, D.Pele, "Improved disparity estimation for the coding of stereoscopic television," Proc. of *SPIE Visual Communications and Image Processing*, vol. 1818, pp. 1155-1166. 1992.
[3] M.Doma ski, S. Ma kowiak, "Modified MPEG-2 video coders with efficient multi-layer scalability," Proc. of *ICIP*, vol. 2, pp. 1033-36, Oct. 2001.
[4] A.Puri, R.V.Kollarits, and B.G.Haskell, "Basics of Stereoscopic Video, New Compression Results with MPEG-2 and a Proposal for MPEG-4," Proc. of *Image Communications*, vol. 10, pp. 201-234, 1997.
[5] Y.Song, "Improved Disparity Estimation Algorithm with MPEG-2 Scalability for Stereoscopic Sequences," Proc. of *IEEE Tr. on CE*, vol. 42, no. 3, Aug. 1996.
[6] S.Sethuraman, M.W.Siegel, A.G.Jordan, "A multiresolution framework for stereoscopic image sequence compression," Proc. of *ICIP*, vol. 2, pp. 361-365, Nov. 1994.
[7] S.Sethuraman, A.G.Jordan, M.W.Siegel, "Multiresolution based hierarchical disparity estimation for stereo image pair compression," Proc. of *the Symposium on Application of subbands and wavelets*, 1994.
[8] W.Woo, "http://vr.kjist.ac.kr/~3D/".
[9] M.Domanski, A.Luczak, S.Mackowiak, R.Swierczynski, "Hybrid coding of video with spatio-temporal scalability using subband decomposition," Proc. of *SPIE*, vol. 3653, pp. 1018-1025, 1999.
[10] U.Benzler, "Spatial Scalable Video Coding using a Combined Subband-DCT Approach," Proc. of *IEEE Tr. Circuits and Systems for Video Tech.*, vol. 10, pp. 1080-1087, Oct. 2000.
[11] M.Narroschke, "Functionalities and costs of scalable video coding for streaming services," Proc. of *Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, Nov. 2002.