Multi-view Video Coding based on the Lattice-like Pyramid GOP Structure

Kwan-Jung Oh and Yo-Sung Ho

Department of Information and Communications
Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Republic of Korea
{kjoh81,hoyo}@gist.ac.kr

Abstract. With the advancement of computer graphics and computer vision technologies, the realistic visual system can come true in the near future. The multi-view video system can provide an augmented realism through selective viewing experience. However, because of the increased number of cameras, the multi-view video generates a huge amount of data to be processed. Consequently, we need efficient compression algorithms to reduce the amount of data without sacrificing visual quality. In this paper, we present a new multi-view video coding (MVC) scheme using a lattice-like pyramid group of pictures (GOP) structure with variable search range. The proposed algorithm includes a lattice-like GOP arrangement, a pyramid GOP structure, and a variable search range scheme.

Index Terms—MPEG, 3DAV, H.264/AVC, multi-view video coding (MVC)

1. INTRODUCTION

In recent years, various multimedia services have become available and the demand for realistic multimedia systems is growing rapidly. A number of three-dimensional (3D) video technologies, such as holography, two-view stereoscopic system with special glasses, 3D wide screen cinema, and multi-view video have been studied to satisfy these demands.

Among them, multi-view video coding (MVC) is the key technology for various applications including free-viewpoint video (FVV), free-viewpoint television (FVT), 3DTV, immersive teleconference, and surveillance. The traditional video is a two-dimensional (2D) medium and only provides a passive way for viewers to observe the scene. However, MVC can offer arbitrary viewpoints of dynamic scenes and thus allow more realistic video. The multi-view video includes multi-viewpoint video sequences captured by several cameras at the same time, but different positions. Because of the increased number of cameras, the multi-view video contains a large amount of data. Since this system has serious limitations on information distribution applications, such as broadcasting, network streaming services, and other commercial applications, we need to compress the multi-view sequence efficiently without sacrificing visual quality significantly [1] [2].

In the past, MVC has been studied in several video coding standards. The MPEG-2 multi-view profile (MVP) proposes a block-based stereoscopic coding to encode the stereo video. Motion-compensated prediction (MCP) is used to reduce temporal redundancy and disparity compensated prediction (DCP) is used to reduce spatial redundancy. The MPEG-4 multiple auxiliary component (MAC) is also related to MVC. In addition, H.263 and H.264 are used for MVC. However, none of them efficiently supports MVC [3].

Recently, ISO/IEC/JTC1/SC29/WG11/MPEG/adhoc group (AHG) on 3-D audio and visual (3DAV) group has started the work on 3DAV standard. More and more people being interested in multi-view video coding. Some preliminary study and experimental results of multi-view video coding have been reported. There are three main coding methods. The most straightforward method is to encode the multiple video sequences separately. In this method, only temporal correlation within one view is used. Another method is to utilize spatial correlation only. In this case, images of one view are only predicted from their left view images. The third method utilizes both temporal and spatial correlation. Experimental results show that the third method is not so efficient when only simple block-based motion compensated prediction is used to exploit spatial correlation. Due to the high similarity and the little displacement between two adjacent views, global motion information can be used to improve the coding efficiency [3].

In this paper, we propose a multi-view video coding based on the lattice-like pyramid group of pictures (GOP) structure. The proposed algorithm applies three sub-algorithms: the lattice-like GOP arrangement, the pyramid GOP structure, and the variable search range scheme. The lattice-like GOP arrangement scheme alternately arranges the I-frame between odd and even views. It can efficiently reduce the number of bits for coding without sacrificing the visual quality. The pyramid GOP structure is a hierarchical GOP structure composed of several I-frames, B-frames, and RB-frames. Pictures in low level are used for prediction of pictures in high level. It would be more efficient than the other GOP structures since past and future pictures are used as reference frames. Finally, the variable search range scheme solves the global disparity problem caused by the long distance between adjacent views. We employ a larger search range for the spatial reference which is not in the same view but in the adjacent views.

2. MULTI-VIEW VIDEO CODING (MVC)

2.1 General MVC System

MVC system contains the process from the acquisition to the display of multiple video sequences. Figure 1 shows the general MVC system.

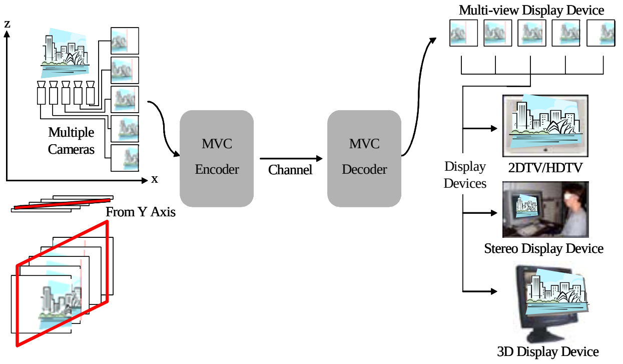


Fig. 1. General MVC system

At first, we acquire multi-views sequences by using several cameras. And then, MVC encoder compresses the multi-view video data. The encoded bitstream is transmitted through the channel. The MVC decoder converts encoded bitstream to multi-view video sequences. Finally, one display device is chosen by its application among the several devices.

2.2 Requirements for MVC

MVC algorithms should satisfy some requirements. In the following, we use “shall” if a certain requirement is mandatory, and “should” if a certain requirement is desirable, but not necessarily required. Requirements for MVC are largely divided into compression related requirements and system support related requirements.

In the case of compression related requirements, MVC shall provide high compression efficiency relative to independent coding of each view of the same content. View scalability shall be supported. In addition, SNR scalability, spatial scalability, and temporal scalability should be supported. MVC shall support low encoding and decoding delay modes, and shall support robustness to errors (also known as error resilience). MVC should enable flexible quality allocation over different views. MVC shall support random access in the time dimension and in the view dimension. For example, it shall be possible to access a frame in a given view with minimal decoding of frames in the time dimension or view dimension.

In the case of system support related requirements, MVC shall support accurate temporal synchronization among the multiple views and should enable robust and efficient generation of virtual views or interpolated views. Also, MVC should support efficient representation and coding methods for 3D display including IP (integral photography) and non-planar image (e.g. dome) display systems. Finally, MVC should support transmission of camera parameters [4].

2.3 Anchor Coding Method

MPEG 3DAV adhoc group has led the research about MVC and made an anchor for evaluation for several MVC algorithms. The anchor coding method is very simple. The anchor just encodes each view independently by using H.264. The anchor does not consider the spatial correlation between adjacent views. The purpose of anchor is to help proponents to develop their algorithms. They will be able to compare their results with the results of anchor. Table 1 shows the H.264 parameters for anchor coding.

Table 1. H.264 parameters for anchor coding

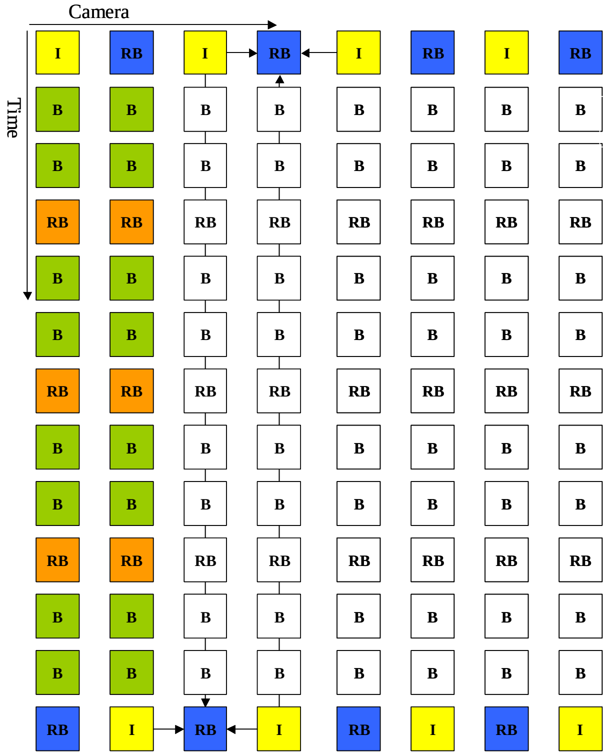
Feature / Tool / Setting	H.264 Parameters
Rate Control	Yes, Basic Unit =1 MB row
RD Optimization	Yes
Specific Settings	Loop Filter, CABAC
Search Range	±32 for VGA/XVGA
# of Reference Pictures	5
Temporal Random Access (Open GOP)	1sec(15fps), 0.5sec(25/30fps)
GOP Structure	IBBP…
Direct Mode	Spatial
FRExt Tools (Adaptive Block Transform)	Yes

3. PROPOSED ALGORITHMS

The previous anchor coding algorithm just uses the temporal correlation of multi-view sequences. However, the core of MVC is how to use spatial correlation efficiently. The proposed algorithms try to find the suitable condition with respect to coding efficiency and encoding time. In this section, we describe the three proposed sub-algorithms such as the lattice-like GOP arrangement, the pyramid GOP structure, and the variable search range scheme in detail [5] [6].

3.1 Lattice-like GOP Arrangement

In video coding, overall coding efficiency of a given sequence highly depends on coding efficiency of I-frames in general. I-frames are usually encoded with high quality since it is referred by P-frames and B-frames and it influences overall video quality. Moreover, I-frame plays an important role with respect to display, because it used as the basic access unit. Despite of these structural features, frequent usage of it may yield the lower coding efficiency since I-frame requires much more bits than P-frame or B-frame in general.



I	RB	I	RB	I	RB	I	RB
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
RB	RB	RB	RB	RB	RB	RB	RB
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
RB	RB	RB	RB	RB	RB	RB	RB
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
RB	RB	RB	RB	RB	RB	RB	RB
B	B	B	B	B	B	B	B
B	B	B	B	B	B	B	B
RB	I	RB	I	RB	I	RB	I

Fig. 2. Lattice-like GOP arrangement

In the case of the anchor coding scheme, it currently uses I-frames in 0.5 second. For example, if we use the 25fps sequence, we periodically insert one I-frame per 12 frames. Since the anchor scheme only depends on temporal correlation, it needs many I-frames. However, if we consider spatial correlation, the number of I-frames for the anchor coding method is so many.

To reduce the number of I-frames without any loss, we propose the lattice-like GOP arrangement that can locate I-frames alternately in even and odd views and substitute RB-frames for I-frames. The RB-frame in the previous sentence is similar to B-frame however it can be referred by other frames. RB-frames play an important role in the lattice-like GOP arrangement. Since the position of RB-frames is originally same with the position of I-frames, RB-frames should have similar characteristics with I-frames.

So, we have tried to encode the RB-frame having same quality to I-frame. In the proposed algorithm, RB-frames are encoded by referring pictures located in the left and the right view and located in the same view but temporally far. We can efficiently reduce the bit rate without quality degradation. Figure 2 shows the proposed lattice-like GOP arrangement in the case of 8 views and a 25fps sequence.

3.2 Pyramid GOP Structure

Coding efficiency of one GOP is highly influenced by the structure of the GOP and the efficient bits allocation. The anchor has the I-B-B-P structure and uses five reference frames. Even if the anchor uses multiple reference frames, there is minor difference with respect to coding efficiency compare to the case of using just one reference. Rather, it causes additional coding time because of useless references and the non-efficient GOP structure [7].

In the proposed algorithm, P-frames in the anchor coding structure are replaced by RB-frames and encoded through the hierarchical structure of the pyramid GOP to improve coding efficiency. Also, we reduce the number of reference frames to just one or two. Figure 3 shows the pyramid GOP structure. In Fig. 3, the RB-frames in the level 2 are encoded using I-frame and RB-frame in the level 1. Then, the RB-frames in the level 3 are encoded using the RB-frames in the level 2 and I-frame or RB-frame in the level 1. Finally, B-frames in the level 4 are encoded using the frames in previous levels. Arrows in Fig. 3 indicate reference relationships.

Since the frames located in the high level are more important than the frames locate in the low level, frames in the high level must be coded with high quality. By using the pyramid GOP structure, we can efficiently improve coding efficiency for a given GOP and reduce useless encoding time. If we expand this scheme to spatial dimension, we can achieve better coding efficiency but it has a problem with respect to random access and decoding delay at decoder part. Thus, we just employ the pyramid GOP structure for a temporal dimension. In addition, this hierarchical structure has advantages in terms of the temporal scalability.

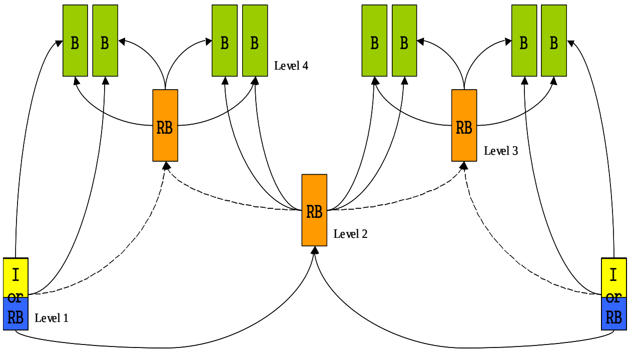


Fig. 3. Pyramid GOP structure

3.3 Variable Search Range

The main difference between MVC and 2D video coding is that MVC can take spatial references between adjacent views. According to the previous research, performance of spatial reference between adjacent views is inferior to that of temporal reference. This is mainly due to variant characteristics of motions and illuminations depending on camera positions. The search range of the anchor coding scheme is ±32. However, the larger image size is the larger its global disparity is. Eventually, the global disparity is larger than ±32 pixels. Figure 4 shows the global disparity between adjacent views.

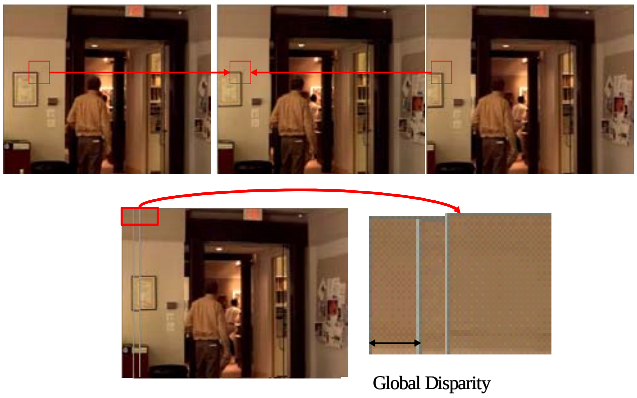


Fig. 4. Global disparity

In the case of “Exit” (640×480), the global disparity is 40~50 pixels. It means that the current ±32 search is too small to compensate the global disparity, which is over 32 pixels in large images such as 640×480 or 1024 ×768 video sequences. To solve this problem, we use ±64 as the search range temporarily for spatial references. By using this scheme, we can improve the motion compensation. However, encoding time is increased. Figure 5 easily explains the problem of global disparity and the effect of the variable search range scheme.

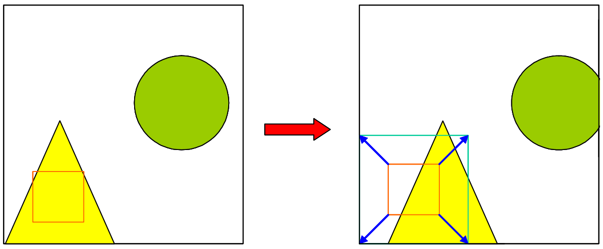


Fig. 5. Effect of the variable search range scheme

In Fig. 5, we assume that the left picture and the right picture are adjacent views. While the inner box cannot cover the proper region, the outer box can cover the proper region in spite of the global disparity.

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed algorithm, we have experimented with “Exit” and “Ballroom” sequences (640× 480, 8 views) provided by MERL. We simulate the anchor and the proposed algorithm by using H.264 reference software JM 9.5. And then, we compare coding results of the anchor and the proposed algorithm with respect to bit rate and the PSNR [8].

We employ ±64 search range for the spatial reference and we just use one or two references. Instead of the rate control scheme for H.264, we manually adjust the quantization parameter (QP) to set target bits. Also we use the pyramid GOP structure instead of IBBP… structure. Since we use the lattice-like GOP arrangement, temporal random access is 2 sec (15fps) and 1 sec (25/30 fps). Except for above conditions, other test conditions follow anchor coding conditions in Table 1. We experiment for three target bits 192, 256, and 384 kbps recommended. QP is the quantization parameter used for the setting of target bits. For example, RB2 means the QP for the RB-frame in the level 2. Since the proposed algorithm does not yet have a proper rate control scheme, we manually regulate the QP for target bits. Table 2, Table 3, and Table 4 show experimental results for 192, 256, and 384 kbps.

Table 2. Experimental results for “Exit” (192kbps)

	Anchor		Proposed		QP
View	Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)	I1-RB1-RB2-RB3-B4
1	192.19	36.39	192.86	37.36	33(30)-32-33-34
2	192.18	36.40	188.07	37.30	33(30)-32-33-34
3	192.15	36.62	189.79	37.53	33(30)-32-33-34
4	192.22	36.26	187.21	37.21	33(30)-32-34-35
5	192.32	35.74	191.95	36.57	34(31)-32-34-36
6	192.33	35.66	194.14	36.58	34(31)-33-34-36
7	192.27	35.00	191.18	35.72	35(33)-34-35-36
8	192.25	34.63	196.62	35.54	35(33)-34-35-37
Avg.	192.24	35.84	191.48	36.73	

Table 3. Experimental results for “Exit” (256kbps)

	Anchor		Proposed		QP
View	Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)	I1-RB1-RB2-RB3-B4
1	256.28	37.31	250.60	38.12	31(28)-29-31-33
2	256.19	37.24	256.05	38.09	31(28)-29-31-33
3	256.20	37.50	255.06	38.32	31(28)-29-30-33
4	256.34	37.24	257.95	38.16	31(28)-29-32-33
5	256.33	36.74	256.68	37.48	32(29)-30-32-34
6	256.42	36.71	257.78	37.50	32(29)-31-32-34
7	256.29	36.01	261.43	36.74	33(30)-32-33-34
8	256.41	35.8	260.75	36.63	33(30)-32-33-35
Avg.	256.31	36.82	257.04	37.63	

Table 4. Experimental results for “Exit” (384kbps)

	Anchor		Proposed		QP
View	Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)	I1-RB1-RB2-RB3-B4
1	384.14	38.42	388.05	39.07	28(26)-27-29-30
2	384.39	38.30	389.28	38.94	28(26)-27-29-30
3	384.19	38.60	381.93	39.18	28(26)-27-29-30
4	384.21	38.46	389.12	39.15	28(26)-27-29-30
5	384.48	37.95	379.04	38.55	29(27)-28-30-31
6	384.62	37.98	391.90	38.63	29(27)-28-30-31
7	384.32	37.23	373.75	37.71	30(28)-29-31-32
8	384.29	37.27	386.74	37.92	30(28)-29-31-32
Avg.	384.33	38.03	384.98	38.64	

The proposed algorithm achieved about 0.62~0.89 dB quality improvement compared to the anchor. Figure 6 shows the rate and distortion curve for “Exit”. As you can see, the rate and distortion curve of the proposed algorithm is located upper than the rate and distortion curve of the anchor.

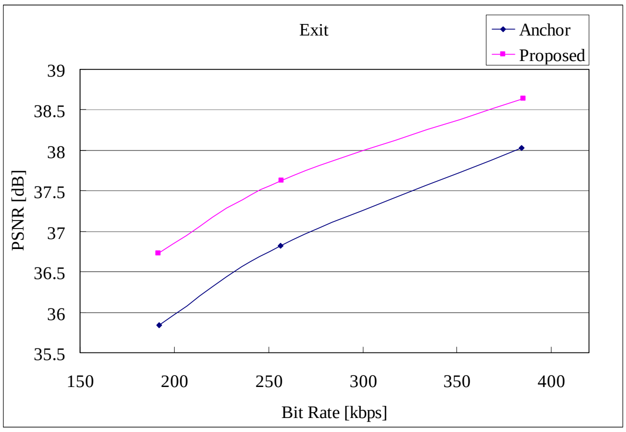


Fig. 6. Rate and distortion curve for “Exit”

Next, we experiment “Ballroom” sequence for three target bits 256, 384, and 512 kbps. Except for target bits, other experimental conditions are the same to the previous experiment. Since “Ballroom” sequence is more complex than “Exit”, its target bits are more than that of “Exit”. Table 5, Table 6, and Table 7 show experimental results for 256, 384, and 512 kbps.

Table 5. Experimental results for “Ballroom” (256kbps)

	Anchor		Proposed		QP
View	Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)	I1-RB1-RB2-RB3-B4
1	256.45	31.06	250.21	31.82	39(36)-37-38-40
2	256.28	31.04	261.57	32.08	39(36)-37-38-39
3	256.47	31.43	259.28	32.40	39(36)-37-38-39
4	256.42	30.62	257.78	31.71	39(36)-38-39-40
5	257.04	30.49	256.27	31.56	39(37)-38-39-40
6	256.77	31.15	254.53	32.31	39(36)-37-39-40
7	256.69	30.96	251.22	31.94	39(37)-38-39-40
8	256.49	29.83	251.53	30.97	39(38)-39-40-41
Avg.	256.58	30.82	255.30	31.85	

Table 6. Experimental results for “Ballroom” (384kbps)

	Anchor		Proposed		QP
View	Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)	I1-RB1-RB2-RB3-B4
1	384.54	32.84	384.14	33.64	36(33)-34-35-36
2	384.38	32.93	388.13	33.82	36(33)-34-35-36
3	384.63	33.38	380.36	34.17	36(33)-34-35-36
4	384.63	32.50	388.36	33.49	36(33)-34-35-37
5	384.83	32.41	382.10	33.32	36(34)-35-36-37
6	384.88	33.23	383.80	34.21	36(33)-34-35-37
7	384.65	32.83	383.60	33.76	36(33)-34-36-37
8	384.62	31.82	380.57	32.89	36(34)-35-37-38
Avg.	384.65	32.74	383.89	33.66	

Table 7. Experimental results for “Ballroom” (512kbps)

	Anchor		Proposed		QP
View	Rate (kbps)	PSNR (dB)	Rate (kbps)	PSNR (dB)	I1-RB1-RB2-RB3-B4
1	512.27	34.13	514.53	34.78	34(32)-32-33-34
2	512.40	34.21	518.39	35.03	34(31)-32-33-34
3	512.45	34.68	518.19	35.48	34(31)-31-33-34
4	512.66	33.83	506.81	34.72	34(31)-32-33-35
5	512.84	33.71	518.74	34.71	34(31)-32-34-35
6	512.88	34.70	516.33	35.59	34(31)-32-34-34
7	512.49	34.13	511.21	34.96	34(31)-32-33-35
8	512.67	33.22	510.28	34.16	34(32)-33-34-36
Avg.	512.58	34.08	514.31	34.93	

The proposed algorithm achieved about 0.85～1.03 dB quality improvement compare to the anchor. Figure 7 shows that the rate and distortion curve for “Ballroom”. As you can see, the rate and distortion curve of the proposed algorithm is located upper than the rate and distortion curve of the anchor.

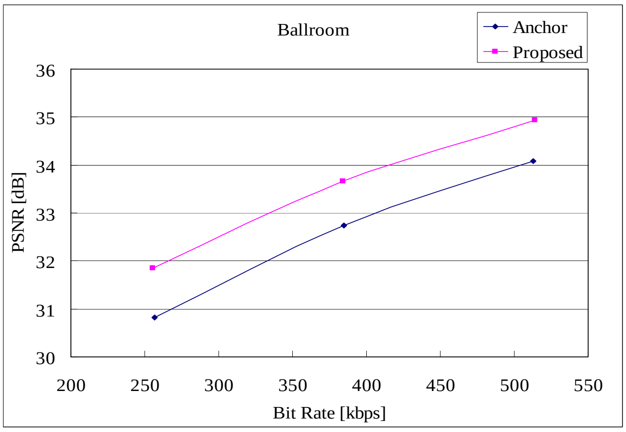


Fig. 7. Rate and distortion curve for “Ballroom”

Table 8 shows the encoding time comparison between the anchor and the proposed algorithm for the second view of “Exit”. As you can see in Table 8, encoding time of the proposed algorithm is about 5 times faster than the anchor and time for ME is about 14 times faster that the anchor. We show only one result for experiments of encoding time comparison, since other experimental results are similar to the result in Table 8.

Table 8. Encoding time comparison

2nd View of “Exit”	Anchor	Proposed
ME Time (sec.)	26682.335	1941.178
Total Encoding Time (sec.)	28527.225	5527.562

5. CONCLUSIONS

In this paper, we have proposed a multi-view video coding based on the lattice-like GOP structure. For the efficient coding structure, we have proposed the lattice-like GOP arrangement method and we can efficiently reduce bit rate by using this scheme. The pyramid GOP structure shows the good encoding performance for each GOP. We also have proposed the variable search range scheme that has larger search range for the spatial reference. The proposed algorithm achieved about 0.6～1.0 dB quality improvement compared to the anchor coding scheme. In addition, the proposed algorithm is approximately 5 times faster than the anchor.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Information and Communication (MIC) through the Realistic Broadcasting Research Center (RBRC) at Gwangju Institute of Science and Technology (GIST).

REFERENCES

1. Smolic, A., Kauff, P.: Interactive 3D video representation and coding technologies. Proceedings of the IEEE, Spatial Issue on Advances in Video Coding and Delivery. Volume 93 (2005) 99-110
2. Smolic, A., Muller, P., Rein, T., Eisert, P., Wiegand, T.: Free viewpoint video extraction, representation, coding, and rendering. Proceeding of IEEE International Conference on Image Processing. Volume 5 (2004) 3287-3290
3. Wang, R. -S., Wang, Y.: Multiview video sequence analysis, compression, and virtual viewpoint synthesis. IEEE Transactions on Circuit and System for Video Technology, Volume 10 (2000) 397-410
4. ISO/IEC JTC1/SC29/WG11 n6501.: Requirements on multi-view video coding (2004)
5. ISO/IEC JTC1/SC29/WG11 m12542.: Multi-view video coding based on lattice-like pyramid GOP structure (2005)
6. ISO/IEC JTC1/SC29/WG11 n6494.: Preliminary call for evidence on multi-view video coding (2005)
7. ITU-T SG16/Q.6 VCEG-N19.: H.264/MPEG-4 AVC reference software enhancement (2005)
8. ISO/IEC JTC1/SC29/WG11 n7327.: Call for proposal on multi-view video coding (2005)