Adaptive Multiview Video Coding Scheme Based on Spatiotemporal Correlation Analyses

Yun Zhang, Gang Yi Jiang, Mei Yu, and Yo Sung Ho

In this paper, we propose an adaptive multiview video coding scheme based on spatiotemporal correlation analyses using hierarchical B picture (AMVC-HBP) for the integrative encoding performances, including high compression efficiency, low complexity, fast random access, and view scalability, by integrating multiple prediction structures. We also propose an in-coding mode-switching algorithm that enables AMVC-HBP to adaptively select a better prediction structure in the encoding process without any additional complexity. Experimental results show that AMVC-HBP outperforms the previous multiview video coding scheme based on H.264/MPEG-4 AVC using the hierarchical B picture (MVC-HBP) on low complexity for 21.5%, on fast random access for about 20%, and on view scalability for 11% to 15% on average. In addition, distinct coding gain can be achieved by AMVC-HBP for dense and fast-moving sequences compared with MVC-HBP.

Keywords: Multiview video coding, correlation analyses, random access, view scalability, in-coding mode switching.

I. Introduction

Multiview video (MVV) is a three-dimensional (3D) extension of traditional single-view video. The MVV system not only enables us to navigate a visual scene freely from different viewing angles. It also provides us with the capability of 3D perception. With these features, MVV can be used for new multimedia applications, such as photorealistic rendering of 3D scenes, 3DTV, and free-viewpoint video (FVV) communications [1], [2]. However, since MVV is generated by simultaneous recording of a moving scene with multiple cameras located at different positions, it has a huge amount of data which is very challenging to process efficiently. To encode and transmit MVV data, the development of an efficient multiview video coding (MVC) algorithm is needed [2], [3].

MVC has been studied in relation to several video coding standards. MPEG-2 multiview profile (MVP) proposed block-based stereoscopic coding to encode stereo video. The MPEG-4 multiple auxiliary component is also related to MVC. In addition, H.263 and H.264 have also been tried for MVC. However, none of them support MVC efficiently [4]. Since ISO/IEC JTC1/SC29/WG11 Moving Picture Experts Group (MPEG) has recognized the importance of MVC technologies, an ad hoc group on 3D audio and visual (3DAV) was established in December 2001, and four main exploration experiments on 3DAV were performed. In response to “Call for Comments on 3DAV,” a large number of companies have expressed their need for standards that enable FVV and 3DTV. Since the MPEG MVC group moved to the Joint Video Team (JVT) in April 2006, MPEG has started the standardization of MVC for MVV which is the input of FVV and 3DTV systems.

As all MVV data originates from the same scene, the

Manuscript received June 19, 2008; revised Jan.14, 2009; accepted Feb. 10, 2009.

This work was supported by the Natural Science Foundation of China (60672073, 60832003) and the Program for New Century Excellent Talents in University (NCET-06-0537).

Yun Zhang (phone: + 86 574 87600017, email: zhangyun_8851@163.com) is with the Institute of Computing Technology, Chinese Academy of Sciences and Graduate School of Chinese Academy of Sciences, Beijing, P.R. China.. He is also with the Faculty of Information Science and Engineering, Ningbo University, Zhejiang, P.R. China.

Gang Yi Jiang (phone: + 86 574 87600935, email: jianggangyi@126.com) and Mei Yu (email: yumei@nbu.edu.cn) are with the Faculty of Information Science and Engineering, Ningbo University, Zhejiang, P.R. China.

Yo Sung Ho (email: hoyo@gist.ac.kr) is with the Department of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Rep. of Korea.

inherent dependencies include inter-view relationships between adjacent views and temporal relationships between temporally successive images of each view. Obviously, the simplest approach is to encode the individual video sequences independently, but for high compression efficiency, similarities among views must also be taken into account. Like motion compensated prediction (MCP) which is employed to eliminate temporal redundancies, disparity compensated prediction (DCP) has been introduced to eliminate inter-view redundancies among neighboring views [5].

MPEG has surveyed some of the MVC schemes [2], such as sequential view prediction, Group-of-GOP prediction (GoGOP), checkerboard decomposition, and so on. The sequential view prediction structure can achieve relatively high compression efficiency; however, it is time-consuming and has poor random access (RA). Oka and others proposed an MVC scheme using multidirectional picture structure [6]. It improves compression efficiency for dense MVV sequences by using multidirectional pictures, but it leads to enormous computational complexity and poor RA. Yamamoto and others proposed a view-interpolation prediction scheme that synthesizes pictures at a given time and a given position by using view interpolation and using them as reference pictures [7]. They tried to compensate for geometry to obtain precise predictions. Kimata and others proposed the GoGOP scheme to improve the RA performance by using multiple intra frames in 2D-GOP at the cost of compression efficiency [8]. Here, 2D-GOP denotes a 2D picture array, in which each row holds temporally successive pictures of one view, and each column consists of spatially neighboring views captured at the same time instant. Mueller and others proposed an MVC scheme based on H.264/MPEG-AVC using hierarchical B pictures (MVC-HBP) [9], [10]. Since MVC-HBP is superior in compression efficiency and temporal scalability, it has been adopted by JVT and used in the joint scalable video model (JSVM) reference software and the joint multiview video model (JMVM) [10], [11]. JSVM and JMVM have been developed as an extension of H.264/MPEG-4 AVC to support the development of MVC.

Figure 1 shows the framework of an interactive FVV system. Since only a portion of the data needs to be displayed at every time step depending on the actual choice of user, only this data needs to be rendered, decoded, and probably transmitted. Therefore, view scalability (VS) is one of the most important requirements enhancing interoperability [12]. It is defined as the functional ability for the same bitstream to be displayed on a multitude of different terminals and over networks with various performance attributes. Also, VS enables partial decoding and view rendering while limiting transmission bandwidth cost. Yu and others proposed a transmission bandwidth distortion model of the point of view to allow view-dependent partial decoding and rendering [13]. Shimizu and others proposed a VS-supporting MVC algorithm in which input views are divided into a base view and enhanced views [14]. Interactive multimedia applications, such as FVV systems, allow the user to freely change viewing position and direction. Therefore, fast RA is another key requirement of MVC [1], [3], [12]. Liu and others introduced three methods to improve RA, namely, SP/SI frame in view dimension, coding images by multiple representations, and interleaved view coding [15]. Unfortunately, some of MVC's requirements conflict with one another. For instance, MVC with DCP requires a trade-off between compression efficiency and random accessibility.

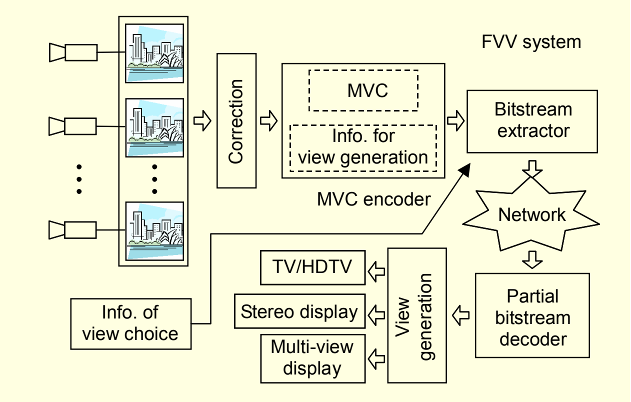


Fig. 1. Framework of interactive FVV system.

In this paper, we propose an adaptive multiview video coding scheme based on spatiotemporal correlation analyses using hierarchical B picture (AMVC-HBP) to achieve better integrative encoding performance, including high coding efficiency, low complexity, as well as fast RA and VS.

The remainder of this paper is organized as follows. Section II presents MVV correlation analyses and describes the problems of previous MVC algorithms. Section III presents the framework of AMVC-HBP in detail. Section IV presents experimental results of the proposed framework. Finally, conclusions are given in section V.

II. Problem of Multiview Video Coding

Temporal and inter-view correlation characteristics of MVV sequences have been analyzed by using a block-matching method as shown in Fig. 2. In the figure, the current coded frame is marked as F, while T denotes the frames temporally preceding the F-frame, and V denotes frames taken at the same time instant as the F-frames in the neighboring views. Blocks in an F-frame are predicted from the V- and T-frames by block matching, in which sum of absolute difference (SAD) is used to measure error. Then, the numbers of most highly matching

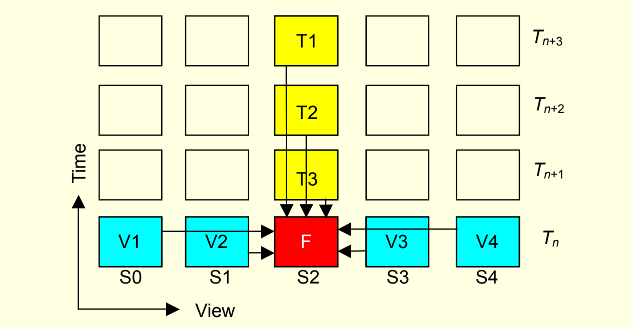


Fig. 2. Correlation analysis of MVV sequences.

Table 1. Temporal and inter-view correlations for MVV sequences.

MVV sequences		Temporal correlation R_T	Spatial correlation R_S
KDDI	Flamenco1	74.94%	25.06%
	Golf1	95.35%	4.65%
	Race1	89.91%	10.09%
	Race2	67.57%	32.43%
	Objects1	87.34%	12.66%
Nagoya Univ.	Aquarium	90.97%	9.03%
	D-Xmas	19.68%	80.32%
MERL	Exit	89.12%	10.88%
	Ballroom	83.79%	16.21%
MSR	Breakdancers	62.75%	38.25%

blocks, that is, the blocks with minimal SAD predicted from the T-frames and V-frames are counted, to measure correlations of different sequences. We implemented experiments with MVV sequences provided by KDDI [16], Nagoya University, MERL [17], and Microsoft Research (MSR).

Table 1 gives the temporal correlation (R_T) and spatial correlation (R_S) of some MVV sequences. Here, R_T is the ratio of blocks predicted from the T-frames to total blocks in the F-frame. Similarly, R_S is the ratio of blocks predicted from the V-frames to total blocks in the F-frame. Thus, R_T+R_S equals 1. The R_T of the Golf1, Race1, Objects1, and Aquarium sequences ranges from 87.34% to 95.35%. These results demonstrate that if the objects move slowly and the camera distance is large, the temporal correlation is dominant in the sequence. By contrast, R_T decreases to 19.68% for the D-Xmas sequence for which the camera distance is quite short. Note that the D-Xmas sequence is made up of downsampled views (one in every ten views) from the original Xmas sequence, so for D-Xmas, the camera distance is 30 mm. In addition to these two kinds of MVV sequences, there is another kind in which the temporal correlation and the inter-view correlation are almost balanced. For example, Breakdancers and Race2 have high temporal correlation and inter-view correlation. Some temporal and inter-view correlation distributions along the time-axis are shown in Fig. 3. In the figure, the horizontal-axis is the frame number, while the vertical-axis indicates the percent of blocks in the F-frame that referenced from the frames V1, V2, V3, V4, T1, T2, and T3, respectively. The spatial correlation of Flamenco1 sometimes increases due to illumination and chroma change, as shown as the red curve in Fig. 3(a), while the spatial correlation of Race2 becomes stronger than temporal correlation when cameras move fast with the racing car.

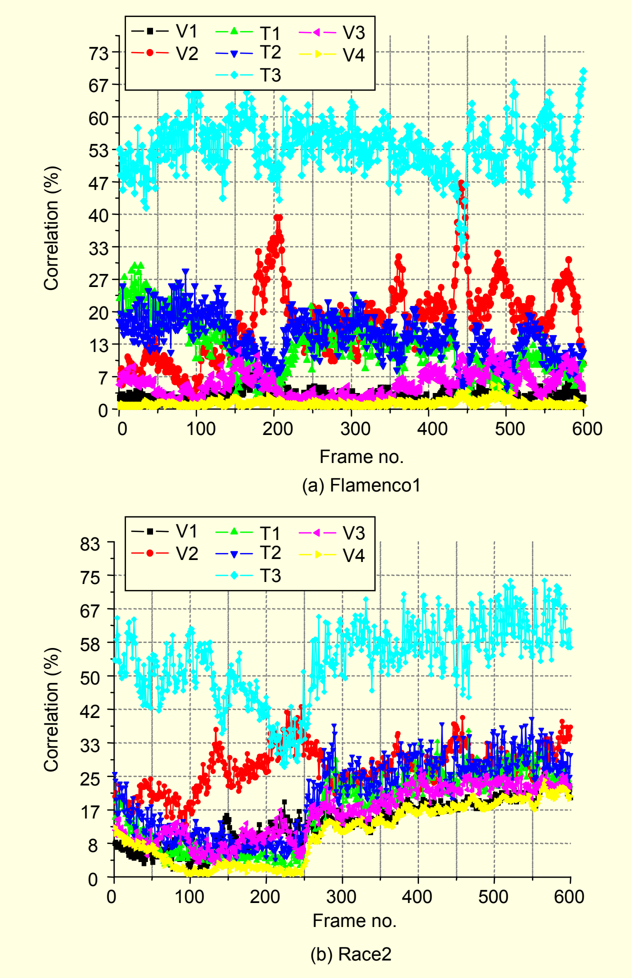


Fig. 3. Results of temporal and inter-view correlation analysis.

From this analysis, we conclude that the correlations of MVV sequences are influenced by the video content, illumination change, velocity of objects and cameras, camera distance, frame rate, and so on. Because of the non-stationary property of video streams, we cannot expect a prediction structure to be universally effective at any time for any scene. In fact, conventional MVC schemes are unable to remove inter-view redundancies efficiently when fast RA and flexible

VS are expected to be achieved because they do not take MVVs' correlation characteristics into consideration.

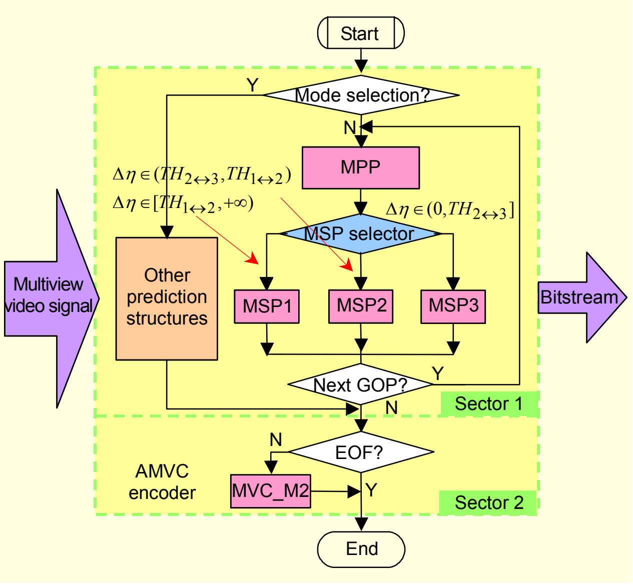
Therefore, to achieve high integrative encoding performance, the proposed MVC algorithm adaptively selects different prediction structures for MVVs with different correlations. We have proposed a coding scheme [18] for fast RA and low-complexity, but its mode updating process lags behind correlation changes and the prediction structures are not as efficient as those of MVC-HBP. In this paper, we propose an AMVC framework and design three kinds of prediction structures using hierarchical B picture, which are referred to as "modes" in this paper. In addition, an adaptive mode switching algorithm is also proposed.

III. Framework of the Proposed AMVC-HBP

1. Framework of AMVC-HBP

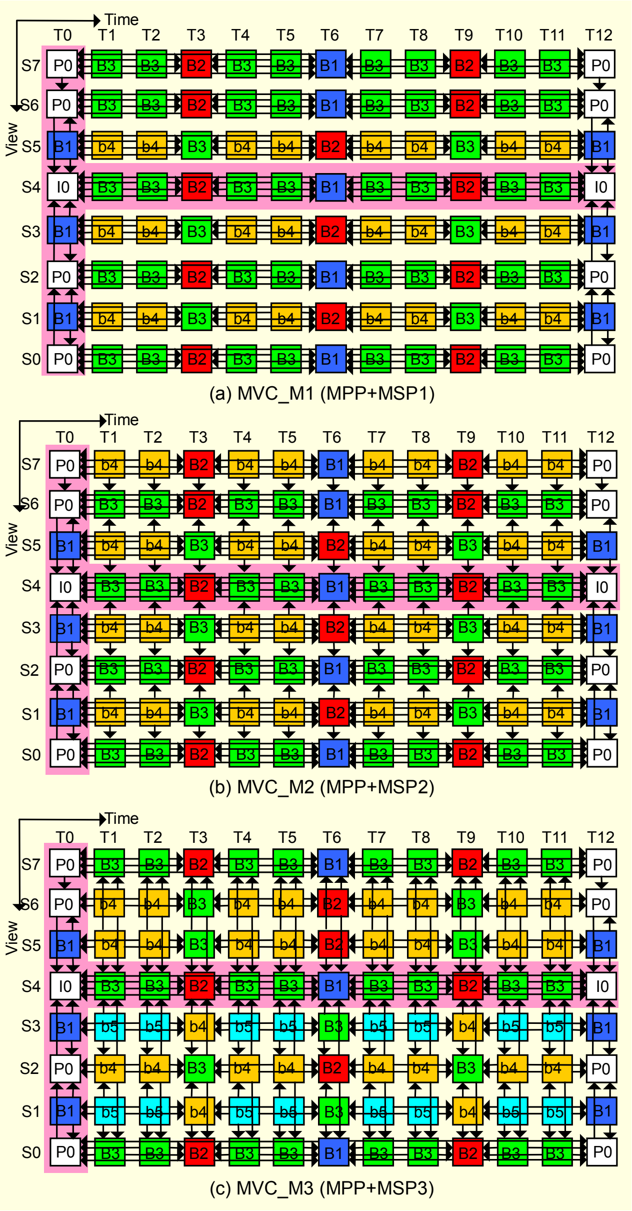
The temporal and inter-view correlations of MVV sequences can be represented by the Lagrangian cost generated by the encoder during the coding process. If the temporal correlation of an MVV sequence is stronger than the inter-view correlation, that is, if $R_T > R_S$, the total Lagrangian cost of the pictures coded by the MCP technique is considered smaller than that of pictures coded with the DCP technique. Therefore, an in-coding mode switching algorithm is proposed. The correlation information is obtained through the Lagrangian cost of the coded pictures during the coding process. The mode switching algorithm will be discussed in detail in subsection 3.3.

Figure 4 depicts the encoding process of the proposed AMVC-HBP framework. The AMVC-HBP encoder is divided into two sectors: sector 1 for the entire 2D-GOP and sector 2 for fractional 2D-GOP. The core part of sector 1 consists of other prediction structures, the mode-public part (MPP), the mode-specific part (MSP), and the MSP selector. The encoding process of the proposed AMVC-HBP framework proceeds as follows. At the beginning, users may select other prediction structures, such as "Simulcast" or "GoGOP," to encode MVV according to their requirements. Otherwise, the AMVC-HBP algorithm is activated. Each 2D-GOP of MVV is initially encoded with MPP, which is a common part for all mode candidates as shown as the pink region in Fig. 5. Meanwhile, the correlation characteristics of the current MVV, represented by Lagrangian cost, $\Delta\eta$, are analyzed. Note that $\Delta\eta$ is a by-



Start
Y
Mode selection?
N
MPP
$\Delta\eta \in (TH_{2\leftrightarrow 3}, TH_{1\leftrightarrow 2})$
$\Delta\eta \in [TH_{1\leftrightarrow 2}, +\infty)$
$\Delta\eta \in (0, TH_{2\leftrightarrow 3}]$
MSP selector
Multiview video signal
Other prediction structures
MSP1
MSP2
MSP3
Bitstream
Next GOP?
Y
N
Sector 1
N
EOF?
AMVC encoder
MVC_M2
Y
Sector 2
End

Fig. 4. Data flow chart of AMVC-HBP.



Time
View
T0 T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12
S7 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0
S6 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0
S5 B1 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 B1
S4 I0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 I0
S3 B1 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 B1
S2 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0
S1 B1 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 B1
S0 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0

(a) MVC_M1 (MPP+MSP1)

Time
View
T0 T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12
S7 P0 b4 b4 B2 b4 b4 B1 b4 b4 B2 b4 b4 P0
S6 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0
S5 B1 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 B1
S4 I0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 I0
S3 B1 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 B1
S2 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0
S1 B1 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 B1
S0 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0

(b) MVC_M2 (MPP+MSP2)

Time
View
T0 T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12
S7 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0
S6 P0 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 P0
S5 B1 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 B1
S4 I0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 I0
S3 B1 b5 b5 b4 b5 b5 B3 b5 b5 b4 b5 b5 B1
S2 P0 b4 b4 B3 b4 b4 B2 b4 b4 B3 b4 b4 P0
S1 B1 b5 b5 b4 b5 b5 B3 b5 b5 b4 b5 b5 B1
S0 P0 B3 B3 B2 B3 B3 B1 B3 B3 B2 B3 B3 P0

(c) MVC_M3 (MPP+MSP3)

Fig. 5. Examples of mode candidates in AMVC-HBP.

product result of encoding MPP frames, which means that no extra efforts should be required to obtain $\Delta\eta$. According to $\Delta\eta$, the MSP selector chooses the most efficient MSP, that is, MSP1, MSP2, or MSP3, to encode the rest of the MVV pictures in the 2D-GOP. Here, the MSP is defined as the individuated part of each mode candidate. Then, for the next 2D-GOP, this encoding process is repeated again. If the remainder of the MVV sequence is less than an entire 2D-GOP, the AMVC encoder moves to sector 2, in which the MVC_M2 prediction mode is adopted to encode the fractional 2D-GOP.

Figure 5 shows three mode candidates designed to encode MVV sequences with different correlation characteristics. Each 2D-GOP has 8 views, and each view has 12 temporally successive pictures. In the 2D-GOP, each column holds temporally successive pictures of one view, and each row consists of spatially neighboring views captured at the same time instant. In the figure, 'B$_i$' or 'b$_i$' represents a hierarchical B picture where i is the level. MPP is defined as the common subpart of the three modes, that is, the frames in the pink-shaded regions. The rest of the frames in the 2D-GOP belong to the so called MSP.

2. Mode Candidates for AMVC-HBP

Unlike conventional single-view video coding, in addition to compression efficiency, other areas of encoding performance, such as abilities of RA and VS, memory requirement, and decoding delay, also have to be considered in the design of mode candidates in MVC.

We have designed three mode candidates as shown in Fig. 5. MVC_M2, which is similar to MVC-HBP, is selected for MVV sequences whose spatial and temporal correlations are almost balanced because it has a good proportion of hierarchical B frames that eliminate both inter-view and temporal redundancies. Some experiments have proved its compression efficiency [9], [10]; however, although the prediction structure of MVC_M2 can achieve high compression efficiency, its high complexity and poor random accessibility limit its applications. For MVV sequences whose temporal correlation is dominant, little coding gain can be achieved by inter-view prediction for non-key pictures [19]. For this reason, we developed MVC_M1, in which inter-view prediction is only kept for key pictures. If an MVV sequence is chiefly temporally correlated, omitting inter-view prediction for non-key pictures leads to little loss of coding efficiency but remarkable reduction of encoding complexity and improvement of RA. For FVV applications, there are also some dense videos in which spatial correlations are dominant, such as D-Xmas and AKKO. The camera distance of this kind of video is small. It is better to apply a hierarchical prediction structure in the view axis for higher compression efficiency because, in this case, DCP works more efficiently than MCP. Therefore, we also developed MVC_M3, which exploits redundancies mainly by using DCP.

Setting the intra frame in the center view (that is, view S4 for 8 views) can be beneficial because two inter-view key pictures directly use the intra frame as a reference. Compression gains can be achieved by reusing intra frame [20]. In addition, this configuration improves RA performance in comparison with the structure in which the intra frame is assigned to the S0 view. These three mode candidates were implemented on JSVM and JMVM by changing the reference picture list reordering and memory management control operation command [21] in SequenceFormatString [12]. It is a pure encoder optimization; therefore, the syntax of the video stream and the decoder do not need to be modified.

The proposed AMVC-HBP framework is flexible. The mode candidates in Fig. 5 are chiefly proposed for fast RA, VS, and low-complexity while maintaining high compression efficiency. For interactive systems, modes like the prediction structure group in [18] can also be adopted in AMVC-HBP to improve RA when it is the most essential aspect of performance.

3. In-Coding Mode Switching

A. Definition of Mode Switching Parameter

AMVC-HBP adaptively selects the most suitable prediction structure for MVV according to its temporal and inter-view correlations. However, for low-complexity, we define a mode switching parameter, which comes from the by-product data of the encoding process of previous frames, to estimate the temporal correlation of the MVV of current 2D-GOP instead of time-consuming block-matching-based correlation analyses. Many of today's video codecs, such as H.264/AVC, are based on predictive coding between one or more reference pictures and the currently encoding picture. Motion estimation is conducted by minimizing a Lagrangian cost function:

$$\min J(\mathrm{MB}_i), \quad \text{where} \quad J(\mathrm{MB}_i) = D + \lambda \cdot R. \qquad (1)$$

The Lagrangian cost function $J(\mathrm{MB}_i)$ is the sum of distortion D and rate R, weighted by the Lagrange parameter λ, for macroblock MB$_i$.

Let n and m denote the numbers of views and pictures in each view in a 2D-GOP, respectively; f_{vt} denotes the picture in the v-th view and the t-th time instant of a 2D-GOP; $iGOP$ is ordinal number of a 2D-GOP; V_{iGOP}= $\{f_{vt} \mid v \in (1, n)$; t = 'T0'; tp(f_{vt}) = 'B'$\}$ is defined as a set of anchor B pictures at time T0 of the $iGOP$-th 2D-GOP, where tp(f_{vt}) indicates the frame type of picture f_{vt}. The B picture f_{vt}, $f_{vt} \in V_{iGOP}$, is encoded using inter-view bi-prediction. Similarly, a set of B pictures in the view with the intra frame is defined as $\boldsymbol{T}_{iGOP}$= {f_{vt} | v = 'S4'; $t \in$ (1 + $iGOP \times m$, $m + iGOP \times m$); tp(f_{vt}) = 'B'}. The B picture f_{vt}, $f_{vt} \in \boldsymbol{T}_{iGOP}$, is coded by using temporal bi-prediction. Here, $\boldsymbol{V}_{iGOP}$ and $\boldsymbol{T}_{iGOP}$ are subsets of total pictures in MPP, that is, $\boldsymbol{V}_{iGOP} \cup \boldsymbol{T}_{iGOP} \subset$ MPP. Let $|\boldsymbol{V}_{iGOP}|$ and $|\boldsymbol{T}_{iGOP}|$ denote the numbers of elements in $\boldsymbol{V}_{iGOP}$ and $\boldsymbol{T}_{iGOP}$, respectively. Because $\boldsymbol{V}_{iGOP} \cap \boldsymbol{T}_{iGOP} = \varnothing$, $|\boldsymbol{V}_{iGOP} \cup \boldsymbol{T}_{iGOP}| = |\boldsymbol{V}_{iGOP}| + |\boldsymbol{T}_{iGOP}|$. The average Lagrangian cost of $\boldsymbol{V}_{iGOP}$ and $\boldsymbol{T}_{iGOP}$ can be expressed as function $g(\boldsymbol{V}_{iGOP}, \boldsymbol{T}_{iGOP})$:

$$\begin{cases} g\left(\boldsymbol{V}_{iGOP},\boldsymbol{T}_{iGOP}\right) = \dfrac{1}{\left|\boldsymbol{V}_{iGOP}\right| + \left|\boldsymbol{T}_{iGOP}\right|} \displaystyle\sum_{f_{vt}\in \boldsymbol{V}_{iGOP}\cup \boldsymbol{T}_{iGOP}} FJ\left(f_{vt}\right), \\ FJ\left(f_{vt}\right) = \displaystyle\sum_{MB_i \in f_{vt}} J\left(MB_i\right), \end{cases} \tag{2}$$

where $FJ(f_{vt})$ denotes the total Lagrangian cost of the picture at position (v, t). From (2), the Lagrangian costs of temporal predicted and inter-view predicted pictures, $\eta_{T,iGOP}$ and $\eta_{V,iGOP}$, are defined as

$$\begin{cases} \eta_{\boldsymbol{T},iGOP} = g\left(\varnothing, \boldsymbol{T}_{iGOP}\right), \\ \eta_{V,iGOP} = g\left(\boldsymbol{V}_{iGOP}, \varnothing\right). \end{cases} \tag{3}$$

By substituting (3) for (2), we obtain

$$\begin{cases} \eta_{\boldsymbol{T},iGOP} = \dfrac{1}{\left|\boldsymbol{T}_{iGOP}\right|} \displaystyle\sum_{F_{vt}\in \boldsymbol{T}_{iGOP}} FJ\left(f_{vt}\right), \\ \eta_{V,iGOP} = \dfrac{1}{\left|\boldsymbol{V}_{iGOP}\right|} \displaystyle\sum_{F_{vt}\in \boldsymbol{V}_{iGOP}} FJ\left(f_{vt}\right). \end{cases} \tag{4}$$

The correlation of an MVV sequence varies with camera distance, lights, camera arrangement, and movement of cameras, objects, and so on. However, the variation is slight during a short time, so we can suppose that the correlation characteristic of an MVV sequence within a 2D-GOP is consistent. In other words, all pictures in a 2D-GOP are assumed to have similar inter-view and temporal correlations.

Fortunately, the correlation characteristics of an MVV sequence can be presented by $\eta_{T,iGOP}$ and $\eta_{V,iGOP}$. For instance, $\eta_{T,iGOP}$ will be smaller than $\eta_{V,iGOP}$ after encoding MPP when the temporal correlation of the current 2D-GOP is stronger than the inter-view correlation. Therefore, we define a parameter, $\Delta\eta$, to represent the relationship between temporal and inter-view correlations:

$$\Delta\eta = \frac{\eta_{V,iGOP}}{\eta_{\boldsymbol{T},iGOP}} = h\left(R_T\right). \tag{5}$$

Obviously, $\Delta\eta$ is an increasing function of temporal correlation R_T, and it will be experimentally determined in section IV. According to $\Delta\eta$ obtained from MPP, pictures in the MSP region are encoded with MSP1, MSP2, or MSP3 to improve integrative encoding efficiency.

B. Thresholds for Mode Switching

The MSP selector chooses the optimal MSP to encode the remaining pictures of the current 2D-GOP according to $\Delta\eta$. We define two thresholds for the MSP selector, $TH_{1\leftrightarrow2}$ and $TH_{2\leftrightarrow3}$. They are the threshold for MSP1 and MSP2 and the threshold for MSP2 and MSP3, respectively. When $\Delta\eta$ is larger than $TH_{1\leftrightarrow2}$, that is, $\Delta\eta \in [TH_{1\leftrightarrow2}, +\infty)$, the MSP selector chooses MSP1 as the optimal prediction structure. If $\Delta\eta \in (0, TH_{2\leftrightarrow3}]$, this indicates that pictures of the current 2D-GOP are chiefly inter-view correlated; therefore, MSP3 is better for AMVC-HBP. Otherwise, MSP2 is the best choice because both the temporal and inter-view dimensions hold a high number of correlations.

Different prediction structures achieve different levels of performance in areas including abilities of random access (RA_k), compression efficiency represented by rate distortion efficiency (RD_k), view scalability (VS_k), and coding complexity (CC_k), where k indicates the different prediction structure. These performance areas are evaluated with the methods introduced in [22]. The performance cost of AMVC-HBP, $\boldsymbol{P}_{\text{AMVC-HBP}}$, is defined as

$$\begin{cases} \boldsymbol{P}_{\text{AMVC-HBP}} = \displaystyle\sum_{k=1}^{3} \boldsymbol{P}_k \cdot \rho_k, \\ \boldsymbol{P}_k = \left[RA_k \quad RD_k \quad CC_k \quad VS_k\right]^T, \end{cases} \tag{6}$$

where $\boldsymbol{P}_k$ denotes the performance cost of an MVC prediction structure, $k \in \{1, 2, 3\}$, and ρ_k denotes the percentage of 2D-GOP whose $\Delta\eta$ is located in the range of $(TH_{k\leftrightarrow k+1}, TH_{k-1\leftrightarrow k})$.

$$\begin{cases} \boldsymbol{P}_{\text{MVC}} = \left[\xi_1 \quad \xi_2 \quad \xi_3 \quad \xi_4\right] \boldsymbol{P}_{\text{AMVC-HBP}}, \\ \qquad \arg\min_{TH} \left\{\boldsymbol{P}_{\text{MVC}}\right\} \\ \boldsymbol{TH} = \left[TH_{1\leftrightarrow2} \quad TH_{2\leftrightarrow3}\right]. \end{cases} \tag{7}$$

The performance cost of the MVC system, $\boldsymbol{P}_{\text{MVC}}$, directly relates to $\boldsymbol{P}_{\text{AMVC-HBP}}$ weighted by coefficient ξ_j, $j \in \{1, 2, 3, 4\}$. When defining thresholds, we have to minimize the cost $\boldsymbol{P}_{\text{MVC}}$ to maximize system performance. However, different application scenarios emphasize different requirements so that ξ_j is increased with the importance of the corresponding aspect of performance. Here, compression efficiency is considered the most important aspect of performance, so ξ_2 is set to be the largest value among the ξ_js.

In our experiments, T_1, the temporal correlation boundary between MVC_M1 and MVC_M2, is set to be 85%, and T_2, the temporal correlation boundary between MVC_M2 and MVC_M3, is set to be 55%. According to (9) and correlation boundaries, the thresholds for $\Delta\eta$ are

$$\boldsymbol{TH} = \left[h(T_1) \quad h(T_2) \right] = \left[128\% \quad 60\% \right]. \tag{8}$$

IV. Experimental Results and Analysis

1. Relationship between Temporal Correlation and $\Delta\eta$

In order to evaluate the similarity between $\Delta\eta$ and the actual correlation of MVV sequences, experiments were implemented on JSVM 7.12 software with test MVV sequences. We did not use a higher version JMVM software because it does not support SequenceFormatString. The test MVV sequences, "Exit" and "Ballroom" were provided by MERL. "Breakdancers" was provided by Microsoft Research, "Alt Moabit" was provided by HHI, and "D-Xmas" and "Aquarium" were provided by Nagoya University.

Figure 6 shows the temporal correlation and $\Delta\eta$ maps of the test MVV sequences. We used the block matching methods discussed in Section II to obtain R_T in Fig. 6(a), and $\Delta\eta$ was obtained in the encoding process of MPP. The quantization parameter was set to 31, and there were 12 temporally successive pictures in each view and 8 views for each 2D-GOP. As seen in Fig. 6(a), D-Xmas has little temporal correlation, while Exit and Ballroom are chiefly temporally correlated for each 2D-GOP. The R_T of Breakdancers is smaller than that of Ballroom and larger than that of D-Xmas. For Ballroom, temporal correlation decreases sharply for a frame because of instantaneous flash for photography. The correlation of an MVV sequence varies along with the time. Figure 6(b) presents $\Delta\eta$ of each 2D-GOP for the same four MVV sequences. As seen in Fig. 6, the variation of $\Delta\eta$ coincides with that of the actual correlation along the time axis; therefore, $\Delta\eta$ can be utilized to estimate the correlation of an MVV sequence.

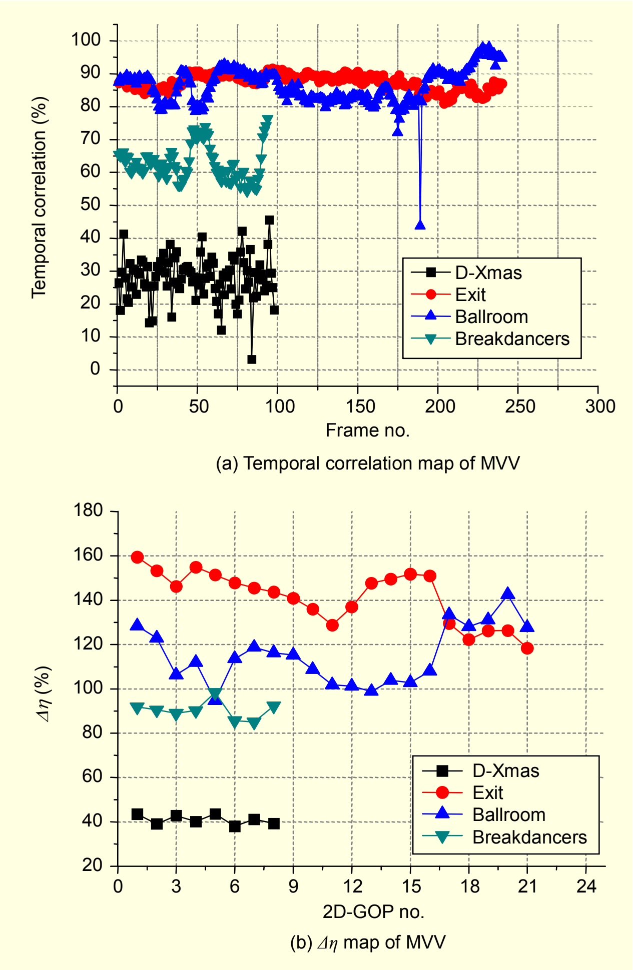


Fig. 6. Temporal correlation map and $\Delta\eta$ map of MVV sequences.

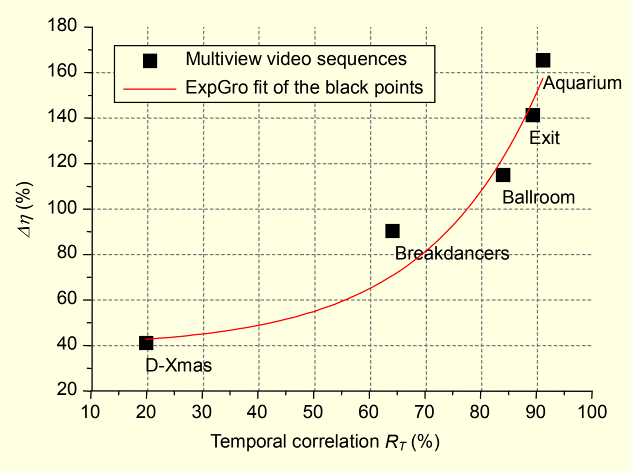


Fig. 7. Map of the relationship between R_T and $\Delta\eta$.

Figure 7 shows the relationship between $\Delta\eta$ and temporal correlation R_T. Its horizontal axis is R_T, and the vertical axis denotes the average value of $\Delta\eta$. The black points are the relationships between R_T and $\Delta\eta$ for different MVV sequences. The red curve is the exponential growth fit of the black points. The more temporal correlation the MVV sequence has, the larger $\Delta\eta$ is. Therefore, we obtain the function $\Delta\eta = h\,(R_T)$ with growth exponential fit, and it can be expressed as

$$\Delta\eta = h(R_T) = \left(y0 + Ae^{\left(\frac{R_T}{\psi}\times 100\right)} \right)\%, \tag{9}$$

where $y0 = 39.0$, $A = 1.414$, and $\psi = 20.5$.

2. Compression Efficiency Comparison

Table 2 specifies coding parameters of the proposed AMVC-HBP and MVC-HBP in the experiments. Figure 8 shows the encoding mode switching process for each 2D-GOP in each MVV sequences. For Aquarium, Breakdancers, and D-Xmas, the AMVC-HBP adaptively uses MVC_M1, MVC_M2, and MVC_M3, respectively. For Exit, Ballroom, and Alt Moabit,

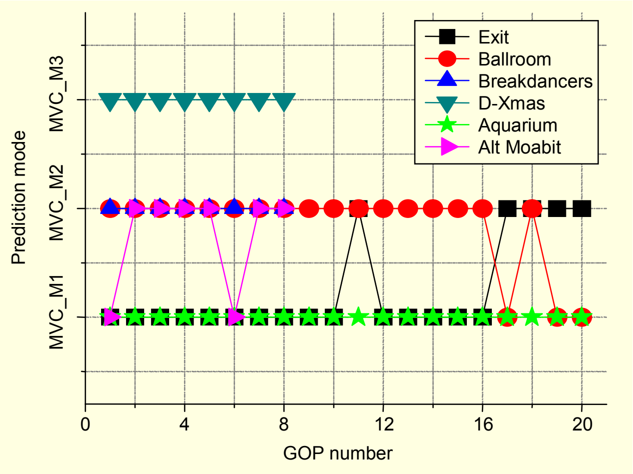


Fig. 8. Mode switching map.

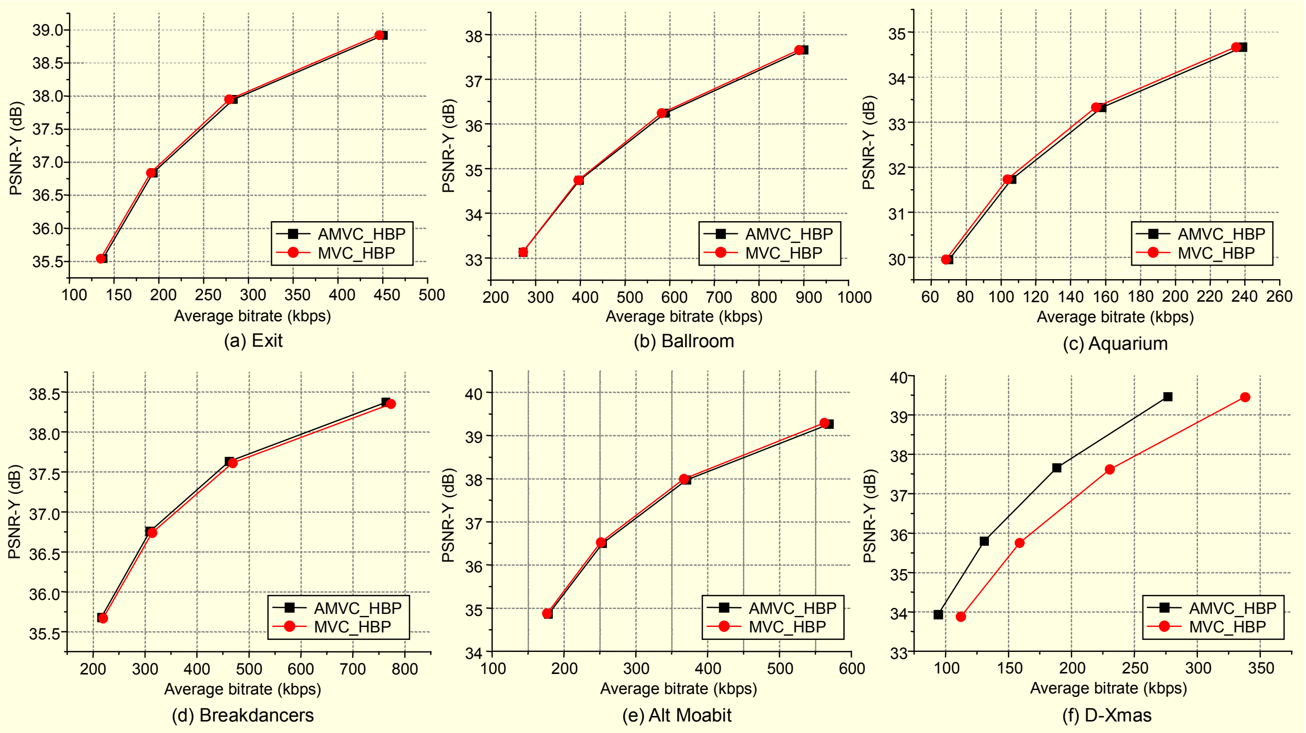
Table 2. Coding parameters.

Software platform	JSVM 7.12
Views × GOP length	8×12
Search range	±96
BiPredIter & IterSearchRange	4 & 8
MaxRefIdxActiveBL0 & MaxRefIdxActiveBL1	1 for MVC_M1 2 for MVC_M2 & MVC_M3
BasisQP	25, 28, 31, 34

the prediction mode switches between MVC_M1 and MVC_M2 along 2D-GOPs. Seventy-five percent of the 2D-GOPs of Exit are coded by MVC_M1, while the other 25% of the 2D-GOPs are coded by MVC_M2. For Ballroom, 85% of the 2D-GOPs are coded by MVC_M2 due to its fast motion. For Alt Moabit, 75% of the 2D-GOPs are coded by MVC_M2 because the short camera distance and low capturing frame rate result in high inter-view correlation.

Figure 9 illustrates the compression efficiency of AMVC-HBP and MVC-HBP. For Exit, Ballroom, and Alt Moabit, the rate-distortion curves of AMVC-HBP and MVC-HBP almost overlap. For Ballroom and Alt Moabit, most of the 2D-GOPs are coded using both DCP and MCP like MVC-HBP does. However, for Exit, most of the 2D-GOPs are highly temporally correlated, so omitting DCP for these 2D-GOPs does not significantly decrease the coding efficiency. For Aquarium, AMVC-HBP is inferior to MVC-HBP due to the lack of DCP, but the gap is less than 0.1 dB. However, AMVC-HBP significantly improves coding performance in terms of RA, complexity, and VS. Coding gain for Breakdancers is achieved because more inter-view key pictures directly use the high quality intra frame as a reference. For D-Xmas, AMVC-HBP outperforms MVC-HBP at about 1.0 dB because DCP is more efficient for such highly inter-view correlated sequence. Figure 10 shows the decoded frames of D-Xmas at position S2T1.

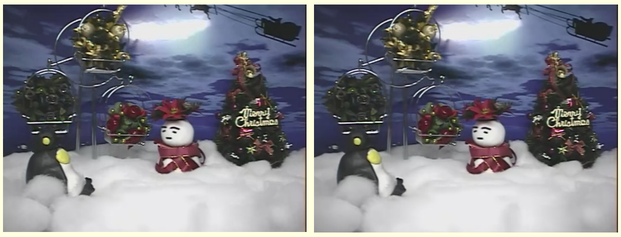
Though Figs. 10(a) and (b) have almost the same objective



(a) Exit (b) Ballroom (c) Aquarium

(d) Breakdancers (e) Alt Moabit (f) D-Xmas

Fig. 9. Compression efficiency comparisons between AMVC-HBP and MVC-HBP.



(a) MVC-HBP (b) Proposed AMVC-HBP

Fig. 10. Decoded images of D-Xmas: (a) level 3, QP=33, PSNR=37.42 dB, 21600 bits and (b) level 4, QP=34, PSNR=37.55 dB, 4592 bits.

and subjective quality, the encoding bits are reduced significantly because AMVC-HBP adopts both inter-view and temporal prediction to encode the S2 view for the highly inter-view correlated D-Xmas sequence instead of temporal prediction only.

3. RA, VS, and Encoding Time Comparisons

In evaluating the random accessibility of the proposed AMVC-HBP, we use F_{AV} and F_{MAX} to indicate the average and maximum frames that need to be decoded to randomly access a frame in a 2D-GOP. Let $x_{i,j}$ be the number of frames that have to be decoded before the frame at (i, j) position is decoded in a 2D-GOP with n time instants and m views, where i and j are the temporal and inter-view positions in one 2D-GOP, respectively. Let $p_{i,j}$ be the probability of the frame at (i, j) being selected by a user, then the RA cost F_{AV} and the maximum number of pre-decoded frames F_{MAX} are defined by

$$F_{\mathrm{AV}} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}\, p_{i,j}, \qquad (10)$$

$$F_{\mathrm{MAX}} = \max\left\{x_{i,j} \,\middle|\, 0 < i \le n, 0 < j \le m\right\}. \qquad (11)$$

In evaluating view scalability, we define F_{SV} and F_{DV} to represent the average number of compulsorily decoded frames of a 2D-GOP when single view and double views are displayed, respectively. Let O_n be a set of the frames in a 2D-GOP, and let $X_{i,j}$ be a set of the compulsorily decoded frames when the frame at (i, j) is displayed; thus, $X_{i,j} \subseteq O_n$. Suppose ϕ_j is the probability that the user will watch the j-th view, and $\phi_{j,k}$ is the probability that both the j-th view and the k-th view will be accessed. Here, F_{SV} and F_{DV} are defined as

$$F_{\mathrm{SV}} = \sum_{j=1}^{m}\left[\Theta\left(\bigcup_{i=1}^{n} X_{i,j}\right)\cdot \phi_j\right], \qquad (12)$$

$$F_{\mathrm{DV}} = \sum_{j=1}^{m}\sum_{k=j+1}^{m}\left[\Theta\left[\bigcup_{i=1}^{n}\left(X_{i,j} \cup X_{i,k}\right)\right]\cdot \phi_{j,k}\right], \qquad (13)$$

where Θ is the cardinality of a set. Additionally, coding complexity is measured by encoding time.

Table 3. RA and VS of the three mode candidates.

	RA (F_{AV}/F_{MAX})/frame	VS (F_{SV}/F_{DV})/frame
MVC_M1	6.4 / 10.0	14.63 / 29.79
MVC_M2	8.7 / 16.0	25.88 / 45.96
MVC_M3	9.9 / 16.0	33.75 / 53.93

Table 3 shows the RA and VS performance of the three mode candidates, where the probabilities in (10), (12), and (13) are uniform probabilities, that is, $p_{i,j}$=1/(n×m), ϕ_j=1/m, and $\phi_{j,k}$=m(m-1)/2. However, the proposed AMVC-HBP adaptively utilizes different prediction modes to encode MVV according to its temporal and inter-view correlations of each 2D-GOP. Thus, we use an average value of F_{AV}, F_{MAX}, F_{SV}, F_{DV}, and encoding time to evaluate the performance in terms of RA, VS, and coding complexity for AMVC-HBP.

Coding experiments were performed on a Dell PowerEdge 2800 (Intel Xeon CPU 3.20 GHz, 3.19 GHz, 4.0 GB DDR-2 memory and Windows server 2003 operating system). Table 4 gives comparisons of AMVC-HBP and MVC-HBP in terms of RA, encoding time, and VS. In the table, “IMPV” denotes improvements achieved by AMVC-HBP in comparison with MVC-HBP. Both the RA and VS of AMVC-HBP were calculated according to (6), Fig. 8, and Table 3. Taking Exit as an example, 75% of the GOPs were coded by MVC_M1, and 25% were coded by MVC_M2, that is, ρ_1=0.75, ρ_2=0.25, and ρ_3=0. Therefore, the average cost of RA is calculated as F_{AV}=6.4×0.75+8.7×0.25+9.9×0=7.0.

The proposed AMVC-HBP outperforms MVC-HBP in terms of encoding time, fast random accessibility, and view scalability for averages of 21.5%, 20%, and 11% to 15%, respectively, while maintaining high compression efficiency. MVC_M1 achieves low-complexity with fast RA and VS by limiting DCP from neighbor views. MVC_M2 improves RA and VS by using a centered intra frame. Therefore, complexity can be reduced by using MVC_M1. Both RA and VS can be improved if 2D-GOPs are coded by MVC_M1 or MVC_M2. MVC_M3 is poor in VS and RA in the time dimension; however, it improves compression efficiency significantly for dense sequences. In addition, MVC_M3 is suitable for virtual view synthesis and displaying special visual effects. AMVC-HBP and MVC-HBP have the same coding delay that derives from the frame reordering after decoding and require the same minimal decoded picture buffer size of the coder.

V. Conclusion

In this paper, we presented a framework of AMVC-HBP for

Table 4. RA, encoding time, and VS for AMVC-HBP and MVC-HBP.

MVV sequences	RA (F_{AV}/F_{MAX})/frame			Encoding time (s)			VS (F_{SV}/F_{DV})/frame		
	MVC-HBP	AMVC-HBP	IMPV (%)	MVC-HBP	AMVC-HBP	IMPV (%)	MVC-HBP	AMVC-HBP	IMPV (%)
Exit	10.1/18.0	7.0/11.5	30.7/36.1	61,328	34,555	43.7	27.4/46.4	17.4/33.8	36.6/28.9
Ballroom		8.4/15.1	16.8/16.1	68,431	59,898	12.5		24.2/43.5	11.7/6
Breakdancers		8.7/16.0	13.9/11.1	76,333	77,046	-0.9		25.9/46.0	5.5/0.9
Aquarium		6.4/10.0	36.6/44.4	9,995	3,964	60.3		14.6/29.8	46.7/35.8
Alt Moabit		8.1/14.5	19.8/19.4	87,635	75,599	13.7		23.1/41.9	15.7/9.7
D-Xmas		9.9/16.0	2.0/11.1	17,802	17,857	-0.3		33.75/53.93	-23.2/-16.2
Average IMPV			20.0/23.0			21.5			15.5/10.9

better integrative encoding performance, in which multiple prediction modes have been integrated. An in-coding mode switching algorithm was also presented for adaptive mode switching. According to temporal and inter-view correlation of MVV, AMVC-HBP is able to adaptively select a better prediction mode from a set of MVC mode candidates in the encoding process without any additional complexity. In our experiments, the proposed AMVC-HBP scheme outperforms the MVC scheme based on H.264/MPEG-4 AVC using hierarchical B picture in terms of low-complexity, fast random accessibility, and view scalability for averages of 21.5%, 20%, and 11% to 15%, respectively, while maintaining high compression efficiency. Additionally, AMVC-HBP obtains 0.1 dB to 1.0 dB coding gain for dense and fast-moving sequences. Thus, the integrative encoding performance is improved significantly.

The proposed AMVC-HBP incurs long delay and high memory consumption because of the adopted prediction modes; therefore, in future works, we will try to create an MVC algorithm with shorter delay and low memory requirement for real-time applications while maintaining high compression efficiency.

References

[1] M. Tanimoto et al., "Proposal on Requirements for FTV," *ISO/IEC MPEG & ITU-T VCEG*, JVT-W127, California, USA, Apr. 2007.

[2] *Survey of Algorithms Used for MVC,* ISO/IEC JTC1/SC29/WG11, N6909, Hong Kong, China, Jan. 2005.

[3] Y.S. Ho and K.J. Oh, "Overview of Multi-view Video Coding," *14th Int'l Workshop Syst., Signals and Image Processing (IWSSIP)*, Maribo, Slovenia, June 2007, pp. 5-12.

[4] R. Koenen, "Overview of the MPEG-4 Standard," *ISO/IEC JTC1/SC29/WG11*, N4030, Singapore, Mar. 2001.

[5] J.-W. Kang et al., "Graph Theoretical Optimization of Prediction Structure in Multiview Video Coding," *Proc. of IEEE Int'l Conf. Image Proc (ICIP)*, vol. 6, San Antonio, Sept. 2007, pp. 429-432.

[6] S. Oka, T. Endo, and T. Fujii, "Dynamic Ray-Space Coding Using Multi-directional Picture," *IEICE Technical Report*, Dec. 2004, pp. 15-20.

[7] K. Yamamoto et al., "Multiview Video Coding Using View Interpolation and Color Correction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, Nov. 2007. pp. 1436-1449.

[8] M. Kitahara et al., "Multi-view Video Coding Using View Interpolation and Reference Picture Selection," *Proc. IEEE Int'l Conf. Multimedia & Expo*, Toronto, Canada, July 2006, pp. 97-100.

[9] P. Merkle et al., "Efficient Prediction Structures for Multiview Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, Nov. 2007. pp. 1461-1473.

[10] *Description of Core Experiments in MVC*, ISO/IEC JTC1/SC29/WG11, w8019, Montreux, Switzerland, Apr. 2006.

[11] H. Schwarz, M. Wien, and J. Vieron, JSVM Software Manual, Joint Video Team, *ISO/IEC MPEG & ITU-T VCEG*, JVT-S070, Geneva, Switzerland, Apr. 2006.

[12] *Requirements on Multi-view Video Coding v.8*, ISO/IEC JTC1/SC29/WG11, N9163, Lausanne, Switzerland, Jul. 2007.

[13] M. Yu et al., "Bandwidth Distortion Model for MVC in Interactive System," *ISO/IEC MPEG & ITU-T VCEG,* JVT-Y027, Shenzhen, China, Oct. 2007.

[14] S. Shimizu et al., "View Scalable Multiview Video Coding Using 3-D Warping with Depth Map," *IEEE Trans. Circuits and Syst. for Video Technol.*, vol. 17, no. 11, Nov. 2007, pp. 1485-1495.

[15] Y. Liu et al., "Low-Delay View Random Access for Multi-view Video Coding," *Proc. IEEE Int'l Symp. on Circuits and Syst. (ISCAS)*, New Orleans, USA, May 2007, pp. 997-1000.

[16] R. Kawada, "KDDI Multiview Video Sequences for MPEG 3DAV," *ISO/IEC JTC1/SC29/WG11*, M10533, Munich, Germany, Mar. 2004.

[17] A. Vetro et al., "Multiview Video Test Sequences from MERL for the MPEG Multiview Working Group," *ISO/IEC JTC1/SC29/WG11*, M12077, Busan, Korea, Apr. 2005.

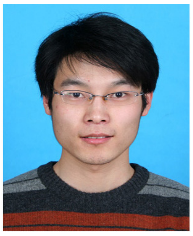
[18] Y. Zhang et al., "An Approach to Multi-modal Multi-view Video Coding," *Proc. of Int'l Conf. Signal Processing (ICSP)*, vol. 2, Guilin China, Nov. 2006. pp. 1405-1408.

[19] U. Fecker and A. Kaup, "Statistical Analysis of Multi-Reference Block Matching for Dynamic Light Field Coding," *Proc. Vision, Modeling and Visual. (VMV)*, Erlangen, Germany, Nov. 2005, pp. 445-452.

[20] K. Sohn et al., "Results on CE1 for Multi-view Video Coding," *ISO/IEC MPEG & ITU-T VCEG*, JVT-T102, Klagenfurt, Austria, July 2006.

[21] H. Schwarz, D. Marpe, and T. Wiegand, "Hierarchical B Pictures," *ISO/IEC MPEG & ITU-T VCEG*, JVT-P014, Poznan, Poland, July 2005.

[22] Y. Zhang, M. Yu, and G.Y. Jiang, "Evaluation of Typical Prediction Structures for Multi-view Video Coding," *ISAST Trans. Electronics and Signal Processing*, vol. 2, no. 1, 2008, pp. 7-15.



Yun Zhang received his BS and MS degrees in information science and engineering from Ningbo University, China, in 2004 and 2007, respectively. He is currently a PhD candidate with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), China. His research interests mainly include digital video compression and communications, multi-view video coding and content-based video processing.

Gang Yi Jiang received his MS degree from Hangzhou University in 1992, and received his PhD degree from Ajou University, Korea, in 2000. He is now a professor with the Faculty of Information Science and Engineering, Ningbo University, China. His research interests include digital video compression and communications, multi-view video coding, and image processing.

Mei Yu received her MS degree from Hangzhou Institute of Electronics Engineering, China, in 1993, and her PhD degree from Ajou University, Korea, in 2000. She is now a professor with the Faculty of Information Science and Engineering, Ningbo University, China. Her research interests include image and video coding as well as video perception.

Yo Sung Ho received the BS and MS degrees in electronic engineering from Seoul National University, Korea, in 1981 and 1983, respectively, and the PhD degree in electrical and computer engineering from University of California, Santa Barbara, in 1990. He joined the Electronics and Telecommunications Research Institute (ETRI), Korea, in 1983. From 1990 to 1993, he was with Philips Laboratories, Briarcliff Manor, New York. In 1993, he rejoined the technical staff of ETRI. Since 1995, he has been with Gwangju Institute of Science and Technology (GIST), where he is currently a professor with the Information and Communications Department. His research interests include digital image and video coding, image analysis and image restoration, advanced source coding techniques, 3-D television, and realistic broadcasting.