Content-Aware Prediction Algorithm With Inter-View Mode Decision for Multiview Video Coding

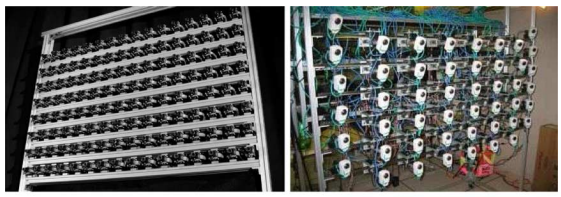
Li-Fu Ding, Pei-Kuei Tsung, Shao-Yi Chien, *Member, IEEE*, Wei-Yin Chen, and Liang-Gee Chen, *Fellow, IEEE*

***Abstract*—3-D video will become one of the most significant video technologies in the next-generation television. Due to the ultra high data bandwidth requirement for 3-D video, effective compression technology becomes an essential part in the infrastructure. Thus multiview video coding (MVC) plays a critical role. However, MVC systems require much more memory bandwidth and computational complexity relative to mono-view video coding systems. Therefore, an efficient prediction scheme is necessary for encoding. In this paper, a new fast prediction algorithm, content-aware prediction algorithm (CAPA) with inter-view mode decision, is proposed. By utilizing disparity estimation (DE) to find corresponding blocks between different views, the coding information, such as rate-distortion cost, coding modes, and motion vectors, can be effectively shared and reused from the coded view channel. Therefore, the computation for motion estimation (ME) in most view channels can be greatly reduced. Experimental results show that compared with the full search block matching algorithm (FSBMA) applied to both ME and DE, the proposed algorithm saves 98.4–99.1% computational complexity of ME in most view channels with negligible quality loss of only 0.03–0.06 dB in PSNR.**

***Index Terms*—3-D video, disparity estimation, H264/AVC, motion estimation, multiview video coding.**

I. Introduction

MULTIVIEW video can provide users with a sense of complete scene perception by transmitting several views to the receivers simultaneously. It can give users a vivid information about the scene structure. Moreover, it can also provide the capability of 3-D perception by respectively showing two of these frames to the eyes. With the technology of 3D-TV [1], [2] and free viewpoint TV (FTV) [3]–[5] getting more and more mature, multiview video coding (MVC) draws more and more attention. Besides, some multiple camera arrays have also been proposed for 3D video applications [6], [7] as shown in Fig. 1. In recent years, JVT/MPEG 3D auido/video (3DAV) group has worked toward the standardization for MVC [8], which also advances the multiview video applications. From the discussion in JVT/MPEG 3DAV meetings, the developed coding scheme for multiview video settings mainly uses H.264/AVC with exploiting temporal and inter-view dependencies [9]. That is, many coding tools of MVC in the related research area are based on the hybrid coding scheme and highly related to H.264/AVC [10].



(a) (b)

Fig. 1. Multiple camera arrays that have been built. (a) 128-camera array [6]. (b) Self-configurable camera array [7].

Although MVC is an emerging technology, huge amount of video data and ultra high computational complexity make it difficult to be realized. An H.264/AVC encoder requires computing power of about 1.3 tera-operations/second (TOPS) on a general-purpose processor to encode single-view HDTV720p videos (1280×720, 30 frames/second) in real time [11]. Different from mono-view video coding, disparity estimation (DE) is also utilized to reduce inter-view redundancy in MVC. Many DE algorithms have been proposed [12]–[15] to enhance the quality of the depth map for view synthesis or other intelligent video processing. Taking coding efficiency into consideration, block-based DE, like motion estimation (ME), is more appropriate for MVC because it has better compatibility with the existing video coding standards. Consequently, the prediction part, which consists of ME and DE, becomes the most computationally intensive part in an MVC system. Taking a three-view coding structure shown in Fig. 2 as an example, an instruction profiling of this coding structure is analyzed, as shown in Table I. It shows that the prediction part occupies 95% computational complexity in an MVC system. The proportion is even higher in some MVC systems with more complex coding structures. Therefore, ultra high computational complexity is a critical design challenge for MVC, especially in the prediction part.

In an MVC system, ME removes the temporal redundancy while DE removes the inter-view redundancy. Because of the setup structure of multiple cameras, there is close relation between motion vectors and disparity vectors in neighboring

Manuscript received November 14, 2007; revised June 20, 2008. Current version published December 10, 2008. This work was supported in part by the National Science Council, Taiwan, under Grant NSC96-2622-E-002-012-CC3 and by the scholarship of Hsing Tian Kong Culture and Education Development Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kiyoharu Aizawa..

L.-F. Ding, P.-K. Tsung, W.-Y. Chen, and L.-G. Chen are with the DSP/IC Design Lab, Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: lifu@video.ee.ntu.edu.tw; iceworm@video.ee.ntu.edu.tw; wychen@video.ee.ntu.edu.tw; lgchen@video.ee.ntu.edu.tw).

S.-Y. Chien is with the Media IC and System Laboratory, Graduate Institute of Electronics Engineering, Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: sychien@cc.ee.ntu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2008.2007314

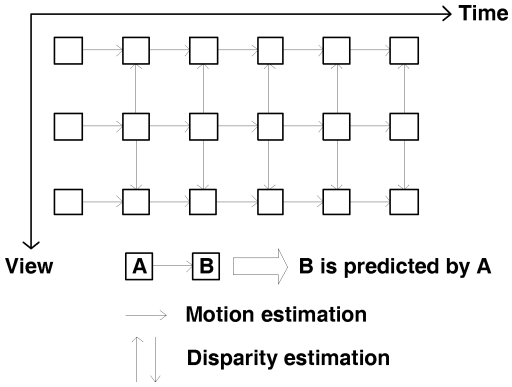


Fig. 2. Illustration of an example of a three-view coding structure. The white blocks represent frames, and the arrows represent prediction directions.

frames. The correlation is shown in Fig. 3. It can be described as [16], [17]

$$DV_{k-1} + MV_R = MV_L + DV_k. \tag{1}$$

By utilizing the correlation between motion and disparity fields, some new coding methods for MVC have been proposed [18]–[20]. Guo *et al.* have proposed an inter-view motion model to model the temporal motions at different views [18]. "Inter-view direct mode" has been introduced to enhance the coding efficiency. Although the target of 4%–6% bit-rate saving is achieved, the complexity is increased due to additional global DE. On the other hand, according to the correlation between views, another kind of redundancy called "computational redundancy" exists in addition to temporal and inter-view redundancies. Based on this concept, a fast prediction algorithm has been proposed to save the computation of ME for stereo video coding in our previous work [19]. However, the coding structures in MVC are more complex than that in stereo video coding. Besides, the previous work cannot deal with variable-block-size ME and complex mode decision. Moreover, Lai *et al.* have proposed predictive fast motion and disparity search. They track along the first estimated field (disparity/motion field) to get candidate vectors for the other field (motion/disparity field) [20]. Great computational complexity is saved with quality loss of 0.1–0.2 dB in PSNR. However, the ME with variable block size is not taken into consideration in their predictive search as well. In summary, all of the previous work only reuses motion or disparity vectors from other views or time slots. There are still some inter-view coding information, such as rate-distortion cost and coding modes. They can be further adopted for complex mode decision and complexity reduction.

TABLE I
INSTRUCTION ANALYSIS OF AN MVC ENCODER WITH THE CODING STRUCTURE SHOWN IN FIG. 2

Functions[a]	MIPS[b]	Percentage
Integer-pel ME	229180.6	75.4%
Integer-pel DE	38386.8	12.6%
Fractional-pel ME/DE	21396.6	7.0%
Others[c]	15183.1	5.0%
Total	304147.1	100.0%

[a]The encoding parameters are QVGA, 30 frames/s, ME with search range of [-32, +31] in both vertical and horizontal directions, DE with search range of [-32, +31]/[-8, +7] in the horizontal/vertical direction, and QP=20.

[b]MIPS stands for million instructions per second.

[c]Other modules include Lagrangian mode decision, intra prediction, variable length coding, transform & quantization, and deblocking filter, and so on.

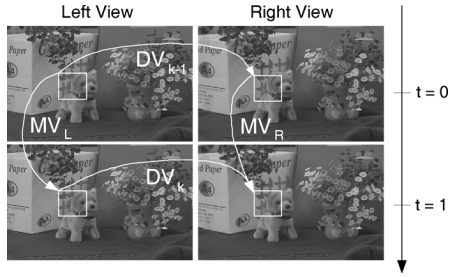


Fig. 3. Rrelation between disparity vectors and motion vectors.

In this paper, a new fast prediction algorithm, content-aware prediction algorithm (CAPA) with inter-view mode decision, is proposed for MVC. Based on the fact that the video contents are highly related between view channels, the proposed algorithm greatly reduces computational complexity while maintains video quality. The remainder of the paper is organized as follows. Section II describes the analysis of macroblock prediction modes in MVC. Next, the proposed CAPA is presented in Section III. Section IV shows the simulation results. Finally, Section V concludes this paper.

II. ANALYSIS OF MACROBLOCK PREDICTION MODES IN MVC

First, the coding structure of MVC is introduced. Many coding structures have been evaluated [21]. However, there

$$\mathrm{Cost}_{\mathrm{ME}} = \min_{B'_{r,\mathrm{ME}} \in SW_{r,\mathrm{ME}}(B_r)} \left\{ \sum_{(k,k') \in (B_r, B_{r,\mathrm{ME}})} |I_{r,t}(k) - I_{r,t-1}(k')| + \lambda \cdot Rate \right\} \tag{2}$$

$$\mathrm{Cost}_{\mathrm{DE}} = \min_{B'_{l,\mathrm{DE}} \in SW_{l,\mathrm{DE}}(B_r)} \left\{ \sum_{(k,k') \in (B_r, B'_{l,\mathrm{DE}})} |I_{r,t}(k) - I_{l,t}(k')| + \lambda \cdot Rate \right\} \tag{3}$$

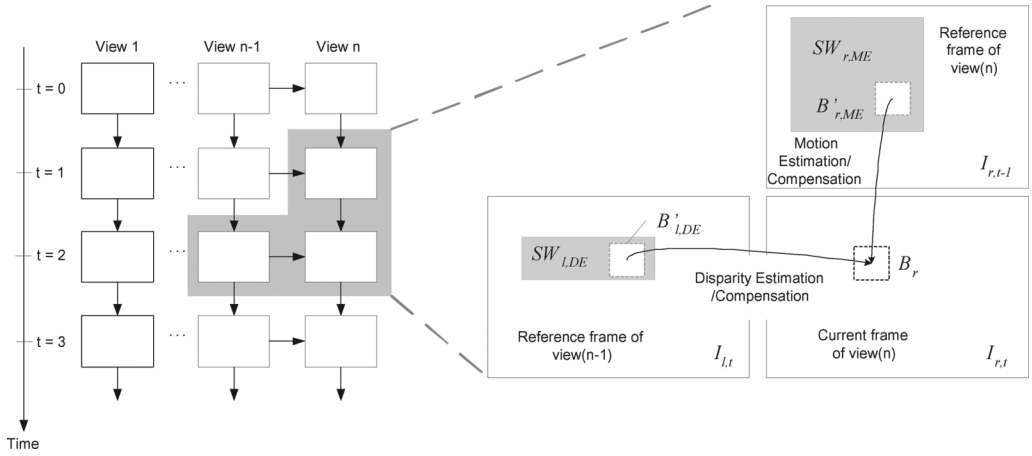


Fig. 4. Illustration of an MVC structure. The arrows represent the prediction directions, and the gray regions are the search windows for B_r.

is no unique coding structure which is appropriate for every video sequence. The selection of coding structures highly relies on the video contents and the corresponding camera setup. Fig. 4 shows the illustration of one coding structure, where the prediction directions of ME and DE are represented by arrows. For convenience of interpretation, the view channel $n-1$ is regarded as the left channel, and the view channel n is regarded as the right channel. There are two types of compensated blocks. They are the motion-compensated blocks and the disparity-compensated blocks, which are illustrated as $B'_{r,\mathrm{ME}}$ and $B'_{l,\mathrm{DE}}$ in Fig. 4, respectively. According to Lagrangian mode decision, the best type of compensated blocks is selected. For each macroblock in the current frame, the costs of ME and DE are computed by (2)–(3), as shown at the bottom of the previous page, where $\mathrm{Cost}_{\mathrm{ME}}$ and $\mathrm{Cost}_{\mathrm{DE}}$ are the minimum costs of motion-compensated and disparity-compensated blocks, respectively. B_r is the current block in the right channel. $B'_{r,\mathrm{ME}}$ is a block of the reference frame in the right channel. $B'_{l,\mathrm{DE}}$ is a block of the reference frame in the left channel. $SW_{r,\mathrm{ME}}(B_r)$ and $SW_{l,\mathrm{DE}}(B_r)$ are the search windows for the current block B_r. After ME and DE, the best matched blocks in the two search windows can be derived. Then the final prediction mode can be decided by selecting the one with lower cost.

There are several prediction modes defined in H.264/AVC standard. In our analysis of mode distribution, the prediction modes are classified into four categories, that is, INTER_ME, INTER_DE, INTRA, and SKIP modes. As shown in Fig. 5, the current macroblock can be predicted by ME from the reference frame in the same view channel, where INTER_ME mode can remove temporal redundancy. On the other hand, INTER_DE mode can remove inter-view redundancy by DE from the reference frame in the neighboring view channel. If the inter prediction cannot predict well, INTRA mode can predict the current macroblock by utilizing boundary pixels in the neighboring macroblocks. Moreover, SKIP mode utilizes the motion vector predictor to predict the current macroblock without performing inter prediction. It not only reduces the computational complexity but also saves the coding bits for motion vectors. The mode decision between INTER_ME and INTER_DE is closely

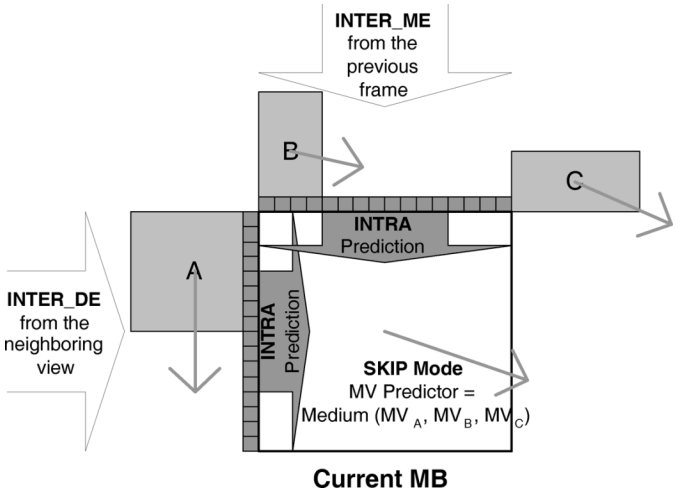
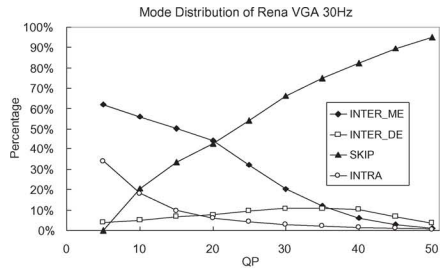


Fig. 5. Current macroblock can be predicted by various prediction modes. These prediction modes are classified into four categories: INTER_ME, INTER_DE, INTRA, and SKIP modes.

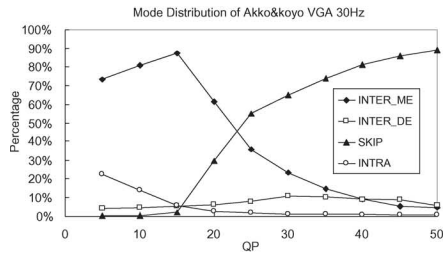
related to video contents [19]. Therefore, the mode classification can reflect the features of video contents.

According to the classification, Fig. 6 shows the mode distribution with various quantization parameters (QPs). It shows that the distribution of INTER_ME and SKIP mode has larger variation with various QPs. SKIP mode is the dominant mode at lower bit-rates (high QPs), while INTER_ME and INTRA modes are dominant at higher bit-rates (low QPs). The distribution is similar to that in mono-view video coding. The main difference between mono- and multiview video coding is INTER_DE mode, which is used in 5–10% macroblocks in a frame. The percentage of INTER_DE-mode macroblocks relies on the video contents. As shown in Fig. 6(c), the moving objects are usually predicted by INTER_DE because INTER_ME cannot predict well in the areas.

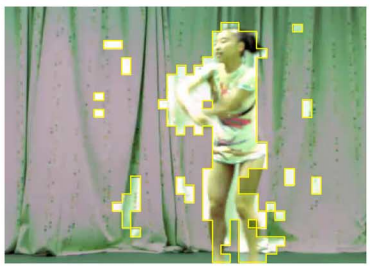
It is observed that certain types of macroblocks which are originally encoded by INTRA mode in mono-view video coding are encoded by INTER_DE mode in MVC. Fig. 7 shows the statistics of the ratio that INTRA mode is replaced



(a)



(b)



(c)

Fig. 6. Mode distribution analysis of two test sequences. Four main prediction modes are analyzed for sequences (a) "Rena" and (b) "Akko & Kayo." (c) Illustration of the compensated block types. The highlighted blocks with yellow boundaries are predicted by INTER_DE mode.

by INTER_DE mode after applying DE. In the cases of median and low bit-rates, over 50% macroblocks which are originally INTRA-coded in mono-view video are instead predicted by INTER_DE mode in MVC. The video contents of these macroblocks usually contain objects with fast motion or occlusions, as shown in Fig. 7(b) and (d). In summary, in an MVC system, the most general types of video contents are predicted by INTER_ME and SKIP modes, while INTRA mode can predict the macroblocks which contain more homogeneous or textural video contents. In addition, some complex video contents with fast moving objects or occlusions can be coded with INTER_DE mode.

It is also observed that the mode distribution of two views are very similar. In mono-view video coding, there are many coding tools adopted to extract data redundancies and remove them. When the video contents are extended from single view to multiple views, another data redundancy appears. On the conditions that cameras are setup with close parallelized structure, the video contents of different views are usually similar. This inter-view similarities exist not only in the video contents but also in the prediction modes. An example is shown in Fig. 8. Two views are encoded separately by an H.264/AVC encoder. The macroblock partition is marked on the reconstructed frames. Black and white blocks represent inter- and intra-predicted blocks respectively. It shows that the inter-view correlation is high. In other words, if the correlation is successfully explored, the computational complexity of the prediction part in MVC can be greatly reduced.

The analysis of macroblock prediction modes for MVC is summarized as follows.

- SKIP mode provides good coding performance and requires little computational complexity in both mono- and multiview video coding.
- INTRA mode tends to be replaced by INTER_DE mode in MVC when the current macroblock contains objects with fast motion or occlusions.
- In most cases, the video contents are similar between view channels. It results in the similar mode distribution between view channels. Therefore, there exists computational redundancy, and the inter-view information can be obtained with DE to predict the coding information. An efficient prediction algorithm with content-aware functionalities can effectively save the unnecessary computation.

III. Proposed Content-Aware Prediction Algorithm (CAPA) With Inter-View Mode Decision

According to the analysis in the previous section, the content-aware prediction algorithm (CAPA) with inter-view mode decision is proposed to save unnecessary computational load by exploiting the correlation between view channels. By utilizing various features of video contents in the coded view channels, macroblock coding modes and their corresponding motion vectors can be predicted with the aid of DE and the coding information of neighboring views. Therefore, ME computational complexity can be greatly reduced. In this section, the system architecture of the multiview hybrid coding system is introduced first, followed by the details of the proposed algorithm.

A. System Architecture

The block diagram of the multiview video encoder with the proposed CAPA is shown in Fig. 9. The encoder adopts the coding tools defined in H.264/AVC standard. Input views are classified into two types of view channels, the primary channel and the secondary channel. A view channel is regarded as a primary channel if no reconstructed frames in other view channels are used for reference when performing mode decision. Therefore, there are no DE operations in primary channels. The coding flow of a primary channel is identical to the flow of mono-view video coding. The block engine includes quantization, transform, and deblocking filter, etc. After the Lagrange mode decision, the coding information, including the rate-distortion costs, the optimum macroblock coding mode, and the corresponding motion vectors of the primary channel are stored for the proposed CAPA. The main difference between the primary and secondary channels is the CAPA part, which contains DE, twin-MB selection, inter-view mode decision, and content-aware ME. Each of them is introduced in the following subsections. In CAPA, DE is performed prior to ME. The purpose of performing DE first is to extract the correlation between

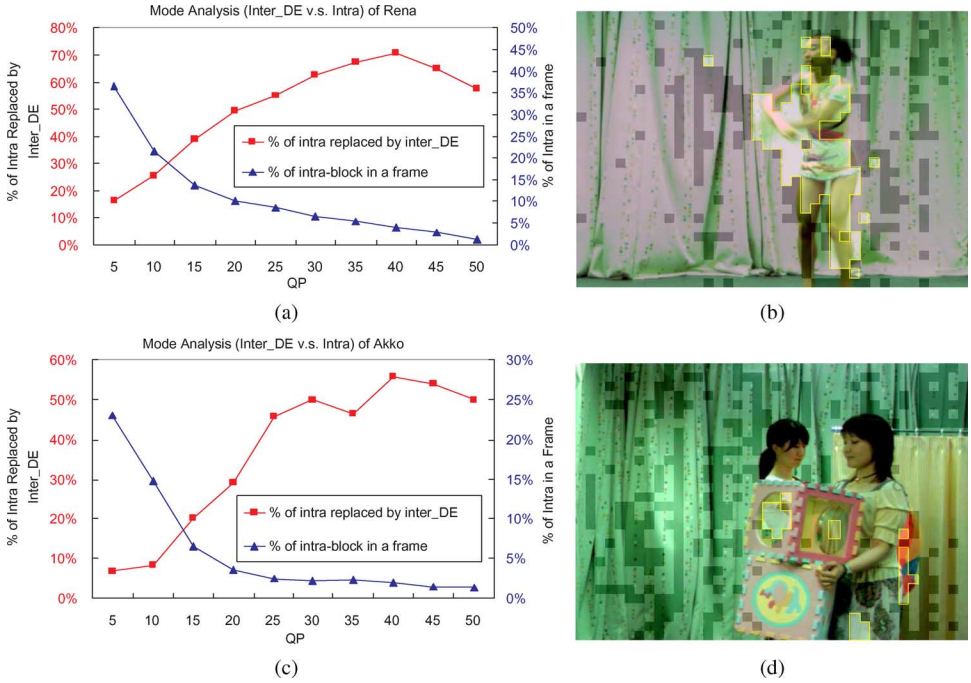


Fig. 7. Statistics of ratio that INTRA mode is replaced by INTER_DE mode after applying DE. Disparity compensated macroblocks which are originally coded with INTRA mode in mono-view video coding are highlighted in subjective views (areas with yellow boundaries). Dark macroblocks are coded with INTRA mode in both cases of mono- and multiview video coding. (a) Mode analysis of "Rena." (b) Coding mode illustration of (a) with $QP = 10$. (c) Mode analysis of "Akko&Kayo." (d) Coding mode illustration of (c) with $QP = 5$.

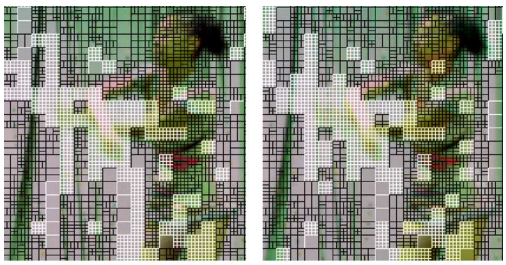


Fig. 8. Illustration of macroblock partition after variable-block-size ME. Two views are independently encoded without DE. (a) Left view. (b) Right view.

views, and the coding information of the corresponding neighboring coded view can be retrieved. With the corresponding coding information, the inter-view mode decision part decides the most probable coding mode for the current macroblock. The most probable coding mode is one of the modes described in Section II, that is, INTER_ME, INTER_DE, SKIP, or INTRA mode. If INTER_ME is chosen as the most probable coding mode by inter-view mode decision, it means ME is further required for better coding efficiency of the current macroblock, and thus proposed content-aware ME is performed. Content-aware ME is a predictor-centered ME algorithm, and it also utilizes the inter-view coding information. Note that, the numbers of primary and secondary channels are decided according to the coding structure. The numbers of secondary channels are normally much more because DE can effectively enhance coding efficiency [9]. After all views are encoded, the compressed bitstream of each channel is assembled and transmitted.

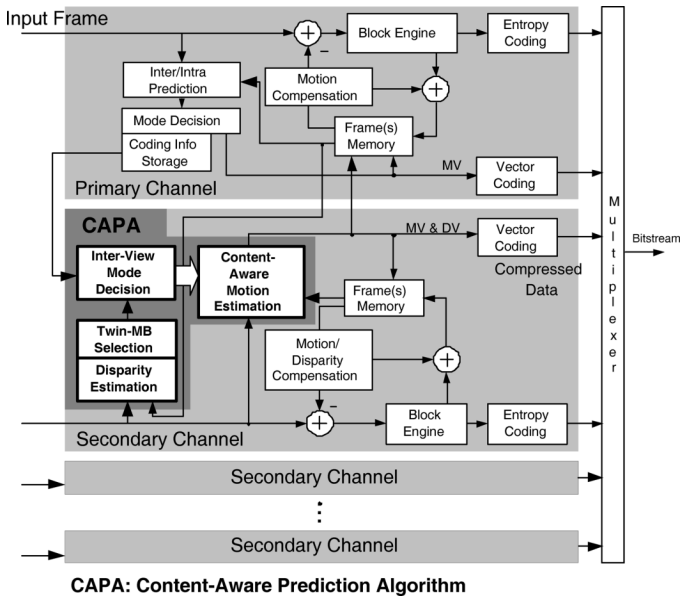


Fig. 9. Block diagram of the multiview video encoder with the proposed CAPA.

B. Disparity Estimation and Selection of Twin-Macroblock

DE is performed between the reference frame in the primary channel and the current macroblocks in the secondary channel. The minimum rate-distortion cost, $\mathrm{Cost}_{\mathrm{DE}}$, is decided by (3). The macroblocks in the primary channel are overlapped by the corresponding best matched block indicated by a disparity vector. And among the macroblocks the one with the largest overlapped area is called the "twin-macroblock," as shown

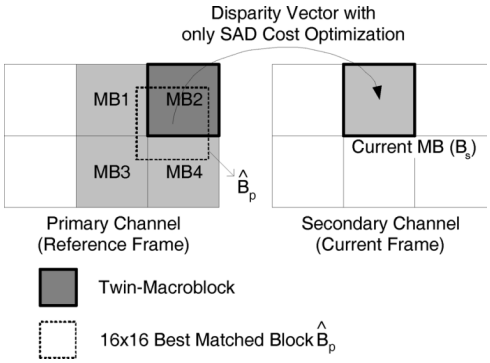


Fig. 10. Twin-macroblock is the macroblock in the primary channel which is overlapped by the corresponding best matched disparity compensated block with the largest overlapped area. In this case, MB2 is the twin-macroblock.

in Fig. 10. To predict the most probable coding mode for the current macroblock in a secondary channel, it is required to retrieve the related coding information from the twin-macroblock in the coded primary channel. As illustrated in Fig. 10, when performing DE, the corresponding best matched block indicated by the disparity vector is derived by

$$\widehat{B_p} = \arg \min_{B'_p \in SW_p(B_s)} \left\{ \sum_{(k,k') \in (B_s, B'_p)} |I_s(k) - I_p(k')| \right\} \tag{4}$$

where B_s represents the current macroblock in a secondary channel I_s, and B'_p represents the search candidates located in the search window in the primary channel I_p. Therefore, MB2 is regarded as the corresponding twin-macroblock of the current macroblock in Fig. 10. Note that the rate part of the block-matching cost is not considered here for deriving the twin-macroblock in the primary channel. However, the best disparity compensated block for coding can still be searched by Lagrangian mode decision at the same time without additional computation. The coding information of the twin-macroblock is then stored for the following inter-view mode decision and content-aware ME.

C. Inter-View Mode Decision

After the selection of a twin-macroblock, the coding information of the twin-macroblock, which includes the rate-distortion cost, the optimum macroblock coding mode, and the corresponding motion vectors, is retrieved. The purpose of inter-view mode decision is to choose the most probable coding mode among INTER_ME, INTER_DE, SKIP, and INTRA modes. SKIP mode is a useful and simple coding tool in H.264/AVC, where the motion vector predictor is adopted for the current macroblock to generate a compensated block. Therefore, the ME computation of a macroblock can be entirely saved if SKIP mode can be pre-decided. SKIP mode is also effective in the multiview video encoder. On the other hand, INTRA mode is chosen by a lot of macroblocks in a frame in the condition of high bit-rate, as shown in Fig. 6. Therefore, if INTRA mode can be pre-decided, many computation operations for ME can be saved by utilizing the correlation between views. In short, the unnecessary computation for ME can be saved if SKIP, INTRA, or INTER_DE mode is chosen. Therefore, inter-view

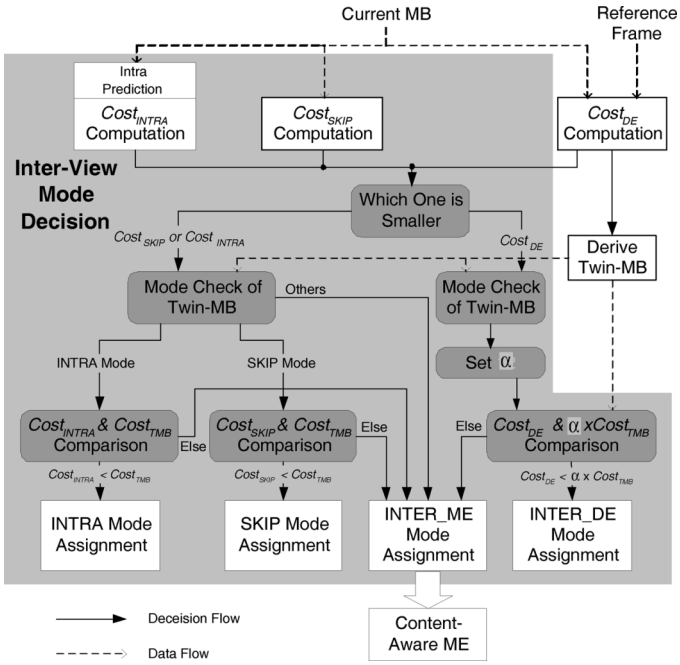


Fig. 11. Decision and data flows of inter-view mode decision.

mode decision can be regarded as an early termination scheme for the following ME.

Fig. 11 shows the decision and data flow of the proposed inter-view mode decision. Solid lines and grey blocks represent the decision flow, and dotted lines represent the data flow. First, intra prediction is performed, and the optimum rate-distortion cost among several modes in intra prediction, $\mathrm{Cost_{INTRA}}$, is derived. It is followed by the cost computation of SKIP mode, and $\mathrm{Cost_{SKIP}}$ is then derived. $\mathrm{Cost_{INTRA}}$ and $\mathrm{Cost_{SKIP}}$ are compared with $\mathrm{Cost_{DE}}$, which is derived from DE. If $\mathrm{Cost_{SKIP}}$ or $\mathrm{Cost_{INTRA}}$ is the smallest, the mode of the twin-macroblock is checked. If the twin-macroblock is also coded by the same mode, it implies that it has high possibility to be the best coding mode for the current macroblock. Then the rate-distortion cost of the twin-macroblock, $\mathrm{Cost_{TMB}}$, is compared with $\mathrm{Cost_{SKIP}}$ or $\mathrm{Cost_{INTRA}}$. Finally, the current macroblock is predicted by SKIP/INTRA mode if $\mathrm{Cost_{SKIP}}$/$\mathrm{Cost_{INTRA}}$ is still smaller than $\mathrm{Cost_{TMB}}$. Otherwise, the current macroblock is assigned to INTER_ME mode and the following content-aware ME is performed.

On the other hand, if $\mathrm{Cost_{DE}}$ is the smallest among three coding modes, the data flow is different from the other cases. According to the analysis introduced in Section II, macroblocks contain objects with fast motion tends to be encoded by INTER_DE mode. In addition, INTRA mode tends to be replaced by INTER_DE mode in MVC when the current macroblock contains objects with fast motion or occlusions. Based on these two concepts, the coding modes of the twin-macroblock and the neighboring coded macroblocks are utilized and checked. A parameter, $\mathrm{TMB_{INTRA}}$, is defined as follows,

$$\mathrm{TMB_{INTRA}} = \begin{cases} 1, & \text{if the twin-macroblock is INTRA-coded,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

TABLE II
DEFINITION OF THE PARAMETER SET α

TMB_{INTRA}	$NMBC_n$	α
$TMB_{INTRA} = 1$	0	α_0
	1	α_1
	2	α_2
	3	α_3
$TMB_{INTRA} = 0$	0	α_4
	1	α_5
	2	α_6
	3	α_7

$\alpha_0 < \alpha_1 < \alpha_2 < \alpha_3, \alpha_4 < \alpha_5 < \alpha_6 < \alpha_7$;
$\alpha_0 > \alpha_4, \alpha_1 > \alpha_5, \alpha_2 > \alpha_6, \alpha_3 > \alpha_7$.

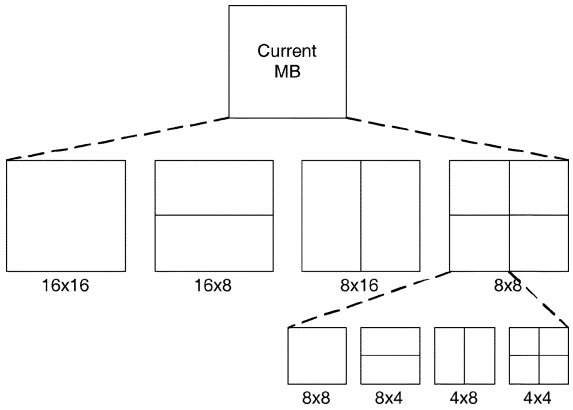


Fig. 12. Macroblock partition rule defined in H.264/AVC. It follows the hierarchical and symmetric manner.

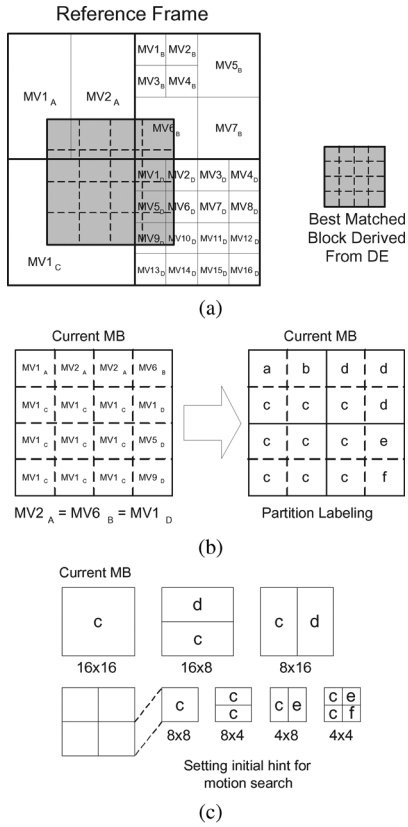


Fig. 13. Illustration of inter-view motion vector prediction. (a) Coding information of the reference frame to be used is stored in the memory. The 16×16 area of the best matched block derived from DE is split into sixteen 4×4 subblocks. (b) Each subblock is assigned a motion vector. (c) An initial guess of the motion vector, CAPA motion vector predictor, is set for each macroblock type according to partition labeling.

$\mathrm{TMB_{INTRA}}$ shows whether the twin-macroblock is INTRA-coded or not. Then INTER_DE mode assignment is described by the following equation:

$$Mode = \begin{cases} \text{INTER\_DE}, & \text{if } \mathrm{Cost_{DE}} < \alpha \times \mathrm{Cost_{TMB}} \\ \text{INTER\_ME}, & \text{otherwise.} \end{cases} \quad (6)$$

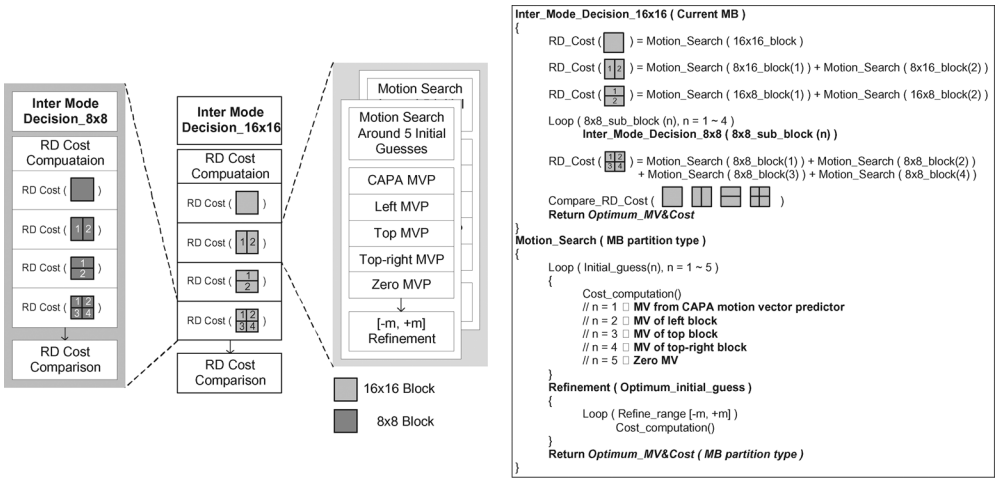
In the above equation, a parameter set, $\alpha \in \{\alpha_0, \alpha_1, \ldots, \alpha_7\}$, is defined to adjust the threshold of the early termination of ME, as shown in Table II. $NMBC_n$ stands for the count of neighboring macroblocks which are encoded by INTER_DE mode. The neighboring macroblocks are the left, top, and top-right macroblocks relative to the current macroblock. Therefore, $NMBC_n \in \{0, 1, 2, 3\}$. The value of α relies on $\mathrm{TMB_{INTRA}}$ and $NMBC_n$. α is bigger in the case that the twin-macroblock is coded by INTRA mode. Similarly, the more neighboring macroblocks coded by INTER_DE mode are, the bigger α is. In our simulation, the values of $\alpha_0, \alpha_1, \ldots, \alpha_7$ are empirically chosen between 0.1 and 2.0, that is

$$\{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\} = \{0.4, 1.0, 1.5, 2.0, 0.1, 0.7, 1.2.1.7\}.$$

Therefore, the coding mode can be decided after inter-view mode decision. If the current macroblock is assigned to INTRA, SKIP, or INTER_DE mode, the following ME operation can be skipped. Otherwise, content-aware ME is performed.

D. Content-Aware Motion Estimation

To further reduce the computational complexity of ME in the secondary channels, content-aware ME is proposed. As shown in Fig. 12, there are seven macroblock partition types according to their block sizes such as 16×16, 8×8, and 4×4, etc. in H.264/AVC [22]. Content-aware ME is proposed to predict the macroblock partition and the corresponding motion vectors of the current macroblock. The proposed algorithm consists of two parts, inter-view motion vector prediction and motion vector refinement with small search range. The illustration of inter-view motion vector prediction is shown in Fig. 13. After the frame in the primary channel is encoded, the coding information is stored in the memory. The location of the best matched block has already been derived from DE shown as the grey area in Fig. 13(a). The grey area is split into sixteen 4×4 subblocks. Each subblock covers a 4×4 area in the reference frame, and then it is assigned with the motion vector of the 4×4 area in the coded reference frame. Note that if the 4×4 area contains more than one different motion vectors, the assigned motion vector is the motion vector of the coded subblock with the largest overlapped area by the best matched block. To prevent prediction error propagation, there are not any early termination and fast prediction schemes applied in the primary channel, which means all kinds of cost must be calculated in the primary channel. Besides, no matter what kind of mode is selected for coding, the best inter prediction mode and its corresponding motion vectors are stored in the primary channel. Therefore, if the covered macroblock in the reference frame is predicted by INTRA or SKIP mode, the



(a) (b)

Fig. 14. (a) Data flow and (b) pseudocodes of searching the optimum motion vectors. Note that the motion search is performed for each macroblock type.

macroblock partition and its corresponding motion vectors are still available for the proposed algorithm.

After each 4×4 subblock is assigned a motion vector, the process of "partition labelling" begins. The subblocks with the same motion vectors are assigned to the same label, as the labels "a, b, c, d, e, f" shown in Fig. 13(b). Next, an initial guess of the motion vector, which is called "CAPA motion vector predictor," is set for each marcoblock type according to partition labelling. An example is shown in Fig. 13(c). To provide a good initial guess of a motion vector, the most representative value should be chosen. For example, when setting the CAPA motion vector predictor for a 16×16 block, the value of the label which appears the most times is chosen as an initial guess, which is "c" in Fig. 13(c).

In addition to the initial guess provided from inter-view coding information, the motion vectors of the left, top, top-right neighboring macroblocks, and zero motion vector, are also adopted as the initial guesses to enhance the coding efficiency. That is, for each macroblock type, there are five initial guesses of motion vectors. The optimum initial guess is chosen by Lagrange mode decision, and then the refinement with a small search range is performed around the optimum initial guess. Fig. 14 shows the data flow and the corresponding pseudocodes of searching the optimum motion vectors. Because the proposed algorithm is a predictor-centered ME, the required search candidates are much less than that for full search block matching algorithm (FSBMA). Therefore, computational complexity can be greatly reduced.

IV. Simulation Results

The proposed algorithm is implemented by modifying the MVC-configuration in JSVM4.5 [23]. Sequences "Akko&Kayo," "Ballroom," "Exit," and "Rena," with size 640×480 are tested. They are standard sequences released by JVT/MPEG 3DAV Group [9]. Two- and three-view channels of these sequences are chosen for simulation. The four coding structures adopted in our experiments are shown in Fig. 15. The grey blocks represent the frames in the primary channel, and the white blocks represent the frames in the secondary channels. The test condition is shown in Table III. The search ranges of DE and ME are both $[-32, +32]$ in the horizontal direction, while the search range of DE is $[-8, +8]$ in the vertical direction. The search range of DE is not a square because the cameras to capture these sequences are parallel-structured [9]. Thus the candidate blocks can be assumed in a belt-shape region [24]. Note that in the simulation of ME complexity, the index "search candidate per macroblock" is adopted to make the data independent of various hardware platforms.

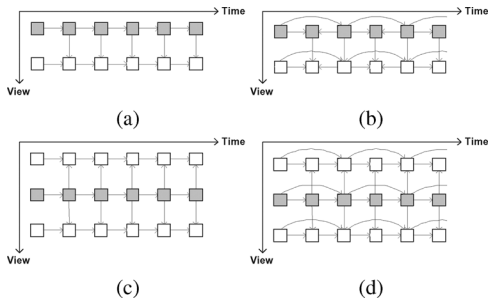


Fig. 15. Four coding structures for experiments. The grey blocks represent the frames in the primary channel, and the white blocks represent the frames in the secondary channels. (a) Two views with IPPP structure. (b) Two views with IBPBP structure. (c) Three views with IPPP structure. (d) Three views with IBPBP structure.

A. Quality and Complexity Analysis of Inter-View Mode Decision

Table IV shows the statistics of complexity reduction and PSNR degradation of the proposed inter-view mode decision. Only the computational complexity of ME in the secondary channel is considered. The computational complexity and coding performance of FSBMA are regarded as references for comparison. It shows that 20.4%–73.4% computation for ME is saved with only 0.015–0.035 dB quality loss in PSNR. It indicates that 20.4%–73.4% macroblocks are assigned to INTRA, SKIP, or INTER_DE mode, so the following ME is then skipped. Therefore, it can be claimed that the proposed

TABLE III
MULTIVIEW VIDEO SEQUENCES FOR THE EXPERIMENTS

Sequences	Frame size	ME search range [x, y]	DE search range [x, y]	Coding structure
Akko&Kayo	640×480	[±32, ±32]	[±32, ±8]	Fig.15 (a)
Ballroom	640×480	[±32, ±32]	[±32, ±8]	Fig.15 (b)
Exit	640×480	[±32, ±32]	[±32, ±8]	Fig.15 (c)
Rena	640×480	[±32, ±32]	[±32, ±8]	Fig.15 (d)

TABLE IV
COMPLEXITY REDUCTION AND QUALITY DROP OF INTER-VIEW MODE DECISION

Sequences	QP	Complexity Reduction	Quality Drop
Akko&Kayo	45	20.4%	–0.017 dB
Ballroom	30	61.8%	–0.032 dB
Exit	35	73.4%	–0.035 dB
Rena	40	47.5%	–0.015 dB

inter-view mode decision effectively executes the mode assignment for each macroblock.

B. Quality and Complexity Analysis of Content-Aware Motion Estimation

The proposed content-aware ME is a predictor-centered ME algorithm. The optimum initial guess is chosen among five initial guesses, and then the refinement with a small search range is performed around the initial guess. Table V shows the distribution of the initial guess which is chosen for further refinement. Over 80% macroblocks choose CAPA motion vector predictor for further refinement. It indicates that the proposed algorithm can provide an accurate initial guess. Table VI shows the analysis of quality and complexity with various refinement ranges. The larger the refinement range is, the less PSNR degradation is. It shows that $[\pm 4, \pm 4]$ refinement keeps PSNR drop within 0.05 dB in average. The motion vector distribution with $[\pm 8, \pm 8]$ refinement range is shown in Fig. 16. Most motion vectors are located within $[\pm 2, \pm 2]$ range. Therefore, $[\pm 2, \pm 2]$ and $[\pm 4, \pm 4]$ are appropriate choices of refinement ranges. In addition, no matter which refinement range is chosen, the required complexity is always much less than that of FSBMA.

C. Rate-Distortion Performance of Content-Aware Prediction Algorithm

The proposed CAPA consists of inter-view mode decision and content-aware ME. It is compared with multicast coding and simulcast coding. Multicast coding means FSBMA is applied to both ME and DE in the coding structures. On the other hand, in simulcast coding, each view channel is encoded independently without DE. Rate-distortion performance of only secondary channels are compared because the ME parts in the primary channels in all cases are implemented with FSBMA. Moreover, the refinement range is $[\pm 4, \pm 4]$. The rate-distortion performance is shown in Fig. 17. It shows that there is almost no quality difference between FSBMA and CAPA, and CAPA provides coding gain of 0.09–1.44 dB over simulcast coding. The comparison of quality and complexity among three coding schemes is shown in Table VII. Compared with multicast coding, CAPA reduces 98.4%–99.1% ME computation in secondary channels with PSNR drop of only 0.03–0.06 dB. In the

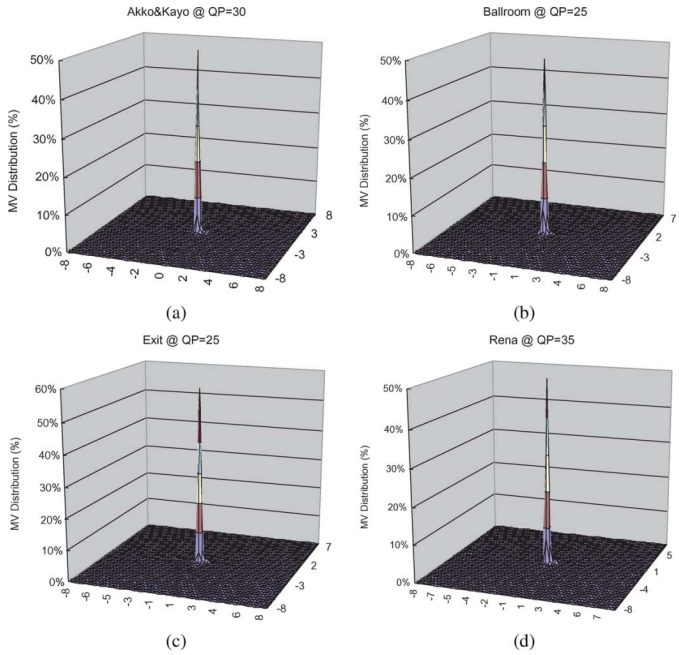


Fig. 16. Motion vector distribution of various test sequences. (a) "Akko & Kayo." (b) "Ballroom." (c) "Exit." (d) "Rena."

TABLE V
DISTRIBUTION OF THE INITIAL GUESS WHICH IS CHOSEN FOR FURTHER REFINEMENT

Sequence	Akko&Kayo	Ballroom	Exit	Rena
CAPA MV predictor	81.6%	88.9%	81.7%	86.2%
Left MV predictor	13.4%	9.1%	9.5%	9.7%
Top MV predictor	2.6%	1.0%	2.2%	1.6%
Top-right MV predictor	0.8%	0.4%	1.3%	0.7%
Zero MV predictor	1.6%	0.6%	5.3%	1.8%

previous work [19], [20], 80% computational complexity is reduced with quality loss of 0.1 dB in [19], and about 95% computational complexity can be reduced with quality loss of 0.1–0.2 dB in [20]. Compared with them, the proposed CAPA effectively removes more computational redundancy while maintains the coding performance.

In addition, taking the computation of ME in the primary channel and DE into consideration, 43.6%–56.2% complexity can be saved. Note that because the computational complexity of DE is 25% of that of ME in our simulation, the total computational complexity of the proposed algorithm is also much less than that of simulcast coding, and only 51.9%–64.1% computational complexity is required. The degree of reduction of the total computational complexity depends on the view numbers. Therefore, it means that the redundant ME computation can be effectively removed by the proposed algorithm. With the proposed CAPA, computational complexity can be greatly saved,

TABLE VI
QUALITY AND COMPLEXITY ANALYSIS FOR VARIOUS REFINEMENT RANGES

Refinement Range	PSNR Drop of Sequences (dB)				Complexity Comparison		
	Akko& Kayo	Ballroom	Exit	Rena	Search Candidates for CAPA	Search Candidates for FSBMA	Complexity Ratio
[±1, ±1]	–0.051	–0.088	–0.055	–0.119	91	29575	0.3%
[±2, ±2]	–0.038	–0.061	–0.052	–0.073	203	29575	0.7%
[±4, ±4]	–0.027	–0.048	–0.044	–0.058	595	29575	2.0%
[±8, ±8]	–0.024	–0.031	–0.030	–0.057	2051	29575	6.9%

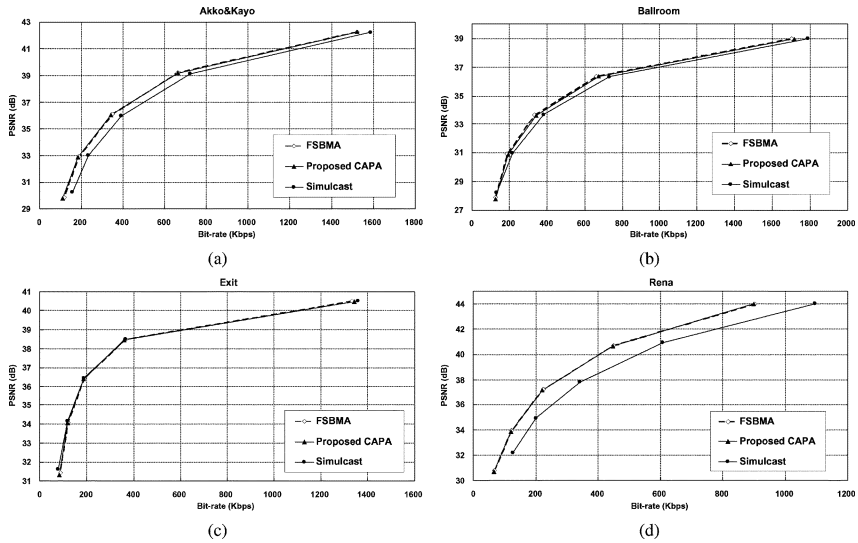


Fig. 17. Comparison of rate-distortion performance among proposed CAPA, FSBMA, and simulcast coding. (a) "Akko&Kayo." (b) "Ballroom." (c) "Exit." (d) "Rena."

TABLE VII
COMPARISON OF QUALITY AND COMPLEXITY REDUCTION RATIO BETWEEN THE PROPOSED ALGORITHM, MULTICAST CODING WITH FSBMA, AND SIMULCAST CODING WITH FSBMA

Sequences	Compare with Multicast with FSBMA			Compare with Simulcast with FSBMA	
	PSNR Drop	ME Complexity Reduction	Total Complexity Reduction	PSNR Gain	Total Complexity Ratio
Akko&Kayo	–0.03 dB	98.4%	44.5%	+0.65 dB	64.1%
Ballroom	–0.06 dB	98.7%	43.6%	+0.47 dB	63.7%
Exit	–0.04 dB	99.1%	56.2%	+0.09 dB	51.4%
Rena	–0.06 dB	98.7%	55.8%	+1.44 dB	51.9%

and near-FSBMA quality is maintained. That is, the inter-view correlation can be effectively exploited by the proposed algorithm, and then the computational redundancy is removed.

V. CONCLUSION

This paper presents an MVC encoder structure with an efficient fast prediction algorithm for the prediction part in MVC. Content-aware prediction algorithm (CAPA) with inter-view mode decision is proposed to overcome the design challenge of ultra high computational complexity in MVC. Based on the concept of high inter-view correlation between views and the feature of mode distribution different from that in mono-view video coding, unnecessary ME computation can be early terminated by inter-view mode decision. Moreover, accurate initial guesses are provided by content-aware ME. Only small refinement ranges, such as $[\pm 2, \pm 2]$ and $[\pm 4, \pm 4]$, are sufficient for maintaining comparable quality to FSBMA. The proposed algorithm effectively reduces 98.4%–99.1% computational complexity for ME in most view channels with negligible quality loss of 0.03–0.06 dB in PSNR. Compared with simulcast coding, the proposed algorithm provides coding gain of 0.09–1.44 dB with only 51.4%–64.1% computational complexity. It indicates that the computational redundancy is effectively removed.

There are still some extensions of the proposed algorithm. After adopting the proposed CAPA in MVC, DE will become the most computation-consuming part. Note that the proposed algorithm is orthogonal to other fast ME search algorithms such as three-step search [25], four-step search [26], and diamond search [27], etc. Therefore, appropriate fast prediction algorithms for DE is also worth developing. In addition, the proposed algorithm can also be further adopted in more complex coding structures, such as hierarchical bidirectional prediction, and eight- and sixteen-view structures. The required number of primary channels in a given coding structure is also an important research issue. Less primary channels reduce

total computational complexity burden while result in quality degradation. Moreover, the proposed CAPA effectively reduces most computational complexity, while it can also provide an accurate MV predictor to enhance bit-rate savings for MV coding. They are challenging research topics and also belong to our future work.

ACKNOWLEDGMENT

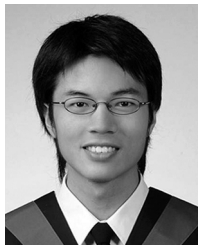
The authors would like to thank Dr. C.-J. Lian for the precious comments and Hsing Tian Kong Culture and Education Development Foundation for supporting the scholarship.

REFERENCES

[1] F. Isgrò, E. Trucco, P. Kauff, and O. Schreer, "Three-dimensional image processing in the future of immersive media," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 3, pp. 388–303, Mar. 2003.
[2] A. Smolic and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proc. IEEE*, vol. 93, no. 1, pp. 98–110, Jan. 2005.
[3] M. Tanimoto, "Free viewpoint television – FTV," in *Proc. 2004 Picture Coding Symp.*, Dec. 2004.
[4] T. Fujii and M. Tanimoto, "Free-viewpoint TV system based on ray-space representation," in *Proc. SPIE*, Mar. 2002, vol. 4864, pp. 175–189.
[5] A. Smolic, K. Mueller, P. Merkle, T. Rein, M. Kautmer, P. Eisert, and T. Wiegand, "Free viewpoint video extraction, representation, coding, and rendering," in *Proc. 2003 IEEE Int. Conf. Image Processing*, Oct. 2004, vol. 5, pp. 3287–3290.
[6] B. S. Wilburn, M. Smulski, H.-H. K. Lee, and M. A. Horowitz, "Light field video camera," in *Proc. Media Processors, SPIE Electronic Imaging*, 2002, vol. 4674, pp. 29–36.
[7] C. Zhang and T. Chen, "A self-reconfigurable camera array," in *Proc. Eurographics Symp. Rendering*, 2004.
[8] *Requirements on Multi-View Video Coding*, ISO/IEC JTC1/SC29/WG11 N6501, 2004.
[9] *Description of Core Experiments in MVC*, ISO/IEC JTC1/SC29/WG11 N6501 W8019, Apr. 2006.
[10] G. Li and Y. He, "A novel multi-view video coding scheme based on H.264," in *Proc. 2003 Joint Conf. 4th Int. Conf. Information, Communications and Signal Processing*, Dec. 2003, vol. 4, pp. 218–222.
[11] Y.-W. Huang, T.-C. Chen, C.-H. Tsai, C.-Y. Chen, T.-W. Chen, C.-S. Chen, C.-F. Shen, S.-Y. Ma, T.-C. Wang, B.-Y. Hsieh, H.-C. Fang, and L.-G. Chen, "A 1.3TOPS H.264/AVC single-chip encoder for HDTV applications," in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2005.
[12] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 7, no. 2, pp. 139–154, Mar. 1985.
[13] N. Grammalidis and M. G. Strintzis, "Disparity and occlusion estimation in multiocular systems and their coding for the communication of multiview image sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 6, pp. 328–344, Jun. 1998.
[14] J.-R. Ohm and K. Müller, "Incomplete 3-D multiview representation of video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 3, pp. 389–400, Mar. 1999.
[15] R.-S. Wang and Y. Wang, "Multiview video sequence analysis, compression, and virtual viewpoint synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 4, pp. 397–410, Apr. 2000.
[16] G. Heising, "Efficient and robust motion estimation in grid-based hybrid video coding schemes," in *Proc. Int. Conf. Image Processing*, 2002, pp. 687–700.
[17] Y. Luo, Z. Zhang, and P. An, "Stereo video coding based on frame estimation and interpolation," *IEEE Trans. Broadcast.*, vol. 49, no. 1, pp. 14–21, Jan. 2003.
[18] X. Guo, Y. Lu, and W. Gao, "Inter-view direct mode for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 12, pp. 1527–1532, Dec. 2006.
[19] L.-F. Ding, S.-Y. Chien, and L.-G. Chen, "Joint prediction algorithm and architecture for stereo video hybrid coding systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 11, pp. 1324–1337, Nov. 2006.
[20] P.-L. Lai and A. Ortega, "Predictive fast motion/disparity search for multiview video coding," in *Proc. SPIE Visual Communications and Image Processing*, 2006.
[21] *Results on CE1 for Multi-View Video Coding*, ISO/IEC JTC1/SC29/WG11 N6501 m13544, 2006.
[22] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with H.264/AVC: Tools, performance, and complexity," *IEEE Circuits Syst. Mag.*, vol. 4, no. 1, pp. 7–28, Jan. 2004.
[23] *AHG on Multiview Video Coding*, ISO/IEC JTC1/SC29/WG11 N6501 N7829, 2006.
[24] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video Processing and Communication*. Englewood Cliffs, NJ: Prentice-Hall, 2001.
[25] T. Koga, K. Linuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion-compensated interframe coding for video conferencing," in *Proc. NTC*, Nov. 1981, pp. C9.6.1–9.6.5.
[26] L. M. Po and W. C. Ma, "A new center-biased search algorithm for block motion estimation," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1995, vol. 1, pp. 23–26.
[27] S. Zhu and K. K. Ma, "A new diamond search algorithm for fast block matching motion estimation," *Inform., Commun., and Signal Process.*, pp. 9–12, Sep. 1997.

Li-Fu Ding was born in Keelung, Taiwan, in 1981. He received the B.S. degree in electrical engineering and the M.S. degree in electronics engineering in 2003 and 2005, respectively, from National Taiwan University, Taipei, Taiwan, where he is currently pursuing the Ph.D. degree in electronics engineering.

His major research interests include stereo and multiview video coding, motion-estimation algorithms, and associated VLSI architectures.



Pei-Kuei Tsung was born in Taipei, Taiwan, in 1984. He received the B.S. degree in electrical engineering and the M.S. degree in electronics engineering in 2006 and 2008, respectively, from National Taiwan University, Taipei, Taiwan, where he currently pursuing the Ph.D. degree in electronics engineering.

His major research interests include stereo and multiview video coding, motion estimation algorithms, view-synthesis algorithms, and associated VLSI architectures.

Shao-Yi Chien (S'00–M'04) was born in Taipei, Taiwan, in 1977. He received the B.S. and Ph.D. degrees from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, in 1999 and 2003, respectively.

During 2003 to 2004, he was a Member of Research Staff with Quanta Research Institute, Tao Yuan Shien, Taiwan. In 2004, he joined the Graduate Institute of Electronics Engineering and Department of Electrical Engineering, NTU, as an Assistant Professor. His research interests include video segmentation algorithm, intelligent video coding technology, image processing, computer graphics, and associated VLSI architectures.

Wei-Yin Chen (S'84–M'86–SM'94–F'01) was born in Penghu, Taiwan, in 1982. He received the B.S. degree in electrical engineering and the M.S. degree in electronics engineering from National Taiwan University, Taipei, in 2005 and 2008, respectively.

In 2007, he was a visiting graduate student at MIT, Cambridge, MA. His major research interests include super high-definition and multiview video coding, associated VLSI architectures, high-level synthesis, and computer architecture.

Liang-Gee Chen (F'01) was born in Yun-Lin, Taiwan, in 1956. He received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from National Cheng Kung University, in 1979, 1981, and 1986, respectively.

He was an Instructor (1981–1986), and an Associate Professor (1986–1988) in the Department of Electrical Engineering, National Cheng Kung University. In the military service during 1987 and 1988, he was an Associate Professor in the Institute of Resource Management, Defense Management College. In 1988, he joined the Department of Electrical Engineering, National Taiwan University (NTU). During 1993–1994, he was Visiting Consultant at DSP Research Department, AT&T Bell Labs, Murray Hill, NJ. IN 1997, he was a visiting scholar of the Department of Electrical Engineering, University of Washington, Seattle. Currently, he is Professor at NTU, and since 2004, he has also been the Executive Vice President and the General Director of Electronics Research and Service Organization (ERSO) at the Industrial Technology Research Institute (ITRI). His current research interests are DSP architecture design, video processor design, and video coding systems.

Dr. Chen is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING, and PROCEEDINGS OF THE IEEE. He was the Associate Editor of the *Journal of Circuits, Systems, and Signal Processing* from 1999 until present and served as the Guest Editor of *The Journal of VLSI Signal Processing Systems* for Signal, Image, and Video Technology in November 2001. He received the Best Paper Award from the R.O.C. Computer Society in 1990 and 1994 and received Long-Term (Acer) Paper Awards annually from 1991 to 1999. In 1992, he received the Best Paper Award of the 1992 Asia-Pacific Conference on Circuits and Systems in VLSI design track. In 1993, he received the Annual Paper Award of Chinese Engineer Society. In 1996, he received the Outstanding Research Award from NSC and the Dragon Excellence Award from Acer. He was elected as an IEEE Circuits and Systems Distinguished Lecturer from 2001–2002. He was the General Chairman of the 7th VLSI Design CAD Symposium and the 1999 IEEE Workshop on Signal Processing Systems: Design and Implementation. He is a member of Phi Tan Phi.