Journal of Electronic Imaging

JElectronicImaging.org

Toward enhancing the distributed video coder under a multiview video codec framework

Shih-Chieh Lee
Jiann-Jone Chen
Yao-Hong Tsai
Chin-Hua Chen

SPIE. IS&T imaging.org

Shih-Chieh Lee, Jiann-Jone Chen, Yao-Hong Tsai, Chin-Hua Chen, "Toward enhancing the distributed video coder under a multiview video codec framework," *J. Electron. Imaging* **25**(6), 063022 (2016), doi: 10.1117/1.JEI.25.6.063022.

Toward enhancing the distributed video coder under a multiview video codec framework

Shih-Chieh Lee,[a] Jiann-Jone Chen,[a,*] Yao-Hong Tsai,[b] and Chin-Hua Chen[a]
[a]National Taiwan University of Science and Technology, No. 43, Section 4, Keelung Road, Taipei 106, Taiwan
[b]Hsuan Chuang University, No. 34, Hsuan-Chuang Road, Hsinchu 300, Taiwan

Abstract. The advance of video coding technology enables multiview video (MVV) or three-dimensional television (3-D TV) display for users with or without glasses. For mobile devices or wireless applications, a distributed video coder (DVC) can be utilized to shift the encoder complexity to decoder under the MVV coding framework, denoted as multiview distributed video coding (MDVC). We proposed to exploit both inter- and intraview video correlations to enhance side information (SI) and improve the MDVC performance: (1) based on the multiview motion estimation (MVME) framework, a categorized block matching prediction with fidelity weights (COMPETE) was proposed to yield a high quality SI frame for better DVC reconstructed images. (2) The block transform coefficient properties, i.e., DCs and ACs, were exploited to design the priority rate control for the turbo code, such that the DVC decoding can be carried out with fewest parity bits. In comparison, the proposed COMPETE method demonstrated lower time complexity, while presenting better reconstructed video quality. Simulations show that the proposed COMPETE can reduce the time complexity of MVME to 1.29 to 2.56 times smaller, as compared to previous hybrid MVME methods, while the image peak signal to noise ratios (PSNRs) of a decoded video can be improved 0.2 to 3.5 dB, as compared to H.264/AVC intracoding. [DOI: 10.1117/1.JEI.25.6.063022]

Keywords: multiview video coding; distributed video coding; block-matching prediction; Wyner–Ziv coder; turbo coder; side information.

Paper 16598 received Jul. 13, 2016; accepted for publication Nov. 17, 2016; published online Dec. 20, 2016.

1 Introduction

Multiview video codec (MVC) design becomes popular,[1] based on which wide-spread applications, such as three-dimensional (3-D) video, free-viewpoint television (FTV), and video surveillance networks, can be developed. The 3-D video provides high quality and immersed multimedia entertainment that can be experienced through various channels, including movies, TV, internet, and so on. The FTV is a MVC system that allows viewpoint switching among different viewpoints, in which the video scene is captured by the camera from a specific view angle. For video surveillance networks, the MVC can be used to monitor and detect unusual events/objects. However, the MVC requires inter-sensor communication, which is expensive and not feasible in some applications. The information amount and required computational loading for a MVC codec is very large, compared to those of monoview. How to efficiently process and compress multiview videos is challenging. The joint video team has been working on the MVC, which captures videos from different video cameras and encodes these signals with reference to each other to yield a single bitstream. To enhance codec performance, most MVC schemes exploit correlations between both intraview and interview frames. At the encoder, it performs block motion compensation (MC) and disparity estimation to remove correlations between images along the intraview/temporal and interview video dimension to achieve high compression efficiency. Under this MVC framework, the time complexity of encoding operations would be high for efficient compression. It cannot provide low complexity encoding for applications like wireless video sensor/surveillance networks and low-power MVC capturing devices. The coding complexity has to be shifted to the decoder to make these applications feasible.

The distributed video coder (DVC)[2,3] was proposed to effectively shift coding complexity to the decoder, which can capture and encode signals from several low-power devices independently and jointly decode these signals. It can be extended to deal with multiview video signals,[4,5] in which the disparity information among images of different views can be exploited for removing correlations, in additional to correlations among intraview images. The DVC[2] was developed based on lossless distributed source coding, also known as the Slepian–Wolf coder (SWC)[6] for lossless coding. An important aspect of the SWC is that separated encoding can theoretically achieve the same compression ratio with joint encoding as long as the correlations among data streams are exploited by a joint decoder. This SWC framework was extended to process lossy compression with side information (SI) at the decoder,[7] as in the case of the Wyner–Ziv (WZ) coder. With the WZ coding algorithm, the DVC treats video compression as a channel coding problem. The input video of DVC is decomposed into odd and even sequences, in which the former is encoded as key frames (KFs) and the latter WZ frames (WZFs). The KFs are encoded with H.264/AVC[8] intramode, H.264/INTRA, and the WZFs are

*Address all correspondence to: Jiann-Jone Chen, E-mail: jjchen@mail.ntust.edu.tw

block-transformed, quantized, and transmitted through error correction codes in a bit-plane by bit-plane approach, in which only part of the parity bits are transmitted. At the decoder, the KFs are utilized to yield the SI a noisy WZF, which is the systematic part of an error correction code that co-operates with the received parity bits to correct channel errors. Compared to current video codec, the DVC effectively shifts a considerable amount of the coding complexity from the encoder to the decoder, which can also be applied to error resilience control[9] that treats the side information frame (SIF) as additional reference information, SI, to correct channel errors. Recently, a new distributed video codec based on modulo operation in the pixel domain has been proposed,[10] which demonstrates lower decoding complexity.

Integrating the MVC with a multiview distributed video coding (MDVC) would allow encoding several low-power capturing and encoding devices independently and decode these signals jointly. A view-synthesis and disparity-based correlation model that exploits interview video correlation is proposed to deliver error-resilient video in a distributed multicamera system.[11] One simple MDVC example with a left-, a right-, and a central-view camera is shown in Fig. 1. The left- (L) and right-view (R) videos are encoded and decoded by the traditional video codec, e.g., H.264/INTRA, to act as KFs (I frames) for the DVC decoding. The central-view video is encoded as interleaved one intra (I) and one WZF, i.e., group of picture ($|\mathrm{GOP}|$) $= 2$. At the decoder, the SI for a WZF can be estimated by exploiting the intra-view and interview image correlations, respectively. The decoded KFs are utilized to jointly reconstruct the WZFs, $\hat{I}_{2t}^{\mathrm{WZ}}$s, based on inter- and intraview image correlations. These correlations are utilized by assigning weights to different estimated motion vectors (MVs) exploited based on the MDVC framework. This decoder-driven fusion method is adopted to improve the codec performances, e.g., peak signal to noise ratios (PSNRs) and time complexity. In addition, the embedded DVC makes it feasible to setup low complexity, mobile encoders for multiview video acquisition to enable low delay and real-time processing of the MDVC. The decoder can consume the shifted computational complexity by setting a high performance computer for central decoding, e.g., large buffers, disk array, and high-speed CPUs.

Researches on improving MDVC SIF quality can be found by many.[12–14] An iterative SIF generation method uses decoded WZF to refine the SIF,[12] based on which the second iteration can enhance the quality of decoded images. By performing interpolation along intra- and interview video dimensions, respectively, to yield candidate SIFs, the final SIF can be fused from these candidate SIFs with a specific reliability measurement.[13] The interview interpolated candidate SIF for fusion can be enhanced by using a perspectively transformed one,[15,16] which can help to fuse better final SIFs and demonstrate better coding performance, as compared to monoview DVC. Three new fusion techniques that exploit signal properties of neighboring residual frames along intra- and interview direction were proposed for robustness and improving SIF quality.[17] The fusion can also adopt a support vector machine to identify a set of features for classifying pixels into either the temporal or the disparity class, by which the fusion can yield better SIF.[18] It provides a good solution for fusing intra- and interview predictions. However, these fusion methods suffer from performance degradation due to low temporally predicted quality and irregular video motion. An adaptive filtering view interpolation method[19,20] was proposed to minimize the difference between SIF and decoded KF, which can compensate for the intercamera mismatches and improve SIF quality. When occlusion exists between interview videos, the temporal frame interpolation is utilized to compensate for the deficiency of interview linear fusion[20] to improve SIF quality. Various SI generation methods are evaluated and compared for better utilization efficiency.

By estimating motion on interpolated frames, the irregular motion artifacts can be eliminated and the SIF quality can be improved.[21] One MDVC codec[22] was designed to transmit a small amount of error control information to replace an untransmitted frame and the information is obtained from a low-dimensional blockwise projection of the frame, i.e., mean-based projection. The most prominent feature of this work is that it is performed as a postprocessing step after decoding and interpolating the received video, which allows easy integration with various video transmission systems.

In the conventional video codec, it usually adopts the coding structure with a GOP size larger than 15, $|\mathrm{GOP}| > 15$, to yield good enough rate-distortion (RD) performances. For the MDVC, the GOP size is usually set to be smaller in that, for the WZ codec to adopt longer GOP sizes, performing ME becomes difficult and less reliable such that the reconstructed SIF quality would be degraded. Previous research[23] investigates the rate-distortion and complexity performance of the feedback-channel based WZ codec as a function of the GOP

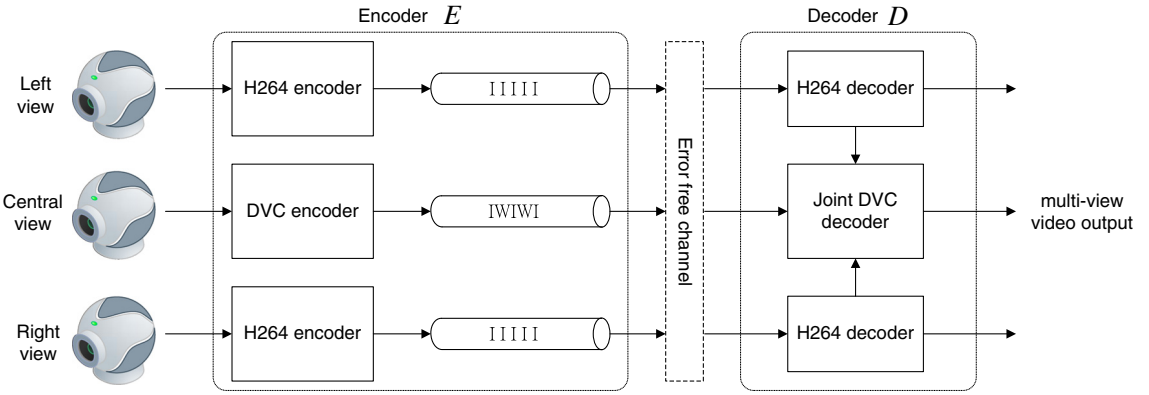

Fig. 1 Multiview distributed video coding with $|\mathrm{GOP}| = 2$.

size and justifies that the lowest encoder complexity, e.g., $|\text{GOP}| = 2$, yields the best RD performance, as compared with the conventional video codec. For the MDVC, the coding structure with $|\text{GOP}| = 2$ is adopted for simplicity and efficiency. Under the MDVC framework, we proposed to process static and nonstatic image regions with different procedures. By exploiting correlations between images along inter- and intraview dimensions, the proposed weighted block-matching prediction (BMP) can yield higher SIF quality. This proposed categorized block matching prediction with fidelity weights method is abbreviated as COMPETE. At the decoder, the scale-invariant feature transform (SIFT)[24] was adopted to find stable key feature points in the first decoded KF images, $\hat{L}_0$, $\hat{R}_0$, and $\hat{I}_0$, which are used for matching correspondent features among interview video images to estimate the homography matrices, $\mathbb{H}_\text{l}$ and $\mathbb{H}_\text{r}$, through a RANSAC[25] algorithm. The SIFT processing time is analyzed to be proportional to image size. The $\mathbb{H}_\text{l}$ and $\mathbb{H}_\text{r}$ are estimated once at the decoder to perspectively transform side-view images to be with central view. The homography matrix can also be estimated with a regular time interval or dynamically according to scene foreground/background change. In the proposed COMPETE algorithm, image blocks are categorized into motion, no-motion, and outlier blocks, with which blocks are processed in different ways. For motion blocks, with both perspectively transformed, $\hat{L}_t'$ and $\hat{R}_t'$, and reconstructed central-view images, $\hat{I}_t$, at the decoder, the block MC procedure can then be performed between adjacent images from these transformed and central-view ones to yield MVs. By combining blocks reached by these MVs with weights proportional to block fidelity, it would generate more smooth and higher quality SIFs. For no-motion blocks, the current block is compensated by the co-located block in the previous frame. For blocks residing on the outlier, resulting from perspective transformation, temporal bidirectional MC is performed between central-view image, $\hat{I}_{2t-1}$ and $\hat{I}_{2t+1}$. The proposed COMPETE algorithm helps to improve the SI confidence and the quality of decoded WZF, $\hat{I}_{2t}^{\text{WZ}}$s, for the MDVC system. The COMPETE also effectively decreases computational load while achieving comparable PSNR performances with other SIF reconstruction methods, e.g., MVME[26] and H.264/INTRA.

For rate control of the MDVC channel coding, the turbo codec is designed to let the decoder receive just enough parity bits from the encoder for signal reconstruction. The rate compatible punctured turbo (RCPT) code is adopted for the MDVC channel coding, which was initialed from unequal error protection for unstable transmission.[27] An automatic repeat request (ARQ) rate control method was developed under RCPT[28] to transmit fewest parity bits for successful decoding. For the turbo decoder to reference more reliable prior probabilities to reduce its iteration times and improve decoding efficiency, the correlation of DCTs between the original and its SIF is modeled as Laplacian distribution.[29] Different puncture patterns were designed for direct and alternate current coefficients, DCs and ACs, to yield the parity bits, based on which the correlation between bit-planes is exploited and utilized to estimate the posteriori probability to provide the priori probability for turbo decoding. Simulations verified that the turbo decoding time can be reduced to 37% as compared to other SIF generation methods.

In what follows, SIF reconstruction methods developed based on the MDVC system and the proposed COMPETE methods are described in Sec. 2. The proposed rate control algorithm to improve the MDVC performance is described in Sec. 3. Section 4 is the simulation study. Section 5 concludes this paper.

2 Multiview Distributed Video Coding Side Information

For one MDVC with $|\text{GOP}| = 2$, half of central-view images are encoded as WZFs and the SIF quality at the decoder would dominate the WZ codec performance. The SIF at the decoder can be considered as a reconstructed image of the original WZF at the encoder transmitted through noise channels. If the SIF quality is high enough, fewer parity bits will be requested during decoding and higher codec efficiency can be achieved. In a monoview video codec, the general approach to yield SIF is performing temporal interpolation/extrapolation from KFs to yield SIF, and there are other approaches adopting motion compensated interpolation to improve SIF quality, such as using an optical flow predictor[30] and hash-based estimator.[31] For the MVC, the same scene is captured from different viewing angles by different cameras, such that the correlation among different view videos can be utilized for SIF generation. Under the MDVC framework, we proposed to utilize the SIFT[24] feature extraction and the RANSAC[25,32] algorithm to exploit feature correspondences among interview video images. The SIFT outperforms other feature descriptors on images with real geometric and photometric transformations,[33] and the RANSAC helps to robustly fit a model to data in the presence of outliers, based on which the homography matrices[34] can be estimated for perspective transform from side-view video to central view. The proposed BMP algorithm can then be carried out to yield high quality SIF and improve the quality of decoded WZF. Different SIF reconstruction methods developed based on the MDVC framework, such as motion compensated temporal interpolation (MCTI),[35] MVME,[26] and hybrid-MVME (H-MVME), will be first reviewed for performance comparisons in the following sections.

2.1 *Side Information Reconstruction*

The MCTI[35] is an image reconstruction/interpolation method, in which block ME and MC are utilized to explore temporal correlation of monoview videos. To interpolate for the current frame, I_{2t}, the MVs estimated from its previous frame I_{2t-1} and the next frame I_{2t+1} are halved for bidirectional MC to yield the interpolated SIF, Y_{SI}. The MVME scheme[26] carried out at the decoder is shown in Fig. 2, in which KFs, Is, are coded with H.264/INTRA and the WZF is to be reconstructed with its SIF. For one WZF, two ME paths can be adopted: the inner path is estimated by performing disparity vector estimation followed by MV estimation, as demonstrated by Fig. 3(a); the outer path can be obtained by reversing the above two vector estimation procedures, as shown in Fig. 3(b). To interpolate for each block with $N \times N$ pixels in the WZF, let the side-view image at time $2^t - 1$, $I^{\text{side}}(2^t - 1)$, be the target image, in which a best matched block, with a disparity vector, $\vec{v}_d$, corresponding to the co-located block in the central-view image, $I^{\text{central}}(2^t - 1)$, is found. The best matched block in $I^{\text{side}}(2^t - 1)$ is then used to find out another best matched block from $I^{\text{side}}(2^t)$

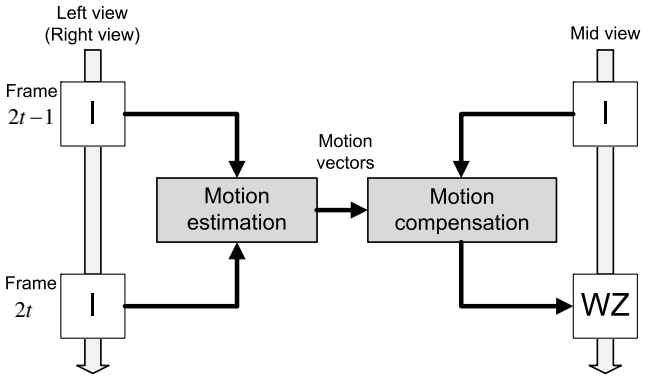

Fig. 2 The general MVME framework.[26]

with a MV $\vec{v}_m$. This procedure would yield one reference $\vec{v}_m$, or one inner path MV, for the co-located block in the current WZF. By applying the same procedure to the other three sets of reference images, three other inner path MVs can be found for the current block in the WZF. The outer path MVs can be obtained by the same procedure but with MV estimation first and then disparity vector estimation.

When all ME paths of the WZF are included, i.e., four inner and four outer paths to perform MVME, it yields eight estimated frames. This SIF can be reconstructed by taking the weighted/nonweighted average of these corresponding blocks of estimated MVs. Although the MVME provides several estimated MVs for reference, it suffers from heavy computation. In addition, it may lead to trivial estimation errors for no-motion blocks. The MVME approach utilizes the general ME operations, designed for intraview video, to estimate disparity vectors among interview images. To bridge this inherent gap between $\vec{v}_m$ and $\vec{v}_d$ estimation, we proposed to estimate the homography matrix to perspectively transform the side-view video to be with central view such that applying ME on interview images would be perfect. This H-MVME approach can yield better PSNR performances than MVME. In addition to handling the MVME in the hybrid approach, we proposed to eliminate trivial ME operations for no-motion blocks and perform BMP based on calculating the weighted sum of MC blocks reached through different MVs, denoted as COMPETE as described above, to improve the MVME to yield high quality SIF. In case the disparity/MV estimation was operated on outlier, i.e., regions without correspondent pixels resulting from performing perspective transformation, the temporal MCTI is adopted to interpolate for the current block in the WZF.

2.2 *COMPETE Side Information Reconstruction*

The COMPETE SIF reconstruction method is proposed to enhance the H-MVME to yield SIF with higher confidence. When homography matrices are not available for perspective transformation, we utilize the SIFT feature extraction and the RANSAC procedure to estimate homography matrices and then utilize BMP to yield high confidence SIF.

2.2.1 *Homography*

The homography relates the pixel coordinates in two images. When it is applied to every pixel, the new image is a warped version of the original one. However, this homography relationship is independent of the scene structure. To be more specific, one homography matrix, $\mathbb{H}$, which is 3×3, can transform one camera view to another.[34] To estimate the $\mathbb{H}_{v\in\{l,r\}}$s, the SIFT[24] algorithm is first applied on the video images of different views, L, R and I, to find stable key feature points. Tentative feature point pairs between two images are selected to provide candidate homography matrices, $\mathbb{H}_{v\in\{l,r\}}$. The feature point pairs and candidate $\mathbb{H}_{v\in\{l,r\}}$s are iteratively selected and justified by finding the maximum consensus set through the RANSAC procedure to yield the best $\mathbb{H}_{v\in\{l,r\}}$. At this stage, it seeks to find all correspondent SIFT points, or matching pairs, between two different view images. Mismatches will occur in that the matching process assumes proximity and similarity, and there are some correspondence located in outliers. In general, the RANSAC outperforms gradient descent

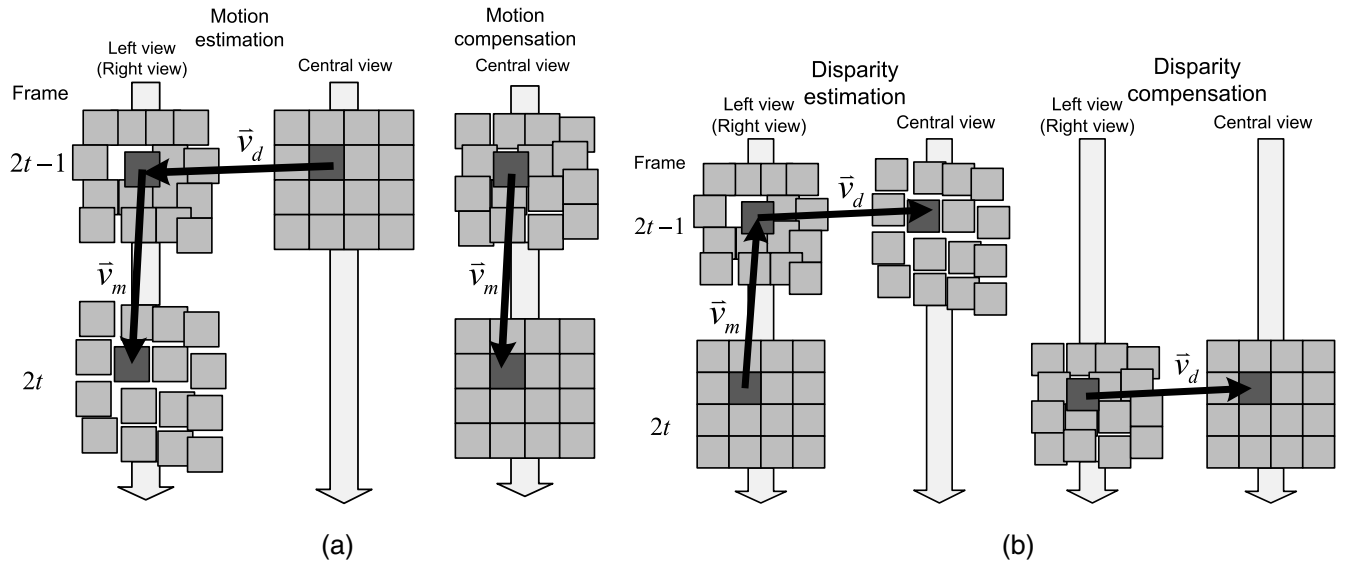

Fig. 3 Practical implementations of MVME.[26] (a) Inner path MVME and (b) outer path MVME.

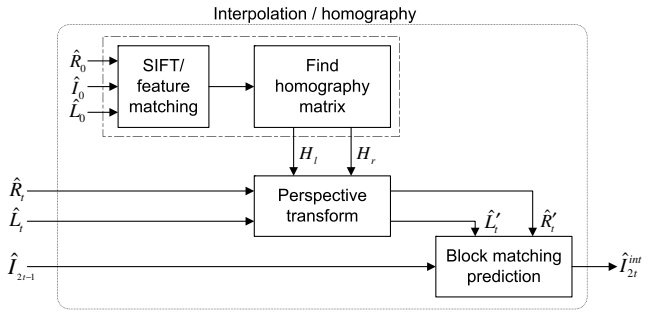

Fig. 4 The block diagram of interpolation/homography.

methods[36] in that too many outliers will prevent the latter from converging to the global optimum.

2.2.2 *Scale-invariant frame transform*

The SIFT[24] procedure helps to represent one image with robust feature points. It transforms one image into scale-invariant feature coordinates corresponding to local features. This procedure would ignore low contrast feature points and eliminate edge response to filter out the remaining stable keypoints.

2.2.3 *Interpolation and homography*

The SIF at the turbo decoder is generated by the "interpolation/homography" module, as shown in Fig. 4. We proposed to exploit correlations among interview images, in addition to intraview ones, to eliminate reference SIFs from having severe disparity. The reference central-view images can be obtained through the homography matrices, $\mathbb{H}_l$ and $\mathbb{H}_r$, from left- and right-view images. To estimate the $\mathbb{H}_l$ and $\mathbb{H}_r$, the first intracoded frames, $\hat{L}_0$, $\hat{R}_0$, and $\hat{I}_0$, received and reconstructed at the decoder, are used as sample images to extract correspondent stable SIFT features between left/right-view and central-view images. To estimate the homography matrix based on the correspondent feature points, the RANSAC procedure was carried out to find the matrices, $\mathbb{H}_l$ and $\mathbb{H}_r$, which yielded maximum inliers. The reference central-view images can then be obtained by performing perspective transform through $\mathbb{H}_l$ and $\mathbb{H}_r$ from the decoded left- and right-view images, $\hat{L}$ and $\hat{R}$, i.e., $\hat{L}' = \mathbb{H}_l(\hat{L})$ and $\hat{R}' = \mathbb{H}_r(\hat{R})$, as shown in Fig. 5(a). With the reference central-view images, the BMP procedure can be carried out to yield the SIF, $\hat{I}_{2t}^{\text{int}}$. For one multiview video, the homography matrix that transforms the side-view video to be with central view has to be estimated only once with reference to $\{\hat{R}_0, \hat{I}_0, \hat{L}_0\}$ at the beginning of decoding. With the homography matrix estimated optimally through the SIFT and the RANSAC procedures, the BMP among L' and R', and the original decoded one $\hat{I}_{2t-1}$ are performed to yield the SIF, described in the following section.

2.3 *Block Matching Prediction*

Performing perspective transformation from side view to central-view frames will result in an outlier, miss transformed area, as shown in Fig. 5(a). The perspectively transformed images, $\hat{L}'$s and $\hat{R}'$s, and the reconstructed central-view images, $\hat{I}_{2t-1}$s, are used to perform block matching to estimate disparity and MVs, denoted as $\vec{v}'_d$ and $\vec{v}'_m$, respectively. The SIF of a central-view image not transmitted can be reconstructed through weighted motion compensated prediction by above $\vec{v}'_m$s and $\vec{v}_m$s, in which the latter were estimated from $\hat{I}_{2t\pm1}$s. This BMP process would reconstruct the SIF, $\hat{I}_{2t}^{\text{int}}$, shown in Fig. 5(b), where B_i is the block in $\hat{I}_{2t-1}$, $\vec{v}'_{d_i}$ and $\vec{v}'_{m_i}$ are the disparity and MVs estimated between reconstructed interview images, e.g., $\{\hat{L}'_{2i-1}, \hat{I}_{2t-1}\}$ and $\{\hat{R}'_{2t-1}, \hat{I}_{2t-1}\}$, and between $\hat{I}_{2t\pm1}$s, respectively. The COMPETE flowchart is shown in Fig. 6. One $\hat{I}_{2t-1}$ is partitioned into M 8×8 blocks, $\{B_i(\hat{I}_{2t-1}) | i \in 1, \cdots, M\}$, and a large block $LB_i(\hat{I}_{2t-1})$ consists of 2×2 blocks, i.e., $LB_i(\hat{I}_{2t-1}) = \{B_i^{11}, B_i^{12}, B_i^{21}, B_i^{22}\}$, in which B_i^{11} is the current block, i.e., $B_i = B_i^{11}$. The four block MVs in LB_i, $(\vec{v}_m^{11}, \vec{v}_m^{12}, \vec{v}_m^{21}, \vec{v}_m^{22})$, are obtained by performing motion estimation (ME) between $\hat{I}_{2t-1}$ and $\hat{I}_{2t+1}$ for the co-located LB_i. If $(\vec{v}_m^{11}, \vec{v}_m^{12}, \vec{v}_m^{21}, \vec{v}_m^{22}) = \vec{0}$, it means B_i in LB_i is a no-motion block and can be reconstructed by direct copy from its previous image, i.e., $B_i^{11}(\hat{I}_{2t}^{int}) = B_i(\hat{I}_{2t-1})$. If $(\vec{v}_m^{11}, \vec{v}_m^{12}, \vec{v}_m^{21}, \vec{v}_m^{22}) \neq \vec{0}$, then B_i is a motion block and the

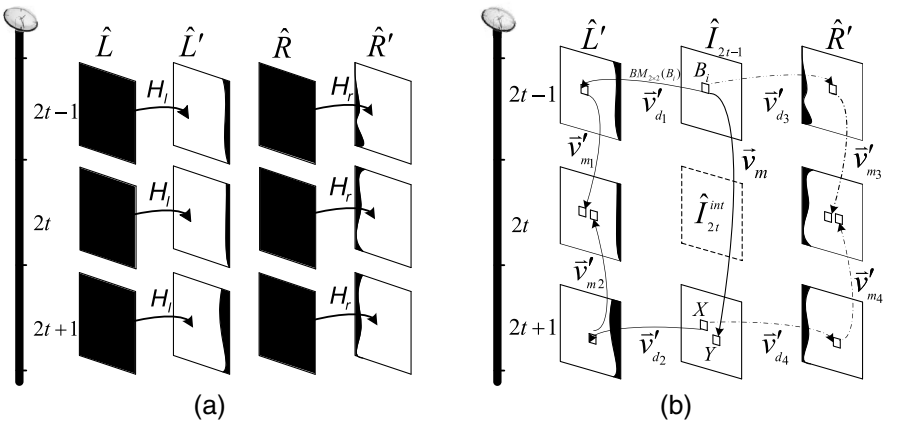

Fig. 5 The ME implementation of a MVC: (a) the perspective transform of left and right views and (b) the block matching search between central view and perspective transformed images: X denotes the colocated block of B_i in $\hat{I}_{2t+1}$, and Y is the best matched block of B_i.

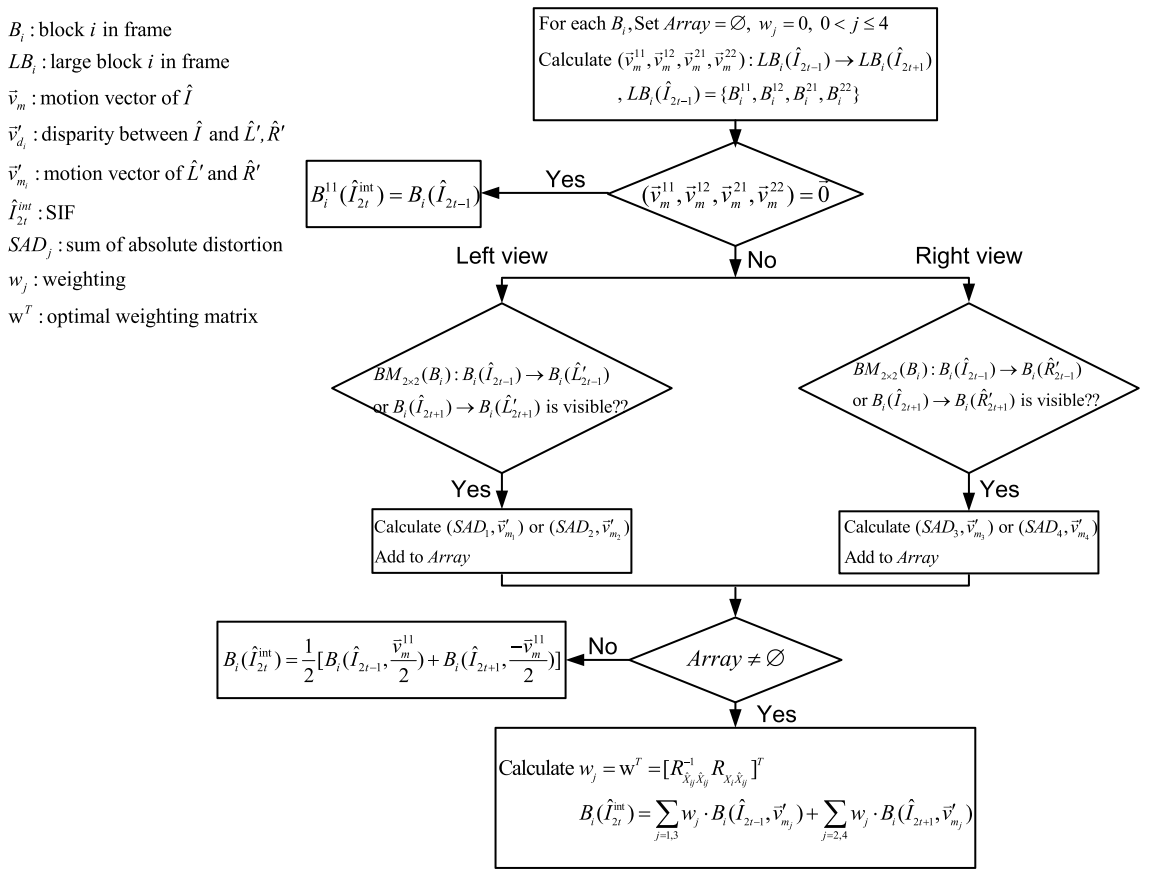

Fig. 6 The flow chart of COMPETE.

corresponding disparity block in side-view transformed images, $\hat{L}'$ and $\hat{R}'$, and B_i's MVs are combined with weights proportional to block fidelity to yield a more accurate compensated block for the B_i in $\hat{I}^{\text{int}}_{2t}$. We take the ME process for a B_i by referencing left- and central-view images as an example and the right-view one can be carried out in the same way. The first-phase block disparity estimation is performed between $\hat{I}_{2\,t-1}$ and $\hat{L}'_{2t-1}$, denoted as $\mathbb{BM}_{2\times2}(B_i) : B_i(\hat{I}_{2t-1}) \to B_i(\hat{L}'_{2t-1})$, which will yield the best matched block from $\hat{L}'_{2t-1}$ with a $\vec{v}'_d$. If the best matched block does not reside on the outlier of $\hat{L}'_{2t-1}$, the second-phase block ME is performed, in which the search range in $\hat{L}'_{2t}$ is two blocks wide along vertical and horizontal directions and centered at the co-located coordinate of B_i on $\hat{L}'_{2t-1}$ with the offset $\vec{v}'_d$. It yields one $\vec{v}'_{m_1}$, and the second $\vec{v}'_{m_2}$ can be obtained by the same procedure $\mathbb{BM}_{2\times2}(B_i) : B_i(\hat{I}_{2t+1}) \to B_i(\hat{L}_{2t+1})$. The other two MVs, $\vec{v}'_{m_3}$ and $\vec{v}'_{m_4}$, are estimated from the right-view video through the same procedure. When performing MC for an $\hat{I}_{2i}$, if any image block reached through the inner-path MV, $\vec{v}'_{m_j}$, resides on the outlier, then its w_j is set zero. Let $B_i(I, v)$ denote the image block obtained from the co-located block on an I with its MV, v, and the B_i reconstruction for the SIF, $\hat{I}^{\text{int}}_{2t}$, can be represented as

$$B_i(\hat{I}^{\text{int}}_{2t}) = \sum_{j=1,3} w_j \cdot B_i(\hat{I}_{2t-1}, \vec{v}'_{m_j}) + \sum_{j=2,4} w_j \cdot B_i(\hat{I}_{2t+1}, \vec{v}'_{m_j}), \quad (1)$$

where the first term yields the weighted central-view image by utilizing MVs of $\hat{I}_{2t-1}$ and the second one from $\hat{I}_{2t+1}$. In general, the w_j should be proportional to the normalized fidelity of the corresponding best matched block with respect to the co-located blocks in central view. The w_j for one MC block reached through $\vec{v}'_{m_j}$ can be computed as

$$w_j = \frac{1}{\text{SAD}_j} \Big/ \sum_{j=1}^{4} \frac{1}{\text{SAD}_j}, \qquad j \le 4, \quad (2)$$

in which SAD_j denotes the sum of absolute distortion whose reciprocal can be used as block fidelity. On the other hand, if all matched blocks reside on the outlier, there would be no prediction result that can satisfy the assumed scenario. Under this condition, only the reference MV, $\vec{v}_m$, estimated between the two reconstructed central-view images, $\hat{I}_{2t-1}$ and $\hat{I}_{2t+1}$, can be used to predict the SIF. The bidirectional MC is used to reconstruct the block of the SIF:

$$B_i(\hat{I}^{\text{int}}_{2t}) = \frac{1}{2}\left[B_i\left(\hat{I}_{2t-1}, \frac{\vec{v}_m^{11}}{2}\right) + B_i\left(\hat{I}_{2t+1}, \frac{-\vec{v}_m^{11}}{2}\right)\right]. \quad (3)$$

To further yield the optimal weight w_j for a MC block $B_i(I, v)$, the linear minimum mean squared error (LMMSE) estimator can be adopted. How to compute the LMMSE weights, w_js, is described in the Appendix. Experiments showed that adopting LMMSE weights can improve the

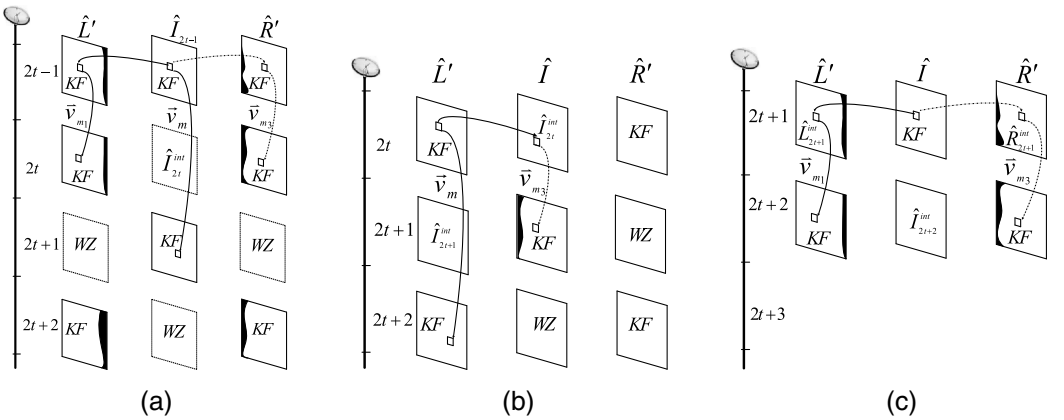

Fig. 7 Perform COMPETE on different GOP structures with KF : WZF = 1 : 1. (a) One $\vec{v}_m$ and two $\vec{v}_d$s for $\hat{I}^{\text{int}}_{2t}$; (b) one $\vec{v}_m$ and two $\vec{v}_d$s for $\hat{L}^{\text{int}}_{2t+1}$; and (c) two $\vec{v}_d$s for $\hat{I}^{\text{int}}_{2t+2}$.

SIF PSNR up to 0.1 and 0.3 to 0.4 dB for low and medium-to-high complexity videos, respectively, compared to those adopting weights proportional to block fidelity presented by Eq. (2).

In our experiments, the COMPETE is operated under the frame ratio KF : WZF = 5 : 1, while the fusion-based homography method is KF : WZF = 1 : 1. The COMPETE can also be adapted to operate under the ratio KF : WZF = 1 : 1. In the COMPETE, it needs to transmit the first KF of each view to estimate homography matrices, as shown in Fig. 7(a), and there are one MV and two disparity vectors that can be used to interpolate for the SIF of $\hat{I}^{\text{int}}_{2t}$. To interpolate for the SI of side-view images, say $\hat{L}^{\text{int}}_{2t+1}$, only one MV and one disparity vector can be referenced, as shown in Fig. 7(b). For the last central-view image, only two disparity vectors can be referenced to interpolate for its SI, as shown in Fig. 7(c). When the WZF/KF ratio is larger than 1, it requires learning-based approaches[37] that apply an expectation maximization algorithm for unsupervised learning of MVs.

3 Multiview Distributed Video Coding Rate Control Algorithm

The internal signal processing flow of the MDVC (Fig. 1) is shown in Fig. 8. The encoder E comprises both H.264 and WZ encoders, in which the left- and right-view images, $\{L_t\}$ and $\{R_t\}$, would be encoded by the former to yield KF bitstreams, s_l and s_r, respectively. The central-view images, $\{I_t\}$, are separated into odd and even image sequences, $\{I_{2t-1}\}$ and $\{I_{2t}\}$. The odd images are encoded by H.264 Intra to provide the KF bitstream s_o and the even ones by the WZ encoder with appended cyclic redundancy check (CRC) checksum to yield parity bits $\tilde{p}_{2t}$. For adaptive rate control, the RCPT[28] code is adopted for channel coding, because it performs near the Shannon limit at low SNR, while providing excellent throughput at high SNR.[28] The WZ encoder will determine whether to send more parity bits or not based on the feedback requested bits NAK from the WZ decoder. The decoder D comprises one H.264 decoder, one WZ decoder, and one interpolation/homography function module. The received bitstreams, s_l, s_r, and s_o, will be decoded by the H.264 decoder to yield reconstructed images of left-, right-, and central-view odd images, $\hat{L}_t$, $\hat{R}_t$, and $\hat{I}_{2t-1}$, respectively. They are inputs of the interpolation/homography modules that will reconstruct the SI, an interpolated central-view image $\hat{I}^{\text{int}}_{2t}$, for the WZ decoder to reconstruct $\hat{I}_{2t}$ with reference to $\hat{I}^{\text{int}}_{2t}$. The multiplexer combines the reconstructed $\hat{I}_{2t-1}$ and $\hat{I}_{2t}$ to yield the final central-view video $\{\hat{I}_t\}$.

3.1 *Wyner–Ziv Coding*

The WZ encoder in the MDVC system is shown in Fig. 9. The input image, I_{2t}, is divided into blocks with 4×4 pixels, which are then transformed to frequency domain coefficients, c_{2t}, through T, and quantized through Q to yield

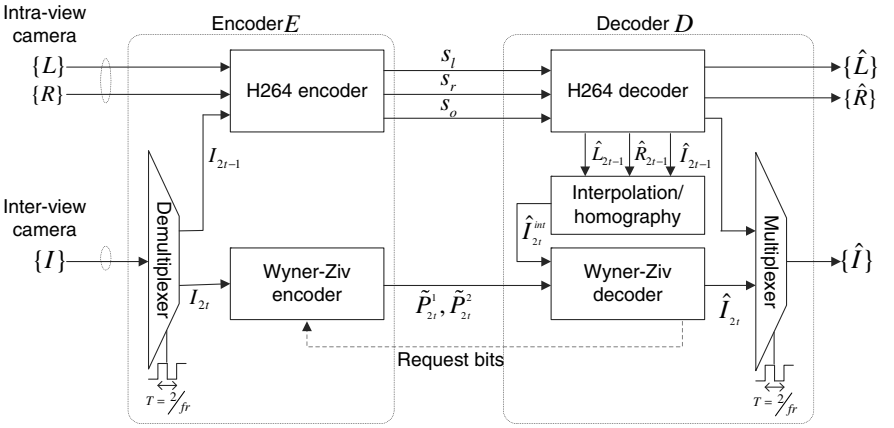

Fig. 8 The MDVC codec framework.

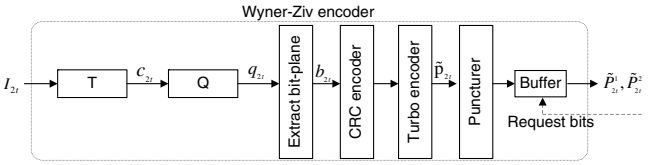

Fig. 9 The WZ encoder.[38]

the quantized coefficients, q_{2t}. To reduce encoding complexity, the integer DCT is adopted for low complexity hardware implementation. In c_{2t}, the DC coefficient comprises most of the block signal energy and will be allocated more bits than other higher frequency ones, ACs. Coefficients in the 4×4 block, c_{2t}, are partitioned into different bands. Each coefficient band is uniformly quantized with a 2^{b_k} level quantizer (Q), where b_k denotes the number of bits assigned to the k'th coefficient. The number of quantization levels, 2^{b_k}s, for a 4×4 DCT coefficient block[38] is determined through an optimal bit allocation procedure on the c_{2t} coefficients.

In practical implementation, the quantization stepsize of the i'th coefficient, Δ_i, was setup with a loading factor, $\sigma = 4$, for a certain coefficient probability density function (PDF),[39] i.e.,

$$\Delta_i = \frac{4\sigma_i}{2^{b_k}}, \quad \text{for } b_k \neq 0. \tag{4}$$

After quantization, each coefficient is represented by its quantization index q_{2t}. For simple demonstration, the parity bits generating process for one 16×16 image is provided. The 16×16 image is decomposed into sixteen 4×4 blocks on which DCT is performed, and the number of bits to represent the quantized indexes of DCs and ACs are 4 and 3, respectively. The DCs and ACs of these sixteen 4×4 DCT blocks are rearranged such that the same frequency coefficients are grouped together and queued with zigzag scan order, i.e., $\{\mathrm{DC}^i\}_{i=1,2,\cdots,n}$ $\{\mathrm{AC}_1^i\}_{i=1,2,\cdots,n}$, $\{\mathrm{AC}_2^i\}_{i=1,2,\cdots,n}$, $\cdots$, $\{\mathrm{AC}_a^i\}_{i=1,2,\cdots,n}$, where n is the number of total blocks in the image and a is the number of ACs for a certain quantization pattern, as shown in the upper image of Fig. 10(a). For turbo encoding, these regrouped 4×4 DCs blocks are subject to bit-plane extraction, as shown in Fig. 10(a), such that the same significant bits are grouped together and transmitted by bit-plane order, i.e., $\mathrm{MSB}_k = \{\mathrm{MSB}_k^i | i = 1, 2, \cdots, 16\}$ for $k = 1, 2, \cdots, K$, where i is the index of the original 4×4 blocks and k is the bit-plane index. For regrouped 4×4 ACs blocks, the above transmission order is reversed, i.e., from the LSB to the MSB. The bit-stream of these reordered bits, b_{2t}, is then used as the input to the CRC encoder, which appends checksum of b_{2t} and passes it to the turbo encoder. After performing interleaving by the turbo encoder, it yields the parity bitstreams, $\tilde{\mathbf{p}}_{2t} = \tilde{\mathbb{P}}_i^1 \cup \tilde{\mathbb{P}}_i^2$, which can be represented as $\tilde{\mathbb{P}}_i^1 = \{\tilde{p}_1^1, \tilde{p}_2^1, \cdots, \tilde{p}_{16}^1, \cdots\}$ and $\tilde{\mathbb{P}}_i^2 = \{\tilde{p}_1^2, \tilde{p}_2^2, \cdots, \tilde{p}_{16}^2, \cdots\}$. Both parity bit streams are punctured with specific patterns of period $\psi = 16$ to form sub-blocks queued in the transmission buffer, denoted as $\tilde{P}_{2t}^1$ and $\tilde{P}_{2t}^2$, which will be sent to the decoder upon request. The puncture pattern is designed to select parity bit according to the specified priority, as shown in Fig. 10(b). For turbo decoding, the skipped systematic bits at E are replaced with the reconstructed SI at D, which would be reconstructed by different methods. The turbo decoder would request more parity bits in case it cannot correctly recover the data. In general, when the SI confidence is high, it would request fewer parity bits and improve the WZF quality. Detailed rate control steps will be described in Sec. 3.2.

To reconstruct the WZF, $\hat{I}_{2t}$, from the received parity bits sub-block, $\{\tilde{P}_{2t}^1, \tilde{P}_{2t}^2\}$, at the WZ decoder shown in Fig. 11, it needs to generate the SI, $\hat{I}_{2t}^{\mathrm{int}}$, by the interpolation/homography module, as shown in Fig. 8. Before turbo decoding, the same T and Q processes will be applied to $\hat{I}_{2t}^{\mathrm{int}}$ to yield $\hat{c}_{2t}^{\mathrm{int}}$ and $\hat{q}_{2t}^{\mathrm{int}}$, respectively. To increase the SI confidence for turbo decoding, the distributions of error between reconstructed SIF and the original WZF are modeled as Laplacian. A transform-domain correlation noise model parameter updating procedure[29] was applied to fit coefficient error distribution for each 4×4 block with the Laplacian model. Since the original image encoded as a WZF is not available at the decoder, the MCTI image, $\hat{I}_{2t}^{\mathrm{int}}$, interpolated from $\hat{I}_{2t\pm1}^{\mathrm{int}}$s, was used instead. After being processed by T and Q, the indexed signals, $\hat{q}_{2t}^{\mathrm{int}}$, are reordered, grouped, and extracted by bit-plane to provide the system bits, $\hat{b}_{2t}^{\mathrm{int}}$, for the turbo decoder. The turbo decoder performs the logarithmic maximum a "posterior" algorithm, Log-Map, with the help of received parity bits sub-blocks, $\{\tilde{P}_{2t}^1, \tilde{P}_{2t}^2\}$, and CRC checksum verification, under a certain confidence measurement[40] to determine either the decoding process is convergent or to request more bits for the next iteration. After b_{2t} being decoded correctly, it is reversely processed by the combining bit-plane module to yield the quantized index, $\hat{q}_{2t}$, which are used as the input of the reconstruction module to refine $\hat{c}_{2t}^{\mathrm{int}}$ for $\hat{c}_{2t}$.

The optimal reconstruction function that exploits the correlation between the original image for WZF and SI[14] is adopted, in which the distribution of the residual signals between the original WZF and the reconstructed SIF is assumed to be Laplacian and it seeks to find the reconstructed samples that demonstrate MMSE. The optimal reconstruction value, $\hat{c}_{2t}$, is the expectation $\hat{c}_{2t} = E[c_{2t} | c_{2t} \in \{\Delta_i^l, \Delta_i^r\}, \hat{c}_{2t}^{\mathrm{int}}]$, where Δ_i^l / Δ_i^r denote the lower/upper boundary of the interval Δ_i that $\hat{c}_{2t}^{\mathrm{int}}$ resides, and the expected value yields the MMSE estimation of the source WZ. This procedure will prevent the reconstructed values from deviating from the original value too much due to low SI confidence. At the last stage, the $\hat{c}_{2t}$ will be inversely transformed to yield the final reconstructed image, $\hat{I}_{2t}$.

3.2 *Rate Control Mechanism*

To improve the decoding efficiency, we proposed to impose specific puncture patterns with transmission order according to signal distribution properties for DCs and ACs, respectively. In the COMPETE framework, we proposed to collect all same order DCs/ACs together, which are then zig-zag scanned for turbo encoding. For block DCT-based video coding, the DC coefficient usually contains most block signal energy. Its MSBs contribute much more signal energy than LSBs, such that the assigned priority of the former is higher than the latter. As shown in Fig. 10(b), the system is designed to transmit the first MSBs of all DCs and then the second MSBs. The magnitude of ACs would be much smaller and around zero magnitude. Since ACs may be positive or negative, by taking its absolute value, it would lead to more skewed magnitude probability distribution. The "sign

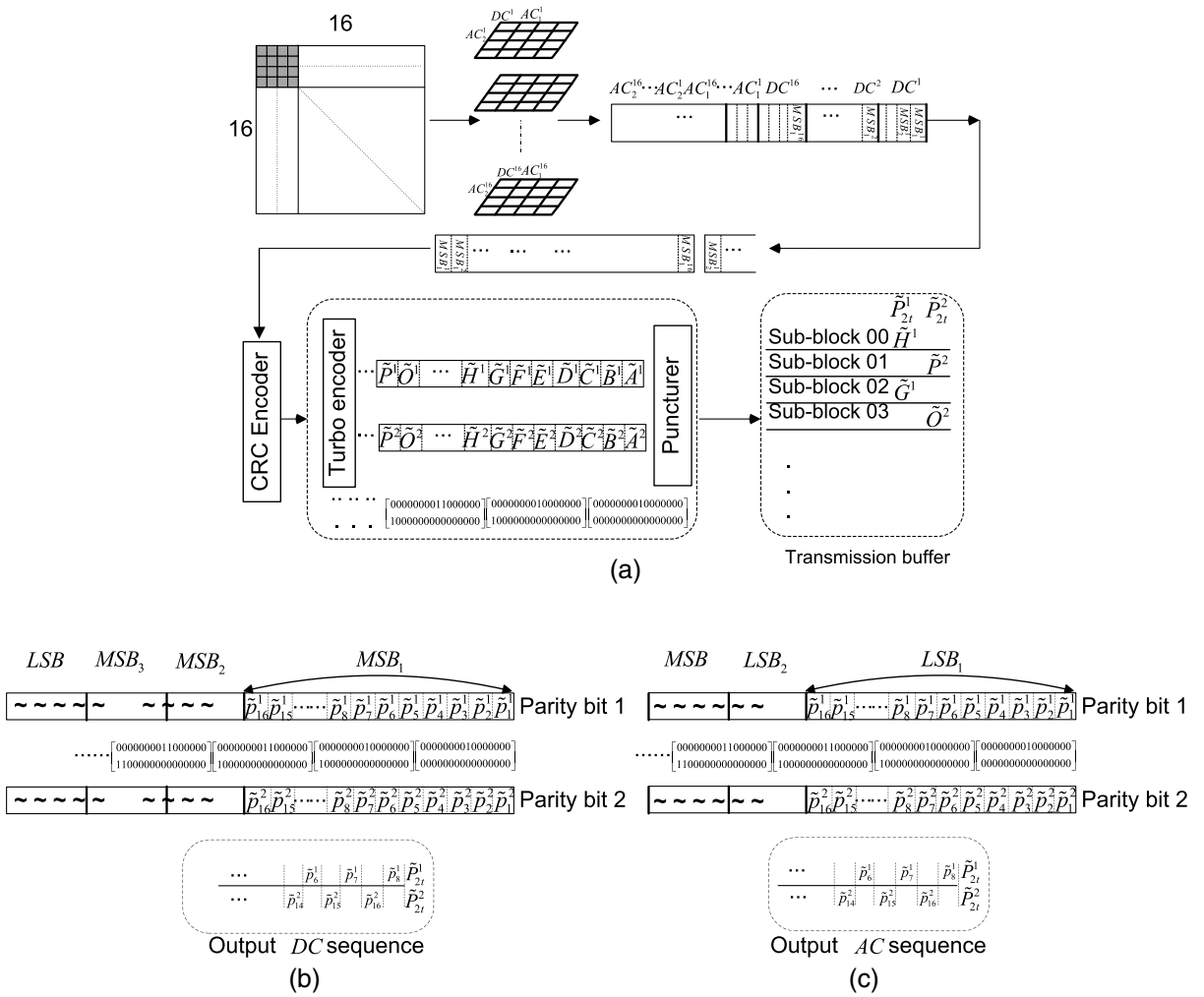

Fig. 10 The parity bit generation and transmission order in the puncture patterns of DCs and ACs with period $\psi = 16$. (a) The parity bits generation process, (b) the sub-block queuing pattern of DC, and (c) the sub-block queuing pattern of AC.

bit" of quantized ACs can be replaced by that of the quantized SIF at the decoder, under which the probability of LSBs to be 0 would be larger than MSBs, when represented with a fixed number of bits. As opposed to DCs, it transmits the LSB first and then the second LSB[41] to speed up turbo decoding, as shown in Fig. 10(c). This transmission strategy for DCs and ACs helps to correct the decoding errors of systematic bits with fewest requested parity bits. Experiments showed that this rate control strategy yields 55% to 59%, fewer requested bits for the turbo decoder.

The proposed rate control algorithm, developed based on the RCPT puncturing mechanism,[28] is demonstrated in Fig. 12.

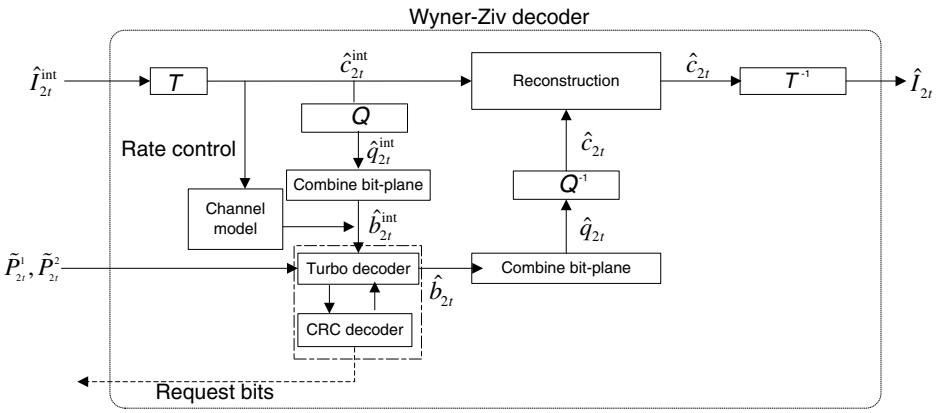

Fig. 11 The WZ decoder.

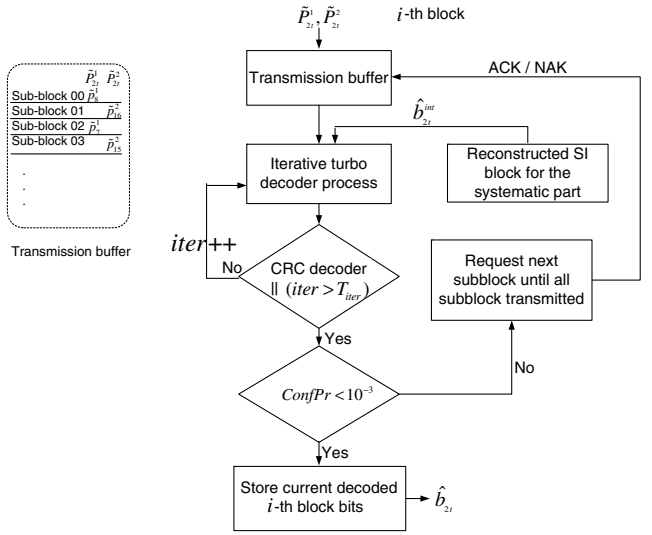

Fig. 12 The proposed RCPT-based rate control algorithm.

In the COMPETE system, the RCPT code is designed to be with rate $1/3$ and puncturing period $\psi = 16$, which is formed from two rate $1/2$ recursive systematic convolutional constituent codes with generator $\frac{1+D+D^2+D^4}{1+D^2+D^3}$. The puncturing table with different rates, $\{\frac{16}{16+V}|V = 0, 1, \cdots, 32\}$, will be generated, in which $V = 0$ will not be used because the systematic bits will be discarded under the DVC framework. Figure 10 demonstrates part of the corresponding puncture table. When the first sub-block parity bits were received, the decoding would be carried out based on the CRC alone.[28] When receiving those of the second sub-block, it would decode the first constituent encoding data and the iterative turbo decoding will start after the third sub-block being received, in which the maximum iteration number, T_{iter}, is set. When decoded results are converged, i.e., an all-zero syndrome of CRC checking or the number of iteration exceeds T_{iter}, the resultant bitstream will be subjected to a second confirmation procedure. Notwithstanding, a larger T_{iter} will lead to heavy computation and the tradeoff between setting T_{iter} and heavy computation should be well manipulated. The value of T_{iter} is determined from experiments on different complexity test videos under different bit rates that can yield convergence. The confidence measurement with the criteria ConfPr $\leq 10^{-3}$,[40] in which

$$\text{ConfPr} = \frac{U}{L_B}, \tag{5}$$

where L_B is the predefined block length for decoding and U is number of uncertain bits whose absolute decoded likelihood ratio, $\frac{Pr(X_i=B|Y)}{Pr(X_i=1-B|Y)}$, is not higher than 0.99. The decoding is successfully completed when both CRC check and confidence measure, ConfPr $\leq 10^{-3}$, are satisfied. When CRC passes but confidence measure fails, i.e., ConfPr $> 10^{-3}$, more sub-block parity bits will be requested by the ARQ mechanism for the next iteration operations until all sub-block bits are sent or the bitstream is decoded successfully.

To improve the turbo decoding performance while requesting fewer parity bits, the correlation among coefficient bit planes was exploited and utilized to estimate the

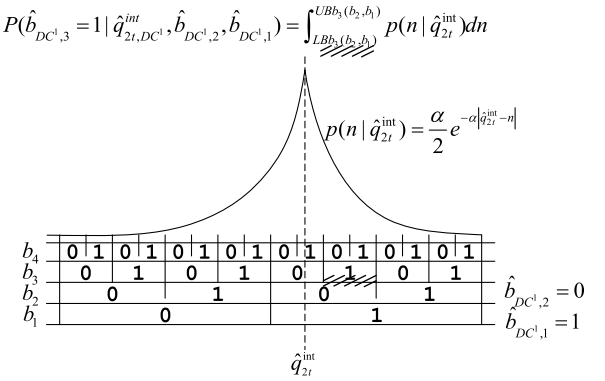

Fig. 13 The probability estimation when decoding $\hat{b}_{\text{DC}^1,3}$, in which the $\hat{q}^{\text{int}}_{2t}$ is obtained from the reconstructed SIF and the PDF is specified in Eq. (6).

posteriori probability, which is used as the priori probability for turbo decoding. The probability distribution of the difference between a SIF and the original image coded as a WZF is assumed to be Laplacian, i.e.,

$$p_{\hat{q}^{int}_{2t}}(n) = \frac{\alpha}{2}e^{-\alpha|\hat{q}^{int}_{2t}-n|}, \qquad \alpha = \sqrt{\frac{2}{\sigma_x^2}}, \tag{6}$$

where σ_x is the variance of residual signal between a WZF and a SIF.[29] The b'th decoded bit of DCs (DC^1) is represented as

$$\hat{b}_b \equiv \arg \max_{i\in(0,1)} Pr_{\text{DC}^1}(i|\hat{q}^{int}_{2t}\hat{b}_{b-1}, \cdots, \hat{b}_2, \hat{b}_1), \tag{7}$$

where $Pr_{\text{DC}^1}(i|\hat{q}^{\text{int}}_{2t}, \hat{b}_{b-1}, \cdots, \hat{b}_2, \hat{b}_1)$ is the posteriori probability of $\hat{b}_b = i$ for DC^1. When decoding $\hat{b}_3$ of DC^1, both $\hat{b}_2$ and $\hat{b}_1$, and the reconstructed SI, $\hat{q}^{\text{int}}_{2t}$, are jointly referenced to specify the probability. Figure 13 shows an example to estimate the probability $Pr_{\text{DC}^1}(\hat{b}_3 = 1|\hat{q}^{\text{int}}_{2t}, \hat{b}_2, \hat{b}_1)$[40] of a quantized DCs represented with four bits, $b_1b_2b_3b_4$ from MSB to LSB. The probability integrated from the shaded interval is for $Pr_{\text{DC}^1}(\hat{b}_3 = 1|\ \cdots)$ and $Pr_{\text{DC}^1}(\hat{b}_3 = 0|\ \cdots)$ can be calculated in the similar way. The turbo decoder will update the priori probability:

$$\log \frac{Pr_{\text{DC}^1}(i = 1|\hat{q}^{\text{int}}_{2t}, \hat{b}_2, \hat{b}_1)}{Pr_{\text{DC}^1}(i = 0|\hat{q}^{\text{int}}_{2t}, \hat{b}_2, \hat{b}_1)}, \tag{8}$$

and performs log-MAP decoding. Experiments verified that this probability estimation and updating method helps the decoder to request fewer parity bits and reduces the turbo decoding time.

4 Simulation Study

The COMPETE encoding performance is compared with other SIF reconstruction methods, such as MCTI, fusion-based homography (F-HOMO), MVME, and H-MVME, for evaluation. The H-MVME is the extension of MVME,[26] in which estimated image blocks that reference to the outlier are obtained through MCTI. In the F-HOMO, both SIFs reconstructed from inter- and intraview images through DCVP[42] and MCTI, respectively, are fused to yield the

final SIF. The quality of $\hat{I}_{2t}^{WZ}$, which is reconstructed with its SIF generated by the above methods, is compared with those from H.264 with inter-, intra-, and inter-no-motion mode. The multiview CIF videos, Race1, Ballroom, Breakdancer, Exit, Ballet and Vassar, provided by ISO/IEC[43] are used as test videos, whose frame rates are 30, 25, 15, 25, 15, and 25 fps, respectively. These videos present different scene complexities rated from high to low, in which the "Race1, Ballroom, and Breakdancer" are classified as high complexity videos, "Exit" as medium and (Ballet and Vassar) as low complexity ones, respectively. Three successions of the six views from a multiview video are used to provide left-, central- and right-view videos. For H.264, the CABAC function is enabled and GOP size is 12 for inter- and inter-no-motion modes. The ME search range for the former is set to be 32 and zero motion is assigned for the latter. For the H.264 coder to yield compromised decoded quality for different videos, different quantization parameters (QPs), QP $\in \{30, 28, 26, 24, 20, 18\}$, are used for different complexity videos. The MDVC codec adopts $|\text{GOP}| = 2$, in which the side-view video and central-view odd frames are encoded with H.264/INTRA to provide KFs for the decoder to reconstruct $\hat{I}_{2t}^{WZ}$s. The quality of reconstructed $\hat{I}_{2t}^{WZ}$s with reference to the four SIF generation methods is compared by image PSNRs for evaluation.

4.1 *Performance Analysis*

To evaluate the performance of the proposed COMPETE, the error analysis based on reconstructed blocks is first carried out to investigate the signal processing behavior. Four SIF reconstruction methods, which comprise MCTI, F-HOMO, MVME, and H-MVME, are also implemented for comparisons. The SI confidence, quality of reconstructed WZFs, $\hat{I}_{2t}^{WZ}$s, and time complexity of different methods are compared and evaluated. The time complexity of SI generation and encode/decode execution time will be discussed in Sec. 4.2.

4.1.1 *Error analysis*

The error distributions of the COMPETE and MVME are investigated to justify how the SI confidence can be improved. In the COMPETE algorithm, by performing intraview ME between central-view images, blocks are classified into motion or no-motion to eliminate unnecessary ME/MC operations. For no-motion blocks, the co-located block of the previous frame is used as the MC blocks with zero motion. For motion blocks, when the search range comprises regions belonging to the outlier, only intraview ME on central-view images is performed. Otherwise, the regular weighted MVME process is carried out. Denote the number of no-motion, motion, and outlier blocks in one frame as K_n, K_m, and K_o, which can be normalized as k_n, k_m, and k_o, respectively, i.e., $k_n + k_m + k_o = 1$. In the COMPETE, the MC interpolated frame can now be represented by

$$\hat{I}_{2t}^{int} = \mathbb{B}_n \cup \mathbb{B}_m \cup \mathbb{B}_o, \tag{9}$$

where $\mathbb{B}_\tau = \{B_\tau(i)|i = 1, \cdots, k_\tau\}$ denote the set of type τ blocks for $\tau \in \{n, m, o\}$ and the number of blocks in the set $\|\mathbb{B}_\tau\|$ is K_τ. The variance of block errors can be represented as

$$\begin{aligned}\sigma_B^2 &= \frac{1}{\sum_\tau K_\tau} E[(I_{2t} - \hat{I}_{2t}^{int})^2] \\ &= \sum_{\tau \in \{n,m,o\}} k_\tau \cdot E[B_\tau(I_{2t}) - B_\tau(\hat{I}_{2t}^{int})^2 | B_\tau(I_{2t}) \in \mathbb{B}_\tau].\end{aligned} \tag{10}$$

For one image block, a specific ME procedure corresponding to its categorization, i.e., motion, no-motion, or outlier, will be imposed. Table 1 shows the percentage of each block category for different videos and Table 2 shows the mean absolute error of block difference, between the original image and its reconstructed SIF, for the six test videos. As shown, the percentage of outlier blocks is very small and their average reconstruction error by the COMPETE is smaller than that of MVME. Both estimated intraview MV and interview disparity vector are utilized to improve the SI confidence, in which the four MVs through inner paths are utilized to perform intraview weighted MC for a central-view SIF. This SIF demonstrated higher confidence than that reconstructed through average MC in both MVME and H-MVME. As shown in Table 2, the average error of reconstructed blocks of the proposed COMPETE is smaller than that of MVME. Table 1 shows that the percentage of no-motion blocks is the highest, which are mostly from the background region or static foreground objects. For no-motion blocks, the proposed COMPETE effectively eliminated the time consuming ME process and prevented noisy MVs resulting from regular ME process of other methods. For example, the MVME method, instead of identifying no-motion blocks and skipping the time-consuming ME process, treats all as motion blocks but does not yield a more accurate estimation, as shown in Table 2. For motion blocks, the MVME does not differentiate interview disparity vector with intraview MV, such that the MC blocks would be more degraded as compared to that of COMPETE. As the COMPETE compensates no-motion blocks by the colocated ones of the previous decoded frame, in addition to reducing time complexity, the ME errors can also be decreased. In total, the proposed COMPETE effectively yielded higher SI confidence while reducing time complexity, as compared to MVME.

Table 1 No-motion, motion, and outlier blocks distribution at QP = 26.

	Block type		
Video	Motion blocks (%)	No motion blocks (%)	Outlier blocks (%)
Race1	72.15	25.67	2.18
Ballroom	37.64	60.06	2.30
Breakdancer	42.48	54.47	3.05
Exit	20.66	78.20	1.14
Ballet	17.17	81.60	1.23
Vassar	3.71	96.08	0.21

Table 2 The comparison of estimation errors.

	Block type					
	Motion Blocks (MAE)		No-Motion Blocks (MAE)		Outlier Blocks (MAE)	
Video	COMPETE	H-MVME	COMPETE	H-MVME	COMPETE	H-MVME
Race1	194.28	423.11	1.46	1.79	255.08	292.30
Ballroom	302.47	356.74	3.36	3.46	310.66	323.96
Breakdancer	694.01	742.19	1.54	1.79	179.59	184.42
Exit	729.95	804.74	1.91	2.73	191.55	206.61
Ballet	134.69	281.63	1.71	2.11	129.74	127.08
Vassar	82.81	364.24	2.59	2.70	182.27	196.24

4.1.2 *Side information confidence*

The SI confidence in PSNR achieved by MCTI, F-HOMO, MVME, the COMPETE with direct linear transform (DLT) homography matrix generation method and the COMPETE performed on all test videos are shown in Fig. 14. As shown, the MCTI performance was severely degraded for high motion videos, Race1, Ballroom and Breakdancer, since it assumes linear motion and interpolates frames only along temporal dimension. For Race1, the SIF by COMPETE is 6.2 to 7.9 dB higher in PSNR than MCTI because it is a panning shot of moving objects such that MCTI cannot find the correct MVs to reconstruct SI. For the F-HOMO, it adopts pixel-based fusion and would lead to image discontinuity artifacts when fusing disparity synthesized and temporal interpolated (MCTI) images. The H-MVME outperforms MVME[26] with 0.5 to 3 dB higher PSNR for both high and low complexity videos. For MVME, it performs ME from both inter- and intraview KFs, which may lead to false/trivial ME and degraded quality, in addition to being time consuming. The H-MVME improves the MVME by eliminating the interview disparity. The proposed COMPETE estimated MVs with reference to perspectively transformed images, $\hat{I}^{v}_{2t-1}$, and detected no-motion blocks to eliminate regular ME operations. The SIFT followed by RANSAC would help to yield more stable matching point pairs, as compared to the COMPETE followed by DLT, as shown in Fig. 14. In comparison, the proposed COMPETE not only achieves the same reconstructed image quality as that of H-MVME but also decreases computation complexity. For the "Ballet," the SIF by COMPETE is 0.1 to 2.3 dB higher in PSNR than H-MVME because the disparity problem of interview ME has been solved by the block prediction through perspective transform. In comparison, the COMPETE effectively reduced computational complexity and well utilized interview and temporal correlations to eliminate disparity block matching noises. Experiments also justified that the proposed COMPETE can yield the best SI confidence, as compared to the others.

4.1.3 *Objective performance evaluation*

The PSNRs of $\hat{I}^{WZ}_{2t}$s coded by the five methods under the MDVC framework and reconstructed images by H.264, with intra-, inter- and inter-no-motion, are calculated for comparisons. The rate-distortion performance is similar to that of the SI confidence. For high-complexity videos, e.g., Race1, Ballroom, and Breakdancer, the SI confidence in PSNRs is comparable to COMPETE and H-MVME, both of which are 0.9 to 7.8 dB higher than MCTI and F-HOMO, as shown in Fig. 14. The reconstructed WZFs with the COMPETE, $\hat{I}^{WZ}_{2t}$s, are 0.8 to 2.9 dB higher in PSNR than those of MCTI and F-HOMO, as shown in Figs. 15(a)–15(c). For high-complexity videos, both MCTI and F-HOMO cannot estimate accurate MVs to compensate for the reconstructed SIFs, which leads to more degraded WZFs. Both COMPETE and H-MVME yield higher SI confidence and hence better reconstructed quality for $\hat{I}^{WZ}_{2t}$. The COMPETE yielded 0.4 to 1 dB higher PSNR than H.264/INTRA for Breakdancer, 1 to 1.5 dB higher than H.264/INTRA for Ballroom and 0 to 0.5 dB higher than H.264/INTRA for Race1. The H.264 intra/inter-no-motion cannot well encode Race1, because the camera was tracking a moving object. For the medium-complexity video, Exit, the SI confidence in PSNRs reconstructed by the COMPETE is 2.4 to 3.9 dB higher than those of MCTI, as shown in Fig. 14. The average PSNRs of $\hat{I}^{WZ}_{2t}$ are 3.5 and 2.5 dB higher than those reconstructed from H.264 intra and MCTI, respectively, as shown in Fig. 15(d). For low complexity videos, Ballet and Vassar, as they demonstrate more static regions, the interpolation and fusion process can perform efficient for all methods and results in smaller difference of PSNR performances. The COMPETE yielded 0.8 to 2 dB higher PSNR than MCTI for $\hat{I}^{WZ}_{2t}$s, and 1.5 to 2.2 dB higher than H.264/INTRA, as shown in Figs. 15(e) and 15(f). In addition, although the MVME-based methods,[26] e.g., MVME, and H-MVME, demonstrate comparable PSNR performances with COMPETE, their time complexity is high. Experiments showed that the COMPETE outperforms the others in SIF and WZF, $\hat{I}^{WZ}_{2t}$, quality, in that it prevents reconstructing blocks in static regions from noise attacks during interpolation and block matching processes. Note that the KF quality setting would impact the SI confidence, and the KF quality depends on QP selection. To justify the COMPETE capability in improving the MDVC codec

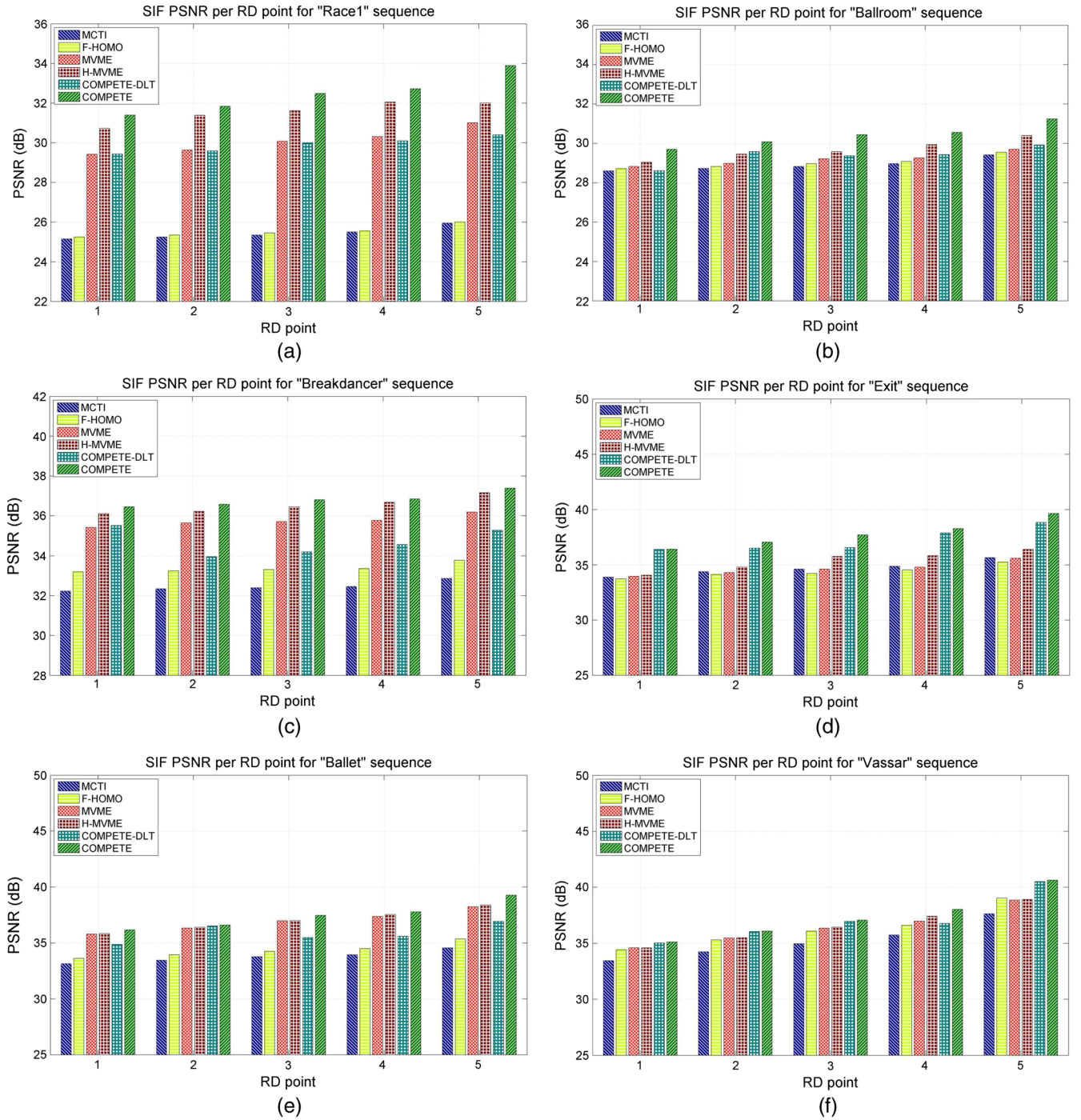

Fig. 14 Comparisons of SIF confidence in PSNR among different reconstruction methods applied on six test videos. (a) Race1, (b) Ballroom, (c) Breakdancer, (d) Exit, (e) Ballet, and (f) Vassar.

performance, the average image PSNR of KFs and WZFs under a fixed bit budget is provided for comparisons. As shown in Fig. 16, the COMPETE outperforms the others in PSNRs from 0.4 to 4 dB under different bitrates for both high and low complexity videos, Race1 and Vassar, respectively.

Experiments revealed that high confidence SI is much more important than the rate control method in DVC coding: (1) When SI confidence is low, the decoding confidence measure, *ConfPr* in Eq. (5), would not satisfy convergence condition, ConfPr $\leq 10^{-3}$. Under this condition, either the rate control procedure was carried out or the decoder requested more parity bits, and the ConfPr could hardly converge. (2) When SI confidence is high enough and the rate control procedure transmits high priority parity bits first, the number of decoding iterations would be reduced and the convergence criteria, ConfPr $\leq 10^{-3}$, would be reached quickly. One practical turbo decoder example[44] shows that when KFs are severely attacked by channel noise, which leads to low confidence SI, the PSNRs of reconstructed WZFs will

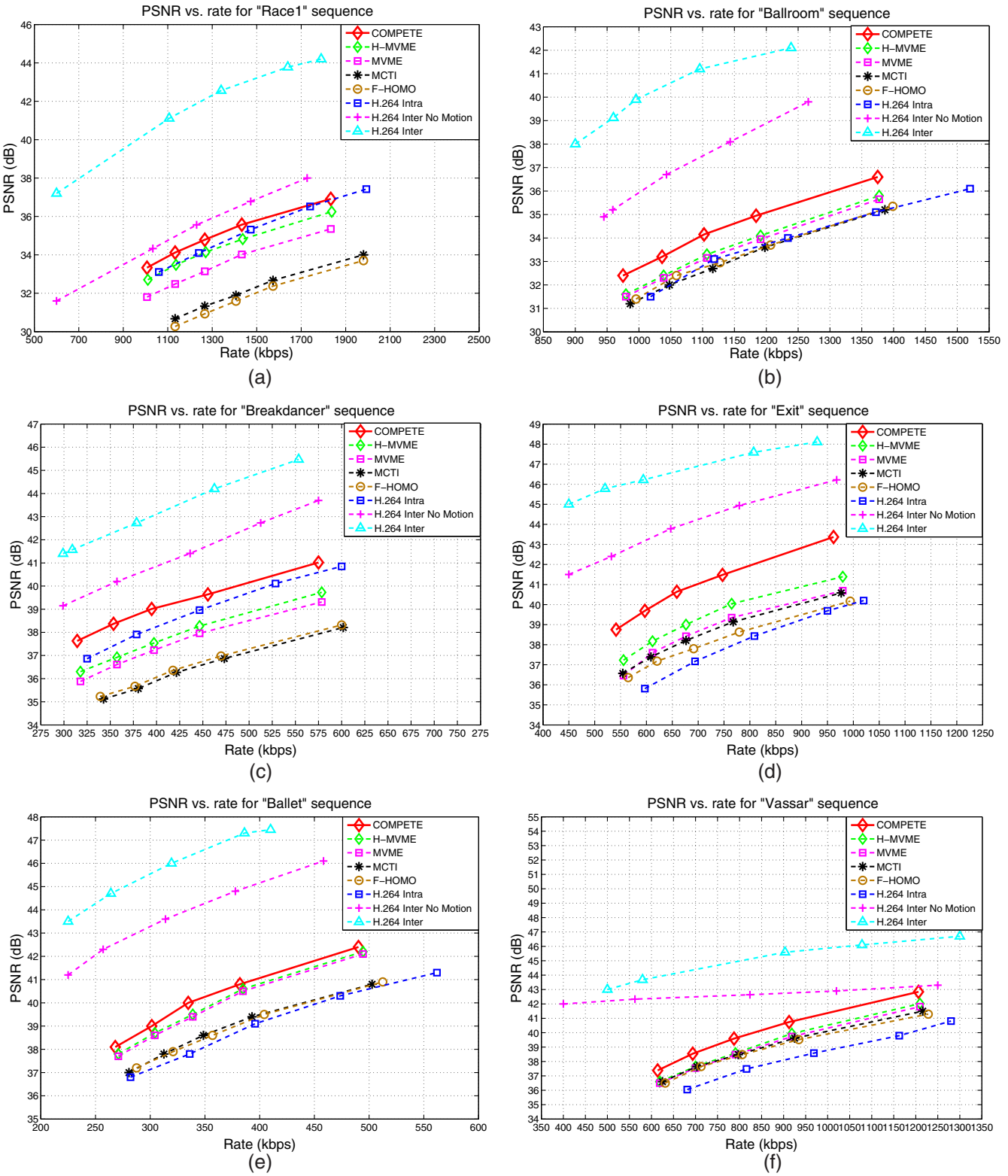

Fig. 15 PSNRs of reconstructed WZFs when encoding (a)–(c) high, (d) medium and (e) and (f) low complexity videos. (a) Race1 WZF, (b) Ballroom WZF, (c) Breakdancer WZF, (d) Exit WZF, (e) Ballet WZF, and (f) Vassar WZF.

degrade rapidly because the turbo decoder cannot recover one WZF from a severely degraded SIF. The number of average requested bits and bit rate saving under different SIF reconstruction methods is provided and compared in Tables 3 and 4, respectively. As shown in Table 3, the proposed COMPETE requested the fewest parity bits among the four methods because it can yield the highest SI confidence. Table 4 shows that the proposed control mechanism enables the four SI reconstruction methods to largely reduce the requested bit rates.

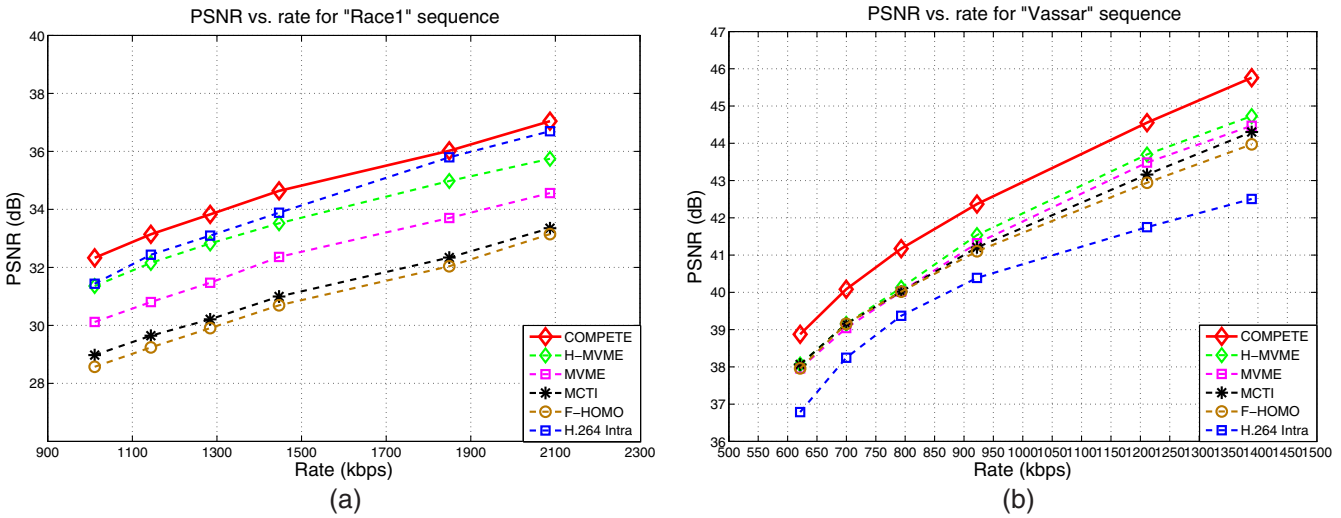

Fig. 16 The average image PSNRs comprising KFs and WZFs under different bitrates on high and low complexity videos: (a) Race1: KF + WZF and (b) Vassar: KF + WZF.

4.2 *Time Complexity Analysis*

The time complexities of the proposed COMPETE, together with the other SI reconstruction methods, are analyzed and discussed. At first, the number of arithmetic operations, addition/subtraction and multiplication/division, required to reconstruct SI is calculated for time complexity analysis. The practical execution time is also measured to justify the time analysis. Denote the image width and height as W and H, respectively, and the block size and search range as B_w and S_r, respectively.

4.2.1 *Motion Compensated Temporal Interpolation*

The MCTI performs intraview ME between images $\hat{I}_{2t-1}$ and $\hat{I}_{2t+1}$ and then performs motion compensated prediction to interpolate SI for WZFs. It performs subtraction and addition operations to yield the absolute difference summation. For one block, it needs B_w^2 subtractions and $B_w^2 - 1$ additions to calculate the block error. As the search area is S_r^2, it requires $(2 \cdot B_w^2 - 1) \cdot S_r^2$ operations for one block to finish ME operations. The number of total operations for one image to finish ME is $(2 \cdot B_w^2 - 1) \cdot S_r^2 \cdot (H \cdot \frac{W}{B_w^2}) \approx 2 \cdot S_r^2 \cdot H \cdot W$. The time complexity of MCTI is denoted as $T_{\text{MCTI}} = 2 \cdot S_r^2 \cdot H \cdot W$.

Table 3 The average requested bit rate of different SIF generation methods (15 FPS QCIF).

	SI generation			
	SIF			
Video	MCTI	F-HOMO	H-MVME	COMPETE
Race1	125.81	128.32	73.63	69.06
Ballroom	89.29	89.56	82.50	79.93
Breakdancer	44.68	40.56	34.14	30.52
Exit	46.22	52.77	45.52	38.54
Ballet	25.99	25.83	22.72	20.65
Vassar	40.19	44.92	37.33	36.32

4.2.2 *Fusion-based homography*

The fusion-based homography was implemented based on the Fusion 1 algorithm in Ref. 15. After performing perspective transformation, the synthesized perspectively transformed images, denoted as $\hat{I}'_v(2t) = \text{synthesis}[\hat{I}'_l(2t), \hat{I}'_r(2t)]$, and the temporarily interpolated image, $\hat{I}_c^{\text{int}}(2t)$, are considered as candidates for the fusion-based central-view image. For each pixel of the SIF to be reconstructed, it seeks to find the one, between $\hat{I}'_v(2t)$ and $\hat{I}_c^{\text{int}}(2t)$, that yields the minimum distance to both the previous and the next central-view image pixel values. Estimation of the initial 3×3 homography matrix can be performed off-line, whose time complexity can be ignored. For perspective transformation, it needs 15 MUL/ADD operations for each pixel and $2 \cdot 15 \cdot H \cdot W$ to yield the two reference central-view images. To obtain the fusion-based image, it needs $2 \cdot H \cdot W$ and $2 \cdot S_r^2 \cdot H \cdot W$ for temporal interpolation. In total, it needs $4 \cdot H \cdot W$ operations to find the pixel that yields the minimum pixel value difference. The number of total operations for the fusion-based homography is $(36 + 2S_r^2) \cdot H \cdot W$. The time complexity of this method is denoted as $T_{\text{homography}} = (36 + 2S_r^2) \cdot H \cdot W$.

4.2.3 *Hybrid Multiview Motion Estimation*

The H-MVME is an improved MVME.[26] In MVME, four ME vectors through inner paths are obtained and averaged to yield the motion compensated prediction image. As the MVME algorithm is designed based on the assumption that when the optical axes of all cameras are orthogonal to the motion. For multiview video, homography transformation is required and there exists an outlier that the MVME may not applicable. In H-MVME, it performs bidirectional temporal MC when the search range resides on the outlier.

Table 4 The turbo decoded bit rate comparisons W/ AND W/O rate control mechanism (15 FPS QCIF).

	SI generation							
	MCTI		F-HOMO		H-MVME		COMPETE	
Video	w/	w/o	w/	w/o	w/	w/o	w/	w/o
Race1	125.81	278.96	128.32	284.52	73.63	167.34	69.06	157.31
Ballroom	89.29	197.54	89.56	198.58	82.50	183.74	79.93	181.65
Breakdancer	44.68	101.78	40.56	93.67	34.14	79.58	30.52	72.32
Exit	46.22	105.05	52.77	118.85	45.52	105.37	38.54	91.54
ballet	25.99	61.73	25.83	61.50	22.72	54.75	20.65	50.37
Vassar	40.19	91.34	44.92	102.55	37.33	87.84	36.32	85.66

The required operations comprise performing the four inner paths ME $4 \cdot 2 \cdot T_{\mathrm{MCTI}}$, calculating weights (three ADDs and eight DIVs) $11 \cdot H \cdot \frac{W}{B_w^2}$ and calculating the average $7 \cdot H \cdot W$. The number of total operations $T_{\mathrm{H\text{-}MVME}}$ is $\left(\frac{11}{B_w^2}+7\right) \cdot H \cdot W + 8 \cdot T_{\mathrm{MCTI}} \approx 8 \cdot T_{\mathrm{MCTI}}$. Its time complexity is smaller than that of T_{MVME}, which is $16 \cdot T_{\mathrm{MCTI}}$.[26]

4.2.4 *COMPETE*

The design target of the proposed COMPETE is to keep high quality reconstruction while reducing computation complexity. At first, it needs to perform perspective transformation from side-view images to be with central view, which requires $6 \cdot 15 \cdot H \cdot W$ operations (three left- and three right-view images). Then, it performs block ME and checks whether it is a motion block or not. It needs at least T_{MCTI} operations. Assume the ratio of motion and no-motion blocks is 1:1. For no-motion block, direct copy from the co-located block of the previous image is adopted, and no operation is required. For motion blocks, the search range for finding disparity vectors can be minimized to $S_r^2/16$ in that the reference frames are perspectively transformed from side-view images. The COMPETE, as well as H-MVME, performs four inner paths ME two times. For inter-view ME, the first disparity vector estimation requires $4 \cdot (2B_w^2 - 1) \cdot \frac{S_r^2}{16} \cdot \left(\frac{H \cdot W}{B_w^2}\right) \approx 0.5 \cdot S_r^2 \cdot H \cdot W$ operations. The second ME after disparity compensation is $4 \cdot T_{\mathrm{MCTI}}$. Finally, by including all the required operations for computing weights and average, the number of total operations is $T_{\mathrm{MCTI}} + 6 \cdot 15 \cdot H \cdot W + 0.5[0.5 \cdot S_r^2 \cdot H \cdot W + 4 \cdot T_{\mathrm{MCTI}} \left(\frac{11}{B_w^2}+7\right) \cdot H \cdot W] \approx 4 \cdot T_{\mathrm{MCTI}}$, which is denoted as T_{COMPETE}. The above time complexity analysis shows that

$$\mathbf{T}_{\mathrm{MCTI}} < \mathbf{T}_{\mathrm{F\text{-}HOMO}} < \mathbf{T}_{\mathrm{COMPETE}} < \mathbf{T}_{\mathrm{H\text{-}MVME}}. \quad (11)$$

Table 5 The average time to encode one image (QCIF) in MDVC, H.264 with intra, inter no motion and intercoding mode (MSEC/FRAME) and CIF ones are provided for comparisons.

	Encoding time			
	Even frame		GOP = 12	
Video	MDVC QCIF (CIF)	H.264 Intra	H.264 Inter no motion	H.264 Inter
Race1	6.70 (23.51)	32.67	73.39	100.06
Ballroom	6.06 (23.07)	33.46	71.97	99.27
Breakdancer	6.06 (22.87)	29.83	68.18	106.38
Exit	6.96 (26.42)	30.15	66.92	90.12
Ballet	5.42 (21.51)	29.36	65.97	91.86
Vassar	6.06 (22.93)	31.57	67.71	89.02
Average	6.17 (23.38)	31.17	69.02	96.12

Table 6 The average time to construct one SIF (MSEC/FRAME).

	Reconstruction time			
	SIF			
Video	MCTI	F-HOMO	H-MVME	COMPETE
Race1	64.1	102.0	498.7	386.5
Ballroom	62.8	102.0	496.8	291.1
Breakdancer	63.8	100.4	495.2	312.5
Exit	62.8	100.8	496.2	242.3
Ballet	62.8	102.0	499.4	238.2
Vassar	63.1	102.4	497.1	193.9

Table 7 The average time saving of turbo decoding with different SI reconstruction methods as compared to MCTI.

	Decoding Δ Time (%)		
	$\Delta\text{Time}(\%) = \frac{T(\text{SI method}) - T(\text{MCTI})}{T(\text{MCTI})}$		
Video	F-HOMO (%)	H-MVME (%)	COMPETE (%)
Race1	1.92	−33.32	−37.45
Ballroom	9.78	−4.59	−6.42
Breakdancer	−4.74	−17.72	−19.89
Exit	13.64	3.31	−11.20
Ballet	12.61	−11.38	−17.44
Vassar	13.30	−6.05	−7.88

Experiments show that the execution time of COMPETE is only half that of H-MVME while achieving the same SI confidence. The execution time for COMPETE is only four times that of MCTI.

4.2.5 *Practical execution time evaluation*

The above time complexity analysis for different SI reconstruction methods is verified by practical execution time. All practical executions are implemented and executed on the same computer for fairness. The execution times of MDVC light encoder and H.264 encoder are first investigated.

Table 5 lists the average encoding time for one frame by MDVC, H.264 intra, H.264 inter no motion and H.264 inter, respectively. As shown, the MDVC light encoder spends about 5 to 15 times less than the others, which justifies the above time analysis. Table 6 lists the average execution time for reconstructing one SIF by MCTI, F-HOMO, H-MVME, and COMPETE, respectively. As shown, the average execution time for reconstructing one SIF of H-MVME is about eight times that of MCTI. For the COMPETE, this average execution time can be largely reduced for lower complexity videos. As the probability to process motion blocks in high complexity videos is high, the percentage of time reduction is limited, which is 1.29 to 2.56 times less than that of H-MVME. Table 7 lists the average turbo decoding time for different SI reconstruction methods. The performance of time reduction was evaluated based on the MCTI execution time for simplicity. Experiments showed that the decoding time would be reduced for higher SI

Fig. 17 Subjective performance comparisons of reconstructed WZFs, whose KFs are encoded with H.264 intra at QP = 26: (a) original; (b) MCTI; (c) F-HOMO; (d) H-MVME; and (e) COMPETE.

confidence, which justifies that the proposed COMPETE can provide better SI than the others.

4.3 *Subjective performance evaluation*

The subjective performance of different methods carried out on test videos is presented in this section. The QP control parameter of H.264 is set to be 26.

4.3.1 *Reconstructed Side Information Frames*

The SIFs reconstructed by MCTI and F-HOMO demonstrate severe block artifacts, which can be smoothed by the proposed COMPETE and modified H-MVME. But the latter suffered block noise in low complexity videos due to performing regular interpolation and block matching that led to static block noises. The proposed COMPETE effectively eliminates this block noise through weighted compensation and prediction.

4.3.2 *Reconstructed Wyner–Ziv Frame*

The SI confidence affects the reconstructed WZF quality. For one reconstructed I_{2t}^{WZ} by MCTI and F-HOMO, due to low SI confidence, many image blocks cannot be well recovered from low confidence SI. In comparison, the COMPETE and H-MVME yield higher SI confidence and hence higher quality for $\hat{I}_{2t}^{\mathrm{WZ}}$. Although COMPETE and H-MVME demonstrate comparable PSNRs for $\hat{I}_{2t}^{\mathrm{WZ}}$, the former consumed less computations. The resultant images are shown in Fig. 17. Reconstructed videos demonstrate that moving objects, cars and persons, are blurred from MCTI and F-HOMO based WZFs, while both COMPETE and H-MVME effectively eliminate this artifact for slow-motion videos, e.g., legs in Breakdancer.

4.4 *Practical Applications*

The WZ decoder combines the SI and the received parity bits to recover the original symbol. Additional parity bits would be requested if the original symbols cannot be reliably decoded. This request-and-decode process is repeated until an acceptable symbol error probability is reached.[2] The rate control performed by the decoder can reduce encoder computational loading. This feedback also enables the decoder to flexibly control SI generation from simple to sophisticated approaches, which can help to adapt to different encoder applications. However, this feedback channel used as an interactive decoding procedure may also hinder practical applications that require independent encoding and decoding. Instead of adopting this "decode-and-request" procedure, the decoder could be implemented with a correlation estimation algorithm, in which the rates of previously reconstructed frames are used to predict the required rates sent to the encoder. Feedback free[45] and unidirection DVC[46] have been proposed to make decoder operations independent of those of the encoder.

5 Conclusions

For a MVC that adopts DVC coding, MDVC, we proposed to utilize interview video correlations and exploit bit value probability distribution of transform coefficients under the block-DCT video codec framework to improve the SIF confidence and accuracy of decoded bits while speeding up the decoder rate control process. Contributions of this paper comprise (1) for specific multiview video applications, such as wireless video sensor and wireless video surveillance networks, the proposed MDVC utilizes the advantage of a DVC and multiview video framework to enable efficient and low complexity video encoding. Simulations verified that the MDVC can reduce encoding complexity to at least five times smaller than H.264/INTRA while enhancing the quality of reconstructed WZFs. (2) To improve the MDVC decoding performance, a multiview SI generation algorithm, COMPETE, was proposed to improve the quality of reconstructed SIF and WZFs. Both temporal correlation among intraview images and disparity correlations among interview images were well utilized to enhance WZF reconstruction. Simulation results showed that the PSNRs of reconstructed WZFs by COMPETE are 0.5 to 3.8 dB higher than those by MCTI when encoding low to high complexity videos. (3) To improve the MDVC rate control performance, we exploit the probability distribution of transform coefficient bits and reorder the transmission priorities of DCs and ACs, such that the turbo decoder would request the fewest bits to decode the WZF. Simulations demonstrate that the PSNRs of decoded WZFs are 0.2 to 3.5 dB higher than those encoded with H.264/INTRA under the same bit rates.

The COMPETE also outperformed H-MVME with 0.15 to 2.93 dB higher image PSNRs, in which the H-MVME outperforms MVME with 0.5 to 1 dB higher PSNR. Besides, the COMPETE effectively reduced the computation complexity, which is 1.29 to 2.56 times smaller than other SI reconstruction methods on average. Some recent research on video coding focus on free-view video codec and transmission. The proposed SI reconstruction method, COMPETE, under the MDVC framework can be extended to enhance the performance of free-view video codec that has to handle dynamic and mobile encoders and view reconstruction, which are considered as our future research. The COMPETE can also be carried out with a pixel-level disparity model. In addition, how to embed a small amount of information at the encoder[22] to improve the decoding efficiency, together with the pixel-level disparity model, are also considered as our future research.

Appendix: Linear Minimum Mean Squared Error

The LMMSE predictor is carried out to compute the w_j for a MC block $B_i(I, v)$ with four observations and can be represented as

$$E_i[e^2] = E_i\{[B_i(I_{2t}) - B_i(\hat{I}_{2t}^{int})]^2 | \forall B_i \in I_{2t}\} = E_i\left[\left(\mathbf{x}_i - \sum_{j=1}^{4} w_j \hat{\mathbf{x}}_{ij}\right)^2 | \forall x_i \in I_{2t}\right], \quad (12)$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_{ij}$ denote $B_i(I_{2t})$ in the original WZF and $B_i(\hat{I}_{2t}^{\mathrm{int}})$ in the reconstructed SIF, respectively. To minimize $E_i[e^2]$, it takes its first derivative as 0, i.e., $\frac{\partial E_i[e^2]}{\partial w_j} = 0$:

$$E_i\left[\hat{\mathbf{x}}_{ij}\cdot\left(\mathbf{x}_i-\sum_{j=1}^{4}w_j\hat{\mathbf{x}}_{ij}\right)\right]=0,\quad \text{or}$$

$$\sum_{j=1}^{4}w_j\mathbf{R}_{\hat{\mathbf{x}}_{ij}\hat{\mathbf{x}}_{ij}}=\mathbf{R}_{\mathbf{x}_i\hat{\mathbf{x}}_{ij}}. \tag{13}$$

The optimal weights, $\mathbf{w}=[w_1w_2w_3w_4]^T$, can be calculated through $\mathbf{w}=\mathbf{R}^{-1}_{\hat{\mathbf{x}}_{ij}\hat{\mathbf{x}}_{ij}}\mathbf{R}_{\mathbf{x}_i\hat{\mathbf{x}}_{ij}}$. This procedure can be carried out entirely at the encoder for higher accuracy but it conflicts with the design target of light encoding. For practical applications, as different videos demonstrate different MVs and the original WZF, I_{2t}, is not available at the decoder, the I_{2t} can be replaced by the MCTI frames, which are interpolated from $\hat{I}_{2t-1}$ and $\hat{I}_{2t+1}$ at the decoder. The LMMSE predictor in Eq. (13) is utilized for the MC to yield optimal weights for individual blocks reached with MV, $\vec{v}'_{m_j}$, instead of assigning the heaviest weight, $w_j=1$, for the block with the minimum SAD. Since only lossy reconstructed KFs are available at the decoder and the block with a MV of minimum SAD cannot always promise a best matched block. When KFs compression ratios are different, the MV prediction results will also be different and unstable. This optimally weighted MC effectively exploited interview disparity correlation for assigning different weights for blocks with different MEs, which can prevent block-based full search from trivial/unstable matching and increase prediction accuracy. Experiments showed that assigning weights obtained from the LMMSE estimator can improve the SIF PSNR up to 0.1 dB and 0.3 to 0.4 dB for low and medium-to-high complexity video, respectively, as compared to that with weights proportional to block fidelity [Eq. (2)]. The PSNR improvement would depend on the accuracy of the four MVs, $\{\vec{v}'_{m_j}\}_{j=1,\cdots,4}$, which would be degraded when encoding higher complexity videos. Under this condition, the difference among the four MVs would be enlarged, and the LMMSE estimator can help to yield stable weights for fusion blocks with different MEs. For low complexity videos, both the LMMSE estimator and normalized fidelity-based weighting strategy demonstrated comparable performances.

Acknowledgments

This work is partially supported by the Taiwan Ministry of Science and Technology with Grant No. MOST 105-2221-E-011-116 and Taiwan Building Technology Center with Grant No. IBRC 105H451709.

References

1. T. W. A. Vetro and G. Sullivan, "Overview of the stereo and multiview video coding extensions of the h.264/MPEG-4 AVC standard," *Proc. IEEE* **99**, 626–642 (2011).
2. A. M. A. B. Girod and S. D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE* **93**, 71–83 (2005).
3. A. D. L. Z. X. Ziong and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Process. Mag.* **21**(5), 80–94 (2004).
4. M. Flierl and B. Girod, "Coding of multi-view image sequences with video sensors," in *Int. Conf. Image Processing*, pp. 609–612 (2006).
5. X. Guo and Y. Lu, "Distributed multiview video coding," *Proc. SPIE* **6077**, 60770T (2006).
6. D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory* **19**, 471–480 (1973).
7. A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory* **22**(1), 1–10 (1976).
8. ISO, "Information technology-coding of audio-visual objects-part 10: advanced video coding," ISO/IEC Std 14496-10 (2004) https://www.cmlab.csie.ntu.edu.tw/~cathyp/eBooks/14496_MPEG4/iso14496-10.pdf.
9. R. Z. A. Aaron, S. Rane, and B. Girod, "Wyner-Ziv coding for video: applications to compression and error resilience," in *Proc. IEEE Data Compression Conf.*, pp. 93–102 (2003).
10. C. Z. Y. Cao, S. Gao, and G. Qiu, "Towards practical distributed video coding for energy-constrained networks," *Chin. J. Electron.* **25**(1), 121–130 (2016).
11. C. Yeo and K. Ramchandran, "The theory of a general quantum system interacting with a linear dissipative system," *Ann. Phys.* **19**(4), 995–1008 (2010).
12. F. D. M. Ouaret and T. Ebrahimi, "Iterative multiview side information for enhanced reconstruction in distributed video coding," *EURASIP J. Image Video Process.* **2009**, 591915 (2009).
13. E. A. X. Artigas and L. Torres, "Side information generation for multiview distributed video coding using a fusion approach," in *Proc. of Nordic Signal Processing Symp.*, pp. 250–253 (2006).
14. D. Kubasov, J. Nayak, and C. Guillemot, "Optimal reconstruction in Wyner-Ziv video coding with multiple side information," in *Proc. Multimedia Signal Process. (MMSP) Workshop*, 183–186 (2007).
15. F. D. M. Ouaret and T. Ebrahimi, "Fusion-based multiview distributed video coding," in *Proc. of ACM Video Surveillance and Sensor Networks*, pp. 139–144 (2006).
16. Y. W. H. Yin, M. Sun, and Y. Liu, "Fusion side information based on feature and motion extraction for distributed multiview video coding," in *Visual Communications and Image Processing Conf.*, pp. 414–417 (2014).
17. M. C. T. Maugey, W. Miled, and B. Pesquet-Popescu, "Fusion schemes for multiview distributed video coding," in *Signal Processing Conf.*, pp. 559–563 (2009).
18. F. Dufaux, "Support vector machine based fusion for multi-view distributed video coding," *Int. Conf. Digital Signal Process. (DSP)* 1–7 (2011).
19. S. Shimizu et al., "Improved view interpolation for side information in multiview distributed video coding," in *Int. Conf. on Distributed Smart Cameras*, pp. 1–8 (2009).
20. G. Petrazzuoli et al., "Novel solutions for side information generation and fusion in multiview dvc," *EURASIP J. Adv. Signal Process.* **2013**, 154 (2013).
21. S. Shimizu and H. Kimata, "View synthesis motion estimation for multiview distributed video coding," in *European Signal Processing Conf.*, pp. 2057–2061 (2010).
22. M. Makar et al., "Quality-controlled view interpolation for multiview video," in *Int. Conf. Image Processing*, pp. 1805–1808 (2011).
23. F. Pereira, J. Ascenso, and C. Brites, "Studying the GOP size impact on the performance of a feedback channel-based Wyner-Ziv video codec," in *Pacific-Rim Symp. Image Video Technol.* pp. 801–815 (2007).
24. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision* **60**(2), 91–110 (2004).
25. R. C. Bolles and M. A. Fischler, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM* **24**(6), 381–395 (1981).
26. E. A. X. Artigas and L. Torres, "A comparison of different side information generation methods for multi-view distributed video coding," in *Proc. SIGMAP*, pp. 450–455 (2007).
27. A. S. Barbulescu and S. S. Pietrobon, "Rate compatible turbo codes," *Electron. Lett.* **31**, 535–536 (1995).
28. D. N. Rowitch and L. B. Milstein, "On the performance of hybrid FEC/ARQ system using rate compatible punctured turbo (RCPT) codes," *IEEE Trans. Commun.* **48**(6), 948–959 (2000).
29. C. Brites and F. Pereira, "Correlation noise modeling for efficient pixel and transform domain Wyner-Ziv video coding," *IEEE Trans. Circuits Syst. Video Technol.* **18**(9), 1177–1190 (2008).
30. M. Salmistraro et al., "Joint disparity and motion estimation using optical flow for multiview distributed video coding," in *European Signal Processing Conf.*, pp. 286–290 (2014).
31. N. Deligiannis et al., "Side-information-dependent correlation channel estimation in hash-based distributed video coding," *IEEE Trans. Image Process.* **21**(4), 1934–1949 (2012).
32. P. Márquez-Neila et al., "Improving RANSAC for fast landmark recognition," in *Proc. Computer Vision and Pattern Recognition Workshop*, pp. 1–8 (2008).
33. K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(10), 1615–1630 (2005).
34. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, England (2004).
35. C. B. J. Ascenso and F. Pereira, "Content adaptive Wyner-Ziv video coding driven by motion activity," in *Proc. IEEE Int. Conf. on Image Processing*, pp. 605–608 (2006).
36. A. Z. P. H. H. Torr, "MLESAC: a new robust estimator with application to estimating image geometry," *Comput. Vision Image Understanding* **78**(1), 138–156 (2000).
37. M. F. D. Varodayan, D. Chen, and B. Girod, "The theory of a general quantum system interacting with a linear dissipative system," *EURASIP Signal Process. Image Commun.* **23**(5), 369–378 (2008).

38. E. S. A. Aaron, S. Rane, and B. Girod, "Transform-domain Wyner-Ziv codec for video," *Proc. SPIE* **5308**, 520–528 (2004).
39. K. Sayood, "Introduction to data compression," 4th ed., Morgan Kaufmann (2012).
40. K. L. D. Kubasov and C. Guillemot, "A hybrid encoder/decoder rate control for Wyner-Ziv video coding with a feedback channel," in *IEEE Workshop on Multimedia Signal Processing*, pp. 251–254 (2007).
41. S. K. Y. Vatis and J. Ostermann, "Inverse bit plane decoding order for turbo code based distributed video coding," in *IEEE Int. Conf. on Image Processing*, Vol. 2, pp. 1–4 (2007).
42. F. D. M. Ouaret and T. Ebrahimi, "Mulitiview distributed video coding with encoder driven fusion," in *Proc. of European Signal Processing Conf.* (2007).
43. ISO, "The theory of a general quantum system interacting with a linear dissipative system," Call for proposals on multi-view video coding, ISO/IEC JTC1/SC29/WG11, N7327 (2005).ftp://ftp.merl.com/pub/avetro/mvc-testseq/
44. J. J. Chen et al., "A multiple description video codec with adaptive residual distributed coding," *IEEE Trans. Circuits Syst. Video Technol.* **22**(5), 754–768 (2012).
45. J. L. Martinez, "Feedback free DVC architecture using machine learning," in *Proc. IEEE Int. Conf. Image Processing*, pp. 1140–1143 (2008).
46. W. A. C. F. M. Badem and A. M. Kondoz, "Unidirectional distributed video coding using dynamic parity allocation and improved reconstruction," in *Int. Conf. Info. Automation for Sustainability*, pp. 335–340 (2010).

Shih-Chieh Lee received his PhD from the National Taiwan University of Science and Technology in 2013 in electrical engineering. He is currently working at Nokia Networks as a network planning and optimization engineer. His research interests include image/video processing and the related topics in multimedia communications.

Jiann-Jone Chen received his PhD from the National Chiao-Tung University in 1997 in electronic engineering. He was a researcher with the Advanced Technology Center, Information and Communications Research Laboratories, Industrial Technology Research Institute (ITRI), Hsinchu. He is currently an associate professor in the Electrical Engineering Department of National Taiwan University of Science and Technology. His research interests include image/video processing, cloud video processing/streaming, image retrieval, and several topics in multimedia communications.

Yao-Hong Tsai received his PhD in information management from the National Taiwan University of Science and Technology (NTUST), Taipei, Taiwan, in 1999. He was a researcher with the Advanced Technology Center, Information and Communications Research Laboratories, Industrial Technology Research Institute (ITRI), Hsinchu. He is currently an associate professor with the Department of Information Management, Hsuan Chuang University, Hsinchu. His current research interests include image processing, pattern recognition, and computer vision.

Chin-Hua Chen received his MSEE degree from the National Taiwan University of Science and Technology in 2010. He is an engineer with the alpha network since 2012. His research interests comprise image/video processing, coding, and channel coding.