Efficient Multi-view Video Coding using 3D Motion Estimation and Virtual Frame

Manoranjan Paul*

Centre for Research in Complex Systems, School of Computing & Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia

1. Introduction

Capturing a scene using multiple cameras from different angles is expected to provide the necessary interactivity in the three-dimensional (3D) space to satisfy end-users’ demands for observing objects and actions from different angles and depths. Multi-view broadcast is becoming increasingly popular in commercial television networks for the added user-level interactivity. Multi-view also improves effectiveness of video surveillance systems. In addition to providing different perspectives, multiple views can offer a natural solution to the occlusion/de-occlusion problem [24][34][41], which often leads to incorrect object recognition and tracking. An occluded object is one that is obscured from view as there is another object in front of it. To encode multi-view video captured by multiple cameras, requires an efficient multi-view video coding (MVC) technique. MVC covers a wide range of active viewing experience, including stereoscopic (two-view) video (popularly known as 3DTV), free viewpoint television (2D view from any viewing angle can be generated from 3D scene modelling) and multi-view 3DTV. FutureSource predicts that 60% of US households will have a 3DTV by 2015 and 75% of households will have 3D Blu-ray players in a similar time frame [29]. More importantly, 3DTV adoption is outpacing HDTV adoption by around 50%. However, 3DTV viewer satisfaction does not meet customer expectations due to lack of content and interactivity. Otherwise, the 3DTV adoption rate would have been much higher. Considering the significant overlapping of the views and, more importantly, the availability of a rich set of relations on the geometric properties of a pair of views from camera properties, known as the epipolar geometry[1], joint encoding/decoding of views can achieve significant compression by exploiting inter-view correlation, in addition to the traditional intra-view correlation.

Usually multi-view video can be defined as the simultaneous multiple video streams from multi-view cameras. Multi-view cameras can be set up in different ways to capture a distinct scene simultaneously e.g., cameras may be trained on (i) a spherical shape to cover all $360^\circ$ angles in $x$-, $y$-, $z$-axes, (ii) inwards semi-circular arrangement at the same height position (i.e., convergent), or (iii) straight line parallel to the scene. The convergent position (Fig 1 (a)), is the most popular for its wide application in movies, advertising, educational video (e.g., surgical instructions), sports and general event broadcasting. However, the multi-view content captured from convergent viewpoints is more difficult for an encoder than that captured from a parallel camera setup, which can be regarded as a simplified form of the convergent setup once the angular difference between two adjacent cameras’ viewing directions is decreased to zero. The reason is that the disparity [1][5] (i.e., the translational displacement of a frame in one view into the frames of the same temporal instance in other views) intensity and disparity directions for the scenes are often much more intensive and heterogeneous across the multi-view frames.
captured using a convergent camera setup [1]. Fig 1(b) depicts the concepts of multi-view video with eight views (S), nine temporal (T) images of each view.

![Scene of interest](image)

Fig 1: (a) Convergent camera setup [1] and (b) Images from views and temporal positions with prediction structure recommended by the HEVC-3D extension standard for referencing different views (S) and different temporal (T) images for MVC [2]-[4].

Obviously, transmission and storing of multi-view video requires a huge amount of computation and data manipulation compared to single view video, although there is a significant amount of data redundancy among views. Recently, H.264/MVC and HEVC-3D [2]-[4] proposed a reference structure i.e., a mechanism for encoding an image using other already encoded images to exploit intra- and inter-view redundancy for improving rate-distortion (RD) gain in MVC. Fig 1 (b) depicts the hierarchical bi-predictive (HBP) reference structure [6][7], for example, T_k-th frame of S_1 view (i.e., B_1) uses two B_1 frames of the same view and two B_1 frames from S_0 and S_2 views respectively as reference frames. This structure clearly exploits intra- and inter-view redundancy from neighbouring frames for maximum compression gain which provides 20% more bitrate reduction compared to the simulcast technique where no-interview redundancy is exploited i.e., each view is encoded separately [2]. However, it introduces huge computational complexity and more importantly random access frame delay (RAFD) problem due to the dependency on other inter/intra-frames. The enormous requirement of computational time limits the scope of MVC applications especially for electronic devices with limited processing and battery power such as smart phones. Moreover, RAFD problem limits the interactivity capacity of the coding scheme results in long time intra-frame and inter-view video switching in immersive interaction system, 3D video on demand, etc. [8]. The RAFD for the highest hierarchical order is given by: 

\[ F_{\text{max}} = 3 \times T_{\text{max}} + 2 \times \left[ \frac{(N-1)}{2} \right] \]

where \( T_{\text{max}} \) is the highest hierarchical order and \( N \) is the total number of views [4]. For instance, in order to access a B-frame in the 4th hierarchical order (B4-frames in Fig 1 (b)), 18 frames must be decoded. By sacrificing compression gain, simulcast coding can remove inter-view switching and uni-directional intra-view referencing can remove intra-view image switching. Huayi et al. [8] modified the exhaustive referencing scheme (Fig 1 (b)) by encoding third view and sixth view independently and other views using those views. By doing this, they reduced inter-view and intra-view image switching time 30% and 0% respectively. Liu et al. [42] proposed three approaches using SP/SI frame coding, interleaved view coding, and secondary representation coding to provide low-delay RAFD. The method [42] should not outperform the MVC as it requires extra frames to encode and could not exploit nearest frame correlations. Abreu et al. [43] analyzed two types of frame reference prediction structure using one key view and two key views. Obviously this procedure could not provide better rate-distortion performance compared to the MVC structure (see Fig 1(b)) as it exploits less number of references compared to the conventional MVC prediction structure. In the proposed scheme, our aim is to reduce inter-view latency and achieve the lowest possible intra-view random access latency using a 3D video compression technique.

HEVC video coding standard improves the coding performance by reducing up to 50% bitstreams compared to its predecessor H.264/AVC by increasing computational complexity around 4 times [9]-[11] for a single view video. In addition, when the HEVC encodes multi-view videos, it requires multiple amounts of computational time compared to the HEVC. The enormous requirement of computational time limits the scope of 3D video coding applications especially for electronics devices with limited processing and battery power such as mobile, iPhone etc.

Although simulcast coding technique (where each view is encoded individually) using HEVC for multi-view videos is inferior compared to HEVC-3D in terms of RD performance, it does not have a RAFD problem. Recently, a video coding technique was proposed where a dynamic background frame (i.e., McFIS- the most common frame in a scene) was used as an extra reference frame [12][13][41] for encoding the current frame assuming that the motion part of the current frame would be referenced using the immediate previous frame and the static background part would be referenced using McFIS. McFIS is generated using Gaussian mixture model [14]-[16]. Other background frame generation techniques [30][31] are also available for video coding purpose; however, Gaussian-based McFIS is better due to its capability to capture uncovered background area. In this paper the first proposed scheme is a simulcast video coding approach (named HEVC-McFIS) based on HEVC using HBP prediction structure is proposed where McFIS of each view is used as an extra reference frame to encode corresponding view. This technique improves RD performance compared to HEVC-3D and simulcast HEVC schemes on multi-view videos which have significant amount of background areas.

The first proposed scheme HEVC-McFIS has no RAFD problem and provides better RD performance compared to HEVC-3D scheme, however, it may suffer RD performance for motion-active video sequences as it does not exploit any inter-view redundancy. Moreover, it takes relatively more computations. To improve RAFD problem of HEVC-3D and better compression, Li et al. [34] propose a MVC scheme where they form a 3D cube by arranging multiple views into a sequence similar with video sequence based on the view coherence. Then they apply 3D-DCT to exploit inter-view redundancy. Due to the applying the 3D transformation directly to the source images without any disparity or view adjusted, Li et al. cannot fully exploit the energy compaction property of the DCT transform [35][36]. An adaptive 3D-DCT technique is proposed in [35]. To overcome the limitation of the scheme [34], Zamarin et al. [36] form a 3D video by warping different views into a reference view using depth information. To
get the maximum benefit of energy compaction, Zamarin et al. apply 3D-DCT to the 3D warped block. The experimental results show that the technique outperforms H.264/MVC significantly at low bit rate. However, the RD performance of the scheme is inferior compared to H.264/MVC at high bit rate. The possible reasons of this are (i) it does not include any motion estimation, (ii) it does not include variable block-size for encoding, and (iii) it does not exploit any strategy for efficient coding of background and foreground. Note that Zamarin et al. use a strategy to handle occlusion. However, they could not exploit any explicit strategy for efficient coding of background and foreground including occluded areas. An improved 3D-warping technique is proposed in [37] by considering temporal displacement of the camera. Tian et al. [38] proposed an improved warping technique by considering block-based warping rather than frame-based warping to improve the computational performance.

To exploit the inter-view redundancy and improve RAFD problem and reduce computational time, in this paper the second new technique is proposed using 3D motion estimation (3DME). In the proposed 3DME technique, a 3D frame is formed using the same temporal frames (i.e., \( t^h \) frames) of all views and motion estimation (ME) is carried out for a macroblock (MB) or coding unit (CU) of the current 3D frame using the immediate previous 3D frame as a reference frame (which is formed by the \((i-1)^{th}\) frames of all views). To exploit the static background referencing the proposed 3DME scheme also uses dual frame referencing concept [25] where a 3D I-frame of the current GOP (i.e., Group of pictures) is used as the long term reference (LTR) frame (i.e., the second reference frame) in the dual reference frame concept. However, to control the computational time we do not use any ME using the LTR frame. As the correlation among the intra-view images is higher than the correlation among the inter-view images, the proposed 3DME technique does not degrade the RD performance significantly, but reduces the overall computational time and reduces the RAFD problem compared to HEVC-3D which enables interactive real time communications.

The proposed cuboid coding technique (i.e., 3DME) may suffer compression gain by avoiding hierarchical bi-prediction structure [7][26], which is the key factor in increasing random access latency. A number of existing techniques use view synthesis prediction which provides an effective way to reduce inter-view redundancy of multi-view video in addition to conventional disparity compensated prediction by generating a virtual view and then using it as a reference view [39]-[41]. Purica et al. [39] constructed multiple temporal predictions of the synthesized frame using optical flow techniques. Stefanoski et al. [40] generated an automatic synthesizes new view using image-domain-warping. Paul et al. [41] generated a virtual frame using dynamic background modelling and used it as a long term reference frame in hierarchical bidirectional prediction structure for a single video coding. To overcome RD problem, we extend the McFIS modelling idea to model a background in each view and we propose the third technique (named 3DME-McFIS) in this paper where an extra reference 3D matrix comprising McFIs of all views is used for 3DME (instead of 3D I-frame). Experimental results reveal that the proposed 3DME-McFIS technique outperforms the HEVC-3D by improving RD performance and reducing computational time without RAFD problem for most of the video sequences.

The contribution/novelty of the proposed methods can be summarized as (i) the proposed 3D motion estimation technique can exploit inter-view redundancy implicitly and intra-view redundancy explicitly to reduce inter-view switching with improved rate-distortion performance, (ii) the proposed integration of McFIS into the 3D cuboid compression structure can handle occlusion problem in a better way compared to the HEVC-3D scheme by referencing uncovered background areas from the McFIS, (iii) the proposed simulcast technique (HEVC-McFIS) eliminates the inter-view switching problem (i.e., inter-view dependency) with improved rate-distortion, (iv) the proposed 3D formation can exploit better redundant correlation using higher dimensional transformation, (v) analysis the number of views is needed to form 3D cuboid, and (vi) no motion estimation for the McFIS and overall fewer number of reference frames in the proposed methods provide improved computational time.

The rest of the paper is organized as follows. Section 2 describes the first proposed HEVC-McFIS coding technique with details of McFIS generation steps. Section 3 explains the second proposed 3DME video coding techniques with experimental rationalities. Section 4 describes the third proposed 3DME-McFIS coding techniques. Section 5 analyses computational complexity of the proposed techniques against the state-of-the-art method. Section 6 describes experimental set up and analyses experimental results, while Section 7 concludes the paper.

2. HEVC-McFIS: The First Proposed Technique

Simulcast coding technique using HEVC is inferior compared to MVC technique using HEVC-3D in terms of RD performance, however, simulcast technique has less RAFD problem. Moreover, McFIS-based coding technique outperforms the HEVC technique by exploiting static and uncovered background areas. Thus, a new simulcast technique is proposed where the current frame is encoded using three reference frames — two from bi-directional frames and a McFIS. Note that a McFIS of a view is generated using the frames of the corresponding view. The ultimate reference frame is selected at block and sub-block levels using the Lagrangian multiplier [32][33][27]. McFIS-based coding technique outperforms the HEVC-3D by exploiting mainly static and uncovered background areas of a scene.

Generally, we consider a pixel as a part of the background if it keeps its intensity for a number of frames. Based on this assumption, dynamic background modeling (DBM) [12]-[16] is formulated. We assume that \( k^\theta \) Gaussian at time \( t \) representing a pixel intensity with mean \( \mu_k \), standard deviation (STD) \( \sigma_k \), recent value \( \gamma_k \), and weight \( o_k \) such that \( \sum_{k=1}^{n} o_k = 1 \). The learning parameter \( \alpha \) is used to balance the current and past values of parameters such as weight, STD, mean, etc. After initialization, for every new observation \( X \) (pixel intensity at time \( t \)) is first matched against the existing models in order to find one (e.g., \( k^\theta \) model) such that \( |X - \mu_k| < 2 \sigma_k \). If such a model exists, update corresponding recent value parameter \( \gamma_k \) with \( X \). Other parameters are updated with the learning rate as:

\[
\mu_k' = (1-\alpha)\mu_k + \alpha X' ;
\]

\[
\sigma_k'^2 = (1-\alpha)\sigma_k^2 + \alpha(X' - \mu_k')(X' - \mu_k') ;
\]

\[
o_k' = (1-\alpha)o_k + \alpha .
\]

The weights of the remaining Gaussians (i.e., \( l \) where \( l \neq k \)) are updated as \( o_l' = (1-\alpha)o_l \). After each iteration, the weights are normalized. If the model does not exist, a new Gaussian model is introduced with \( \gamma = \mu = X', \sigma = 30, \) and \( o = 0.001 \) by evicting the \( K^\theta \) (based on \( w/\sigma \) in descending order) model if it exists. For more details in modeling and model updating, please refer [12]-[16]. To get the background pixel intensity from the above mentioned models for a particular pixel, we take the average of the mean pixel intensity and recent pixel value of the model that has the highest value of weight/standard deviation among the models of a pixel.
find the motion similarity (i.e., similar pixel-displacement of an object with respect to its previous frame due to rotation, translation or other kinds of movement), we have investigated the motion vector relationship among the views of the multi-view video sequences using four standard video sequences such as Exit, Ballroom, Vassar, and Breakdancing. First we determine the motion vectors of all MBs of each frame of a view using a 16×16 block size full search ME technique of ±15 search length. Then find the similarity of the motion vectors of a view with other views.

Four examples of McFISes are shown in Fig 2 using frames of Vassar, Ballroom, Exit, and Breakdancing video sequences respectively. Fig 2 (a), (b), (c), & (d) show the original frames of corresponding videos and (e), (f), (g), & (h) show McFISes. The red-dotted ellipses/rectangles in (e), (f), (g), & (h) indicate the uncovered/occluded background captured by the corresponding McFIS. To capture the uncovered background by any single frame is impossible unless this uncovered background is visible for one frame and that frame is used as the extra virtual reference frame as a dual frames reference coding technique [25]. The experimental results of RD performance and computational time requirements are analysis in Section 4 and Section 5 together with other schemes. The simulcast HEVC-McFIS can outperforms the MVC technique if a video has significant amount of background areas.

3. 3DME: The Second Proposed Technique

A scene is captured by a number of cameras which are placed in different angles in the multi-view system. As the same scene is captured by all cameras, there are inter- and intra-view redundancies. In general, we can assume that relative object movement within a view is very similar to that of other views. To find the motion similarity (i.e., similar pixel-displacement of an object with respect to its previous frame due to rotation, translation or other kinds of movement), we have investigated the motion vector relationship among the views of the multi-view video sequences using four standard video sequences such as Exit, Ballroom, Vassar, and Breakdancing. First we determine the motion vectors of all MBs of each frame of a view using a 16×16 block size full search ME technique of ±15 search length. Then find the similarity of the motion vectors of a view with other views.

Fig 2: Examples of McFISes and uncovered/previously occluded background using Vassar, Ball Room, Exit, and Break Dancing video sequences, (a), (b), (c), and (d) an original frame of Vassar, Ball Room, Exit, and Break Dancing sequences respectively; (e), (f), (g), and (h) corresponding McFISes of the videos respectively.

Fig 3: Motion relationship among multi-view frames: (a) average similarity of the motion vectors among different views for four standard multi-view video sequences where first 10 frames are used for each view of each sequence; (b) (c) (d) and (e) motion vector differences among two views for Exit, Ballroom, Vassar, and Break Dancing video sequences respectively.

Fig 3 (a) shows average similarity of motion vectors among different views where the first 10 frames are used for each view of each sequence. The figure confirms that the similarity is 51% to 93%. The experimental data indicate that the motion vector of the MB at the $i^{th}$ frame of the $j^{th}$ view has 51% to 93% of similarity with the co-located MB at the $i^{th}$ frame of other views. We have also plotted absolute motion vector differences for four standard multi-view video sequences in Fig 3 (b), (c), (d), and (e) respectively using first two frames. The figures confirm that a significant amount of MBs have zero motion vector difference between two views. To be more specific, Vassar and Ballroom sequences have more zero motion vector differences. We can exploit this relationship to avoid RAFD and computational time problems of the existing HEVC-3D prediction structure.

In the proposed 3DME technique, we can make a 3D frame comprising $i^{th}$ frames of a number of views and ME can be carried for a 3D MB (another dimension is formed using co-located MBs from different views) where the reference 3D frame would be formed using the immediate previous i.e., ($i-1$)$^{th}$ frames of all views. To get higher motion estimation and compensation accuracy on repetitive motions, non-integer motion replacement,
occlusion, multiple reference frames (MRFs) are used in the video coding techniques [41]. In MRFs-based coding technique more than one previously encoded/decoded frames are used as reference frames for motion estimation and compensation of the current frame, the ultimate reference frame from those frames is selected based on the Lagrangian Multiplier [33] to optimize the required bits and distortion. In the proposed 3DME scheme we also consider MRFs concept by using an LTR frame to get the reference for static background areas. In the implementation we have used the 3D I-frame of the current GOP as an LTR frame. For the LTR frame we do not use any ME because we assume that only static background area of the LTR frame will be used as reference for the current 3D MB. We do not apply ME for the LTR frame due to two reasons (i) to reduce the computational time requirement, (ii) to provide benefit for error resilience for the static areas as the static area is copied from the 3D I-frame. We do not use any explicit experimental results for error resilient using the proposed scheme in this paper, however, it is generally believed that using portion of I-frame provides better error resilient compared to that of P- or B-frame as I-frame is more robust compared to P- or B-frame in error propagation. In the proposed 3DME technique, we do not exploit inter-view redundancy explicitly, due to the following three reasons: (i) the correlation among the intra-view images is higher than the correlation among the inter-view images [1]-[4], (ii) to avoid RAFD problem for better interactivity, and (iii) to reduce computational time.

The proposed 3DME method does not require any disparity estimation [17] for inter-views as we do not explicitly use any inter-view relationships. Instead of multiple ME for each reference frame (e.g., B4-frame of S3 view at T3 position in Fig 1(b) requires 4 times ME using 4 reference frames), the proposed method requires only one ME. A significant amount of computational time reduction can be achieved using the proposed method as the proposed method does not need disparity estimation and ME for MRFs. The proposed method does not require any frame delay for random access which is another benefit of the proposed method against the existing prediction structure as all frames at \( T_t \) are available for encoding/decoding \( T_{t+1} \) frames (see Fig 1(b)).

4. 3DME-McFIS: The Third Proposed Technique

Although the proposed 3DME method successfully overcomes two limitations such as computational time and RAFD problem, it (with its current state) could not outperform HEVC-3D in terms of RD performance as the experimental results (in Fig 3) reveal that the motion vector similarity is not 100%. The experimental results also reveal that some cases such as motion active video sequences (Exit and Breakdancing), the motion vector similarity is around 50%. In results, the proposed method (i.e., 3DME) degrades the RD performance for those cases (see Fig 11). It is also worthy to note the utilization of the computational gain of the proposed method for improving the RD performance without sacrificing computational gain and RAFD problem.

As mentioned earlier, McFIS can successfully capture a static background including occluded background areas (if expressed once) from a scene of a video sequence. The 3D McFIS is formed using the McFISes of all views and then used it as the second reference frame when 3D ME was carried out for the current 3D frame. Note that a McFIS is two dimensional and when we pile-up all McFISes from a number of views, it forms 3D McFIS where the third dimension is the number of views. The proposed 3DME-McFIS technique uses McFIS as a second reference frame whereas the proposed 3DME technique uses 3D I-frame of the current GOP as a second reference frame. Obviously the 3DME-McFIS technique requires additional computational time compared to the 3DME technique due to the McFIS modelling, however, better RD performance is achieved due to foreground (using the immediate 3D previous frame) and background (using 3D McFIS) referencing. Note that similar to the 3DME scheme, the proposed 3DME-McFIS does not carry out ME on the McFIS.

4.1. Encoding and Decoding

For actual coding in both schemes, at first for a 3D MB, ME is carried out using intermediate previous 3D reference frame and find the distortion (in terms of SSD i.e., sum of square differences) between the current 3D block and motion compensated 3D block of the reference frame. We also calculate distortion between 3D block of the current frame and 3D block of the McFIS (or LTR frame for the proposed 3DME scheme). If the distortion using McFIS (or LTR) is less compared to the distortion using motion compensated 3D block from the immediate previous reference frame, we just treat the current 3D block as a skipped mode and copy 3D block from the McFIS (or LTR) for the current 3D block. This process is also applicable in sub-block level. If other distortion is less, then we encode the current 3D block using transform and entropy coding such as apply transformation, quantization, zigzag, variable length code etc.

It is well understood that compressibility in the frequency-domain increases at higher dimension through joint exploitation of spatial and temporal/disparity correlations. It has been demonstrated that preferential treatment of low-frequency 3D-DCT coefficients derived from 3D inter-frame blocks (co-located 2D-blocks of several successive frames) can achieve comparable compression without any motion compensation for low-motion
single-view videos [18]-[22]. Recently, the same principle is applied to inter-view 3D-blocks (co-located 2D-blocks of several views of the same instance) to achieve reasonable compression without any disparity compensation [23]. Motion and disparity compensation, however, is the key compressibility factor in multi-view coding. In fact, traditional techniques spend 50-70% of encoding time on just finding the best compensated reference blocks. Our preliminary proof-of-concept results suggest that any technique capable of using motion and disparity compensation on inter-view 3D-blocks is expected to rival the compression efficiency of the state-of-the-art 2D-block based techniques with much faster random access due to minimal frame- and view-dependency. According to Shannon’s famous source coding theory [28], coding efficiency tends to reach the achievable entropy-bound limit as more symbols are encoded jointly. Recently, higher dimensional channel codes such as Turbo and low density parity codes [28] have been successfully used for syndrome-based source coding. In the implementation 3D DCT on 4x4x4 block is applied, however, after getting DCT coefficients we encode them using HEVC zigzag and CAVLC for each view. Comprehensive analysis is needed to apply 3D Zigzag and 3D CAVLC code, thus we do not consider them in this paper.

Decoding is quite straightforward and similar to any residual-coded decoding. As all the views in a group will be decoded jointly, instantaneous switching among these views will be possible, resulting in zero inter-view latency. Moreover, absence of hierarchical bi-directional referencing will keep decoding dependencies to the minimum number of frames to achieve the lowest possible intra-view random access latency. One of the challenges of the proposed scheme was to establish that the improvement in compression efficiency due to volumetric coding can reasonably compensate for the absence of HBP referencing, which is the key factor in increasing random access latency.

4.2. McFIS updating

In the 3DME-McFIS technique, after encoding a 3D frame, we have updated 3D McFIS by updating individual McFIS (for each view) using the latest encoding frame of the corresponding views. For example (i-1)th 3D McFIS is used while ith 3D current frame is encoded and the ith 3D McFIS is updated using the ith encoded frames. The benefit of the updated 3D McFIS is to keep the McFIS relevant in terms of referencing.

After a scene change, the McFIS needs to be reset as the current McFIS is not relevant for the new scene of the video. In the proposed scheme a scene change is detected based on sum of absolute difference (SAD) and percentages of referencing using the McFIS. Thus, two different ways are used (i) based on the ratio of SAD, and SAD,i where SAD is calculated between 3D McFIS and 3D ith frame (i.e., the current frame) and SAD,i between the 3D McFIS and 3D (i-1)th frame (i.e., previous frame), and (ii) based on the percentage of the McFIS references in encoding. As the McFIS contains stable portion of a scene, SAD, between the current frame and the McFIS is a good indicator for scene change. In the proposed scheme, a scene change is considered if (SAD / SAD,i) > 1.7 . Paul et al. [12] mentioned that the percentage of McFIS referencing is a good indication to test the relevancy of the current McFIS as a reference frame. Thus, a new McFIS is generated if the percentage of the McFIS reference is below a threshold (e.g., in this implementation 5% is used).

4.3. Number of Views in 3D Frame Formation

A multi-view video may have a number of views. Thus, it is worthy to investigate how many views would be reasonable in the 3D frame formation of the proposed schemes for better performance in terms of RD, computational time, and random access latency. If a scene is captured using a large number of cameras (by maintaining reasonable angular distance among them), the overlapping regions of the scene among all cameras are less, thus, the proposed schemes need to perform 3D ME using a subset of cameras so that reasonable RD performance can be achieved. For better RD performance and higher interactivity (i.e., reducing inter-dependency among views for coding/decoding), a small number of camera views will be preferable in the proposed schemes for 3D ME as they will perform well due to the larger overlapping regions (i.e., redundancy) among views. However, if the number of views in performing 3D ME is less, more computational time is needed for coding the given number of total views. The RD performance and computational time are compared using three and four views. Fig 5 shows that the proposed methods with three camera views outperform the corresponding methods with four camera views in terms of RD performance. The proposed methods with large number of views in 3D frame could not perform better due to lack of exploitation of inter-view redundancy and less number of overlapping regions among the views.

![Fig 5: RD performance by the proposed schemes using three or four camera views in 3D frame formation for two standard video sequences namely Ballroom and Exit where 3DME-3C and 3DME-4C indicate the proposed 3DME scheme with three and four camera views respectively and 3DME-McFIS-3C and 3DME-McFIS-4C indicate the proposed scheme with three views and four views respectively.](image)

Fig 6 shows the percentage of time saving of the proposed schemes with different camera views in the 3D frame formation against the HEVC-3D for a wide range of QPs using Exit video sequence. The figure reveals that the proposed methods (3DME and 3DME-McFIS) with four views in 3D frame formation perform better compared to that of three views in terms of computational time. For example, the proposed scheme 3DME with four views (3DME-4V) saves 71 to 73% of computational time against the HEVC-3D whereas the proposed scheme with...
three views (3DME-3V) saves 59 to 65% of computational time against the HEVC-3D. The figure also reveals that other proposed scheme (3DME-McFIS) with four views in 3D frame formation (3DME-McFIS-4V) saves 58 to 62% of computational time against the HEVC-3D whereas the proposed scheme with three views (3DME-McFIS-3V) saves 48 to 52% of computational time against HEVC-3D. Obviously the computational time saving and RD performance of the proposed schemes are varied from sequence to sequence due to the amount of background, motion complexity, etc.; however, the normal trend is that if we increase the number of views in 3D formation, the computational time saving will increase but the RD performance will decrease. Thus, for compromising these two trends, we have selected three views in 3D frame formation in the proposed schemes.

5. Computational Complexity

One of the objectives of the proposed methods is to reduce the computational time of the existing MVC standard to enhance the scope of 3D video coding applications. Fig 5 and Fig 6 show two different performance scenarios of the proposed 3DME-McFIS and 3DME schemes. Fig 5 confirms that the proposed schemes (both 3DME and 3DME-McFIS) with small number of views in the 3D formation perform better in terms of RD performance; however, Fig 6 confirms that the proposed schemes perform better with large number of views in terms of computational time. Fig 7 shows computational time comparisons among the proposed methods and the existing video coding standard using 31 search lengths on four standard video sequences where three views are used in 3D frame formation of the proposed 3DME and 3DME-McFIS schemes. Fig 7(a) reveals that the proposed methods (3DME, 3DME-McFIS, and HEVC-McFIS) reduce around 62%, 50%, and 19% of the computational time on average using a wide range of bit rates compared to HEVC-3D standard when a search length 31 is used. Fig 7(b) shows detail computational performance for the individual video sequence in a certain quantization parameter (QP=32) of the proposed scheme against HEVC-3D. It can easily observed from the figure that the computational gains of the proposed schemes in Exit and Vassar video sequences are higher compared to other two sequences. The main reason for this would be larger amount of background areas in those sequences compared to others. Thus, a large number of blocks use reference blocks from McFIS of the proposed schemes compared to the second reference frame of the HEVCs. As any reference from McFIS requires short motion search, the computational gain is higher in the proposed schemes for those video sequences with higher background areas. Experiments are conducted on a PC with Intel(R) Core (TM) 2 CPU 6600@2.40 GHz, 2.39 GHz, and 3.50 GB of RAM.

The proposed HEVC-McFIS scheme saves 19% computational time compared to HEVC-3D. Fig 1(b) shows that HEVC-3D requires ME for four references frames for a significant number of frames; however, the proposed HEVC-McFIS scheme requires ME for two reference frames and only checking the McFIS frame (no ME is performed using the McFIS). Thus, we have computational saving of the proposed scheme compared to HEVC-3D. On the other hand, the proposed scheme (3DME and 3DME-McFIS) save huge amount of computational time due to only one reference frame for ME and processing all views at a time in encoding. In the proposed scheme we do not use any ME for the virtual reference frame (i.e., McFIS) as we assume that if the current MB is a part of normal background or uncovered background, it does not have motion. Due to the fixed amount of operations requirement for the McFIS modelling, the proposed 3DME-McFIS method reduces computational time slightly less compared to that of the proposed 3DME scheme. When large search length is used, the computational time requirement for the McFIS modelling is negligible compared to the ME. The proposed 3DME-McFIS scheme does not use ME using the McFIS as the McFIS is only used for referencing the background which has no motion. Thus, the computational time reduction for both methods is almost the same for large search length.
6. Experimental Results

To compare the performance of the proposed schemes (HEVC-McFIS, 3DME and 3DME-McFIS), all algorithms are implemented based on the HEVC-3D recommendations [44] with 25 frames per second, 32 coding unit size, Lagrangian multiplier for rate-distortion optimization, symmetric and asymmetric block-partitions, smallest transformation block, ±31 search length, quarter-pel accuracy, and 16 GOP size. In the proposed schemes, the IBP prediction format is considered, whereas hierarchical B-picture predication structure is used for HEVC-3D and HEVC-McFIS schemes. Obviously the proposed HEVC-McFIS and 3DME-McFIS techniques take extra operations to generate McFISes. The same technique for modeling McFIS at the encoder and decoder is used so that the McFISes are not encoded and transmitted to the decoder. For the experiment four standard multi-view texture videos namely Breakdancing (384x512), Exit (244x320), Ballroom (244x320), and Vassar (244x320) are used. Appropriate common test condition is applied for comparison with different state-of-the-art methods [45].

Fig 9 shows MB reference distributions using the proposed 3DME-McFIS scheme where MBs with black area is referenced using the immediate previous frame and MBs with normal scene area (added 25 for better visibility) is referenced using the McFIS.

Fig 8 shows the average percentages of MBs references using the LTR frame or McFIS of the proposed schemes. In the HEVC-McFIS scheme McFIS is used as the third reference frame. The figure reveals that 35 to 53% of CUs use McFIS as a reference frame. The percentage of references decreases with the bit rates increases. In the proposed 3DME-McFIS scheme 3D McFIS is used as the second reference frame. The figure reveals that the percentage of references almost the same for different bit rates. The percentages are within 24 to 29%. In the proposed 3DME scheme a 3D LTR frame (the first 3D I-frame of the current GOP) is used as the second reference frame. The figure reveals that the percentages are within 12 to 14% and slightly increasing with the bit rates. As the percentages of references using background frame (McFIS) or LTR frame of the proposed schemes for referencing static background and uncovered background are significant (i.e., 12 to 53%), this indicates that the proposed schemes would improve the RD performance with reduced computational time as the referencing LTR or McFIS does not require any ME cost.

Fig 9 shows an original frame and its MB referencing distributions of the proposed 3DME-McFIS scheme using Exit video sequence. Fig 9(a) shows original 36-th frame of the first view of the Exit video sequence and Fig 9(b) shows reference distribution using the proposed scheme where CUs with black area is referenced using the immediate previous frame and MBs with normal scene area is referenced using the McFIS. It is clear from the Fig 9 that static background and uncovered background areas are normally referenced using the McFIS. Note that the corresponding block from the McFIS is copied if the distortion using McFIS (i.e., between the current block and the McFIS) is less compared to the distortion using the immediate previous reference frame (i.e., between the current block and motion compensated block of the immediate previous frame).

Fig 10 shows the MB reference distribution for the first 145 frames of the sequence almost at the same bit rates. Note that for clear visualization only first 145 frames are given in the figure; however, for other experimental results (such as computational time, rate-distortion performance, etc.) all frames of a sequence are used. The figure demonstrates that the proposed scheme provides more stable PSNRs for frames within a GOP compared to that of HEVC-3D. In the hierarchical B-picture predication structure, HEVC-3D uses different QPs within a GOP, thus, the PSNR fluctuation within GOP is higher, which creates poor perceived video quality compared to the same PSNR output with less PSNR fluctuation. In the proposed scheme we use the same QPs for all frames within a GOP (except two extreme positioned frames). Thus, the proposed scheme provides better perceived video quality as the proposed scheme exhibits less PSNR fluctuation. Moreover, the proposed scheme provides better average PSNR (around 0.60dB almost at the same bit rates) as well as frame level PSNRS. Initially, the proposed scheme does not provide better PSNR as the McFIS generation is not stable with few frames. Note that the PSNR increases with the frames for the proposed scheme compared to HEVC-3D due to better McFIS. A number of frames (e.g., 25 frames or more) are required to get better uncovered background for the McFIS.
Fig 11: Rate-distortion performance by HEVC-3D and the proposed schemes (3DME, 3DME-McFIS, and HEVC-McFIS) using four standard video sequences namely Exit, Ballroom, Vassar, and Breakdancing.

Fig 11 shows RD performance using HEVC-3D and three proposed schemes such as HEVC-McFIS, 3DME and 3DME-McFIS using four standard multi-view video sequences. The figure reveals that the RD performance of the proposed 3DME scheme is inferior to the HEVC-3D. However, the proposed 3DME scheme outperforms HEVC-3D by reducing computational time by 62% (see Fig 7) and reducing RAFD problem. The proposed 3DME-McFIS scheme outperforms HEVC-3D in terms of RD performance by improving 1.0dB, 0.5dB, 0.0dB, and 2.2dB PSNR for Exit, Ballroom, Breakdancing, and Vassar respectively. The proposed 3DME-McFIS scheme also saves computational complexity by 50% compared to the HEVC-3D. Due to the huge motions (less effectiveness of background modeling), the proposed 3DME-McFIS method does not outperform HEVC-3D for the Breakdancing sequence significantly. The proposed HEVC-McFIS scheme outperforms the HEVC-3D in terms of RD performance and saves 19% computational time compared to HEVC-3D. All proposed schemes (3DME, 3DME-McFIS, and HEVC-McFIS) reduce RAFD problem compared to the HEVC-3D for better interactivity.

Li et al.’s MVC coding technique [34] is one of the relevant technique with the proposed method as they applied 3D-DCT after forming 3D frames. Moreover, Jacob et al. [35] improved the 3D-DCT technique. Thus, we combine both techniques (Li + Jacob) and compare the RD performance with the proposed 3DME-McFIS method. Zamarin et al.’s MVC technique [36] is another relevant technique with the proposed method as they form 3D frame using warping based on depth information and applied 3D-DCT. To make it more effective and compare with the proposed method in equal platform, we incorporate Sangappa et al.’s 3D warping [37] for improving motion vector and McFIS to improve the quality of warped image. A number of methods [8][42][43] tried to reduce RAFD problem by restructuring the reference frame in MVC. Among them Abreu et al. [43] proposed the latest one. Thus, a comparative study against the proposed 3DME-McFIS method in terms of RD performance is also included. Note that the whole interactive streaming system of [43] is not considered for comparison, the IBP prediction structure proposed in [43] with one key view is only used as the interactive streaming system is not integrated neither in the proposed methods nor in HEVC-3D scheme. Fig 12 shows the RD performance of the best performed proposed method i.e., 3DME-McFIS against four state-of-the-arts methods (HEVC-3D, Li + Jacob, Zamarin + Sangappa, and Abreu et al.). The figure reveals that the proposed method (3DME-McFIS) outperforms other relevant state-of-the-art methods. The proposed 3DME-McFIS method saves bits compared to HEVC-3D and other state-of-the-art methods [8][34][36][42][43] by avoiding explicit encoding static background and uncovered background regions by referencing McFIS. Moreover, the ME using 3D inter-view block and searching in the different temporal areas in the proposed 3DME-McFIS scheme can implicitly exploit intra-view redundancy and explicitly exploit inter-view redundancy respectively, thus, it has the ability to outperform the low-delay RAFD techniques [8][42][43]. In addition, the proposed technique is able to outperform techniques in [34][36] by exploiting ME from the reference frames to reduce the residual. It also avoids complicated and inefficient warping problem.
7. Conclusions

This paper outlines three proposed schemes including 3-Dimensional motion estimation and motion compensation schemes to reduce the computational time and reduce the random access frame delay of the existing HEVC-3D multi-view video coding standard. To reduce random access frame delay, firstly a simulcast McFIS-based technique on the HEVC platform was proposed. The proposed technique outperforms HEVC-3D in terms of RD performance with 19% reduced computational time. However, it could not exploit inter-view redundancy at all, thus, it could not provide significant RD performance improvement. In the proposed 3DME technique, a 3D frame is formed using the same temporal frames of all views and motion estimation is carried out for a block of the current 3D frame using the immediate previous 3D frame as reference frame. A second reference frame is also used as long term reference frame for referencing static or de-occluded background areas without motion estimation in the 3DME scheme. This technique outperforms the existing standard by reducing computational time by 62% and reduce random access frame delay by sacrificing rate-distortion performance compared to the state-of-the-art methods and HEVC-3D standard.

The paper proposed the final technique (i.e., 3DME-McFIS) where an extra 3D reference frame is used in addition to the immediate previous 3D frame. The extra 3D frame is formed using dynamic background frames of each view which are popularly known as McFISes (the most common frame in a scene) based on Gaussian mixture modelling. The experimental results reveal that 3DME-McFIS outperforms the HEVC-3D coding standard by improving 0.90dB PSNR on average, by reducing computational time by 50%, and by reducing random access frame delay compared to the existing HEVC-3D multi-view video coding standard. The proposed technique also outperforms other three state-of-the-art methods. The proposed techniques enhance the 3D video coding application scopes at the interactive real time video communications.

Acknowledgement

This work was supported in part by the Australian Research Council under Discovery Projects Grant DP130103670.

References


![Fig 12: Rate-distortion performance comparison using the proposed 3DME-McFIS scheme against four state-of-the-art methods.](image-url)


[26] HTM 6.2 Software for HEVC-3D
https://hevc.hhi.fraunhofer.de/svn/svn_3D/CSoftware/tags/HTM-6-2/