Efficient Multiview Video Coding using 3D Coding and Saliency-based Bit Allocation

Manoranjan Paul, Senior Member, IEEE

Abstract— Capturing a scene using multiple cameras from different angles is expected to provide the necessary interactivity in the 3D space to satisfy end-users’ demands for observing objects and actions from different angles and depths. Existing multiview video coding (MVC) technologies face trade-off among rate-distortion performance, random access frame delay i.e., interactivity, and computational time. To address above mentioned trade-offs, a novel cuboid MVC strategy is proposed with 3D frame referencing structure to improve interactivity and computational time, an additional reference frame to improve rate-distortion performance for occluded areas, and visual attention-based bit allocation to provide better perceptual video quality. The experimental results reveal that the proposed scheme provides better interactivity, reduced computational time, and better perceptual quality compared to the 3D-HEVC implementation, HTM 15.0.

Index Terms—3D motion estimation, 3D DCT, uncovered background, visual attention modeling, variable bit allocation.

I. INTRODUCTION

MULTIVIEW broadcast, where an event is captured with multiple cameras from different viewing angles, is becoming increasingly popular in commercial television networks for the added user-level interactivity. Considering the significant overlapping of the views and, more importantly, the availability of a 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. 3D-HEVC [56][54] and H.264/MVC [2]-[4][60] supported hierarchical B picture (HBP) prediction frame reference structure (see Fig 1) [5][6] among views (S) and temporal (T) texture images to obtain maximum rate-distortion (RD) performance. This prediction provides 20% more bitstream reduction compared to the simulcast technique with only intra-view prediction [2].

Fig 1 shows the HBP structure where 8 texture views are used with 8 group of picture (GOP) size. In the figure, arrows are used to show the encoding frame with its reference frames and subscription is used to show different levels of I, P, and B-frames. Here, a frame can use up to 4 reference frames from inter/intra-views. To encode/decode a frame, we need to encode/decode a number of frames in advance, thus, the structure introduces random access frame delay (RAFD) problem and restricts the interactivity. RAFD is measured based on the maximum number of frames that must be decoded in order to access a B-frame in the hierarchical structure. The access delay for the highest hierarchical order is given by:

\[ F_{\text{max}} = 3 \times l_{\text{max}} + 2 \times (N - 1)/2 \]

where \( l_{\text{max}} \) is the highest hierarchical order and \( N \) is the total number of views [4].

Due to the RAFD problem, some applications e.g., interactive real-time communication may not be possible using the existing prediction structure. The scenario is worse if we consider joint texture and depth encoding in the HBP prediction. From now on, we refer 3D-HEVC as multiview HEVC (MV-HEVC) as we only consider texture video in this paper rather than both texture and depth videos.

This work is partially supported by the Australian Research Council (ARC) Discovery Project (Project id: DP130103670).

M. Paul is with the School of Computing and Mathematics, Charles Sturt University, Australia (phone: +61-2-6338 4260; Fax: +61-2-63384260; email: Mpaul@csu.edu.au).
In this paper a technique is proposed using a novel cuboid MVC framework with disparity-adjusted (DA) 3D frame formation, and 3D coding to overcome the interactivity and computational time problems of the existing MVC technique. In the proposed Cuboid technique, inter-view disparity is effectively used to align scene-overlapped regions of views to achieve strong correlation among the 2D-blocks in the 3D-blocks.

Recently dynamic background frame [12]-[15] is used in video coding techniques [16]-[29]. Paul et al. [25][26] used the most common frame in a scene (McFIS) using Gaussian mixture-based dynamic background modelling (DBM) for video coding. The McFIS is used as an additional reference frame in dual-frame referencing scheme [30]-[32][60] assuming that the motion part of the current frame would be referenced using the immediate previous frame and the static/uncovered background part would be referenced using the McFIS. The ultimate reference area is selected at block and sub-block levels using the Lagrangian multiplier [8][10]. To improve the performance of the Cuboid technique, we also propose Cuboid-McFIS technique where a 3D McFIS is used as the second reference frame.

Yuan, et al. proposed a method [61] to improve overall RD performance by allocating bits based on the inter-view dependency between reference view and non-reference view for MVC. The results reveal that the method provides better RD performance and inter-view quality consistency compared to the existing methods. On the other hand, perceptual video coding (PVC) allocates bits for encoding a given frame non-uniformly [41]-[43] using visual attention modelling (VAM) [39][40]. The basic idea is that the PVC techniques improve the perceived video quality for a given bit budget by allocating more bits to encode attentive areas and fewer bits for non-attentive areas. For better perceptual quality, we also proposed Cuboid-PVC-McFIS technique for variable bit allocation through innovative quantization approach based on the combined saliency map of moving region and central tendency.

The contributions of the paper are (i) 3D motion estimation and compensation for reducing interactive problem of the MV-HEVC, (ii) 3D frame formation using DA views for determining better motion vectors through motion estimation (ME) and motion compensation (MC) based on object alignments in different views, (iii) 3D video coding using 3D DCT and 3D Zigzag for better compression, (iv) incorporating McFIS for better exploitations of uncovered areas, (v) incorporating variable bit allocation based on the new saliency map with moving region and central tendency for better perceive quality, (vi) analysis of RAFD problem of the proposed scheme against MV-HEVC scheme, and (vii) quality assessment in subjective and objective tests.

The preliminary idea is published in [58] based on H.264/MVC. The novelty of the proposed schemes can be summarized as (i) fundamentally different coding framework using cuboid or volumetric 3D frame, (ii) alignment of frames based on the object location using disparity-adjusted 3D formation, (iii) joint intra and inter-view redundancy exploitation through 3D DCT and Zigzag, (iv) alignment and exploitation of uncovered background areas through 3D McFIS formation and referencing respectively, and (v) determination of variable QP values corresponding to coding unit (CU) using saliency map and central tendency for variable bit allocation through variable quantization.

The rest of the paper is organized as follows. Section II describes the proposed disparity-adjusted Cuboid with McFIS. Section III explains proposed perceptual video coding technique. Section IV describes the experimental results with set up, while Section V concludes the paper.

II. PROPOSED DISPARITY-ADJUSTED MVC TECHNIQUE

Existing MVC technologies either sacrifice interactivity to improve computational time and RD performance or sacrifice RD performance to improve computational time and interactivity. In general, we can assume that object movement of a view is very similar to that of other views. Thus, relative movement of objects between two temporal positions on different views should have strong correlation. To exploit the correlation, reduce the computational time, and improve the interactivity firstly we have formed 3D frame comprising all the same temporal positioned frames of all views. In the 3D frame, 2D is formed using frame and the additional dimension is formed using frames from different views. After formation of 3D frames, we estimate motion for the current 3D CU using the variable size blocks (e.g., 32×32×8 to 8×8×8, etc. where last dimension is for the number of views) using the previous 3D reference frame. Then, we encode 3D CU using 3D coding. Details are described in the sub-sections.

![3D Frame Formulation]

Fig 2: Example of a disparity-adjusted 3D-block formation from 3 views that were captured synchronously for Breakdancing video.

A. 3D Frame Formation

According to Kaup et al. [59] and our investigation, a major portion of references (70-90%) in MVC are coming from inter-view. Thus, using 3D formation and 3D ME we may sacrifice up to 30% RD performance, however, we can achieve more desirable two properties such as better interactivity and computational time that enhance the scope of the MVC. We can form 3D frame in different ways. In
this paper we have discussed two ways: (i) just plain stacking the same temporal positioned frames of all views and (ii) plain stacking the scene overlapped regions of the same temporal positioned frames of all views. In the second case, we need to find disparity [33] between two adjacent views and then we can form 3D frame using the scene overlapped regions. As the objects are not normally aligned in different views (i.e., objects are not at the same position in different views), thus, the first approach of 3D formation may produce misleading motion vectors while 3D ME is carried out. An example of object locations captured synchronously in different views for Breakdancing video sequence is shown in Fig 2. This figure inspired us to form 3D frame using disparity adjusted views as the positions of the same object are different in different views. We can align the object at the same position by removing disparity among the views.

To find the scene overlapped regions, we need to conduct disparity estimation between two same temporal positioned frames from two adjacent views. When we find all disparities between adjacent views, we only take the common scene or overlapped regions of the scene to form disparity adjusted 3D frame. The formation processed is shown in Fig 3. For 3D formation, first we need to find disparity among different views. The determination of disparity is performed in frame level. In the proposed scheme we take a frame from one view and the same temporal frame from other view. We find the matching point (i.e., disparity) based on the minimum value using sum of absolute difference by shifting one frame against other frame horizontally assuming that there is no disparity in the vertical direction as the videos are captured using the cameras with the same height i.e., co-planner. We need to calculate the disparity once for an entire video sequence as we assume that the disparity is the same for all frames of a video, thus, 3D formation in both encoder and decoder is simple. If the calculated disparities of the first view and the second view against the third view are $D_1$ and $D_2$ respectively and to make the overlapping areas among three views are multiple of CU size $\Gamma$, we need to form a 3D frame in reduced size by considering an offset ($F$) (see Fig 3 (b) for 3D formation using three views). Thus, an overlapping area would be $H \times (W-D_1-D_2-F)$ where $H$ and $W$ are the height and width of a video frame respectively and $F = \Gamma - D_1 - D_2$. For each video sequence, we need to send the information of overlapped and non-overlapped region only once for an entire video sequence.

3D ME is carried out using the scene overlapped regions and the rest parts are coded separately. In the above description with Fig 3 (b), the size of the non-overlapping regions of a frame is $Hx\Gamma$. This area is normally a background area of a video sequence as moving regions occupied in the middle area of a frame. We encode the non-overlapping areas using HEVC [57] by considering only intra-view residual coding. Motion vector relationship is also investigated among the views of the multiview video sequences using the first approach of 3D formation. The experimental data indicate that the motion vector of the CU at the $j^{th}$ frame of the $j^{th}$ view has $51\%$ to $93\%$ of similarity with the co-located CU at the $i^{th}$ frame of other views. Using the proposed cuboid formation approach, the average motion vector relationship is far better (i.e., $70\%$ to $95\%$).

We also compare the motion vectors by observing if $\min \left( d_{x_1}, d_{x_2}, d_{x_3}, d_{x_4} \right) \leq d_{x_j} \leq \max \left( d_{x_1}, d_{x_2}, d_{x_3}, d_{x_4} \right)$ and, if true, 3D ME is considered acceptable for that block. In the condition, $d_{x_j}$ is the $x$ motion vector obtained by performing ME on the 3D CU and $d_{x_k}$ is the $x$ motion vector estimated using just the $k^{th}$ view. This evaluation method we considered, with the thoughts behind it being that if the motion vector found using 3D CU is outside of the range of the set of motion vectors obtained with individual views, then the motion vector is almost certainly not viable. The average acceptance is $0.98$ for Exit video sequence, which indicates Cuboid creates reasonable motion vector. Fig 4(left) shows the distribution of motion vectors of Frame 11 for Exit sequence where, ‘1’ indicates the 3D motion vector is between the maximum and minimum motion vectors obtained by each view individually.

We also compare error ratio between 3D ME using all views and total errors using individual ME of each view, i.e., $\alpha = E_r \left( E_r + E_1 + E_2 + E_3 \right)$ where $E_r$ is the error of 3D ME and $E_n$ is the error of $n$ view. If $\alpha$ is less than $1.0$, it indicates that 3D ME provides better ME and hence 3D ME will provide better coding performance. We considered those with error ratio under $1.05$ to be blocks in which 3D ME is viable in terms of errors (had no more than $5\%$ additional errors than doing it separately). We have considered extra $5\%$ because we assume that if we can use one 3D motion
vector instead of \( n \) motion vectors (one for each view), then we can save at least 5\% of total bits. The saved bits can then be used to encode extra 5\% residual error so that eventually we do not sacrifice any compression. Fig 4(middle) shows average error ratio comparison against each frame for Exit sequence with DA (i.e., ExitErrorRatio) and without DA (i.e., ExitErrorRatioND) 3D ME. The figure indicates that 3D ME performs better for DA 3D formation compared to just plain staking 3D formation. The error ratio comparison curves for other video sequences such as Vassar and Ballroom are shown in Fig 4(right). We can infer that RD performance for the Vassar would be better compared to both Exit and Ballroom sequences.

B. 3D Motion Estimation

In the proposed Cuboid technique, we form a 3D frame comprising \( n \) frames of all views using scene overlapped regions only and ME can be carried out for a 3D CU where the reference 3D frame would be formed using the immediate previous i.e., \((i-1)\)th frames of all views using scene overlapped regions. An extra 3D reference frame is also used using \((i-2)\)th frames of all views to enable dual frame referencing. The 3D ME process is similar to the existing 2D ME using sum of absolute difference criteria. The motion search length is \( \pm 31 \) with quarter-pel accuracy. As the resolution of the video sequences is low to mid-range, we use 32\times32 block as a CU size. We also explore exhaustive block-partitioning up to 8\times8 similar to HEVC. We do not use predictive motion vector in this implementation. We may further reduce the computational time of the proposed scheme using predictive motion vector. The HEVC recommended Lagrangian multiplier based cost function is used in the proposed scheme for ultimate mode selection in the variable block size ME. However, we do not do any ME using the second 3D reference frame. But we use variable size block partitioning. We have selected the second 3D reference frame if the distortion between the co-located block in the second reference frame and the current block is less than the distortion between the motion estimated match-block in the immediate previous reference frame and the current block.

![3D Motion Estimation](image)

We do not perform ME using the second reference frame for two reasons: (i) to use the second reference frame if the current block is a background block, and (ii) to reduce the computational time for ME. In the proposed 3D ME 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, and (iii) to reduce computational time. Instead of multiple ME for each reference frame (e.g., B4-frame of S3 view at T3 position in Fig 1 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 multiple reference frames in the actual coding phase. For better understanding we also provide Table 1 where we show the differences between 3D and 2D ME.

### Table 1: Comparison between 3D and conventional 2D motion estimation

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Proposed Cuboid Model</th>
<th>MV-HEVC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CU Size ( H\times W \times 3 )</td>
<td>( H\times W \times 3 )</td>
<td>( H\times W )</td>
</tr>
<tr>
<td>Motion Estimation</td>
<td>Using only one reference frame</td>
<td>Maximum 4 including bi-directional</td>
</tr>
<tr>
<td>Transform</td>
<td>3D DCT</td>
<td>2D DCT</td>
</tr>
</tbody>
</table>

We do not perform ME using the second reference frame for two reasons: (i) to use the second reference frame if the current block is a background block, and (ii) to reduce the computational time for ME. In the proposed 3D ME 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, and (iii) to reduce computational time. Instead of multiple ME for each reference frame (e.g., B4-frame of S3 view at T3 position in Fig 1 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 multiple reference frames in the actual coding phase. For better understanding we also provide Table 1 where we show the differences between 3D and 2D ME.
The proposed method has reduced RA FD problem which is another benefit of the proposed method against the existing prediction structure as all frames at T are available for encoding/decoding T+1 frames (see Fig. 1). The proposed Cuboid scheme distributes the coding delay uniformly to avoid the RA FD problem. However, in the proposed scheme we encode all frames in the same time from different views together, thus the proposed scheme does not have dependency on the future frames, which is the case for MV-HEVC if we consider Fig. 1 referencing. In this sense, the proposed scheme reduces RA FD problem.

C. 3D Coding

Research articles including [34][35] reported that huge computational gain can be achieved by transform coding (without ME) compared to the ME-compensation-transform coding while a 3D-block is formed with temporal images and 3D-DCT [34]-[37] is applied on 3D-block. This technique works well for low motion video sequences; however, a significant compression loss occurs due to the lack of correlation among temporal images in 3D-block for high motion video sequences. As the proposed technique forms the 3D-block from DA different views which have more correlations among them, 3D-DCT can concentrate the image energy in the upper-top-left areas in 3D-block more perfectly so that 3D zigzag [37] scanning has been applied. For zigzag (i.e., conversion of 3D to 1D) we have applied reshape Matlab function which rearranges the elements in column wise. Further investigation is needed to find optimal zigzag scan order. After 3D-DCT, the distributions of a majority of the significant AC coefficients can be modelled by the Gamma distribution and the distribution of the DC coefficient can be approximated by a Gaussian distribution in most cases. This knowledge can enable the design of optimal quantizers for 3D-DCT coefficients that produce minimum distortion and thus achieve close to optimal compression efficiency [35]. The proposed technique uses following quantization

\[
q(k) = \left[ \frac{QP}{4} \left( 1 + k_1 + k_2 + k_3 \right) \right]
\]

where \( q(k) \) is the quantization value at position \( k \), \( QP \) is the quantization parameter, and the value of \( \phi \) should be 0 to 1 where ‘0’ provides same quantization for all coefficients and ‘1’ provides coarse quantization for high frequency components. As full list of CAVLC/CABAC codes are not available for 3D DCT coefficient, we have generated variable length codes by dividing all coefficients into allowable number of coefficients by H.264 [7]-[11].

D. McFIS Generation

General assumption is that a pixel can be considered as a part of a background if it has almost same intensity for a number of consecutive frames. Based on this assumption, DBM [5][12]-[15][25][26] is formulated for each pixel independently. A number of Gaussian mixture models are used to capture different background and foreground for a pixel position. We assume that \( k^{th} \) Gaussian at time \( t \) represents a model with mean \( \mu_k \), standard deviation (STD) \( \sigma_k \), and weight \( \omega_k \) such that \( \sum_{k} \omega_k = 1 \) where \( K \) is the maximum Gaussian models. The learning parameter \( \alpha \) is used to update different parameters such as weight, STD, mean, etc. All parameters are initialized with predefined values such as \( \mu_k = X \), \( \sigma_k = 30 \), and \( \omega_k = 0.001 \) where \( X \) is the pixel intensity at time \( t \). For each new pixel intensity \( X' \) at time \( t \) is first matched against the existing models in order to find suitable model (e.g., \( k^{th} \) model) such that \( |X' - \mu_k| \leq 2.5 \sigma_k \). If such a model exists, then we update corresponding parameters with the learning rate as follows:

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

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

\[
\omega_k = (1-\alpha)\omega_k + \alpha.
\]

The weights of the remaining Gaussians (i.e., \( l \) where \( l \neq k \)) are updated as \( \omega_l = (1-\alpha)\omega_l + \alpha \). The weights are normalized among models for each iteration. If the recent pixel intensity \( X' \) does not match with any existing model, a new Gaussian model is introduced with \( \mu_{k+1} = X' \), \( \sigma_{k+1} = 30 \), and \( \omega_{k+1} = 0.001 \) if \( K > (k+1) \), otherwise the new model is created by evicting the \( K^{th} \) (based on \( w/\sigma \) in descending order among existing models) model. For more details in modelling and model updating, please refer [26]. To get the dynamic background pixel intensity from the above mentioned models for a particular pixel position at time \( t \), we take the \( \mu_k \) of the \( k^{th} \) model that has the highest value of \( w/\sigma \) among the models.

Fig 5: Examples of McFIS of Exit video where (a) an original frame with a number of moving objects and (b) corresponding McFIS where uncovered background areas marked by yellow dotted rectangles.

For better understanding of the capturing uncovered occluded background areas using McFIS, Fig 5 shows an example of McFIS using Exit video sequence where Fig 5(a) shows an original frame with a number of moving objects and Fig 5(b) shows its corresponding McFIS where uncovered background areas are marked with yellow dotted. In the traditional referencing approach using previously encoded frames could not provide the uncovered background areas unless these uncovered background areas are once visible in a frame and that frame is used as the
reference frame. Selection of that kind of frame for encoding frames is almost impossible due to (i) one frame might not have all uncovered areas, (ii) the reference frame might be far from the current frame so that we need to store all encoded frames, (iii) finding a particular frame among all frames requires huge computational time as well a large number of bits to signal the index of the frame. Thus, McFIS is an easier solution to capture uncovered areas and to be used as a reference frame for all frames.

E. Proposed Cuboid Technique with McFIS

Although the proposed method successfully reduces two limitations such as computational time and interactivity, it (with its current state) could not outperform MV-HEVC in terms of RD performance because we do not fully exploit inter-view redundancy which contributes around 15% references. The experimental results also reveal that some cases such as very motion active video sequences (e.g., Breakdancing), the motion vector similarity is low and error ratio is high. In results, the proposed method degrades the RD performance for those cases. It is also worthy to investigate the utilization of the computational gain of the proposed method for improving the RD performance without sacrificing computational gain and RAFD.

McFIS can be formed using DBM based on the Gaussian Mixture [12]-[15]. McFIS can successfully capture a static background including occluded background areas (if expressed once) from a scene of a video. We have formed 3D McFIS using the McFISes of all views and then used it as a second reference frame when 3D ME is carried out for the current 3D frame. The McFIS method is integrated in the proposed cuboid coding framework. The proposed scheme uses two different 3D reference frames (they are formed using immediate previous frames and McFISes from different views). The frames are stored similar to List_0 of HTM for MC.

The proposed Cuboid-McFIS technique uses McFIS instead of using the immediate second previous frame in the Cuboid technique. Obviously the Cuboid-McFIS technique requires additional computational time compared to the Cuboid technique due to the McFIS modelling, however, better RD performance is achieved due to uncovered background (using 3D McFIS) referencing.

MV-HEVC requires \( F_{\text{max}} = 3l_{\text{max}} + 2 \times \lfloor (N-1)/2 \rfloor \) number of frames to decode first to see a particular frame of a view where \( l_{\text{max}} \) is the highest hierarchical order and \( N \) is the total number of views [4]. As the proposed scheme encodes/decodes all frames in a particular time \( t \) from \( n \) views, then maximum \( n(n-N) \) number of frames to decode first to see a particular frame of a view at \( t+1 \) time if we assume that \( n \) number of views are used to form a 3D frame. Note that \( n \) is normally a subset of \( N \) number of views to be encoded altogether. For example if we form a 3D frame using \( n \) number of views, the proposed scheme requires \( n \) number of I/P frames at \( t \) time to decode first to see any frame at \( t+1 \) time (see Fig 6(a)). As when we see any frame at \( t+1 \) time, all frames at \( t \) time are already decoded, there is no interactivity (frame delay) problem by the proposed scheme assuming that all frames of all views at \( t \) time are already decoded at a time.

By removing a number of reference links among interviews and intra-view in Fig 1, we can make the same delay of the existing reference approach of MV-HEVC with the proposed scheme. This makes MV-HEVC towards simulcast with fewer levels of hierarchical B-references. By making simulcast and low-level hierarchical references, we increase bitstream compared to the existing MV-HEVC [2]. Thus, the proposed scheme outperforms the modified MV-HEVC (towards more interactivity compared to the existing one) in terms of RD performance using the same frame delay with greater margin.

Fig 6(b)&(c) show number of reference frames needed in the existing MV-HEVC scheme and the proposed scheme respectively while we encode 9 views with GOP = 8 frames. If we compare the computational time between the proposed scheme and MV-HEVC based on the number of reference frames by ignoring search pattern differences, reference frame difference, and other criteria, theoretically we can conclude that the proposed scheme reduces 39% computational time compared to 3D-HEVC.

3D-HEVC or MV-HEVC can change the flexibility in the horizontal direction i.e., in the time direction by relaxing bi-direction hierarchical intra-view and vertical direction i.e., in the view direction by relaxing inter-view dependency. The proposed Cuboid coding framework has also flexibility in both directions although we emphasis in the vertical direction to avoid dependency in terms of time for reducing RAFD problem. We can make the Cuboid coding framework flexible in both directions. For example, we can introduce IBBP or IBP or HBP format for better coding gain by increasing dependency in the horizontal direction, we can change the number of frames for 3D formation to make it flexible in the vertical direction. Even we can add ME using 3D frame in the vertical direction. Thus, the proposed scheme can be flexible in different applications.

Fig 7 reveals that the proposed Cuboid and Cuboid-McFIS reduce computational time by 59% and 56%
The most of the video coding techniques exploit both spatial and temporal redundancy within a video to compress the video data. The existing video coding standards H.264, HEVC, and MPEG used a number of innovative techniques including variable block size ME, variable block size transformations, entropy coding to reduce spatial and temporal redundancy based on the motion patterns. However, all contents/regions of a video might not be equally important or equally perceived by an observer. Thus, more compression and/or better perceived quality can be achieved by encoding important regions with superior quality and non-important regions with inferior quality. Perceptual video coding using VAM [39][40] allocates bits for a given frame non-uniformly [41]-[43] based on the importance of the regions. The basic idea is that the perceptual video coding techniques improve the perceived video quality for a given bit budget by allocating more bits for encoding more visually attentive (i.e., important) areas and fewer bits for less attentive areas. The existing techniques [41][42] mainly create salience map and allocate bits by applying VAM on entire image. We incorporate variable bit allocation in the proposed method (Cuboid-McFIS) based on the salience map of the moving region instead of the entire image as we believe that moving regions are more attentive areas compared to static areas although static area may appeal more attention due to intensity variations, colour contrast, and neighbourhood information.

The visual acuity decreases with increased distance, or eccentricity from the fixation point [45]. Therefore, a foveation model considered for video compression will lead to a more intelligent representation of visual scenes. We can easily observe that when a video is captured, the normal intention is that the visual attention areas are placed at the middle point of the capturing devices. As a result, a fixation point can be found at the middle of an image for the most of the cases. Thus, in the proposed method, we also apply central tendency based on the eccentricity to further improve the moving region-based salience map. The eccentricity is calculated based on the approach proposed in [44]. The final salience map is calculated by fusing the salience map based on the moving regions and the salience map based on the eccentricity (considering the middle point of an image as a fixation point). For final salience map, we take certain percentages of the salience map from the moving regions and the eccentricity. Note that we calculate moving regions of an image by taking absolute difference between the image and the corresponding McFIS. McFIS is generated using Gaussian-based DBM [25].
visual attentive areas using entire frame [39], visual attentive areas using the moving regions, and QP adjustment (\(\delta QP\)) for the salience map using the moving regions. If we look at the salience map using the entire image, the bigger areas including the wall (left of the image) and the roof (brightest areas at the top of the people) are identified as the salience areas. On the other hand, if we see the saliency map using the moving regions, the salience map is relatively smaller and covers only the moving regions. Normally, moving regions of a video attract more attention of an observer rather than brightest static areas. Thus, better quality in the moving regions should provide better perceived quality for a video. We have determined a \(\delta QP\) matrix for a frame in block level based on the moving region and central tendency salience maps. The last sub-figure shows an example of \(\delta QP\) where minimum possible value is -2 for the most saliency areas and maximum possible value is 2 for the least saliency areas. The \(\delta QP\) is 2D matrix and the size of the \(\delta QP\) is the same as the number of coding unit blocks in a frame. We calculate \(\delta QP\) for each frame based on the salience map. In the proposed multiview cuboid perceptual video coding with McFIS (Cuboid-PVC-McFIS) scheme has replaced the original QP by adding corresponding block-position \(\delta QP\) with the original QP. Thus, different bits will be allocated based on the salience map of the frame. As the visually attentive areas are encoded with smaller QPs (i.e. provides superior quality) and visually less attentive areas are encoded with larger QPs (i.e. provides inferior quality), ultimately better perceived quality is obtained by allocating variable bit allocation.

Fig 9: Visual attentive areas and corresponding \(\delta QP\) mapping using visual attention modelling focusing on moving regions; (left to right and top to bottom) the original 22nd frame of Exit video, McFIS, difference between McFIS and the frame i.e., moving regions, saliency map using the frame, salience map using the moving region, and QP mapping using the moving region saliency map.

We calculate fused salience map, \(M(i, j)\) by combining the salience map \(G(i, j)\) from moving regions and the salience map \(C(i, j)\) from eccentricity where \(i\) and \(j\) indicate pixel positions. Note that all salience maps are normalized to 0 to 1.0 range. The fusing parameter \(\omega\) (0 to 1.0) balances the contribution of each salience map into the fused salience map. The fused salience map \(M(i, j)\) is defined as:

\[
M(i, j) = \omega G(i, j) + (1-\omega)C(i, j).
\]

Then, an intermediate map \(\Psi(i, j)\) is formed based on the value of \(M(i, j)\) in each pixel position where two thresholds \(T_1\) and \(T_2\) (where \(T_1 > T_2\)) are used:

\[
\Psi(i, j) = \begin{cases} 
-2 & \text{if } M(i, j) > T_1 \\
0 & \text{if } T_2 < M(i, j) \leq T_1 \\
2 & \text{otherwise}
\end{cases}
\]

Finally, the adjusted QP (i.e. \(\delta QP\)) is calculated as follows:

\[
\delta QP(r, c) = \frac{1}{\beta^2} \sum_{i=(r-1)\beta+1}^{r\beta} \sum_{j=(c-1)\beta+1}^{c\beta} \Psi(i, j)
\]

where \(\beta\) is the CU size (for example if the CU size is 16x16 pixels then \(\beta = 16\)), \(r\) is the row, and \(c\) is the column number of a CU location. An example of \(\delta QP\) is given in the last sub-figure in Fig 9. First, we have assigned QP adjustment in pixel level on fused saliency value using two thresholds T1 and T2 (see (2)), then we determine the integer average for a CU to find the small variation of QP (i.e. \(\delta QP\)) in block level. As perceived quality of a video with huge variations of qualities within a frame is inferior, we keep the variations of \(\delta QP \approx \pm 2\) from the original QP within a frame [42]. Based on the saliency map a \(\delta QP\) value is determined from the set \{-2, 0, 2\} for each CU of the current frame and a CU is quantized for encoding using an adjusted QP where the adjusted QP is calculated as \(QP_{\text{adjusted}} = QP_{\text{original}} + \delta QP\). It may be important to make the adaptive QP adjustment under the bit-budget constraints. However, If the adaptive procedure fails, then it not only increases the requirements of bits with inferior quality of the current frame but also costs to the future frames as the current frame can be used as reference frame. Moreover, any adaptive modelling requires extra computational time. To avoid these, after a number of trials with all video sequences, we have fixed the QP values.

In the proposed scheme we do not need to transmit saliency map to the decoder, we only use it for QP adjustment for encoding. As we compare the performance of the proposed cuboid coding structure with perceptual coding against MV-HEVC and the cuboid coding with McFIS, the rate-distortion performance comparisons among three schemes should provide clear understanding about the strength of the proposed perceptual cuboid coding scheme. Moreover, the change of QP is very small i.e., \(\pm 2\) compared to the original QP. This kind of small QP adjustment is also available in MV-HEVC, HEVC, H.264/AVC, and H.264/MVC encoders, e.g. for different level B-frames in hierarchical bi-prediction structure, texture and depth coding, different type frames (I, P, and B-frames), and block level for controlling rate-distortion performance under limited bit budgets, etc.

IV. EXPERIMENTAL RESULTS
To compare the performance of the proposed schemes (Cuboid, Cuboid-McFIS, and Cuboid-PVC-McFIS), we have implemented all the algorithms based on the MV-HEVC recommendations and HTM 15.0 [54] implementation with 25 Hz, ±31 as the search length with quarter-pel accuracy, with 16 as the GOP size. In the proposed schemes, we have considered the IPP...I prediction format whereas we have used the hierarchical B-picture predication structure for MV-HEVC. Fig 10 shows RD performance using MV-HEVC and three proposed schemes for four standard multiview video sequences. The figure reveals that the RD performance of the proposed Cuboid is comparable with that of MV-HEVC; however, the proposed Cuboid-McFIS outperforms the MV-HEVC comprehensively by the margin 0.25dB to 2.0dB. The proposed Cuboid-PVC-McFIS also outperforms the MV-HEVC with the same margin. Note that in the experiment we have formed 3D frame using three views. We also use $\omega = 0.7$, $\beta = 16$, $T_1 = 0.4$, $T_2 = 0.1$, and $\phi = 0.3$ for the experiments and analysis. Sometimes, better rate-distortion performance can be obtained using adaptive thresholds. Some cases, the adaptive model may fail, then it not only increases the required bits with inferior quality of the current frame but also increases bits for the future frames as the current frame can be a reference frame. Moreover, an adaptive modelling normally requires extra computational time. To avoid these, after a number of trials with all video sequences, we have fixed the values of those parameters.

Experimental setup should follow common test condition [55] if a proposed approach is a modification of the current coding framework. Although the proposed coding scheme is fundamentally different from the MV-HEVC recommendation, in the experimental setup we follow common test condition in the most cases. However, we use GOP size 16 instead of 8 (recommended in [55]) and different video sequences (captured by semi-circular co-planner camera setup rather than parallel co-planner camera setup) instead of test sequences suggested in [55]. As we use the same GOP and video sequences for the proposed and the MV-HEVC schemes, the performance comparison should provide clear understanding of the strength of the proposed scheme against MV-HEVC. 3D video coding for videos captured by parallel co-planner camera setup might be easier to exploit inter-view redundancy compared to semi-circular as there is no orientation difference between two views. In the proposed scheme, we particularly test the performance of the proposed cuboid coding framework against MV-HEVC for those videos as they are more challenging.

The Structural Similarity (SSIM) index is a method for measuring the similarity between two images [46]. As SSIM exploits human visual systems to measure the quality of a reconstructed image against the original reference image, we have used SSIM to compare the performance of the proposed schemes. Weighted peak signal-to-noise ratio (WPSNR) based on contrast sensitivity function is also used to measure the quality of reconstructed image against the original image by exploiting human visual system [47]. Fig 10 shows that the RD performances of the proposed Cuboid-McFIS and Cuboid-PVC-McFIS are almost same.

![Fig 10: Rate-distortion performance by MV-HEVC and the proposed schemes (Cuboid, Cuboid-McFIS, and Cuboid-PVC-McFIS) using standard video sequences namely Exit, Ballroom, Vassar, and Breakdancing.](image1)

![Fig 11: Rate-distortion performance by the proposed scheme Cuboid-McFIS scheme against Li et al.’s, and the mimic of Zamarin et al.’s (i.e., Zamarin* et al.), Abreu et al.’s, and MV-HEVC schemes using four standard video sequences.](image2)

We compare the performance of the proposed scheme with Li et al.’s scheme and Abreu et al.’s scheme. We could not directly compare the proposed scheme with Zamarin et al.’s scheme in its current form as in the propose scheme we do not consider any depth information. To see the effectiveness of the proposed scheme we compare the proposed scheme with a modified version of the proposed scheme without motion estimation. The rationality of the comparison is that the modified version of the proposed...
scheme is the mimic of Zamarin et al.’s scheme (named as Zamarin* et al.) as where depth information based 3D warping and occlusion handle are replaced by disparity-adjusted view alignment and background modelling respectively. Abreu et al. uses less number of frame-reference dependency to reduce the RAFD problem. The experimental results show that the proposed scheme outperforms Li et al.’s, Abreu et al.’s, Zamarin* et al.’s, and MV-HEVC schemes (see Fig 11) in terms of RD performance using four video sequences. One of the main reason is the ability of the proposed scheme to exploit uncovered background areas for encoding.

We also verify the performance of the proposed schemes using SSIM and WPSNR. Fig 12 shows the perceptual quality performance comparison between the proposed schemes (Cuboid-McFIS and Cuboid-PVC-McFIS) using middle part of the images (we do not consider outer 32 pixels). In the figure, we can easily observe that the proposed Cuboid-PVC-McFIS scheme outperforms the Cuboid-McFIS scheme in terms of SSIM and WPSNR. This demonstrates that the proposed Cuboid-PVC-McFIS provides better perceive quality compared to the proposed Cuboid-McFIS scheme (see Fig 12) although the overall RD performance is similar (see Fig 10).

Table 2: BD-PSNR and BD-Bit Rate comparison among methods including the proposed method (Cuboid-McFIS) against MV-HEVC

<table>
<thead>
<tr>
<th></th>
<th>Li et al.</th>
<th>Abreu et al.</th>
<th>Zamarin* et al.</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ballroom</td>
<td>-0.94</td>
<td>-0.36</td>
<td>0.39</td>
<td>0.80</td>
</tr>
<tr>
<td>Breakdancing</td>
<td>-0.63</td>
<td>-0.14</td>
<td>-0.48</td>
<td>0.03</td>
</tr>
<tr>
<td>Exit</td>
<td>-0.86</td>
<td>-0.33</td>
<td>0.61</td>
<td>1.37</td>
</tr>
<tr>
<td>Vassar</td>
<td>-0.28</td>
<td>-0.05</td>
<td>2.55</td>
<td>2.80</td>
</tr>
</tbody>
</table>

Table 2 shows performance comparison in terms of BD-PSNR and BD-BitRate of the proposed Cuboid-McFIS scheme against other relevant techniques. It reveals that the proposed technique outperforms all relevant techniques.

Fig 13: The results of the subjective quality test to compare the quality of the videos by Cuboid-PVC-McFIS and Cuboid-McFIS where (a) shows the scale used in the experiment and (b) shows the average comparative results.

To evaluate the perceptual video quality performance, we perform a subjective test using Double-Stimulus Continuous Quality Scale (DSCQS) assessment using the test conditions of [50][51]. In this method, we used a number of video sequences serially. In each test, the viewers are asked to rate the quality of two video sequences known as “A”, and “B” on a continuous scale ranging between “Excellent” and “Bad” (see Fig 13(a)). Either A or B (chosen at arbitrarily) was a reconstructed video sequence using the proposed Cuboid-PVC-McFIS algorithm while the other was a reconstructed sequence using the proposed Cuboid-McFIS algorithm. Fig 13 (b) shows the numerical average results among 15 viewers. The figure shows that 60% cases viewers rate the video reconstructed by the Cuboid-PVC-McFIS scheme better compared to the Cuboid-McFIS scheme.

To compare the perceptual performance of both algorithms, we reconstruct the images using both techniques. The bits per pixel (bpp) and PSNR of these four sequences using both algorithms are equal 0.55bpp and 41.76 dB, 0.43 bpp and 42.44 dB, 0.22 bpp and 42.97 dB, and 0.19 bpp and 42.25 dB for Ballroom, Breakdancing, Exit, and Vassar respectively. In all examples, reconstructed frames using the proposed scheme Cuboid-PVC-McFIS can be readily perceived better compared to the reconstructed frames using the proposed scheme Cuboid-McFIS. We also test visual attention area of the video sequences using Tobii X120 eye tracker. The experimental results on fixations and
saccades indicate that the most of the cases the points of gaze are in the motion area. These results demonstrate that motion areas of a video sequence are more visually attractive and eye catching areas compared to the no-motion areas. Thus, the proposed Cuboid-PVC-McFIS scheme outperforms the proposed Cuboid-McFIS scheme by exploiting variable bit allocation strategy although both provide almost same rate-distortion performance.

![PSNR Distribution](image)

**Fig 14**: Frame level PSNR distributions for the first 145 frames of the second view of Exit video sequence using the proposed Cuboid-PVC-McFIS scheme and the MV-HEVC to demonstrate PSNR fluctuations.

Fig 14 shows PSNR distribution against the first 145 frames using the proposed Cuboid-PVC-McFIS scheme and the MV-HEVC respectively for the second view of the exit video sequence at almost same bit rates. We have selected the second as according to the hierarchical structure in the MV-HEVC, the images of the second view take the maximum reference frames for encoding. The figure demonstrates that the proposed scheme provides more stable PSNRs for frames within a GOP compared to the MV-HEVC. In the hierarchical B-picture prediction structure, MV-HEVC uses different QPs within a GOP, thus, the PSNR fluctuation within a GOP is higher, which creates inferior perceived video quality compared to the same PSNR output with less PSNR fluctuation. In the proposed scheme we use the same QPs (with ±2 variations in CU level within a frame based on salience map) 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 0.90dB at almost 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 a few number of frames. Note that the PSNR increases with the frames for the proposed scheme compared to MV-HEVC due to better McFIS. We need to wait a number of frames (e.g., 25 frames or more) to get better uncovered background for the McFIS.

**Fig 15**: Reconstructed frames by different schemes to demonstrate the quality difference marked by dotted rectangles.

![Original 14th frame](image)

![Frame by Cuboid-McFIS](image)

![Frame by Cuboid-PVC-McFIS](image)

For the video sequences, where the baseline between two cameras is bigger, the overlapping areas should be smaller. As the proposed technique applies cuboid concept on the overlapping areas, the performance of the proposed method will be lower. However, this also lowers the performance of the MV-HEVC as it does not get handsome amount of redundant information due to less overlapping areas. Thus, the relative performance difference remains almost the same.
between the proposed methods and the MV-HEVC.

V. CONCLUSIONS

In this paper, we propose a new 3D motion estimation and motion compensation scheme using disparity-adjusted 3D video formation to reduce the computational time and interactivity problem of the existing MV-HEVC. In the proposed technique, a 3D frame is formed using the same temporal frames of all or a subset of disparity-adjusted views and motion estimation is carried out for a coding unit of the current 3D frame using the immediate previous two 3D frames as reference frames. This technique outperforms the existing standard by reducing computational time by more than 59% and interactivity problem compared to MV-HEVC with comparable RD performance. This paper also proposes another technique (Cuboid-McFIS) where the second 3D reference frame is used in addition to the immediate previous 3D frame. The second 3D reference frame is formed using dynamic background frames of each view which are popularly known as McFis (the most common frame in a scene) based on Gaussian mixture modeling. The experimental results reveal that the proposed Cuboid-McFIS method outperforms the MV-HEVC in terms of improving image quality by 0.2dB ~2.0dB and reducing computational time by 56%, and reducing random access frame delay time. Finally, the third method Cuboid-PVC-McFIS, an extension of the proposed Cuboid-McFIS method, is proposed using variable bit allocation based on the visual attentive areas for better perceivable video quality. The experimental results show that the proposed Cuboid-PVC-McFIS scheme outperforms the proposed Cuboid-McFIS scheme in terms of visual quality assessment (measured in SSIM and WPSNR) by maintaining the same RD performance.

VI. REFERENCES


[50] Recommendation ITU-R BT 500-10, Methodology for the Subjective Assessment of the Quality of Television Pictures.


[54] HTM 15.0 Software for HEVC-3D https://hevc.hhi.fraunhofer.de/ svn/svn_3DVCSoftware/tags/HTM-15.0/


Manoranjan Paul (M’03 SM’13) received PhD from Monash University in 2005. He has worked as a Post-Doctoral Research Fellow in the University of New South Wales, Monash University, and Nanyang Technological University. He is currently working as an Associate Professor in the School of Computing and Mathematics, Charles Sturt University where he is a Director of E-Health. His major research interests are in the fields of image/video coding, medical imaging, EEG signal analysis, and computer vision. He has published more than 140 refereed papers in international journals and conferences. A/Prof Paul regularly published journal articles in the IEEE Transactions. He was a keynote speaker in the IEEE DICTA-17, WoWMoM-14 Workshop, DICTA-13, and ICCIT-10. He is an ICT Researcher of the Year (Finalist), 2017 selected by Australian computer Society (ACS). A/Prof Paul is a Senior Member of the IEEE and ACS. He has served as a guest editor of Journal of Multimedia and Journal of Computers for five special issues, Program Chair of PSIVT 2017 and Publicity Chair of DICTA 2016. He is an Associate Editor of EURASIP Journal on Advances in Signal Processing since 2013. More information is available in http://csusap.csu.edu.au/~rpaul/.