DISPARITY-ADJUSTED 3D MULTI-VIEW VIDEO CODING WITH DYNAMIC BACKGROUND MODELLING

Manoranjan Paul and Christopher J. Evans
School of Computing and Mathematics, Charles Sturt University, Australia
Email: m paul@csu.edu.au, nosivire@gmail.com

Manzur Murshed
Gippsland School of Information Technology, Monash University, Australia
Email: manzur.murshed@monash.edu

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 are not sufficiently agile to exploit the interactivity and inefficient in terms of image quality and computational time. In this paper a novel technique is proposed using disparity-adjusted 3D MVC (DA-3D-MVC) with 3D motion estimation (ME) and 3D coding to overcome the problems. In the proposed scheme, a 3D frame is formed using the same temporal frames of all disparity-adjusted views and ME is carried out for the current 3D macroblock using the immediate previous 3D frame as a reference frame. Then, 3D coding technique is used for better compression. As all the same temporal position frames of all views are encoded at the same time, the proposed scheme provides better interactivity and reduced computational time compared to the H.264/MVC. To improve the rate-distortion (RD) performance of the proposed technique, an additional reference frame comprising dynamic background is also used. Experimental results reveal that the proposed scheme outperforms the H.264/MVC in terms of RD performance, computational time, and interactivity.

Index Terms—McFIS, 3D motion estimation, 3D video coding, 3D Zigzag, and video interactivity.

1. 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. Multiview 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, which often leads to incorrect object recognition and tracking. 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, joint encoding/decoding of views can achieve significant compression by exploiting inter-view correlation, in addition to the traditional intra-view correlation. Recently, H.264/MVC [1]–[3] proposes a frame reference structure among views (S) and temporal (T) images. In the reference structure of MVC, hierarchical B-picture prediction format [4] is used for intra- and inter-view. The technique exploits the redundancy from the neighbouring already encoded frames as reference from both inter- and intra-views to encode the current frame. The inter- and intra-view referencing technique provides 20% more bitstream reduction compared to the simulcast technique where no inter-view redundancy is exploited i.e., each view is encoded separately [1].

1) 18 frames must be decoded. Due to the RAFD problem, some applications e.g., interactive real-time communication may not be possible using the existing prediction structure.

H.264/AVC video coding standard improves the coding performance by reducing up to 50% bitstreams size compared to its predecessor H.263 by increasing computational complexity up to 10 times [5][6] for a single view video. In addition, when the H.264/MVC encodes multi-view videos, it requires multiple amounts of computational time compared to the H.264/AVC. The enormous requirement of computational time limits the scope of 3D video coding applications especially for electronics devices with limited processing and battery power. Thus, reduction of
computational time with better RD performance and greater interactivity is highly desirable for almost all applications.

In this paper a novel technique is proposed using disparity-adjusted 3D MVC (DA-3DMVC) with 3D ME and 3D coding to overcome the interactivity and computational time problems of the existing MVC technique. In the DA-3D-MVC 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. In the proposed technique, a 3D frame is formed using the disparity-adjusted (DA) same temporal frames (i.e., $i^{th}$ frames) of all views and 3D ME is carried out for the current 3D macroblock (MB) using the immediate previous 3D frame as a reference frame (which is formed by the $(i-1)^{th}$ frames of all views). Then, 3D zigzag scan and 3D coding have been used for better compression. As the correlation among the intra-view images is higher than the correlation among the inter-view images, the proposed DA-3D-MVC technique does not degrade the RD performance significantly, however, reduces the overall computational time and eliminates the RAFD problem compared to H.264/MVC which enables interactive real time communications.

Recently dynamic background frame is used in video coding techniques [4][7]. Paul et al. [4][7] used the most common frame in a scene (McFIS) using Gaussian mixture [8]-based dynamic background modeling for video coding. The McFIS is used as an additional reference frame 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 and uncovered background part would be referenced using the McFIS. The ultimate reference is selected at block and sub-block levels using the Lagrangian multiplier [6]. In this paper, we also propose another technique (DA-3DMVC-McFIS) where an additional reference 3D matrix comprising McFISes of all views is used for 3D ME. Experimental results reveal that the proposed DA-3DMVC-McFIS technique outperforms the H.264/MVC by improving RD performance and reducing computational time without RAFD problem.

2. PROPOSED DISPARITY-ADJUSTED MVC TECHNIQUE

Capturing a scene using multiple cameras from different angles is expected to provide the necessary interactivity in the three-dimensional space to satisfy end-users' demands for observing objects and actions from different angles and depths. Existing MVC technologies do not sufficiently support interactivity, computational time and RD performance. 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 the additional dimension is formed using frames from different views. After formation of 3D frames, we estimate motion for the current 3D MB using the variable size blocks (e.g., $16\times16\times8$, $16\times8\times8$, $8\times16\times8$, $8\times8\times8$, etc. where last dimension is for the number of views) using the previous 3D reference frame. Then, we encode 3D MB using 3D coding. Details are described in the sub-sections.

2.1 3D Frame Formation

According to our investigation, 85% references in the H.264/MVC scheme are coming from inter-view. Thus, using 3D formation and 3D ME we may sacrifice up to 15% 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 later case, we need to find disparity between two adjacent views and then we can form 3D frame using the scene overlapped regions. Paul et al. [11] first saw the benefit of 3D ME using the first approach of 3D formation. 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 Ball room 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.

![Fig 2: Example of a disparity-adjusted 3D-block formation from 3 views](image)

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 DA 3D frame. The formation processed is shown in Fig 3. 3D ME is carried out using the scene overlapped regions and the rest parts are coded separately. Paul et al. [11] also investigated motion vector relationship 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 MB at the $i^{th}$ frame of the $j^{th}$ view has 51% to 93% of similarity with the co-located macroblock at the $i^{th}$ frame of other views. Using the proposed DA 3D formation approach, the average motion vector relationship is far better (i.e., 70% to 95%).

![Fig 3: Formation of 3D frame using scene overlapped regions by considering disparity among the views](image)

We also compare the motion vectors by observing if $\min(dx_x,dx_y,dx_z,dx_z) \leq dx_y \leq \max(dx_x,dx_y,dx_z,dx_z)$ and, if true, 3D ME is considered acceptable for that block. In the condition, $dx_y$ is the $x$ motion vector estimated by performing ME on the 3D MB and $dx_z$ is the $x$ motion vector estimated using just the $n^{th}$ view. This
evaluation method we considered, with the thoughts behind it being that if the motion vector found using 3D MB 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 DA-3DME creates reasonable motion vector. Fig 4(a) 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., \( \alpha = \frac{1}{n} \sum \frac{E_i + E_j + E_k + E_l}{E_i} \) where \( E_i \) 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(b) 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(c). We can infer that RD performance for the Vassar would be better compared to both Exit and Ballroom sequences.

2.2 3D Motion Estimation

In the proposed DA-3DMVC technique, we form a 3D frame comprising \( p \) frames of all views using scene overlapped regions only and ME can be carried out for a 3D MB where the reference 3D frame would be formed using the immediate previous i.e., \((t-1)^{th}\) frames of all views using scene overlapped regions. 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]-[3], (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. The proposed method has not RAFD problem which is another benefit of the proposed method against the existing prediction structure as all frames at \( T_i \) are available for encoding/decoding \( T_{i+1} \) frames (see Fig 1).

2.3 3D Coding

Research articles including [12],[13] 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 [14] 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 [15] 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 [13]. The proposed technique uses following quantization \( q(k) = \frac{Q}{4}(1 + k_p + k_p + k_q + k_q) \) where \( q(k) \) is the quantization value at position \( k \), \( Q \) is the quantization parameter, and the value of \( p \) 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 CAVC/LCABAC codes are not available for 3D DCT coefficient, we have generated VLC codes by dividing all coefficients into allowable number of coefficients by H.264.
3. PROPOSED DA-3DMVC TECHNIQUE WITH MCFIS

Although the proposed method successfully overcomes two limitations such as computational time and interactivity, it (with its current state) could not outperform H.264/MVC 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., Break Dancing), 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 random access delay.

McFIS can be formed using dynamic background modelling [8] based on the Gaussian Mixture. 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 proposed DA-3DMVC-McFIS technique uses additional reference frames compared to the proposed DA-3DVC technique. Obviously the DA-3DVC-McFIS technique requires additional computational time compared to the DA-3DVC technique due to the McFIS modelling and extra motion estimation using 3D McFIS, however, better RD performance is achieved due to background (using 3D McFIS) referencing.

Fig 5(e) reveals that the proposed DA-3DMVC and DA-3DMVC-McFIS reduce computational time by 51% and 37% respectively compared to H.264/MVC using 31 search lengths. Due to the fixed amount of operations requirement for the McFIS modelling, the proposed DA-3DMVC-McFIS method reduces computational time less compared to that of the proposed DA-3DMVC scheme. When large search length is used, the computational time for the McFIS modelling is negligible compared to ME. Fig 5(f) shows McFIS which is generated from Exit video by the proposed scheme where only background area is captured in the dotted areas by removing moving objects.

4. EXPERIMENTAL RESULTS

To compare the performance of the proposed schemes (DA-3DVC and DA-3DMVC-McFIS), we have implemented all the algorithms based on the H.264/MVC recommendations 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 IBBP prediction format whereas we have used the hierarchical B-picture prediction structure for H264/MVC. Fig 5(a)-(d) shows RD performance using H.264/MVC and two proposed schemes using 4 standard multiview video sequences. We also compare the RD performance using without DA 3D ME. The figure reveals that the RD performance of the proposed DA-3DMVC-McFIS outperforms all relevant schemes. The proposed scheme performs better for Exit and Vassar video sequences as they have more static and uncovered background areas.

5. CONCLUSIONS

In this paper, we proposed a new 3D ME and compensation scheme using DA 3D video formation to reduce the computational time and eliminate interactivity problem of the existing H.264/MVC. In the proposed technique, a 3D frame is formed using the same temporal frames of all DA views and ME is carried out for an MB of the current 3D frame using the immediate previous 3D frame as reference frame. This technique outperforms the existing standard by reducing computational time by more than 50% and eliminating interactivity compared to H.264/MVC. This paper also proposes another technique (DA-3DMVC-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 McFIS based on Gaussian mixture modelling. The experimental results reveal that the DA-3DMVC-McFIS outperforms the H.264/MVC by improving 0.35dB PSNR, by reducing computational time by 37%, and by eliminating interactivity problem compared to the H.264/MVC.
6. REFERENCES


