# 3D MOTION ESTIMATION FOR 3D VIDEO CODING

*Manoranjan Paul, Junbin Gao, and Michael Anotolovich*
School of Computing and Mathematics, Charles Sturt University, Bathurst, Australia
Email: {mpaul, jbgao, mantolovich}@csu.edu.au

## ABSTRACT

H.264/MVC multi-view video coding provides a better compression rate compared to the simulcast coding using hierarchical B-picture prediction structure exploiting inter- and intra-view redundancy. However, this technique imposes random access frame delay as well as requiring huge computational time. In this paper a novel technique is proposed using 3D motion estimation (3D-ME) to overcome the problems. In the 3D-ME technique, a 3D frame is formed using the same temporal frames of all views and ME is carried out for the current 3D frame using the immediate previous 3D frame as a reference frame. As the correlation among the intra-view images is higher compared to the correlation among the inter-view images, the proposed 3D-ME technique reduces the overall computational time and eliminates the frame delay with comparable rate-distortion (RD) performance compared to H.264/MVC. Another technique is also proposed in the paper where an extra reference 3D frame comprising dynamic background frames (the most common frame of a scene i.e., McFIS) of each view is used for 3D-ME. Experimental results reveal that the proposed 3D-ME-McFIS technique outperforms the H.264/MVC in terms of improved RD performance by reducing computational time and by eliminating the random access frame delay.

***Index Terms***— McFIS, 3D Motion Estimation, 3D Video Coding, uncovered background, hierarchical B-picture, and MRFs.

## 1. INTRODUCTION

A scene captured by different video cameras from different angles (i.e., multi-view videos or 3D videos) provides more realistic experience about the scene compared to a single view captured from a single video camera. Obviously, transmission and storing of multi-view videos requires huge amounts of computations and data manipulations compared to the counter part of single view video, although there is a significant amount of data redundancy among views. Recently, H.264/MVC [1]-[3] proposes a reference structure among views (*S*) and temporal (*T*) images. In the reference structure of multi-view video coding (MVC) or 3D video coding, hierarchical B-picture prediction format [4] is used for intra- and inter-view. The technique exploits the redundancy from the neighbouring frames as reference from both inter- and intra-view 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].

Fig 1 shows the prediction structure of the MVC or 3D video coding recommended by the H.264/MVC standard where eight views are used. According to the prediction structure, a frame may use 4 frames as reference frames and encoding/decoding a frame sometimes requires encoding/decoding a number of frames in advanced, thus, the structure introduces random access delay. The random access delay 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:

$$F_{\max} = 3 \times level_{\max} + 2 \times \lfloor (N-1)/2 \rfloor \quad (1)$$

where $level_{max}$ is the highest hierarchical order and $N$ is the total number of views [3]. For instance, in order to access a B-frame in the 4th hierarchical order (B4-frames in Fig 1), 18 frames ($F_{\max}$ = 18) must be decoded. Due to the random access problem, some applications such as interactive real-time communications may not be possible using the existing prediction structure.

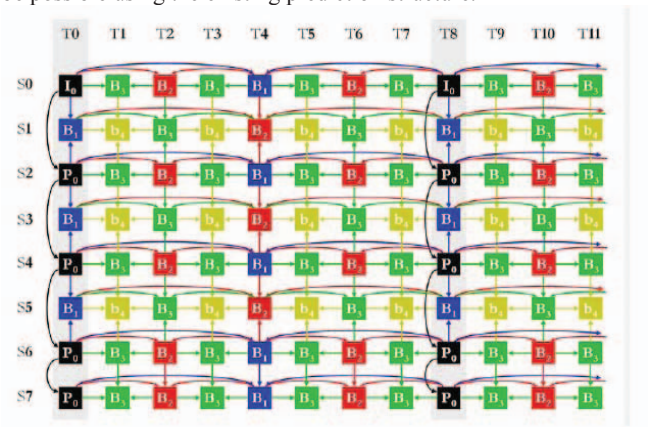


Fig 1: Prediction structure recommended by the H.264/MVC for referencing different views (S) and different temporal (T) images for coding.

H.264/AVC video coding standard improves the coding performance by reducing up to 50% bitstreams 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.

In this paper a novel technique is proposed using 3D motion estimation (3D-ME) to overcome the random access frame delay and computational time problems of the existing 3D video coding technique. In the 3D-ME technique, a 3D frame is formed using the same temporal frames (i.e, $i^{th}$ frames) of all views and motion estimation is carried out for a macroblock of the current 3D frame using the immediate previous 3D frame as reference frame (which is formed by the $(i\text{-}1)^{th}$ frames of all views). As the correlation among the intra-view images is higher than the correlation among the inter-view images, the proposed 3D-ME technique does not degrade the rate-distortion performance significantly, but reduces

the overall computational time and eliminates the random access frame delay compared to H.264/MVC which enables interactive real time communications.

Recently a dynamic background frame termed the McFIS (*the most common frame of a scene*) [4][7] has been developed for video coding using dynamic background modeling based on the *Gaussian* mixture [8][9]. The McFIS is used as a second 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 part would be referenced using McFIS. The ultimate reference is selected at block and sub-block levels using the Lagrangian multiplier. Another technique is also proposed in this paper where an extra reference 3D matrix comprising McFISes of all views is used for 3D-ME. Experimental results reveal that the proposed 3D-ME-McFIS technique outperforms the H.264/MVC by improving rate-distortion performance and reducing computational time without frame delay.

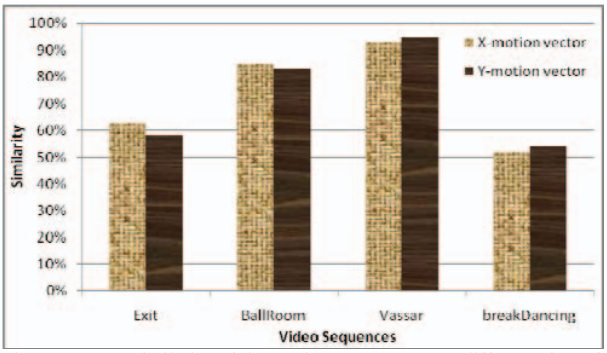

Fig 2: Average similarity of the motion vectors among different views for four standard multi-view video sequences where first 10 frames are used for each view of each sequence.

## 2. PROPOSED 3D MOTION ESTIMATION (3D ME) TECHNIQUE

A scene is captured by a number of cameras which are placed in different positions and angles in the multi-view system. As the scene is the same for all cameras, there are inter- and intra-view redundancies. In general, we can assume that object movement of a view is very similar to that of other views. To find the motion similarity, we have investigated the motion vector relationship among the views of the multi-view video sequences using four standard video sequences such as *Exit*, *Ball Room*, *Vassar*, and *Break Dancing*. First we determine the motion vectors of all macroblocks of each frame of a view using a 16×16 mode. Then find the similarity of the motion vectors of a view with that of other views. Fig 2 shows the 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 macroblock 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. We can exploit this relationship to avoid random access delay and computational time problems of the existing prediction structure.

In the proposed 3D-ME technique, we can make a 3D frame comprising $i^{th}$ frames of all views and ME can be carried out for a 3D macroblock (another dimension is formed using co-located macroblocks from different views) where the reference 3D frame would be formed using the immediate previous i.e., $(i\text{-}1)^{th}$ frames of all views. 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 random access frame delay, and (iii) to reduce computational time.

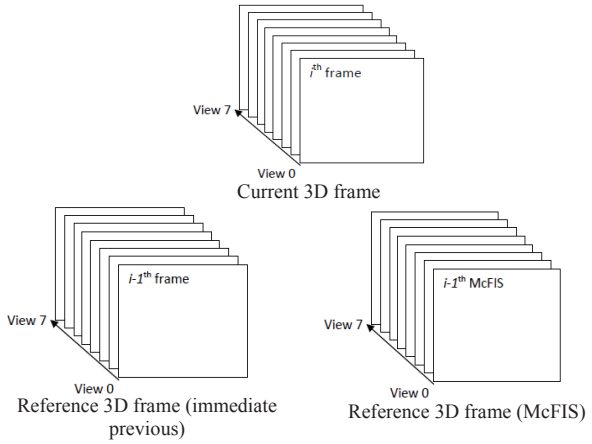

Fig 3: Different 3D frame for motion estimation and compensation for the proposed (3D-ME-McFIS) method where ME&MC of a macroblock in a current 3D frame (comprises $i^{th}$ frames of all views) is carried out using both reference 3D frame (comprises $(i\text{-}1)^{th}$ frames of all views) and 3D McFIS (comprises all McFISes of views up to $(i\text{-}1)^{th}$ frames).

Fig 3 shows the formation of the 3D frame using $i^{th}$ frames and $(i\text{-}1)^{th}$ frames of all views where the first 3D frame is the current 3D frame and the later is the reference 3D frame (where the third dimension is formed using $i^{th}$ and $(i\text{-}1)^{th}$ frames from different views respectively). We will discuss the other 3D reference frame in Section 3. The proposed method (3D-ME) does not require any disparity estimation [10] for inter-views as we do not explicitly use any inter-view relationships. Instead of multiple motion estimation for each reference frame (e.g., B4-frame of S3 view at T3 position in Fig 1 requires 4 times motion estimation using 4 reference frames), the proposed method requires only one motion estimation. A significant amount of computational time reduction can be achieved using the proposed method as the proposed method does not need disparity estimation and motion estimation for multiple reference frames. 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 $T_i$ are available for encoding/decoding $T_{i+1}$ frames (see Fig 1).

## 3. PROPOSED 3D MOTION ESTIMATION TECHNIQUE WITH 3D MACFIS (3D-ME-MCFIS)

Although the proposed method successfully overcome two limitations such as computational time and random access frame delay, it (with its current state) could not outperform H.264/MVC in terms of rate-distortion performance as the experimental results (in Fig 2) reveal that the motion vector similarity is not 100% accurate. The experimental results also reveal that some cases such as very motion active video sequences (*exit* and *Break Dancing*), the motion vector similarity is around 50%. In results, the proposed method degrades the rate-distortion performance for those cases. It is also worthy to investigate the utilization of the computational gain of the proposed method for improving the rate-distortion

performance without sacrificing computational gain and random access delay.

McFIS- (*the most common frame in a scene*) can be formed using dynamic background modelling (DBM) [8][9] based on the *Gaussian Mixture Model*. McFIS can successfully capture a static background including occluded background areas (if expressed once) from a scene of a video sequence. We have formed 3D McFIS using the McFISes of all views and then used it as a second reference frame when 3D motion estimation was carried out for the current 3D frame. The proposed 3D-ME-McFIS technique uses additional reference frames compared to the proposed 3D-ME technique. Obviously the 3D-ME-McFIS technique requires additional computational time (see Section 3.2) compared to the 3D-ME technique due to the McFIS modelling and extra motion estimation using 3D McFIS, however, better rate-distortion performance is achieved due to foreground (using the immediate 3D previous frame) and background (using 3D McFIS) referencing.

In the 3D-ME-McFIS technique, after encoding a 3D frame, we have updated 3D McFIS by updating individual McFIS (for each view) using the latest decoded frame of the corresponding views in the encoder. For example ($i$-1)$^{th}$ 3D McFIS is used while $i^{th}$ 3D current frame is encoded and the $i^{th}$ 3D McFIS is updated using the $i$th encoded frames. The benefit of the updated 3D McFIS is to keep the McFIS relevant in terms of referencing. Same procedure is also applied in the decoder to generate 3D McFIS from the decoded frames so that we do not need to transmit McFIS from encoder to decoder. This technique increases decoder computational complexity compare to that of the existing decoder.

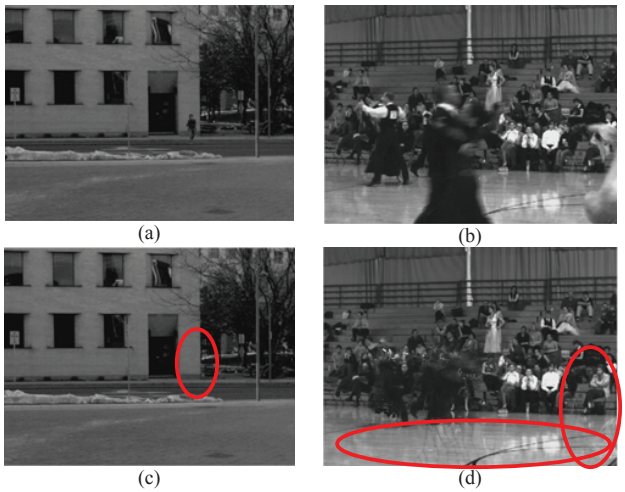


Fig 4: Examples of McFISes and uncovered/previously occluded background using *Vassar* and *Ball Room* video sequences, (a) & (b) an original frame of *Vassar* and *Ball Room* sequences respectively; (c) & (d) corresponding McFISes of the videos respectively.

### 3.1 3D McFIS generation

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, DBM is formulated. We assume that $k^{th}$ Gaussian at time $t$ representing a pixel intensity with mean $\mu_k^t$, *standard deviation* (STD) $\sigma_k^t$, recent value $\gamma_k^t$, and weight $\omega_k^t$ such that $\sum \omega_k^t = 1$. 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 $X^t$ (pixel intensity at time $t$) is first matched against the existing models in order to find one (e.g., $k^{th}$ model) such that $|X^t - \mu_k^{t-1}| \le 2.5\,\sigma_k^{t-1}$. If such a model exists, update corresponding recent value parameter $\gamma_k^t$ with $X^t$. Other parameters are updated with the learning rate as:

$\mu_k^t = (1-\alpha)\mu_k^{t-1} + \alpha X^t$;

$\sigma_k^{t\,2} = (1-\alpha)\sigma_k^{t-1\,2} + \alpha(X^t - \mu_k^t)^T(X^t - \mu_k^t)$;

$\omega_k^t = (1-\alpha)\omega_k^{t-1} + \alpha$; and the weights of the remaining Gaussians (i.e., $l$ where $l \neq k$) are updated as $\omega_l^t = (1-\alpha)\omega_l^{t-1}$. After each iteration, the weights are normalized. If the model does not exist, a new Gaussian model is introduced with $\gamma = \mu = X^t$, $\sigma$=30, and $\omega$=0.001 by evicting the $K^{th}$ (based on $w/\sigma$ in descending order) model if it exists. For more details in modeling and model updating, please refer [4][7]-[9]. 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.

Two examples of McFIS are shown in Fig 4 using frames of *Vassar* and *Ball Room* video sequences respectively. Fig 4 (a) & (b) show the original frames of corresponding videos and (c) & (d) show McFISes. The circles in (c) & (d) 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 second reference frame.

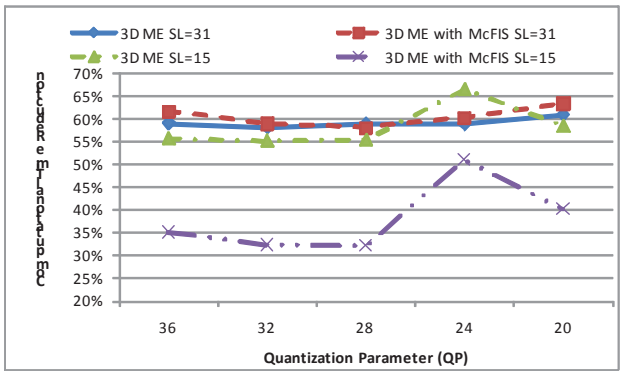


Fig 5: Average computational complexity reduction by the proposed methods (3D-ME and 3D-ME-McFIS) against the H.264/MVC using four video sequences where search length 15 and 31 are used.

### 3.2 Computational Complexity

One of the objectives of the proposed methods is to reduce the computational time of the existing multi-view video coding standard to enhance the scope of 3D video coding applications. Fig 5 shows computational time comparisons among the proposed methods and the existing video coding standard using two search lengths (i.e., 15 and 31) on four standard video sequences. The figure reveals that the proposed methods reduce around 60% of the computational time compared to the video coding standard when a search length 31 is used. Due to the fixed amount of operations

requirement for the McFIS modelling, the proposed 3D-ME-McFIS method reduces computational time slightly less compared to that of the proposed 3D-ME scheme. When large search length is used, the computational time requirement for the McFIS modelling is negligible compared to the motion estimation. The proposed 3D-ME-McFIS scheme uses a small search length (e.g., 2) for the motion estimation 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.

## 4. EXPERIMENTAL RESULTS

To compare the performance of the proposed schemes (3D-ME and 3D-ME-McFIS), we have implemented all the algorithms based on the H.264/MVC recommendations with 25 Hz, ±15 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 predication structure for H264/MVC. Obviously the proposed 3D-ME-McFIS technique will take some extra operations to generate McFIS. We used the same technique for modeling McFIS at the encoder and decoder, thus, we do not need to encode and transmit the McFISes to the decoder.

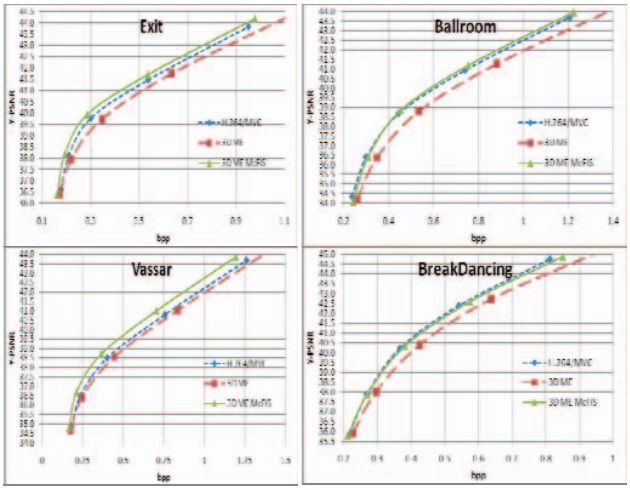


Fig 6: Rate-distortion performance by H.264/MVC and the proposed schemes (3D-ME and 3D-ME-McFIS) using four standard video sequences namely Exit, Ball Room, Vassar, and Break Dancing.

Fig 6 shows rate-distortion performance using H.264/MVC and two proposed schemes such as 3D-ME and 3D-ME-McFIS using four standard multi-view video sequences. The figure reveals that the rate-distortion performance of the proposed 3D-ME scheme is comparable to H.264/MVC. However, the proposed 3D-ME scheme outperforms H.264/MVC by reducing computational time by 60% (see Fig 5) and eliminating random access delay. The proposed 3D-ME-McFIS scheme outperforms H.264/MVC in terms of rate-distortion performance (Break Dancing is an exception) by improving more than 0.25 dB PSNR, computational complexity by reducing more than 60% of time, and eliminating random access frame delay. Due to the huge motions (less effectiveness of background modeling), the proposed 3D-ME-McFIS method does not outperform H.264/MVC for the Break Dancing sequence.

## 5. CONCLUSIONS

In this paper, we proposed a new 3-Dimensional motion estimation and compensation scheme to reduce the computational time and eliminate the random access frame delay of the existing H.264/MVC multi-view video coding standard. In the 3D-ME 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. This technique outperforms the existing standard by reducing computational time by more than 60% and eliminating random access frame delay without degrading the rate-distortion performance significantly compared to H.264/MVC multi-view video coding standard. This paper also proposes another technique (3D-ME-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 of a scene*) based on Gaussian mixture modelling. The experimental results reveal that 3D-ME-McFIS outperforms the H.264/MVC coding standard by improving 0.25dB PSNR, by reducing computational time by 60%, and by eliminating random access frame delay compared to the existing H.264/MVC multi-view (i.e., 3D) video coding. The proposed techniques enhance the 3D video coding application scopes at the interactive real time video communications.

## 6. REFERENCES


[1] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard," Proceedings of the IEEE, 99(4), 626 - 642, 2011.

[2] P. Pandit, A. Vetro, Y. Chen, "Joint Multiview Video Model (JMVM) 7 Reference Software," N9579, MPEG of ISO/IEC JTC1/SC29/WG11, Antalya, Jan. 2008.

[3] M. Talebpourazad, "3D-TV content generation and multi-view video coding, PhD thesis, 2010.

[4] M. Paul, W. Lin, C. T. Lau, and B. –S. Lee, "McFIS in hierarchical bipredictive picture-based video coding for referencing the stable area in a scene," *IEEE International conference on Image Processing* (IEEE ICIP-11), 2011.

[5] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard ," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, 2003.

[6] M. Paul, M. Frater, and J. Arnold, "An efficient Mode Selection Prior to the Actual Encoding for H.264/AVC Encoder," *IEEE Transaction on Multimedia*, vol. 11, no. 4, pp. 581-588, June, 2009.

[7] M. Paul, W. Lin, C. T. Lau, and B. –S. Lee "Explore and model better I-frame for video coding," *IEEE Transaction on Circuits and Systems for Video Technology*, 2011.

[8] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *IEEE CVPR*, vol. 2, 246–252, 1999.

[9] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Transactions on PAMI,* 27(5), pp. 827–832, May 2005.

[10] X. Li, D. Zhao, S. Ma, and W. Gao, "Fast disparity and motion estimation based on correlations for multi-view video coding," *IEEE Transactions on Consumer Electronics*, 54(4), pp. 2037-2044, 2008.