Afsana, Paul, M., Murshed, M., & Taubman, D. (2021). Efficient High-Resolution Video Compression Scheme Using Background and Foreground Layers. *IEEE Access*, 9, 157411–157421.

Available online: https://doi.org/10.1109/ACCESS.2021.3130249

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 26, 2021, accepted November 14, 2021, date of publication November 23, 2021,
date of current version December 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3130249

Efficient High-Resolution Video Compression Scheme Using Background and Foreground Layers

FARIHA AFSANA[1], **(Member, IEEE)**, **MANORANJAN PAUL**[1], **(Senior Member, IEEE)**, **MANZUR MURSHED**[2], **(Senior Member, IEEE)**, **AND DAVID TAUBMAN**[3], **(Fellow, IEEE)**

[1]School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia
[2]Faculty of Science and Technology, Federation University, Churchill, VIC 3842, Australia
[3]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia

Corresponding author: Fariha Afsana (fafsana@csu.edu.au)

This work was supported by the Australian Research Council through the Discovery Projects Grant DP190102574.

ABSTRACT Video coding using dynamic background frame achieves better compression compared to the traditional techniques by encoding background and foreground separately. This process reduces coding bits for the overall frame significantly; however, encoding background still requires many bits that can be compressed further for achieving better coding efficiency. The cuboid coding framework has been proven to be one of the most effective methods of image compression which exploits homogeneous pixel correlation within a frame and has better alignment with object boundary compared to traditional block-based coding. In a video sequence, the cuboid-based frame partitioning varies with the changes of the foreground. However, since the background remains static for a group of pictures, the cuboid coding exploits better spatial pixel homogeneity. In this work, the impact of cuboid coding on the background frame for high-resolution videos (Ultra-High-Definition (UHD) and 360-degree videos) is investigated using the multilayer framework of SHVC. After the cuboid partitioning, the method of coarse frame generation has been improved with a novel idea by keeping human-visual sensitive information. Unlike the traditional SHVC scheme, in the proposed method, cuboid coded background and the foreground are encoded in separate layers in an implicit manner. Simulation results show that the proposed video coding method achieves an average BD-Rate reduction of 26.69% and BD-PSNR gain of 1.51 dB against SHVC with significant encoding time reduction for both UHD and 360 videos. It also achieves an average of 13.88% BD-Rate reduction and 0.78 dB BD-PSNR gain compared to the existing relevant method proposed by X. Hoang Van.

INDEX TERMS Cuboid partitioning, DCT, quality scalability, SHVC, video coding.

I. INTRODUCTION

The emergence of high-quality video including ultra-high-definition (UHD) video, 360-degree immersive video, initiates new and exciting applications in virtual reality (VR), augmented reality (AR) and mixed reality (MR) in education, training, entertainment, and other markets [1], [2]. The bandwidth-intensive nature of these new generation video contents poses challenges in handling transmission and storage burdens while ensuring low latency delivery [3]. Furthermore, based on different user necessities including heterogeneous network capacities, display, power, and computing capabilities, the need for other video formats (e.g. SD, HD) is still in demand for adaptation to N-screen devices where different terminals are involved in managing video contents [4]. In this scenario, scalable video coding has emerged as a viable solution. The Scalable High-Efficiency Video Coding (SHVC) [5], the scalable extension of High-Efficiency Video Coding (HEVC) is the latest standard to the scalable era of video contents. The SHVC bitstream comprises one base layer (BL) and at least one or more enhancement layers (ELs). By leveraging inter-layer predictions (ILP) among BL and ELs, SHVC achieves high coding performance improvement of 30% over simulcast HEVC at the cost of high computational complexity [6]. However, further compression of SHVC is required for

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

compressing UHD/360-degree videos because due to having high resolution and frame rate, these videos incur additional data overhead.

The purpose of this study is to develop a new strategy for video compression of high-resolution video by exploiting the layered concept of SHVC scheme. In a video sequence, the foreground changes over time and the background remains static over time. Thus, a coding gain can be possible using background frame as the background needs to be encoded only once for a scene. So far the existing background-based video coding techniques [7]–[9] encode background frame as an intra-frame by a coding standard or avoid encoding by modeling from the already decoded frames in both encoder and decoder. If encoded, it takes more bits to get a decent quality, and if modelled, it does not provide high-quality image as it uses decoded frames. Thus, they encode the background frame in high quality which requires more bits similar to intra-coded frame. The existing cuboid-based techniques [10], [11] encode original frames where both background and foreground exist; thus, it cannot exploit homogeneity as an expected level due to the dynamic nature of foreground. The background has high spatial homogeneity compared to the original frame where foregrounds exist. It also has high temporal correlation among frames and can be compressed efficiently using cuboid coding as there is more pixel homogeneity than the frame with background and foreground.

In this paper we apply cuboid-based partitioning on background frame where more homogeneity exists due to static nature of the content. Thus, we can partition the frame in a better way for more compression for a given quality. In the proposed scheme, the structure of SHVC scheme has been adopted where foreground and background are encoded in separate layers implicitly. Thus, we are referring BL as a reference layer (RL) which provides EL with Static information (background) from a Sequence of original frames ($S_I S_{OF}$) of the video sequence. By exploiting the availability of the static information in RL, the objective is to combine it with EL stream to form an improved EL prediction stream and thus, further improving coding efficiency of EL. To improve the overall coding efficiency, the RL is externally encoded using cuboid partitioning and further compressed by adopting 2-D Discrete Cosine Transform (DCT) scheme.

The main contributions of the paper are as follows:

- A new video coding strategy has been proposed by adopting the structure of SHVC scheme where externally encoded coarse representation of background is provided as RL in one layer, and the original frame is provided in separate EL layer. The RL is meant for providing the most common information of the frame sequence while the immediate previous frame of EL will provide the motion information.
- The effectiveness of cuboid coding on the Static information frame of a Sequence of original frames ($S_{IF} S_{OF}$) has been studied to improve the overall coding gain as well as reduce time complexity.
- A modified coarse frame generation scheme has been adopted where instead of replacing each cuboid by mean intensity values as adopted in [12], DCT is applied on each cuboid and visually most significant information is exploited by truncating high-frequency components.
- A comprehensive analysis of the Rate-Distortion (R-D) performance based on the reconstructed RL has been provided to understand the applicability of the proposed method in the SHVC video coding scheme.

II. LITERATURE REVIEW

A number of works have been carried out in the field of video coding to improve the performance of video compression. In the scalable coding sector, for achieving better coding efficiency, most of the works focus on Inter-Layer prediction (ILP) based on decoded BL data.

The works in [13]–[15], improved the performance of SHVC by taking an adaptive filtering approach in inter-layer reference prediction. Hoang *et al.* in [16] proposed a joint layer coding mode by linearly combining BL and EL decoded information at pixel level to improve SHVC coding efficiency. [17] is another approach where a joint layer prediction (JLP) method was proposed to improve the performance of SHVC. By applying the decoded information achieved from both BL and EL, the JLP method could create a new prediction picture. The work in [18] proposed an adaptive long-term reference selection algorithm for surveillance cameras using scalable video coding. Based on the content analysis of a video sequence, this approach selected a coded picture as long-term reference picture. Though these mechanisms outperformed SHVC in terms of bit saving, those still required a lot of extra bits to encode the ILP picture.

Apart from improving the coding efficiency of scalable video coding, various efforts have been adopted to reduce the computational complexity of SHVC encoder [19]–[22]. Most of these methods introduced fast CU depth level decisions and applied spatial scalability to reduce overall run time but experienced an increase in the bit rate.

However, these methods are still deficient due to having a trade-off between improved coding efficiency and reduced complexity. In the proposed work, our aim is to improve the coding efficiency of SHVC for UHD and 360-degree videos while preserving the quality as well as controlling the time complexity.

III. PROPOSED FRAMEWORK

The proposed method consists of three steps: (i) Generation of background frame, $S_{IF} S_{OF}$, (ii) externally encoding $S_{IF} S_{OF}$ to reconstruct RL stream and (iii) encoding EL stream using regenerated RL stream. The schematic diagram of RL stream generation is depicted in Fig. 1. In this method, we explore the effectiveness of externally encoded RL with $S_{IF} S_{OF}$ in SHVC encoding scheme. For this, first, few $S_{IF} S_{OF}$s are extracted from corresponding chunks of

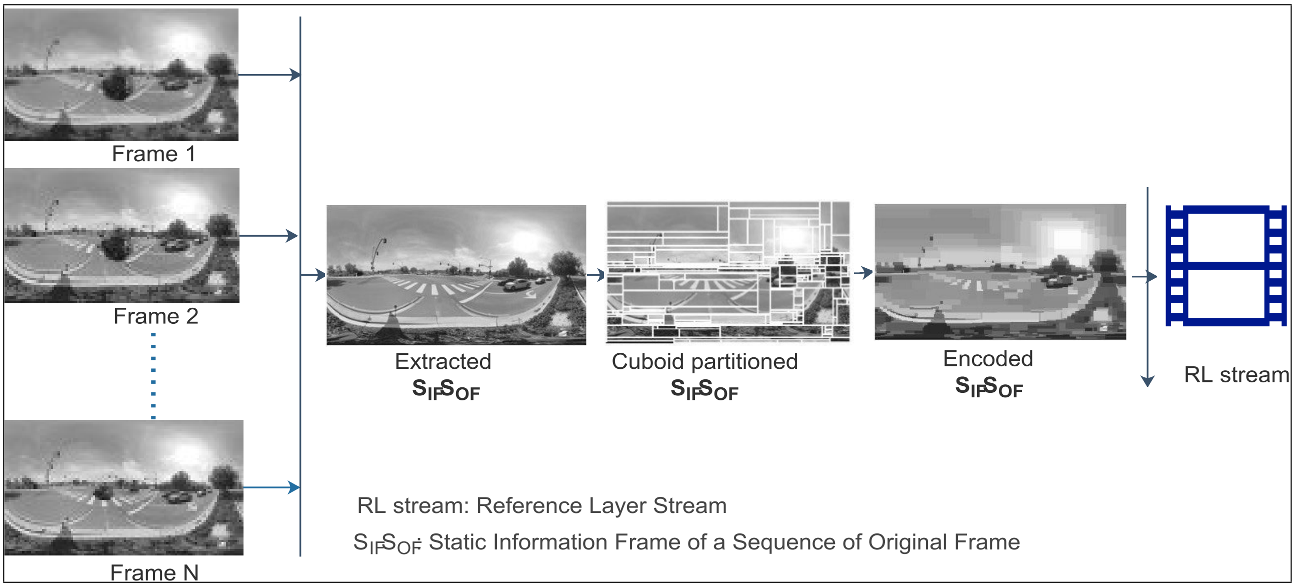

Frame 1
Frame 2
Frame N
Extracted
S_IF S_OF
Cuboid partitioned
S_IF S_OF
Encoded
S_IF S_OF
RL stream
RL stream: Reference Layer Stream
S_IF S_OF: Static Information Frame of a Sequence of Original Frame

FIGURE 1. Block diagram of RL stream regeneration.

original frames of a video sequence using an existing dynamic background modeling scheme [23]. Then the extracted background frames are segmented using cuboid partitioning [10]. To encode the RL stream externally, those cuboid partitioned frames are compressed with DCT scheme. Thus, externally encoded $S_{IF}S_{OF}$s are used to regenerate RL stream and finally used to encode EL stream of UHD and 360-degree video sequences. All the steps are elaborated in the later sections.

A. GENERATION OF BACKGROUND FRAME, $S_{IF}S_{OF}$

Given a video sequence, $\nu \in \mathbb{R}^{P\times Q\times\eta}$, of resolution $P \times Q$ and frame number η, the aim is to extract the most static information of frames over a number of frames, $I_p < \eta$. In a video sequence, scenes captured by a static camera have steady background, β, over the frames except the interference of moving objects and changes of illumination at foreground, ξ. Thus, a video sequence can be decomposed as, $\nu = \beta + \xi$. As the distribution of background pixels differ from that of the foreground ones, the β layer can be formulated by judging the pixel deviation [24]–[26]. The recent state-of-the art methods for β layer subtraction use low-rank subspace learning approaches [27]–[31]. The standard approach used for background subtraction is Low-Rank Matrix Factorization (LRMF) [32]. For r-rank matrix factorization, the LRMF factorizes β into two smaller matrices, $M \in \mathbb{R}^{PQ\times r}$ and $N \in \mathbb{R}^{I_p\times r}$, where $r < min(PQ, I_p)$, such that,

$$\beta = Fold(N^T M) \tag{1}$$

Here, the operation 'Fold' folds up each column of a matrix into the corresponding frame matrix of a tensor. Thus, at each intra-period interval, I_p, a background frame, β, is generated.

TABLE 1. Changes of PSNR and bits with different numbers of cuboids.

Video Sequence	No. of Cuboids n_s	PSNR(dB)	Bits (n_b)
Broadway (HD)	2000	34.45	16169
	3000	35.17	23149
	4000	35.36	29896
Community (UHD)	9000	36.99	76280
	10000	37.41	83835
	11000	37.48	91270

B. CUBOID PARTITIONING

After β extraction, the next step is to encode it externally using cuboid partitioning. The cuboid coding scheme divides β of resolution $P \times Q$ into $n \cong n_s$ cuboids based on a user-defined number of segments n_s. n_s depends on the resolution of video as well as its quality. In the segmentation process [33], $\beta_{P\times Q}$ is first separated into two half cuboids - $\beta^{m1}_{P\times Q}$ of size $m\times Q$ and $\beta^{m2}_{P\times Q}$ of size $(P-m)\times Q$, using a vertical line $x = m+0.5$ in $P-1$ ways with $m \in 1, 2, \cdots P-1$. Similarly, using a horizontal line $y = n + 0.5$ in $Q - 1$ ways, where $n \in 1, 2, \cdots Q-1$, $\beta_{P\times Q}$ can be split into two half-cuboids - $\beta^{(P-1+n)1}_{P\times Q}$ of size $P \times n$ and $\beta^{(P-1+n)2}_{P\times Q}$ of size $P \times (Q - n)$. The pixel intensity contrast distance between a cuboid split pair, denoted as $D^{s1}D^{s2}$, is defined as [34],

$$f(s|\beta_{P\times Q}) = D^{s1}D^{s2}, \tag{2}$$

Then, to maximize the objective function,$f(s|\beta_{P\times Q})$, a greedy optimization heuristic is applied to find the best split

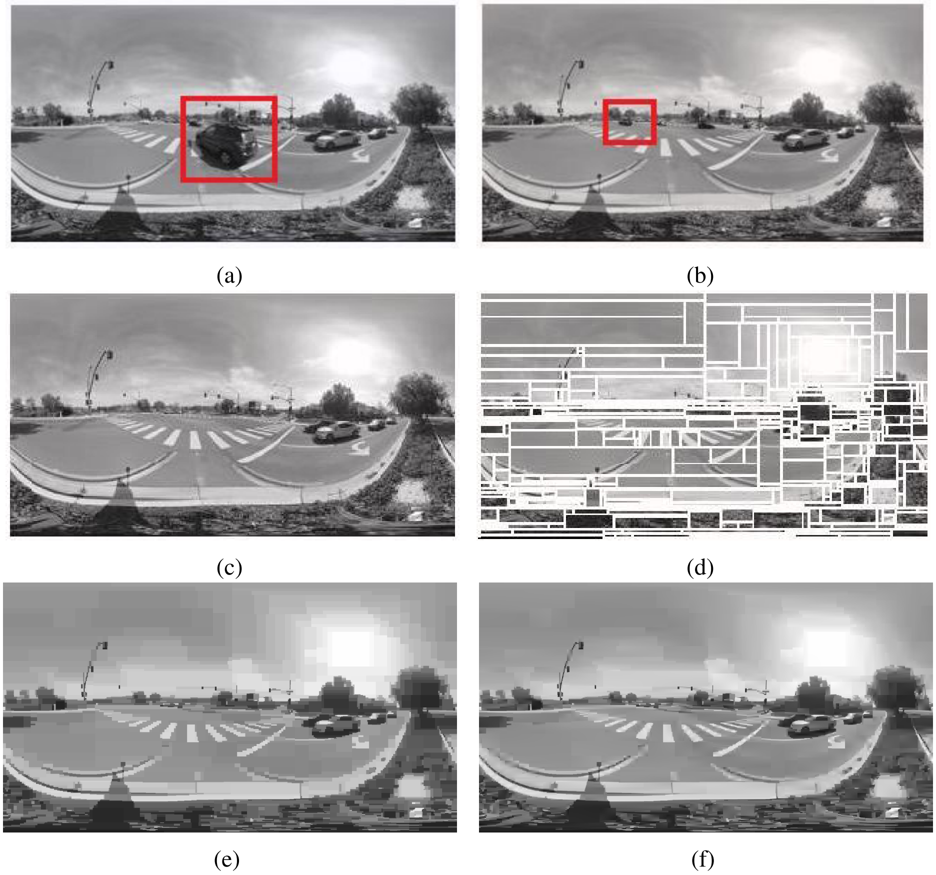


FIGURE 2. **Example of coarse background frame generation from 'Intersection' Video sequence. (a) Original first frame, (b) Original 60th Frame (c) Background frame ($S_{IF}S_{OF}$), (d) Background frame partitioned with cuboids (e) $S_{IF}S_{OF}$ encoded with mean [PSNR 32.69 dB] (f) $S_{IF}S_{OF}$ encoded with DCT [PSNR 33.53 dB] (proposed).**

of $\beta_{P\times Q}$ from the possible $P+Q-2$ ways as:

$$maximize_{1\leq s\leq P+Q-2} f(s|\beta_{P\times Q}) \tag{3}$$

By recursively partitioning one cuboid into two halves using the optimal split, s_*, a hierarchical partitioning algorithm is designed which terminates when all possible ways of splitting get invalid. At the end, the algorithm returns a binary partitioning tree, Γ, of $\beta_{P\times Q}$ with horizontal or vertical split lines and a cuboid map, $\lambda = \{s_*\}_{j=1}^{n_s-1}$, found from the current frame. It is possible to reproduce the previous image using the indices of λ. Thus, it is necessary to send the indices values (λ) from the encoder to the decoder so that the decoder can reconstruct the cuboid map and the frame. To encode these indices, Exponential-Golomb coding technique is used [35]. The encoded indices are then augmented to EL bitstream to be transmitted at decoder.

The number of cuboids, n_s are determined based on the resolution and quality of the video as well as bits, n_b, required to run the cuboid partitioning. This is an optimisation problem. A number of different video sequences with different texture and motion information have been analysed to determine the number of cuboids. In table 1, we presented only two results as an example to visualize how a suitable cuboid number has been chosen. From table 1, it is observed that with the increase of n_s, both the PSNR and n_b increase. The computational time increases as well. So, it is required to decide on a value of n_s, for which both the PSNR and n_b will be suitable. Analysing a number of video sequences, we found that, for HD videos, after $n_s > 3000$, n_b increases but the PSNR increase is insignificant. Same is the case with $n_s > 10000$ for UHD videos. Thus in this study, we used $n_s = 3000$ and 10000 for HD and UHD videos, respectively.

C. COARSE REPRESENTATION OF RL STREAM

At this stage, a set of cuboids, $\mathbb{C} = \{\beta^1_{P\times Q}, \beta^2_{P\times Q}, \cdots, \beta^{n_s}_{P\times Q}\}$ is found. To extract essential information from each cuboid $\beta^i_{P\times Q} \in \mathbb{C}$, we use DCT in the spatial domain that transforms information into the frequency domain. In an image, usually, low-frequency data contains most common information, and high-frequency data contains detailed information. According to the information compaction property of the DCT, the most important information is concentrated in few of the output data points of the 2-D DCT coefficient matrix. The first element (DC value) of the output data points carries the most important information of the original image. The remaining coefficients (AC values) carry the detailed information in the decreasing order if arranged in a zigzag manner [36]. Since the low-frequency data has much bigger effect than the high-frequency data, the least significant high-frequency coefficients are masked off and removed from the coefficient matrix by applying data quantization process. After quantization, most of the high-frequency data points (lower right corner of the matrix) are zero. In order to maintain the balance between the perceptual video quality and the bitrate, we truncated a number of DCT coefficients while keeping the most significant values. It is found that after the top four coefficient values, if we increase the number of coefficients further, the number of bits needed to encode the DCT coefficients increases rapidly compared to the PSNR increase. Again, in the cuboid-based approach, it is found that with the increased number of cuboids, most of the cuboid contains null high-frequency values after four top-left coefficient values. Thus, for the reconstruction of RL frame, we have considered top four significant coefficient values and truncated the rest. Since the large DCT coefficients are concentrated in the low-frequency area [37], the top four significant values reside in the low-frequency zone having high energy compactness characteristics. Finally, the RL frame is generated using inverse DCT that uses the truncated coefficients where the high-frequency components are absent.

Algorithm 1 represents the overall process of RL generation. Fig. 2 represents an example scenario of coarse background frame generation. The red marked block of Fig. 2a and Fig. 2b denotes the only moving object among frame 1 to frame 60. Thus, the background extraction model considers the rest as static scene and generated background as Fig. 2c. Fig 2d represents the cuboid partitioning of generated background frame. Due to applying DCT, the overall number of bits required to compress Fig. 2f is larger than that of Fig. 2e. To control the increase of bits, we have kept only the top four significant DCT components from each cuboids. Thus, the overall PSNR of the generated Fig. 2f has gained 0.9 dB improvement compared to mean reconstructed image (Fig. 2e).

Thus, the coarse background frame is generated using the cuboid map and top four DCT coefficients from each cuboid. A new background frame is updated based on the intra-period interval. Finally, all coarse representations of background frames are accumulated to reconstruct RL stream where the same frame is repeated between two consecutive I-frames. The information of λ and truncated DCT coefficient matrix are then fed into exponential golomb processor to get the desired compressed bit stream and transmit those to the decoder.

Algorithm 1 RL Generation Algorithm

Notations:

ν = Video Sequence,
I_P = Intra-Period,
ν_{I_p} = Sequence of frame within an intra-period,

β = Extracted background from a sequence of frames,
$\mathbb{C}$ = List of cuboids,
$\beta^i_{P\times Q}$ = A single cuboid of β where $P \times Q$ is the dimension of β,
$\Omega^i_{P\times Q}$ = Matrix of DCT data points for cuboid $\beta^i_{P\times Q}$,
$\Omega^i(j,k)$ = Coefficient of DCT matrix from a specific position (j,k),

$\digamma^i_{P\times Q}$ = New cuboid constructed with the value of $\Omega^i_{P\times Q}$,
$\digamma$ = Reconstructed β with new values of cuboids,
ν_N = Reconstructed video sequence;

```
Algorithm RLGereneration($\nu_{I_p}$)
    foreach $\nu_{I_p} \longleftarrow \nu$ do
        $\beta = ExtractBackground(\nu_{I_p})$;
        $\mathbb{C} \longleftarrow CuboidPartitioning(\beta)$;
        foreach $\beta^i_{P\times Q} \longleftarrow \mathbb{C}$ do
            $\Omega^i_{P\times Q} \longleftarrow DCT2(\beta^i_{P\times Q})$;
            $\Omega^i_{P\times Q} \longleftarrow Quantize(\Omega^i_{P\times Q})$;
            $\Omega^i(j,k) = \begin{cases} \Omega^i(j,k) & \text{if } 1 \le j,\ k \le 2 \\ 0 & \text{Otherwise} \end{cases}$
            $\Omega^i_{P\times Q} \longleftarrow DeQuantize(\Omega^i_{P\times Q})$;
            $\digamma^i_{P\times Q} \longleftarrow IDCT2(\Omega^i_{P\times Q})$;
        end
        $\nu_{I_p} = \digamma$;
        Append $\nu_{I_p}$ to $\nu_N$;
    end
    Return $\nu_N$;
```

IV. EXPERIMENTAL RESULTS

In this section, the result for evaluating the performances of the proposed method is presented. To evaluate the coding performance of the SHVC for two-layer SNR scalability, the most recent SHVC reference software, SHM-12.4 [38] is used. The experiments have been carried out as per the common test conditions on an AMD Ryzen 7 processor (PRO 3700Uw) running at 2.30 GHz with 16 GB RAM. To conduct the evaluation, several benchmark HD and UHD/360-degree test sequences of YUV 4:2:0 format with different motion characteristics and content variation [39]–[41] are used. Test sequences are encoded with general coding options for random access (RA) coding structure for all encodings with a hierarchical GOP of size 16 and an intra-period of 32. For studying the encoding performance, 120 frames have been selected from each of the sequences. The specification of the test sequences are summarized in Table 2.

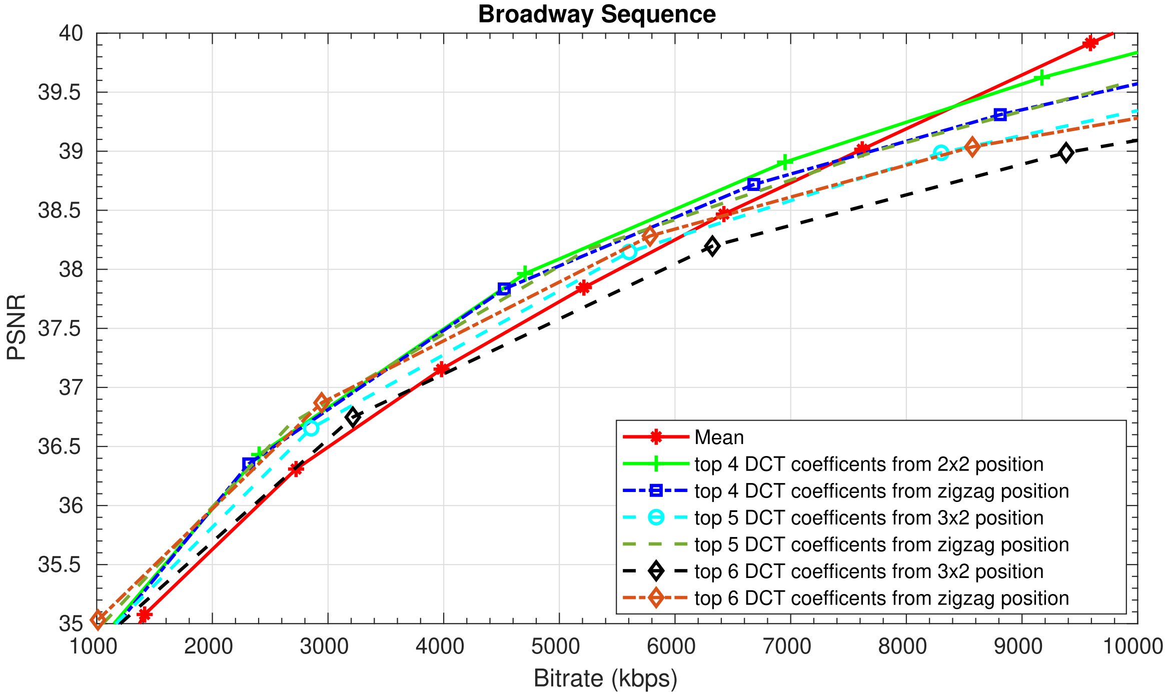


FIGURE 3. Performance comparison of R-D curves between mean and DCT reconstructed image for 'Broadway' Sequence.

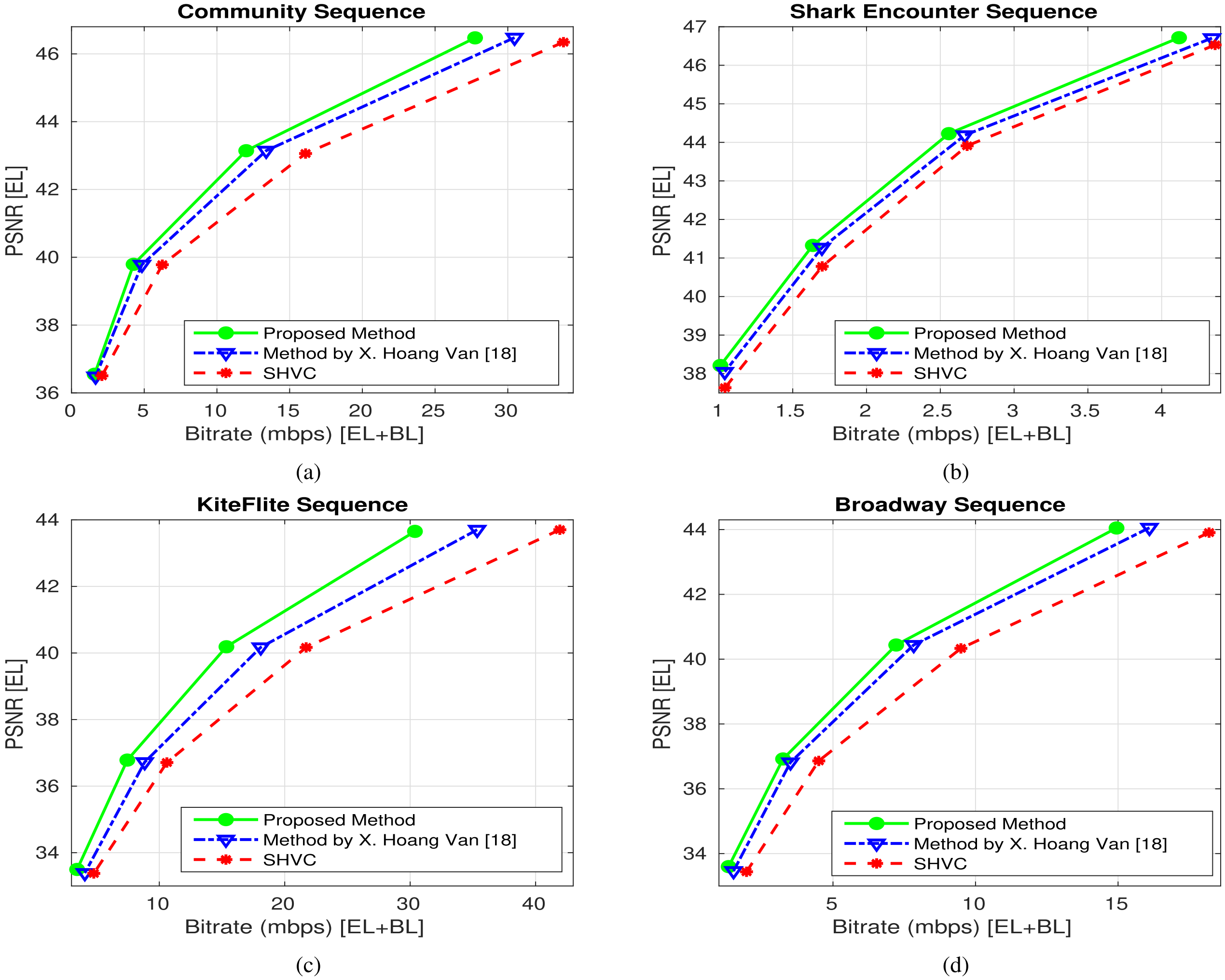
TABLE 2. Specification of Test Sequences.

Video Sequence	Resolution	Video Type	Frame Rate	QP values
Community [39]	3840x2160	UHD/360	30	QP_{RL} ={24,29,34,39} QP_{EL}={QP_{RL}-2}
KiteFlite [39]	3840x2160	UHD/360	30	
Intersection [39]	1280x720	HD/360	30	
Broadway [39]	1280x720	360	30	
Shark Encounter [40]	1280x720	360	30	
Library [41]	1280x720	HD/ Normal	30	

In the proposed method, we have one EL along with the RL for SNR scalability. This method differs from the traditional SHVC as the RL contains only $S_{IF}S_{OF}$. The encoded bitstream of the cuboid map and DCT coefficients of RL are augmented with EL bitstream to evaluate the performance of the proposed scheme. The performance of the proposed method is assessed in terms of Bjontegaard BD-Rate, BD-PSNR [42] and execution time with respect to the original unmodified SHVC scheme.

The performance comparison between mean and DCT for reconstructing a frame is analyzed in Fig. 3 for a background frame of 'Broadway' sequence. The information loss by gradually suppressing a number of high-frequency data points from the DCT coefficient matrix is also observed. From the figure, it is observed that the R-D curve for DCT with top four coefficient values performs best compared to the R-D curve with five or six DCT coefficients and mean. In the proposed method, with the increase of cuboids, many of the values after four coefficients get null values in the DCT matrix (as the block size is not always square in cuboid partitioning). For this reason, DCT with larger number of coefficients cannot outperform DCT with four coefficients. Thus, we selected the top-left 2 × 2 DCT coefficients as a good compromise to outperform mean reconstructed image. From the figure, it is observed that, with the increase of cuboids, for very high-quality image, mean outperforms DCT. Since the cuboids can be rectangular, when a large number of cuboids are used many of the DCT blocks contain only DC value, which gets quantized and loses more information; thus, cannot outperform mean. For the later experiments, to verify the feasibility of the proposed method, DCT with top four coefficients has been used to reconstruct RL.

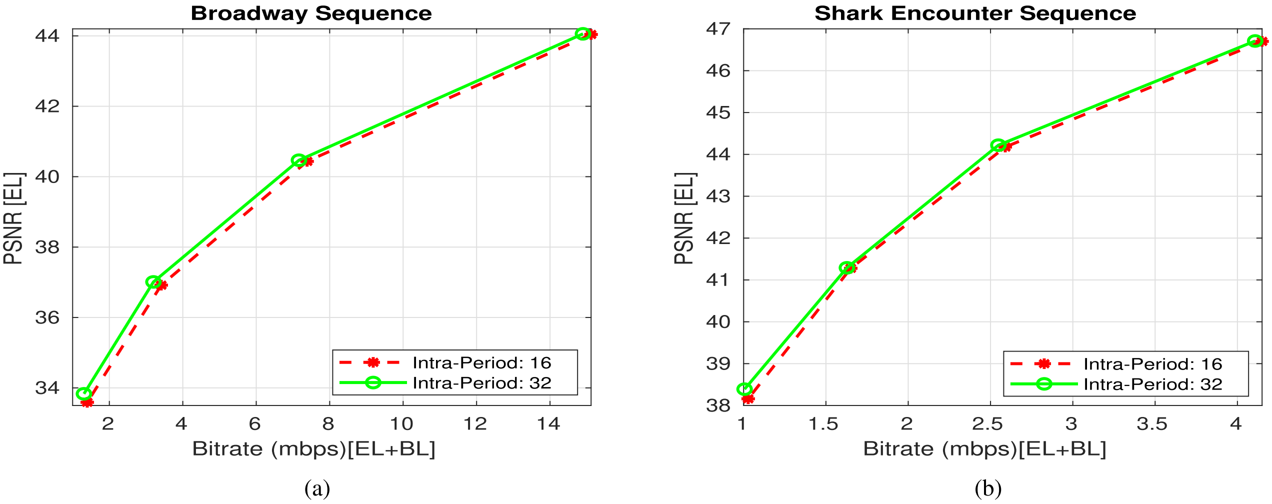
The performance comparison among proposed method, traditional SHVC and reference method [18] is presented



(a) (b)

(c) (d)

FIGURE 4. **Performance comparison of R-D curves.**



(a) (b)

FIGURE 5. **Performance comparison of R-D curves of proposed method for different Intra-Periods.**

in Fig. 4 for 'Community', 'Shark Encounter', 'KiteFlite' and 'Broadway' video sequence. The reason for comparing the proposed method against the reference method [18] is, this method uses background as a long-term reference in the SHVC model. However, in our case, the background frame is further encoded to compress it ensuring the quality. From the figure it is observed that the proposed method outperforms the reference method in terms of bit rate saving but the

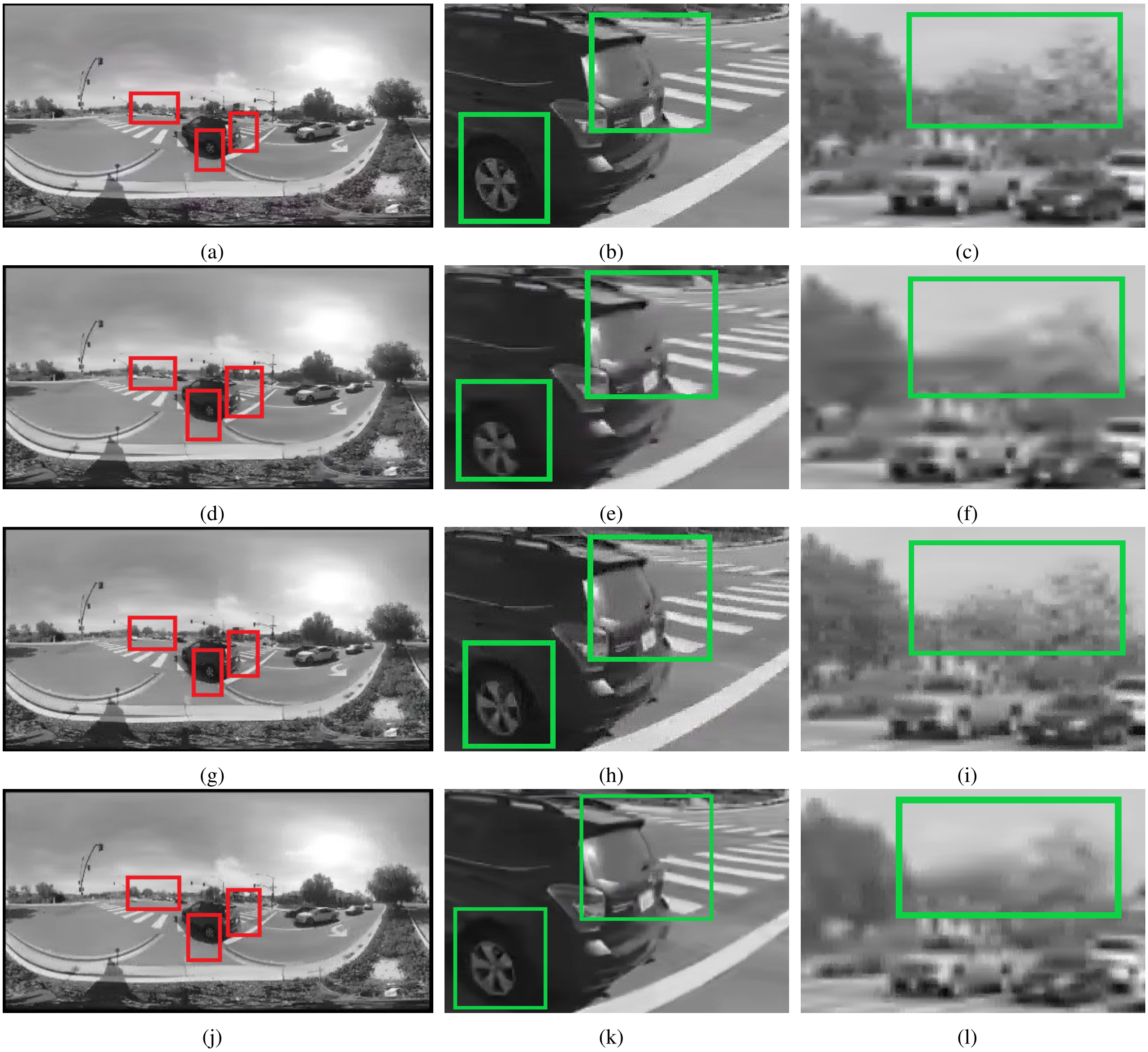


FIGURE 6. Qualitative comparison results of video frame from 'Intersection' sequence (a) Original image, (b) Part of Original image (Boundary of car and wheel objects), (c) Part of Original image (Tree objects) (d) Reconstructed image (SHVC) [PSNR 33.14, bitrate = 5.7 Mbps], (e) Part of Reconstructed image (SHVC) (Boundary of car and wheel objects), (f) Part of Reconstructed image (SHVC) (Tree objects) (g) Reconstructed image (Proposed) [PSNR 37.20, bitrate = 5.1 Mbps], (h) Part of Reconstructed image (Proposed) (Boundary of car and wheel objects) (i) Part of Reconstructed image (Proposed) (Tree objects). (j) Reconstructed image (method proposed by X. HoangVan) [PSNR = 33.18, bitrate = 5.8 Mbps],(k) Part of Reconstructed image (method proposed by X. HoangVan) (Boundary of car and wheel objects), (l) Part of Reconstructed image (method proposed by X. HoangVan) (Tree objects).

PSNR values are almost same. Furthermore, from the figure it is observed that our proposed method outperforms SHVC significantly in terms of PSNR and bit rate.

In Fig.5, the performance comparison of the proposed method for two different Intra-periods 16 and 32, is studied for 'Broadway' and 'Shark Encounter' sequences. From the results, it is observed that there is an overall improvement in the performance of the proposed method for these two video sequences when the intra-period is doubled.

Table 3 represents the BD-Rate reduction, BD-PSNR gain and execution time saving of six video sequences compared to the SHVC scheme and the method proposed by X. HoangVan [18]. From the table, it is observed that, for normal video sequence ('Library'), the proposed method performs best compared to HD/UHD/360 videos. The overall performance indicates that, the proposed approach achieves an average of 26.69% BD-Rate saving and 1.51 dB BD-PSNR gain on top of SHVC. Again, in case of the reference method, the proposed method achieves an average of 13.88% BD-Rate reduction and 0.78 dB BD-PSNR gain against the reference method.

In the case of time complexity reduction, the proposed method outperforms SHVC with an average of 9.95% time-saving. The proposed method also outperforms the method

TABLE 3. BD-rate reduction, BD-PSNR gain and execution time saving of the proposed scheme against SHVC scheme and the method by X. HoangVan [18].

Video Sequence	Against SHVC			Against method by X. Hoang Van [18]		
	BD-Rate (EL+BL) (%)	BD-PSNR (EL) (dB)	Execution Time (%)	BD-Rate (EL+BL) (%)	BD-PSNR (EL) (dB)	Execution Time (%)
Community	-28.87	+1.15	11.9	-10.26	+0.37	12.7
KiteFlite	-28.77	+1.68	12.9	-14.89	+0.82	13.5
Intersection	-30.53	+2.18	7.84	-26.23	+1.97	8.21
Broadway	-26.24	+1.43	12.2	-9.21	+0.42	11.27
Shark Encounter	-10.43	+0.68	5.81	-4.77	+0.27	4.98
Library	-35.32	+1.98	9.1	-17.95	+0.87	9.47
Average	-26.69	+1.51	9.95	-13.88	+0.78	10.02

proposed by X. Hoang Van [18] with an average of 10.02%. As multiple platforms (Matlab, SHM encoder) are used to encode RL and EL separately, the execution time is approximated. We considered the time required to encode EL, construct background, and partition cuboid to calculate the runtime of the proposed method. The computational time to encode the RL was estimated from the ratio of RL encoding time and EL encoding time [43]. The ratio differs with the varying resolution of the test sequences. Finally, it was considered with the EL encoding time to estimate the overall coding time. An estimation of execution time is presented here in Table 3 where the time for SHVC encoding is kept unchanged. But to calculate time for the proposed method, along with the encoder time, time required for background generation as well as cuboid partitioning has been added. The proposed method needs extra time to generate background and cuboid partitioning; however, it saves significant encoding time as it does not need to explore different modes of intra-coding done by the SHVC coding scheme.

A. SUBJECTIVE EVALUATION

Figure 6 represents the visual quality comparison of the proposed method, conventional SHVC and the method by X. HoangVan [18] against the original video frame. The example frame was taken from the 'Intersection' video sequence. Here, frames from the proposed method, SHVC method and the reference method were encoded with adjusted QP values to meet a specific bitrate. The bitrate needed to encode the original frame (Fig. 6a) was 5.1 Mbps for the proposed method, 5.7 Mbps for SHVC and 5.8 for the reference method. To present the comparison in image quality, let us concentrate on the boundary of the car, wheel, and tree objects marked by the red squares in Figure 6a, 6d, 6g, and 6j which were zoomed to larger sizes in the later images. Indeed, the visual quality of the reconstructed frame obtained from the proposed scheme (Fig. 6g) is better than that of the SHVC frame (Fig. 6d) and the reference frame (Fig. 6j). For example, the boundary of the car object in Fig. 6b (marked by the green rectangle) is better defined in the proposed scheme (Fig. 6h) than in SHVC (Fig. 6e). The SHVC method created a blur boundary of the car object and also the boundary of the wheel is unclear where the proposed method defined those clearly with better quality. In the case of the reference method, the visual quality of the proposed method (Fig. 6h) is still better compared to that of the reference method. In the reference method (Fig. 6k), the boundary of the car object is defined better than SHVC but the boundary of the wheel is still unclear. Again, the tree objects in Fig. 6c (marked by the green rectangle) are well defined in the proposed scheme (Fig. 6i) but the SHVC scheme (Fig. 6f) and the reference shceme (Fig. 6l) failed to define them clearly. Both the SHVC method and the reference method could not reconstruct the clear view of the tree objects but the proposed method defined that. Since cuboids are encoded based on the homogeneous pixel intensity of pixels, the proposed method could define objects clearer than that of block-based SHVC.

V. CONCLUSION

In this paper, an efficient video coding strategy for high-resolution video has been proposed using the coarse representation of background image. The impact of cuboid coding on externally encoded background frames is investigated by adopting the structure of SHVC. It is found that cuboid coding works better in aligning homogeneous object boundary, and thus, a lot of compressions have been achieved, especially for the background. Again, the cuboid coded background is further compressed by exploiting the human-visual sensitive information. This method is different compared to the traditional SHVC as RL layer has only the background and the EL layer has the foreground. However, a new scalable coding technique can be designed where one ground layer can be added for the background and other layers for the foreground or multiple layers for foregrounds and backgrounds. The experimental results confirm that the proposed method outperforms SHVC in terms of bit rate saving (26.69%) and PSNR gain (1.51 dB) with significant encoding time reduction for high-resolution videos.

REFERENCES

[1] B.-G. Kim, "Fast coding unit (CU) determination algorithm for high-efficiency video coding (HEVC) in smart surveillance application," *J. Supercomput.*, vol. 73, no. 3, pp. 1063–1084, Mar. 2017, doi: 10.1007/s11227-016-1730-y.

[2] T. Biatek, W. Hamidouche, J.-F. Travers, and O. Deforges, "Optimal bitrate allocation for high dynamic range and wide color gamut services deployment using SHVC," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2016, pp. 299–308, doi: 10.1109/DCC.2016.26.

[3] D. Nguyen, T. Le, S. Lee, and E.-S. Ryu, "SHVC tile-based 360-degree video streaming for mobile VR: PC offloading over mmWave," *Sensors*, vol. 18, no. 11, p. 3728, Nov. 2018, doi: 10.3390/s18113728.

[4] J. M. Boyce, Y. Yan, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable extensions of the high efficiency video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 20–34, Jan. 2016, doi: 10.1109/TCSVT.2015.2461951.

[5] G. J. Sullivan and J.-R. Ohm, *Joint Call for Proposals on Scalable Video Coding Extensions of High Efficiency Video Coding (HEVC)*, document VCEG-AS90 and WG 11 N12957, 45th VCEG Meeting and 101st MPEG Meeting, Stockholm, Sweden, Jul. 2012.

[6] A. A. Ramanand, I. Ahmad, and V. Swaminathan, "A survey of rate control in HEVC and SHVC video encoding," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2017, pp. 145–150, doi: 10.1109/ICMEW.2017.8026268.

[7] M. Paul, W. Lin, C.-T. Lau, and B.-S. Lee, "Explore and model better I-frames for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1242–1254, Sep. 2011, doi: 10.1109/TCSVT.2011.2138750.

[8] Y. Xu, J. Dong, B. Zhang, and D. Xu, "Background modeling methods in video analysis: A review and comparative evaluation," *CAAI Trans. Intell. Technol.*, vol. 1, no. 1, pp. 43–60, 2016, doi: 10.1016/j.trit.2016.03.005.

[9] D. Samaiya, A. Joshi, and K. K. Gupta, "Background modeling for HEVC compressed videos using radial basis network," in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, Chennai, India, Apr. 2019, pp. 37–40, doi: 10.1109/ICCSP.2019.8697986.

[10] S. Tania, M. Murshed, S. W. Teng, and G. Karmakar, "Cuboid colour image segmentation using intuitive distance measure," in *Proc. Int. Conf. Image Vis. Comput. New Zealand (IVCNZ)*, Auckland, New Zealand, Nov. 2018, pp. 1–6, doi: 10.1109/IVCNZ.2018.8634676.

[11] A. Ahmmed, M. Murshed, and M. Paul, "Leveraging cuboids for better motion modeling in high efficiency video coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 2188–2192, doi: 10.1109/ICASSP40776.2020.9053851.

[12] F. Afsana, M. Paul, M. Murshed, and D. Taubman, "Efficient low bit-rate intra-frame coding using common information for 360-degree video," in *Proc. IEEE Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[13] P. Lai, S. Liu, and S. Lei, "Low latency directional filtering for inter-layer prediction in scalable video coding using HEVC," in *Proc. Picture Coding Symp. (PCS)*, San Jose, CA, USA, Dec. 2013, pp. 269–272.

[14] M. Guo, S. Liu, and S. Lei, "Inter-layer adaptive filtering for scalable extension of HEVC," in *Proc. Picture Coding Symp. (PCS)*, San Jose, CA, USA, Dec. 2013, pp. 165–168.

[15] C.-S. Park, T.-J. Kim, J.-H. Lee, and B.-G. Kim, "Adaptive inter-layer prediction algorithm for scalable extensions of high efficiency video coding," in *Proc. IEEE Region Conf.*, Singapore, Nov. 2016, pp. 3063–3066.

[16] X. HoangVan, J. Ascenso, and F. Pereira, "Improving SHVC performance with a joint layer coding mode," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 1145–1149, doi: 10.1109/ICASSP.2016.7471855.

[17] X. Hoangvan and B. Jeon, "Joint layer prediction for improving SHVC compression performance and error concealment," *IEEE Trans. Broadcast.*, vol. 65, no. 3, pp. 504–520, Sep. 2019, doi: 10.1109/TBC.2018.2881355.

[18] X. HoangVan, "Adaptive quantization parameter estimation for HEVC based surveillance scalable video coding," *Electronics*, vol. 9, no. 6, p. 915, May 2020, doi: 10.3390/electronics9060915.

[19] A. J. Diaz-Honrubia, J. L. Martinez, and P. Cuenca, "A fast hybrid scalable H.264/AVC and HEVC encoder," *J. Supercomput.*, vol. 73, no. 1, pp. 277–290, Jan. 2017, doi: 10.1007/s11227-016-1802-z.

[20] X. Lu, C. Yu, Y. Gu, and G. Martin, "A fast intra coding algorithm for spatial scalability in SHVC," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1792–1796, doi: 10.1109/ICIP.2018.8451844.

[21] X. Lu, C. Yu, and G. Martin, "Fast encoding algorithms for SHVC Intra/Inter coding," in *Proc. Data Compress. Conf. (DCC)*, Snowbird, UT, USA, Mar. 2019, p. 595, doi: 10.1109/DCC.2019.00107.

[22] L. Shen, P. An, and G. Feng, "Low-complexity scalable extension of the high-efficiency video coding (SHVC) encoding system," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 2, pp. 44:1–44:23, Jun. 2019, doi: 10.1145/3313185.

[23] M. Li, Q. Xie, Q. Zhao, W. Wei, S. Gu, J. Tao, and D. Meng, "Video rain streak removal by multiscale convolutional sparse coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6644–6653, doi: 10.1109/CVPR.2018.00695.

[24] S. Chakraborty, M. Paul, M. Murshed, and M. Ali, "An efficient video coding technique using a novel non-parametric background model," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Chengdu, China, Jul. 2014, pp. 1–6, doi: 10.1109/ICMEW.2014.6890590.

[25] M. Paul, W. Lin, C. T. Lau, and B.-S. Lee, "Video coding with dynamic background," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, p. 11, Jan. 2013, doi: 10.1186/1687-6180-2013-11.

[26] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Comput. Sci. Rev.*, vols. 11–12, pp. 31–66, May 2014.

[27] D. Meng and F. De La Torre, "Robust matrix factorization with unknown noise," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2013, pp. 1337–1344.

[28] X. Cao, Q. Zhao, D. Meng, Y. Chen, and Z. Xu, "Robust low-rank matrix factorization under general mixture noise distributions," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4677–4690, Oct. 2016.

[29] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang, "Robust principal component analysis with complex noise," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 55–63.

[30] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and Y. Yan, "L_1-norm low-rank matrix factorization by variational Bayesian method," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 829–839, Jan. 2015.

[31] H. Yong, D. Meng, W. Zuo, and L. Zhang, "Robust online matrix factorization for dynamic background subtraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1726–1740, Jul. 2018.

[32] N. Zarmehi, A. Amini, and F. Marvasti, "Low rank and sparse decomposition for image and video applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2046–2056, Jul. 2020, doi: 10.1109/TCSVT.2019.2923816.

[33] M. Murshed, S. W. Teng, and G. Lu, "Cuboid segmentation for effective image retrieval," in *Proc. Int. Conf. Digit. Image Comput., Techn. Appl. (DICTA)*, Sydney, NSW, Australia, Nov. 2017, pp. 1–8, doi: 10.1109/DICTA.2017.8227422.

[34] S. Shahriyar, M. Murshed, M. Ali, and M. Paul, "Cuboid coding of depth motion vectors using binary tree based decomposition," in *Proc. Data Compress. Conf.*, Snowbird, UT, USA, Apr. 2015, p. 469, doi: 10.1109/DCC.2015.43.

[35] A. Ahmmed, M. Murshed, M. Paul, and D. S. S. Taubman, "A commonality modeling framework for enhanced video coding leveraging on the cuboidal partitioning based representation of frames," *IEEE Trans. Multimedia*, early access, Oct. 4, 2016, doi: 10.1109/TMM.2021.3117397.

[36] N. R. Kishor, H. Barman, U. S. N. Raju, S. K. Kanaparthi, and H. Ala, "Content based image retrieval using frequency domain features: Zigzag scanning of DCT coefficients," in *Proc. Int. Conf. Artif. Intell. Smart Syst. (ICAIS)*, Mar. 2021, pp. 1535–1540, doi: 10.1109/ICAIS50930.2021.9396008.

[37] G. Madhuri and C. H. Bindu, "Performance evaluation of multi-focus image fusion techniques," in *Proc. Int. Conf. Comput. Netw. Commun. (CoCoNet)*, Dec. 2015, pp. 248–254, doi: 10.1109/CoCoNet.2015.7411194.

[38] *SHM Reference Software for SHVC*. Accessed: May 1, 2021. [Online]. Available: https://hevc.hhi.fraunhofer.de/

[39] *JVET-G1030: JVET Common Test Conditions and Evaluation Procedures for 360° Video*. Accessed: May 1, 2021. [Online]. Available: https://www.researchgate.net/publication/326504378_JVET-G1030_JVET_common_test_conditions_and_evaluation_procedures_for_360_video

[40] *Shark Encounter 360 Degree Sequence*. Accessed: May 1, 2020. [Online]. Available: https://www.youtube.com/watch?v=WwR1PP-ANzs

[41] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *Proc. 5th Int. Workshop Qual. Multimedia Exper.*, Klagenfurt, Austria, Jul. 2013, pp. 34–35.

[42] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33 ITU-T Q6/16, Austin, TX, USA, Apr. 2001.

[43] F. Afsana, M. Paul, M. Murshed, and D. Taubman, "Efficient scalable UHD/360-video coding by exploiting common information with cuboid-based partitioning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 15, 2021, doi: 10.1109/TCSVT.2021.3113056.

FARIHA AFSANA (Member, IEEE) received the B.Sc. degree (Hons.) and the M.Sc. degree in information technology from Jahangirnagar University, Dhaka, Bangladesh, in 2015 and 2016, respectively. She is currently pursuing the Ph.D. degree with Charles Sturt University, Bathurst, NSW, Australia. Her current research interests include video coding, image processing, e-health, data mining, the Internet of Things, nanotechnology, and wireless sensor networks.

MANORANJAN PAUL (Senior Member, IEEE) received the Ph.D. degree from Monash University, Australia, in 2005. He was a Postdoctoral Research Fellow with the University of New South Wales, Monash University, and Nanyang Technological University. He is currently a Full Professor, the Director of Computer Vision Laboratory, and the Leader of the Machine Vision and Digital Health (MaViDH) Research Group, Charles Sturt University, Australia. He was an Invited Keynote Speaker in IEEE DICTA-17 & 13, CWCN-17, WoWMoM-14, and ICCIT-10. He received more than $3.6 million competitive external grant, including Australian Research Council (ARC) Discovery grants, and Australia-China grant. He has supervised 15 Ph.D. students to completion. He has published around 200 peer-reviewed publications, including 72 journal articles. His major research interests include video coding, image processing, digital health, wine technology, machine learning, EEG signal processing, eye tracking, and computer vision. He was awarded the ICT Researcher of the Year 2017 by the Australian Computer Society. He was a General Chair of PSIVT-19 and a Program Chair of PSIVT-17 and DICTA-18. He is an Associate Editor of three top ranked journals, including the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, and the *EURASIP Journal in Advances on Signal Processing*.

MANZUR MURSHED (Senior Member, IEEE) received the B.Sc.Engg. degree (Hons.) in computer science and engineering from the Bangladesh University of Engineering and Technology, Dhaka, in 1994, and the Ph.D. degree in computer science from Australian National University, Canberra, in 1999.

He was with Federation University Australia (FedUni) as a Robert HT Smith Professor and a Personal Chair in information technology, from 2014 to 2018, and with Monash University as the Head of the Gippsland School of Information Technology, from 2007 to 2013. He is currently a Professor of information technology and the Research Director of the Centre for Multimedia Computing, Communications, and Artificial Intelligence Research, FedUni. He is also leading FedUni's Research Priority Area on Information Forensics and Security. He has published over 230 refereed research articles with over 4000 citations as per Google Scholar and received $3.6 million competitive research funding, including six Federal Government grants—four Discovery Projects grants and the Linkage Infrastructure, Equipment, and Facilities Grant from the Australian Research Council, and the Education Integration Project Grant from the Australian General Practice Training. He has successfully supervised 25 Ph.D. and three M.Phil. students to completion. His research interests include video technology, information theory, wireless communications, distributed and cloud computing, and security and privacy. He received the Vice-Chancellor's Knowledge Transfer Award (commendation) from the University of Melbourne in 2007, the Inaugural Early Career Research Excellence Award from the Faculty of Information Technology, Monash University, in 2006, and the University Gold Medal from the Bangladesh University of Engineering and Technology in 1994. He is a Program Co-Chair of PCS 2015, DICTA 2017, DICTA 2018, and PSIVT 2019, an Area Chair of ICME 2014, and the Awards Chair of ICME 2012. He is an Associate Editor of the IEEE Transactions on Multimedia and the IEEE Transactions on Circuits and Systems for Video Technology, and a Guest Editor of the IEEE Transactions on Cloud Computing.

DAVID TAUBMAN (Fellow, IEEE) received the B.S. and B.E. degrees in electrical engineering from the University of Sydney, in 1986 and 1988, respectively, and the M.S. and Ph.D. degrees from the University of California at Berkeley, in 1992 and 1994, respectively. From 1994 to 1998, he was with Hewlett-Packard's Research Laboratories, Palo Alto, CA, USA. He joined UNSW, in 1998, where he is currently a Professor with the School of Electrical Engineering and Telecommunications. He has authored the book entitled *JPEG2000: Image Compression Fundamentals, Standards and Practice*, with M. Marcellin. His research interests include highly scalable image and video compression, motion estimation and modeling, inverse problems in imaging, perceptual modeling, and multimedia distribution systems. He received the University Medal from the University of Sydney. He has received two best paper awards from the IEEE Circuits and Systems Society for the 1996 paper entitled "A Common Framework for Rate and Distortion-Based Scaling of Highly Scalable Compressed Video," and from the IEEE Signal Processing Society for the 2000 paper entitled "High Performance Scalable Image Compression with EBCOT."