

Multi-resolution extreme learning machine-based side information estimation in distributed video coding

Bodhisattva Dash¹ · Suvendu Rup¹ ·
Anjali Mohapatra¹ · Banshidhar Majhi² ·
M. N. S. Swamy³

Received: 21 September 2017 / Revised: 17 March 2018 / Accepted: 21 March 2018 /
Published online: 27 March 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract *Context:* Encoding of video frames in a traditional video coding architecture involves exhaustive computations due to the motion estimation (ME) task. Hence, it requires a considerable amount of computing aid, battery power, and resource memory. These codecs are not effective and reliable for applications like surveillance systems, wireless sensor networks, wireless camcorders, having scarcity in the availability of resources and computing ability. Therefore, in such scenarios, distributed video coding (DVC) represents a viable solution for power-constrained hand-held devices. DVC empowers the adaptability in distributing the complexity between the encoder and the decoder. *Objective:* Like any other building block, the decoder driven side information (SI) generation module plays a key role in a DVC codec. The efficacy of a DVC codec firmly relies on the quality of the SI generated at the decoder. SI is considered to be the facsimile of the original Wyner-Ziv (WZ) frame. Hence, the superior the quality of SI, improved is the efficiency of the codec. The primary objective of the present work is to enhance the quality of the SI frame so that the overall performance of the DVC is improved. To achieve this objective, this article deals with a hybrid SI generation scheme utilizing the principles of discrete wavelet transform (DWT) and extreme learning machine (ELM) algorithm in a transform domain-based DVC framework. *Results:* Exhaustive simulations have been carried out for some standard video sequences with the proposed and benchmark schemes. The proposed scheme is evaluated with respect to different performance metrics such as rate-distortion (RD), SI peak-signal-to-noise-ratio (PSNR) vs frame number, number of parity requests per SI frame, and so

✉ Bodhisattva Dash
bdash.fac@gmail.com

¹ Image and Video Processing Laboratory, Department of Computer Science and Engineering, International Institute of Information Technology, Bhubaneswar-751003, Odisha, India

² Pattern Recognition Research Laboratory, Department of Computer Science and Engineering, National Institute of Technology, Rourkela-769004, Odisha, India

³ Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada

on. Experimental results and its analyses corroborate that the performance of the proposed technique surpasses as that of the benchmark schemes.

Keywords Distributed video coding (DVC) · Transform domain wyner-ziv video coding (TDWZ) · Discrete wavelet transform (DWT) · Side information (SI) · Extreme machine learning (ELM) · Structural similarity index (SSIM) · Rate-distortion (RD)

1 Introduction

Since decades, traditional video DVC compression methods have been extended to various applications with a large-scale of bit-rate limitations. To achieve high coding efficiency, traditional video codecs, namely, the H.26x series [47] employ an asymmetric complexity partitioning between the encoder and the decoder. To obtain the benefit of temporal correlation between adjacent frames, the motion estimation (ME) approaches are utilized by these codecs. The ME methods help in improving the compression efficiency and the rate-distortion (RD) characteristics as well. However, the high computational burden is one of the major drawbacks of these approaches. In traditional video codecs, the encoder complexity is considered to be extremely high, whereas the complexity of the decoder is really simple. These are typically suitable for applications where the encoding of the video frames is done only once and decoding is carried out many-a-time.

On the other hand, as these encoders need a considerable amount of computational overhead, it becomes arduous for applications with minimal resources (e.g. computing ability, battery power, and memory requirement) like wireless video sensors or surveillance systems, capsule endoscopy, and so on, which need a real-time transferal [7]. Further, considering the computing abilities of the aforementioned scenarios, a complex ME task cannot be performed as it might consume more time. Additionally, there also lies a possibility in draining of the battery power. These resource-constrained applications are perhaps not capable of handling the delay incurred due to the ME task. Moreover, it might not be feasible to load a large amount of computational resources into these sensors and hence designing complex encoders is not a practicable choice.

For all such situations, distributed video coding (DVC) can be considered to be an affirming alternative as it aims at switching the computational complexity from the encoder to the decoder. The concept of DVC has been entrenched in [39, 48], and it can be briefly outlined as contrary to the traditional video codecs. DVC stems on the theory of distributed source coding [39, 48], wherein two inter-dependent sources are individually encoded and jointly decoded. In recent years, various DVC-based video codecs have been presented along with the state-of-the-art architecture [3, 13, 32]. An effective SI refinement framework [6] is presented where a low pass filter is applied initially on both the key frames, followed by a block matching algorithm for the ME task. This coding configuration is also referred to as the Instituto Superior Técnico Transform Domain Wyner-Ziv (IST-TDWZ), which is adopted by many DVC researchers. Another efficient framework called Distributed Coding for Video Services (DISCOVER) has been proposed to generate a better quality of SI frame [3]. Both IST-TDWZ and DISCOVER codec show similar performance, however, they use different encoding techniques.

In these paradigms, initially, the input video sequence is segregated into odd (key) and even (WZ) frames. The odd frames are encoded and decoded using the traditional-based video codecs, whereas the even frames are SW-based encoded and decoded. Further, a frame estimation technique is used to generate the side information (SI) for the corresponding WZ

frames, using the neighboring decoded key frames. The rate-distortion (RD) characteristic of these codecs firmly relies on the quality of the generated SI frame. Moreover, the quality of the SI frames is again dependent on various factors such as the quality of the neighboring decoded odd frames, the motion behavior between frames, and so on. To enhance the quality of the decoded key frames, a Burrow-Wheeler transform (BWT)-based intra-frame coding technique [35] has been proposed. To improve the RD behavior of the codec with longer GOPs, an unsupervised motion learning technique has been presented [45]. Similarly, two different techniques have been presented to improve the RD characteristics considering both the issues, namely, intense motion and longer GOPs [27, 28].

Though various methodologies have been presented to enhance the quality of the SI frame (see Section 3), there still exists a scope to develop efficient SI estimation algorithms for further performance enhancement of the DVC codec. In the present work, a multi-resolution (MR) extreme learning machine (ELM)-based SI estimation in a DVC framework is proposed. The wavelet transform is allied to MR analysis and sub-band decomposition, and hence, it has been effectively utilized in numerous image and video processing applications [25, 37, 38]. On the other hand, unlike other neural networks (NNs), ELM is a straightforward machine learning algorithm [17], which has a better generalization capability with reduced time complexity. In ELM, the parameters (weights, bias) associated with the input and hidden layers are randomly chosen. Additionally, the weights between the hidden and output layers are analytically determined using the least square technique [29].

The contribution of the suggested work is summarized as follows. Initially, a level-3 sub-band decomposition using DWT is employed on a predefined number of frames (both even and odd), to obtain the wavelet coefficients (approximation, and detailed), out of which, only the approximation coefficients are used to create the training (input, target) pattern. Next, the created training pattern is utilized to train the ELM network. Like any other machine learning algorithm, ELM also works in two phases, namely, the training phase and the testing phase. Once the network is trained, in the testing phase, it is used to estimate the approximation coefficient for the remaining even (WZ) frames. Further, using the estimated approximation coefficient and the previously stored detailed coefficients, an inverse discrete wavelet transform (IDWT) is employed, at each of the sub-band levels, to generate the eventual estimated SI frame in the original form (spatial domain).

The remaining sections of the present article are arranged as follows. An outline of the basic Stanford-based transform domain DVC framework is presented in Section 2. Section 3 presents a brief review of the relevant alternate SI generation approaches. A generalized conceptual overview of DWT and ELM techniques are outlined in Section 4. The proposed hybrid SI generation algorithm is critically discussed in Section 5. The detailed experimental setup, comprehensive simulation, and the results are illustrated in Section 6. Finally, in Section 7, the closing remarks along with the scope for future work are presented.

2 Stanford-based transform domain DVC framework

The basic framework of the transform domain-based DVC codec [1], started in the Image group of the Instituto Superior Técnico (IST) is represented in Fig. 1. It is also referred to as the IST-TDWZ framework. The operational workflow is briefly discussed below.

Initially, at the encoder side, the video sequence is segregated into odd (key) and even (WZ) frames. Then, the odd frames are intra-coded using the conventional H.264/AVC encoder. The intra-coded frames are reconstructed (decoded) using the H.264/AVC decoder.

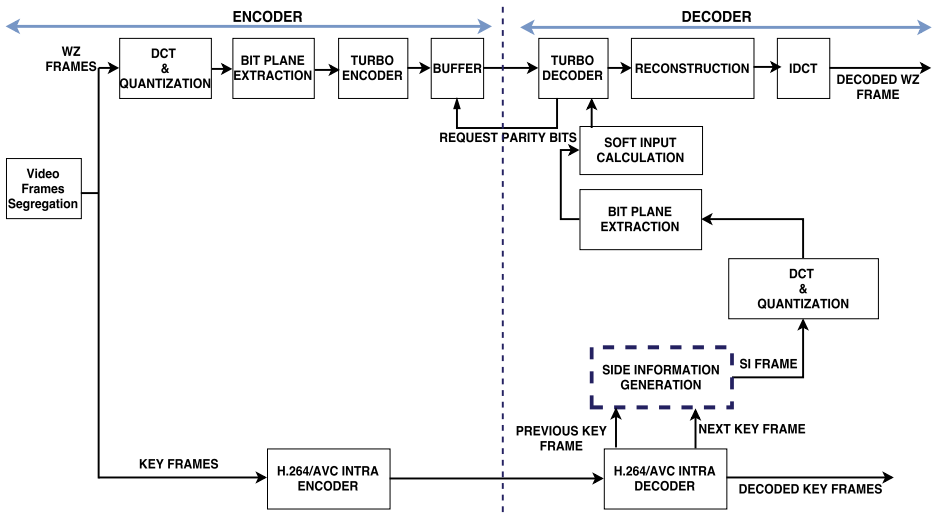


Fig. 1 Stanford-based transform domain DVC architecture

These reconstructed frames are used to estimate the side information (SI) for the corresponding WZ frames. As represented in Fig. 1, a 4×4 block-based integer discrete cosine transform (DCT) is used to compress the WZ frames. Then, depending on the desired output quality, a uniform quantization process using a set of quantization matrices [6] is employed to generate one DC, and sixteen different AC coefficient bands. Next, each of the obtained bands is individually sent to a bit plane extraction module. Here, each of the coefficients present in a band is represented in terms of bits, and bit planes are extracted starting with the most significant bit (MSB) to least significant bit (LSB). Then, the extracted bit planes, starting with MSB to LSB, are sent individually to a half-rate turbo encoder (TE). For each bit of a bit plane, TE produces two bits, namely, a systematic bit and a parity bit, using the following

$$\begin{bmatrix} 1 & \frac{1 + D + D^3 + D^4}{1 + D^3 + D^4} \end{bmatrix}; \text{ where } D \text{ denotes the delay.} \tag{1}$$

As the estimated SI frame for the corresponding WZ frame is already available at the decoder, the systematic bits are rejected. This helps in attaining an additional amount of compression efficiency. A buffer is used to accumulate the parity bits which are further sent to the decoder in a cyclic manner, upon request, chunk-by-chunk, based on a pseudo-random puncturing format with a span of 48. Similarly, at the decoder end, depending on the group-of-picture (GOP) size, SI for the corresponding WZ frame is estimated using the previously reconstructed adjacent key/WZ frames. If GOP is two, the two neighboring key frames will be the immediate past and immediate next of the current WZ frame. For larger GOP's, the already estimated and decoded WZ frames will act as the reference frames for creating and decoding the next SI frames. Upon the estimated SI frame, similar steps are followed, starting with the DCT transformation to bit plane extraction techniques. Then, SI along with the requested parity bits are fed to an iterative decoder (also known as turbo decoder (TD)). Further, using the log-likelihood ratio (LLR), a bit error rate (BER) probability (P_e) for each of the bit planes is computed. If $P_e > 10^{-3}$, requests for the additional amount of parity

bits is made to reduce the error, or else, the current bit plane is considered to be successfully decoded. After obtaining all the decoded streams, the DCT coefficients are regenerated by applying the method presented in [21]. Based on the quantization matrices used, the bands which are not transmitted are replaced directly by the SI coefficients. Upon the regenerated coefficients, an inverse discrete cosine transform (IDCT) is carried out to obtain the pixel values. These pixel values are reorganized to generate the final decoded WZ frame. Finally, all the decoded (key and WZ) frames are ordered to build the video sequence.

3 Alternate approaches for SI generation in DVC

From the architectural description of the DVC framework (refer Section 2), it is noticed that there exist few key modules which affect the overall RD behavior of the codec. Out of these, from many observations, it is learnt that SI frame estimation has a significant influence on the overall codec efficacy. SI creation is a process of estimating the WZ frames utilizing the intra-/inter-information present at the decoder end. The following equation represents the relationship between the estimated SI and its corresponding WZ frame.

$$\hat{Y} = Y + \eta \quad (2)$$

where η is the error that exists due to imprecise SI estimation, Y represents the WZ frame, and \hat{Y} represents the estimated SI frame.

Since past few decades, to reduce the noise ‘ η ’, which in turn minimizes the required bit rate by the SW codec, numerous SI creation techniques have been presented by DVC researchers. Among these, Girod *et al.* at Stanford University, and Ramchandran *et al.* at the University of California, Berkeley, have formulated some of the promising SI generation frameworks. Further, among the presented frameworks, the Stanford-based DVC framework has gained a lot of popularity among the DVC researchers. Hence, in the present article, the Stanford-based DVC architecture is adopted, and some of the relevant literature in the context of SI generation based on Stanford-based DVC framework is highlighted below.

Ascenso *et al.* presented a framework which can adjust dynamically to the inter-coding pattern by controlling the GOP length [4]. To control the GOP size, a rank-based clustering algorithm has been proposed. Moreover, to generate SI, a block-based motion compensated frame interpolation (MCFI) technique has also been proposed. A Kalman-based filtering technique has been presented to model the motion vectors [40]. In this work, a comparison has been made taking different situations, namely, the prediction of the motion vector at the encoder, the inter-/extra-polation of motion field at the decoder. Petrazzuoli *et al.* proposed a high-order motion interpolation approach [31] to create SI using four reference frames, namely, two previous and two next frames for the current WZ frame. An auto regressive (AR) model for generating better SI has been proposed in [51]. In their approach, a window is chosen in the previously decoded $t - 1$ WZ/Key frame, where t denotes the time index. A linear weighted summation operation is performed on the pixels within the selected window to generate each pixel of SI for the considered WZ frame. Further, for final SI generation, a probability-based fusion model along with a centrosymmetric correspondence approach has also been proposed. The reported methodology is able to reduce the gap between the traditional and distributed video codecs.

An integrated frame approximation approach [19] using an optical flow (OF) and overlapped block motion compensation (OBMC) technique has been proposed. In this work, both OF and OBMC techniques are used to create the SI frame, which is further

utilized by a multi-hypothesis-based TDWZ decoder, employing a weighted joint distribution approach. An SI generation technique which involves a combination of global and local MCFI approach [2] has been proposed. Though there is an enhancement in RD characteristics, the encoder complexity increases considerably. A refinement-based framework has been presented which generates SI using an overlapped block motion and multi-hypothesis-based estimation technique for visual sensor applications [9]. A progressive-based DVC architecture has been proposed in which the spatial similarity between the video frames are exploited in-order to enhance the motion-compensated temporal interpolation (MCTI) [41]. Further, to enhance the quality of SI, a side information refinement (SIR) process is also employed. [42] presented a new SI generation scheme which uses variable block size method to create SI. Van *et al.* [43] presented a MORE technique which uses the concept of optical flow. In this work, the proposed MORE scheme is merged with the SING-TDWZ codec [44] to analyze the overall performance.

A continuous learning-based SI generation framework using a multi-resolution and motion refinement (MRMR) technique has been presented in [23]. To satisfactorily estimate the bit probability distribution utilizing the advantage of higher order information of a video sequence, a wavelet-domain MRMR-based context adaptive paradigm [33] for DVC has been proposed. Rup *et al.* [36] proposed an SI generation scheme, where the WZ frames are predicted from two decoded key frames adjacent to it, using a multi-layer perceptron (MLP) network. Dash *et al.* proposed an ensemble of MLP networks for SI generation in a DVC framework [8]. A review of further studies related to SI creation in a DVC framework is presented in [11, 20].

From the above discussions, the observation inferred are as follows. SI generation is one of the most crucial tasks in a DVC framework, and the compression efficiency of DVC heavily depends on the correlation between the SI frame and the corresponding original WZ frame. Further, to generate SI at the decoder, most of the proposed methodologies employ a spatial domain based ME/MC task. These are extremely complex and highly time-consuming. Furthermore, the video sequences constitute non-linear motion patterns and few machine learning-based approaches have been utilized to estimate the SI frame in a DVC paradigm. With this in mind, the authors are motivated to propose a multi-resolution (MR) extreme learning machine (ELM)-based SI estimation in a DVC framework.

4 Preliminaries

4.1 Wavelet transform

The wavelet transform (WT) is considered to be an effective mechanism for analyzing the information in the frequency domain. The primitive usefulness of these transformation techniques is that the concerned signal can be analyzed in multiple scales and resolutions [46], also referred to as MR analysis (MRA). The iterative mathematical formulation of MRA can be depicted as

$$A_j = D_{j+1} \oplus A_{j+1} = D_{j+1} \oplus D_{j+2} \oplus \cdots \oplus D_{j+n} \oplus A_n \quad (3)$$

where A_{j+1} and D_{j+1} denotes the approximation and detailed coefficients for a given signal at scale $j + 1$, respectively, \oplus denotes the summation of two decomposed signals, and ' n ' is the level of decomposition.

One of the major differences between the wavelet and other types of transforms is that WT generates the time-frequency positioning of the signal. The mathematical representation of a wavelet family is

$$\psi_{u,s}(t) = \frac{1}{\sqrt{|s|}} \psi\left(\frac{t-u}{s}\right) \tag{4}$$

where $\Psi(t)$ is a real or complex valued function, ‘s’ indicates the scaling component for shrinking ($s < 1$) or stretching ($s > 1$) the wavelet function (WF), and ‘u’ represents the translation component utilized to shift the position of the WF. The factor $(1/\sqrt{|s|})$ denotes the energy normalization constant along various scales.

To carry out both the MRA decomposition and reconstruction at the same time, it adopts two functions, namely, the wavelet function (φ) and the scaling function (ϕ). The former is used to propagate the detailed variant having high frequency elements of the signal, and the later to produce the approximate variant of the signal which includes low frequency elements. The mathematical definitions of these two functions (in discrete domain) are given below.

$$\varphi_{j,n}[t] = 2^{\frac{j}{2}} \sum_n w_{j,n} \varphi[2^j t - n] \tag{5}$$

$$\phi_{j,n}[t] = 2^{\frac{j}{2}} \sum_n s_{j,n} \phi[2^j t - n] \tag{6}$$

where s_j and w_j represent the scaling and wavelet coefficients at j_{th} scale, respectively. These two functions are required to be orthonormal and satisfy the following conditions.

$$\langle \phi \cdot \phi \rangle = \frac{1}{2^j} \tag{7}$$

$$\langle \varphi \cdot \varphi \rangle = \frac{1}{2^j} \tag{8}$$

$$\langle \phi \cdot \varphi \rangle = 0 \tag{9}$$

Given, the scaling parameter $s = 2^j$ where $j \in \mathbf{Z}$ (set of all integers), a binary or dyadic WT [24] can be generated. Selecting the translation and scaling parameters to be multiples of 2 leads to a discrete wavelet transform (DWT). The DWT of a discrete signal $y(n)$ can be found using the equations

$$A_{2^j} y(n) = \sum_{z \in \mathbf{Z}} l_z A_{2^{j-1}} y(n - 2^{j-1} z) \tag{10}$$

$$D_{2^j} y(n) = \sum_{z \in \mathbf{Z}} h_z A_{2^{j-1}} y(n - 2^{j-1} z) \tag{11}$$

where ‘n’ indicates the sample number, $A_{2^j} y(n)$ & $D_{2^j} y(n)$ denote the approximation and detailed coefficients, respectively. The parameter $l_z, h_z (z \in \mathbf{Z})$ represents the low pass and high pass coefficients, respectively, and are given by

$$L(\omega) = \sum_{z \in \mathbf{Z}} l_z e^{-izw} \tag{12}$$

$$H(\omega) = \sum_{z \in \mathbf{Z}} h_z e^{-izw} \tag{13}$$

4.2 Extreme learning machine (ELM)

Unlike the traditional feed-forward neural networks (FFNNs), which are based on the principle of recursive adjustment of all of its network parameters, Huang *et al.* [17] presented an ELM technique which insists independent and random assignment of the hidden node parameters. Moreover, in ELM, the output parameters (i.e. weights between the hidden and the output layer) of the network can be computed using the least square method [29]. The learning phase of the ELM network can be done effectively with less number of iterations, and can attain a better generalization characteristic.

For ‘ S ’ random distinct samples (p_i, q_i) , where $p_i = [p_{i_1}, p_{i_2}, \dots, p_{i_n}]^T \in \mathbb{R}^n$ and $q_i = [q_{i_1}, q_{i_2}, \dots, q_{i_m}]^T \in \mathbb{R}^m$, a standard single-layer feed forward network (SLFN) having ‘ H ’ hidden nodes can be mathematically represented as

$$o_i = \sum_{k=1}^H \beta_k A(w_k, b_k, p_i), i = 1, 2, \dots, S \tag{14}$$

where w_k and b_k denotes the hidden node parameters (weight, and bias), $\beta_k = [\beta_{k_1}, \beta_{k_2}, \dots, \beta_{k_m}]^T$ denotes the output weight vector between the k^{th} hidden node and the output nodes, $A(w_k, b_k, p_i)$ represents the output of the k^{th} node with respect to p_i , and o_i denotes the true output with respect to p_i . Further, $A(w_k, b_k, p_i)$ can be mathematically modeled as

$$A(w_k, b_k, p_i) = a(w_k^T \cdot p_i + b_k), w_k \in \mathbb{R}^n, b_k \in \mathbb{R} \tag{15}$$

where $a(p) : \mathbb{R} \mapsto \mathbb{R}$ denotes the sigmoid activation function, $w_k = [w_{k_1}, w_{k_2}, \dots, w_{k_n}]^T$ denotes the weight vector between the input and the k^{th} hidden node, and b_k denotes the threshold of the k^{th} hidden node.

With ‘ H ’ hidden nodes, the SLFN can estimate ‘ S ’ samples with no error, i.e., the cost function $C = \sum_{i=1}^S \|(o_i - q_i)\|_2 = 0$. In other words, there exist (w_k, b_k) and β_k such that

$$q_i = \sum_{k=1}^H \beta_k A(w_k, b_k, p_i), i = 1, 2, \dots, S \tag{16}$$

Equation (16) can be represented compactly as

$$\tilde{H}\beta = Q \tag{17}$$

where

$$\tilde{H} = \begin{bmatrix} A(w_1, b_1, p_1) & \dots & A(w_H, b_H, p_1) \\ \vdots & \ddots & \vdots \\ A(w_1, b_1, p_S) & \dots & A(w_H, b_H, p_S) \end{bmatrix}_{S \times H} \tag{18}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_H^T \end{bmatrix}_{L \times m} \tag{19}$$

$$Q = \begin{bmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_s^T \end{bmatrix}_{S \times m} \tag{20}$$

\bar{H} denotes the hidden layer output matrix of SLFN. Equation (17) is referred to as a linear system [15], and the output weights associated with the model can be found using

$$\beta = \bar{H}^\dagger Q \tag{21}$$

where \bar{H}^\dagger denotes the Moore-Penrose generalized inverse of \bar{H} [18]. Further, for better clarity, the Moore-Penrose generalized inverse is discussed in [16]. Algorithm 1 summarizes the steps involved in the ELM approach.

Algorithm 1 ELM algorithm

Input : Training Pattern: $T^r = (p_i^r, q_i^r)_{i=1}^{N^r}$; Testing Pattern: $T^e = (p_i^e, q_i^e)_{i=1}^{N^e}$; Activation Function: $a(p)$; and Hidden Node Size: H

Output: Output weight vector β

- I Set random values for the hidden node parameters $(w_k, b_k, 1 \leq k \leq H)$.
- II Determine the output matrix \bar{H} using (18).
- III Compute the output weight vector β using (21).

5 Proposed methodology

As opposed to the SI generation techniques discussed in the literature which is based on the spatial domain, the novelty of the proposed methodology, is to estimate the SI frames in the frequency domain with the help of the wavelet coefficients. The wavelet coefficients for a 2-dimensional (2D) signal (e.g. images) can be generated by applying 1D-DWT (see Section 4.1) to the rows and columns of the signal, individually, as reported in [26] (see Fig. 2). At each of the individual levels, the 2D-DWT decomposition results in four sub-bands, namely, low-low (LL), low-high (LH), high-low (HL), and high-high (HH). Among the generated sub-bands, LH(D_k^h), HL(D_k^v), and HH(D_k^d) represent the detail coefficients along

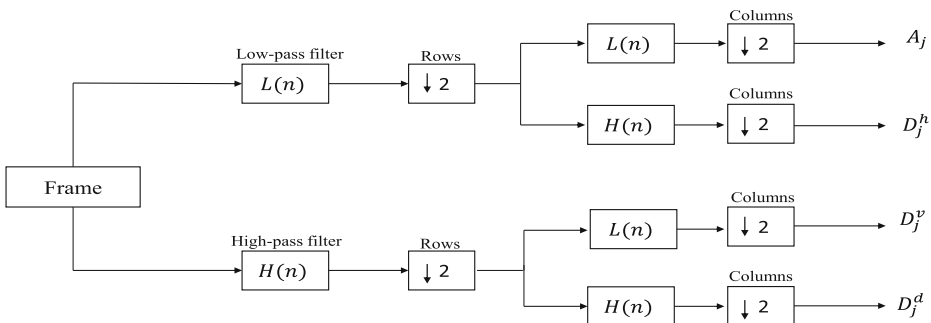


Fig. 2 Schematic Diagram of 2D-DWT for single resolution level

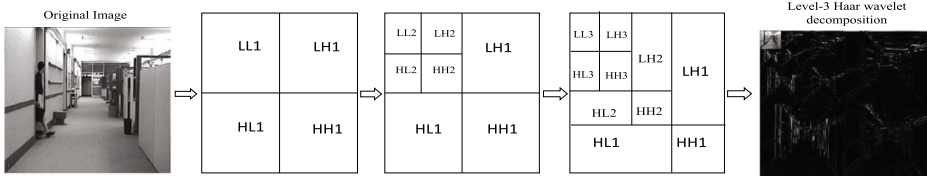


Fig. 3 A frame of the *Hall-Monitor* video sequence and its level-3 wavelet coefficients

the horizontal, vertical, and diagonal directions, respectively. For better clarity, a level-3 wavelet decomposition for a given frame is shown in Fig. 3.

The $LL(A_k)$ represents the approximation coefficient which is further used for higher-level decomposition. The generation of sub-bands at different resolution levels represent a pyramid like structure which is also referred to as the wavelet pyramid [22]. In the present work, the prime motivation behind the use of wavelet pyramid is that it has two major benefits. First, it preserves the detailed coefficients generated at each of the decomposition levels. Secondly, it provides a reduced computational burden. There exist several types of wavelet functions, out of which, Haar wavelet is considered to be one of the elementary and significant wavelet functions used in several applications [12]. Further, it is the simplest type of wavelet, and in discrete domain, it is related to a mathematical operation referred to as the Haar transform. It acts as a prototype for all other wavelet transforms. It has the ability to produce better outcomes in noisy environment and satisfies both the property of conformity and orthogonality. Additionally, it can be utilized to obtain the structural information of the signal.

The objectives of the present work are three-fold. First, a level-3 2D-DWT technique using Haar wavelet is employed to extract the approximation coefficients. Secondly, an ELM neural-network (ELM-NN) is utilized to generate the estimated approximation coefficients for the current WZ frame. The ELM-NN consists of three phases, namely, initialization, training, and estimation. In the initialization phase, the extracted approximation coefficients are used to create the training pattern. In the training phase, the created pattern is used to make the ELM-NN learn the pattern. In the estimation phase, trained ELM-NN is utilized to produce the approximation coefficients for the current WZ frame. Finally, a level-by-level, IDWT technique is used to get back the eventual estimated SI frame in original form (spatial domain). Figure 4 represents the Stanford-based TDWZ codec integrated with the proposed SI generation technique. The working principle of the proposed SI generation block is explained below.

Initially, the switch ‘ $S1$ ’ is closed, and the learning phase of the ELM network is initiated. In the learning phase, depending on the GOP size, the respective neighboring (key and/or formerly reconstructed WZ) frames for the current WZ frame are considered. Upon the neighboring and WZ frames considered, a level-3 DWT is employed, and a set of thirty wavelet coefficients, namely, three approximation and twenty-seven detailed coefficients are obtained. Out of these, only the approximation coefficients are utilized to create the training pattern (input, target). The approximation coefficients of the neighboring, and WZ frames act as the input, and target, respectively. Once the network is trained, switches ‘ $S1$ ’, and ‘ $S2$ ’ is made to open and close, respectively. This initiates the estimation phase of the ELM network. Unlike the training phase, here, the level-3 DWT is applied on the respective neighboring (key and/or formerly reconstructed WZ) frames, only. Similarly, this operation results in twenty wavelet coefficients, namely, two approximation and eighteen detailed coefficients. Next, using both the obtained approximation coefficients and

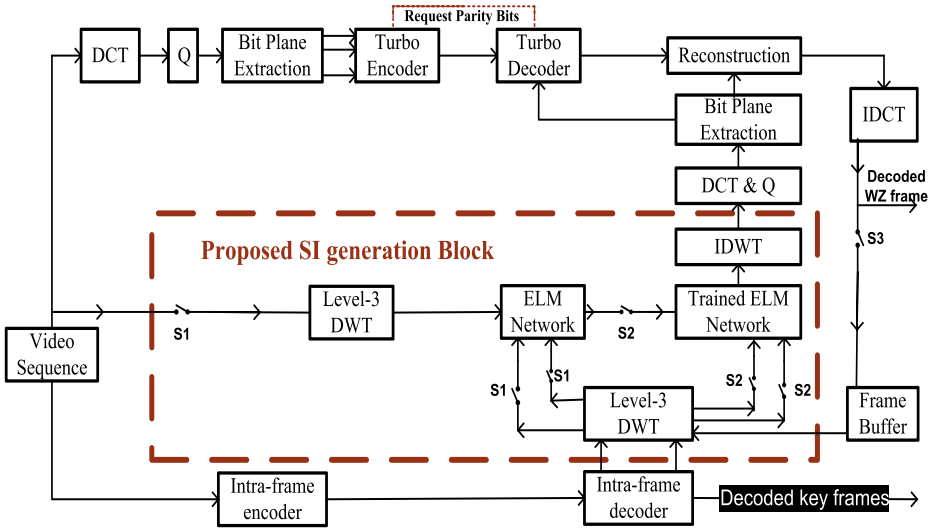


Fig. 4 Stanford-based TDWZ codec integrated with the proposed SI generation technique

the output weight vector ‘ β ’ (refer Algorithm 1), the trained ELM network produces the approximation coefficients (estimated) for the current WZ frame.

Further, upon the estimated approximation coefficients along with the eighteen detailed coefficients, an inverse discrete wavelet transform (IDWT) technique, level-by-level, is employed, to obtain the eventual estimated WZ frame (also known as SI) in the original form (spatial domain). The IDWT technique at any level can be mathematically modeled as

$$A_{j+1} \oplus D_{j+1} = A_j \tag{22}$$

Figure 5 shows the block diagram for the steps involved in the IDWT process. For instance, to obtain the level-2 low-low sub-band (also known as approximation coefficient) for the current WZ frame, a IDWT process is employed considering both the estimated

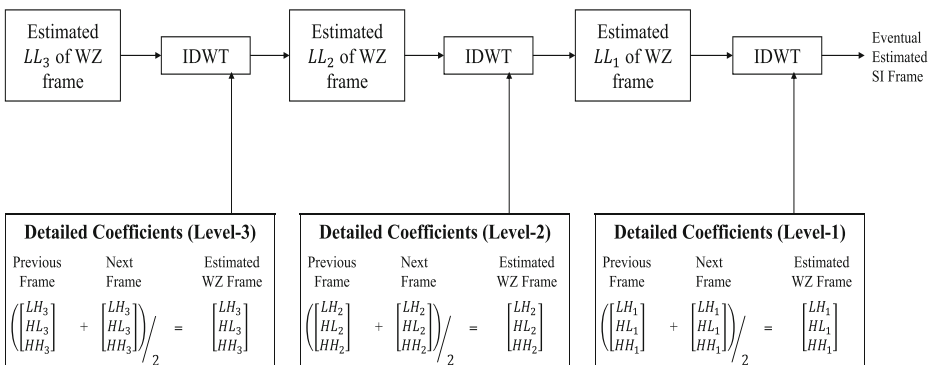


Fig. 5 Block diagram for IDWT

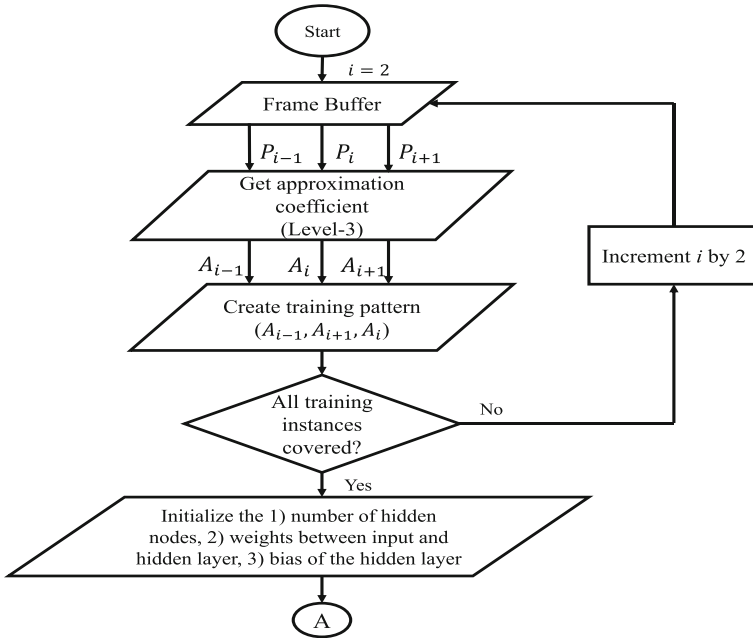


Fig. 6 Flowchart for initialization step

level-3 low-low sub-band of the current WZ frame and the level-3 detailed coefficients (LH, HL, and HL) of the respective neighboring frames, as input. Similar process is repeated until the eventual estimated WZ frame (also known as SI) in the original form (spatial domain) is obtained. Flowcharts shown in Figs. 6, 7 and 8 give a clear insight about the steps (initialization, learning, and estimation) involved in the proposed SI estimation scheme. Additionally, a brief explanation of each of the steps is also presented in Algorithm 2.

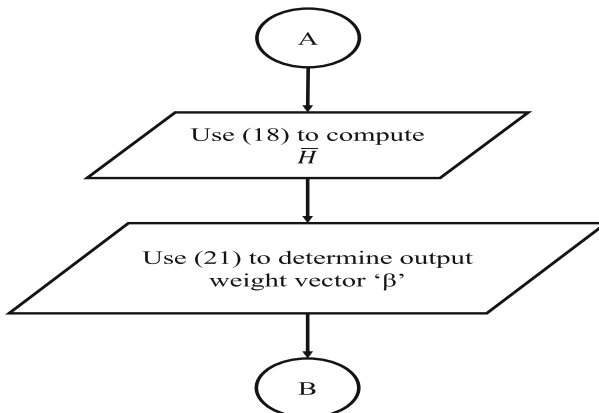


Fig. 7 Flowchart for learning step

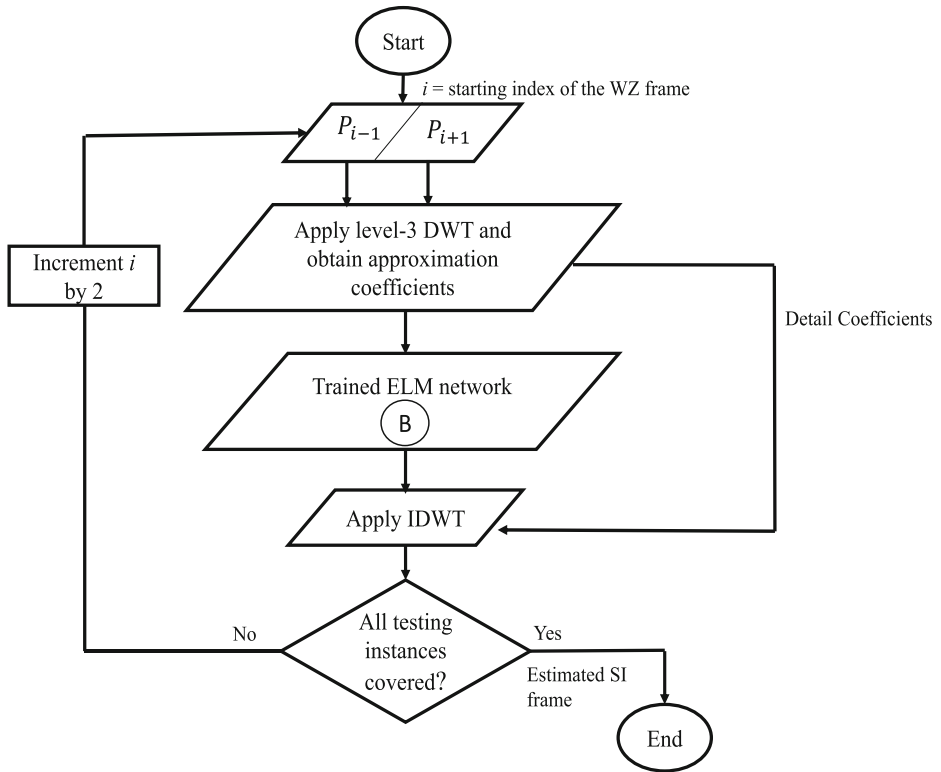


Fig. 8 Flowchart for estimation step

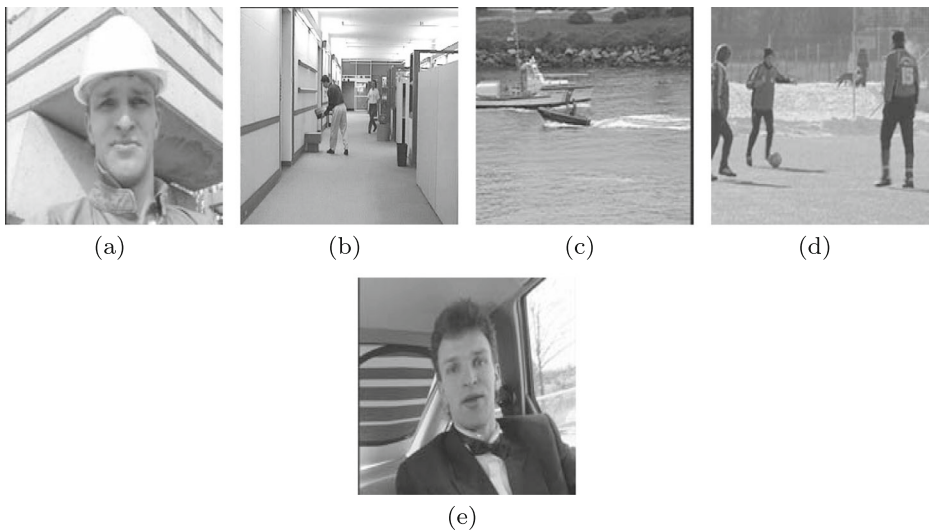


Fig. 9 Prototypical frame of: **a** Foreman, **b** Hall-Monitor, **c** Coastguard, **d** Soccer, and **e** Carphone

6 Discussion and analysis of the results

To appraise the efficaciousness of the proposed SI generation scheme, exhaustive simulations are carried out in MATLAB under certain relevant and specific test environment.

Algorithm 2 Algorithm for proposed SI estimation scheme

I Initialization Phase

- 1) Initialize the total number of frames in the training instance which includes both the key and the WZ frames.
- 2) Set the starting frame index (i) to the initial index of the WZ frame present in the training pattern.
- 3) Depending on the GOP size, select the two immediate previous and next adjacent frames (key-key or key-WZ) for the corresponding WZ frame. For instance, if $GOP = 2$, select $P_{i-1} = i - 1$, $P_i = i$, $P_{i+1} = i + 1$, where P_{i-1} , P_{i+1} represent the key frames, and P_i represents the WZ frame.
- 4) Select the wavelet function and the decomposition level j . In this work, Haar wavelet and $j = 3$ is used.
- 5) Obtain the approximation coefficients (LL_3) (see Fig. 6) followed by the formation of the training pattern (P_{i-1} , P_{i+1} (input); P_i (target)).
- 6) Initialize the ELM network parameters, i.e., total number of hidden nodes, the weights between the input and hidden layer, and the bias associated with the hidden nodes.

II Learning Phase

- 7) Provide the generated training pattern from **Step 5** to the ELM network (**Algorithm 1**).
- 8) Use **(18)** to compute the hidden layer output matrix \vec{H} .
- 9) Use **(21)** to determine the weight vector β between the hidden and output layer.

III Estimation Phase

- 10) From the remaining frames of the video sequence, initialize the starting frame index (i) to the initial index of the WZ frame.
 - 11) Follow **Step 3**, and apply level-3 Haar wavelet decomposition to obtain the approximation and detailed coefficients.
 - 12) Provide the output of **Step 11**, and **Step 9** to the trained ELM network.
 - 13) Obtain the estimated approximation coefficients for the corresponding WZ frame.
 - 14) Perform level-by-level IDWT process using the detailed coefficients of **Step 11** and the output of **Step 13**, to obtain the eventual estimated SI frame.
 - 15) Check if all the frames are covered. If ‘No’, then increment the frame index (i) by 2 and go to **Step 11** or else ‘Exit’.
-

6.1 Test environment

The test environment is illustrated in a briefly described below.

Table 1 Standard test video sequences

Video Sequence	Resolution	Frames per second	Format
<i>Foreman</i>	176 × 144	15	QCIF
<i>Soccer</i>	176 × 144	15	QCIF
<i>Hall-Monitor</i>	176 × 144	15	QCIF
<i>Coastguard</i>	176 × 144	15	QCIF
<i>Carphone</i>	176 × 144	15	QCIF

6.1.1 Video sequences

To carry out exhaustive simulations, some of the standard and widely available video sequences, namely, *Foreman*, *Hall-Monitor*, *Soccer*, *Coastguard*, and *Carphone* are adopted. These sequences constitute diversified motion (low to fast) characteristics, a wide variety of texture contents, and have been used for assessment in various DVC research works [3], [28], [43], [44]. A prototypical frame of each of these video sequences is shown in Fig. 9. Moreover, the properties of each of these sequences are depicted in Table 1.

6.1.2 Group of picture (GOP)

The length of GOP may be 2, 4, or 8 (usually, only 2 in many of the DVC literature)

6.1.3 Rate-distortion parameters

To determine various RD trade-off points, eight distinct quantization matrices (QM) [10], [30] are used. The adopted quantization matrices with different quantization levels are shown in Fig. 10. It illustrates the quantization levels allied with each of the coefficients in a

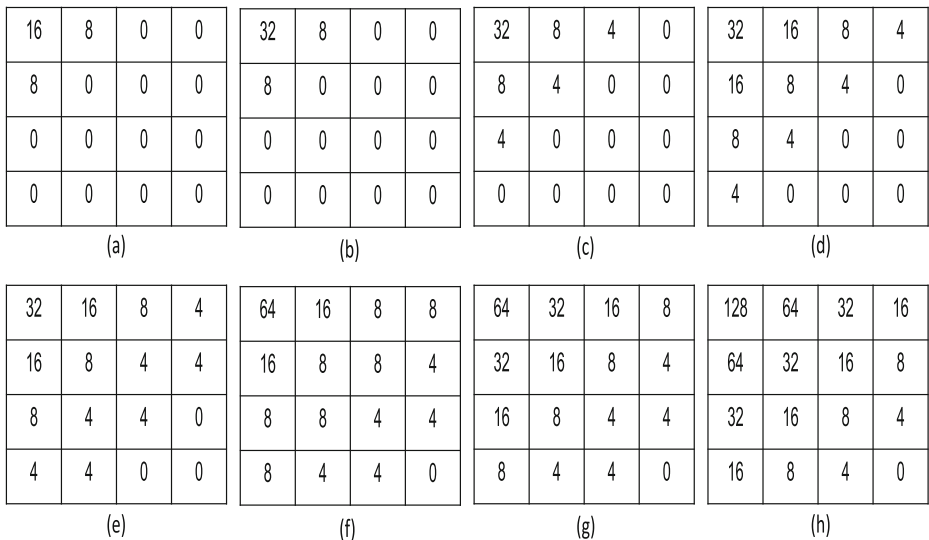


Fig. 10 Eight distinct quantization matrices

DCT band. In Fig. 10, starting with the top-left to bottom-right, depending on the QM used, an improvement in the quality of the decoded frames can be distinguished. However, the bit rate also increases. For each DCT band, the non zero entry in QM depicts the transmission of the parity bits for the corresponding DCT band, whereas the cell with zero denotes non-transmission of the parity bits.

6.1.4 Benchmark schemes

To assess and compare the efficaciousness of the proposed SI estimation technique, the following schemes are adopted as the benchmark. A brief description of these schemes is given below.

H.264/AVC (Intra) This corresponds to a pure intra-frame codec, where the temporal correlation is not being exploited.

H.264/AVC (No motion) Contrary to H.264/AVC (Intra) codec, here, the temporal correlation is exploited. However, the motion estimation between the frames is not adopted. For comparative analysis, this codec has been used in many DVC works ([27]; [5]). In the present work, the result reported in [27] is referred to.

IST-TDWZ [6] It corresponds to a transform-domain-based state-of-the-art DVC architecture and has been widely adopted as a benchmark by many DVC researchers. It has been developed at the Image Group of the Instituto Superior Técnico (IST).

MLP-SI [36] This is a transform-domain-based DVC architecture in which SI is generated using multi-layer perceptron neural-network.

Progressive-DVC [41] It constitutes a transform-domain-based DVC architecture that utilizes the spatial dependency of the video frames to enhance the motion-compensated temporal interpolation (MCTI). In particular, WZ frames are segregated into various spatially correlated groups which are then transmitted progressively to the decoder.

It may be noted that H.264/AVC (Intra) and H.264/AVC (No motion) schemes correspond to the traditional-based video coding solutions, whereas IST-TDWZ, MLP-SI, and Progressive DVC techniques belong to DVC-based video codec framework. Hence, in the present work, IST-TDWZ, MLP-SI, and Progressive-DVC are considered as the benchmark techniques.

6.1.5 Training and testing samples

Generally, in machine learning-based approaches, the input dataset is segregated into two groups, namely, training and testing groups. The present work utilizes the same approach and the video frames are segregated into training and testing samples. Here, in this work, the predefined number of initial frames of each test video sequences is considered as the training sample so as to make the training model to learn all possible motion and texture features of a video sequence. Further, in the testing phase, the remaining frames that are not included in the training phase are considered to validate the performance of the proposed scheme.

6.1.6 Other conditions

In DVC-based video coding solutions, the Y (luminance) feature of the video frames is particularly used to compute the PSNR and RD characteristics. Hence, in the present work, to make the comparison fair, only the luminance feature is considered. Moreover, both odd and even frames are utilized to compute the RD behavior.

6.2 Detailed experimental analysis

To provide a clear insight and better clarity of the performance analysis, the overall simulation is grouped into eight different experiments. Each of the experiments is described below in detail.

6.2.1 Performance analysis of SI estimation with respect to PSNR (in dB)

As aforementioned, the prime objective of the present work is to estimate a better quality of SI using a hybrid approach based on MRA and ELM techniques. To fix the number of hidden units in the hidden layer is one of the essential assignments in an NN architecture, and depends on the application scenarios. However, to have a fair comparison, in the present work, the number of hidden nodes (h_n) is taken to be the same as in MLP-SI scheme ($h_n = 14$).

In this experiment, the quality of the SI frames is assessed in terms of PSNR (in dB) between the estimated SI and the original WZ frame, for the test video sequences considered. A brief comparison of the estimated SI frame quality achieved with different DVC techniques, namely, IST-TDWZ [6], MLP-SI [36], Progressive DVC [41], and the proposed technique, for three distinct GOP sizes (2, 4, and 8) and five distinct video sequences, namely, *Foreman*, *Carphone*, *Coastguard*, *Hall-Monitor*, and *Soccer*, is summarized in Table 2. It is observed that the PSNR values (in dB) with the proposed scheme is notably higher than that of IST-TDWZ, MLP-SI, and Progressive-based DVC schemes. This shows that the proposed scheme can generate quality SI for video sequences with varied resolution and GOP sizes.

Further, Fig. 11a-c show the PSNR (in dB) comparison among the schemes for *Carphone*, *Hall-Monitor*, and *Foreman* video sequences, respectively. It is observed that in a majority number of frames, PSNR (in dB) with the proposed scheme is notably higher than that of MLP-SI, and IST-TDWZ schemes. The results obtained clearly indicates that the proposed SI scheme is able to generate a better quality of SI for video sequences with different resolution.

6.2.2 Analysis of perceptive measure of SI with respect to SSIM

This experiment is to obtain the perceptive measure in terms of SSIM [14] for the proposed and benchmark schemes. SSIM computes the structural similarity between two images and determines the degradation in the picture quality caused by some processing techniques like data compression or transmission. Higher the similarity between the images, closer is the SSIM value to unity. Table 3 depicts the average SSIM obtained with the proposed and benchmark schemes for the test video sequences considered. From the table, it is noticed that the proposed method has the highest average SSIM value.

Table 2 Average PSNR values (in dB) of the estimated SI frames for GOP sizes 2, 4, and 8

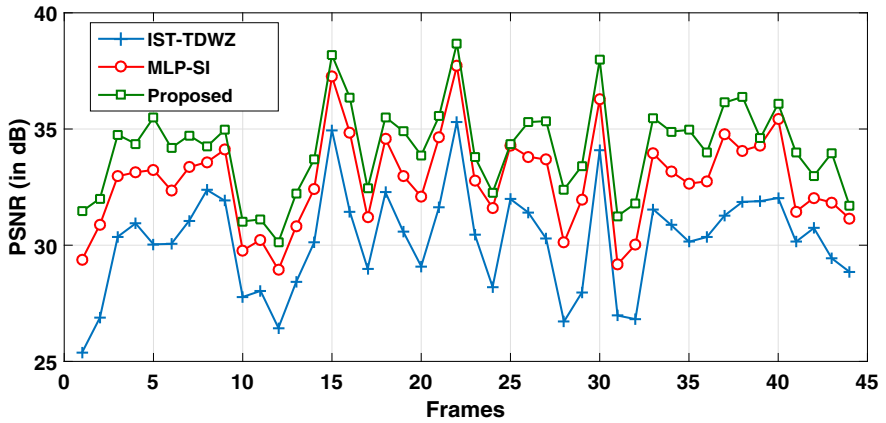
	IST-TDWZ	Progressive DVC (2 groups)	Progressive DVC (3 groups)	Progressive DVC (4 groups)	MLP-SI	Proposed
GOP= 2						
<i>Foreman</i>	29.55	31.39	32.05	32.12	32.52	33.94
<i>Carphone</i>	30.18	30.90	31.46	31.53	32.77	34.16
<i>Coastguard</i>	31.66	32.32	31.71	31.73	31.41	32.05
<i>Hall-Monitor</i>	36.14	36.46	36.37	36.40	36.52	37.24
<i>Soccer</i>	21.69	24.59	25.48	25.54	26.05	27.14
GOP=4						
<i>Foreman</i>	27.72	30.06	30.85	30.95	31.01	32.43
<i>Carphone</i>	28.72	30.45	31.05	31.13	31.89	33.54
<i>Coastguard</i>	30.14	30.93	30.52	30.56	30.85	31.49
<i>Hall-Monitor</i>	35.01	35.42	35.19	35.41	35.43	36.15
<i>Soccer</i>	20.50	23.41	24.30	24.35	24.86	25.96
GOP=8						
<i>Foreman</i>	26.37	28.67	29.50	29.98	30.13	31.55
<i>Carphone</i>	28.07	29.95	30.62	30.71	31.46	32.98
<i>Coastguard</i>	29.04	30.00	28.90	29.73	29.63	30.27
<i>Hall-Monitor</i>	34.23	34.73	34.74	34.48	34.72	35.44
<i>Soccer</i>	19.90	23.06	23.15	23.78	23.90	24.70

The bold font signifies the improvement achieved with our proposed method over other schemes

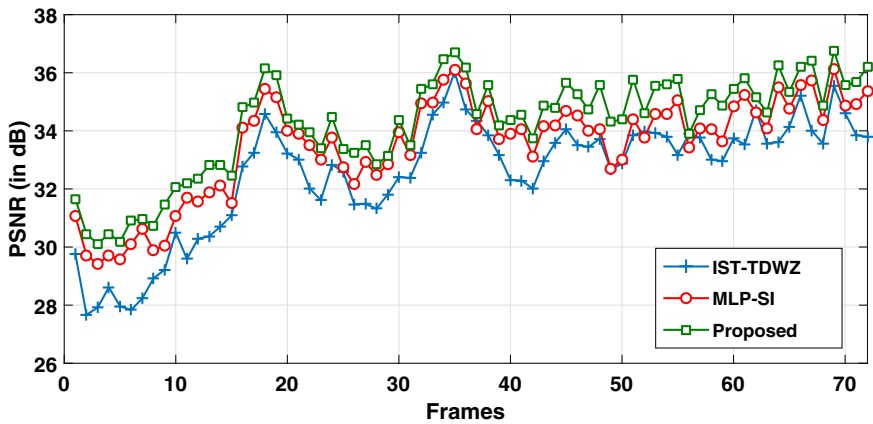
Further, for visual (subjective) analysis, the original 108th frame of *Carphone* and the corresponding estimated SI frames with IST-TDWZ, MLP-SI, and the proposed technique are shown in Fig. 12a-d, respectively. Figure 13a-c represent the binary image (difference image) of the IST-TDWZ, MLP-SI, and the proposed technique with respect to the original 108th frame, respectively. Similarly, Fig. 14a-d represent the original 84th frame of *Hall-Monitor* and the corresponding estimated SI frames with IST-TDWZ, MLP-SI, and the proposed technique, respectively. The difference between the original & IST-TDWZ, original & MLP-SI, and original & the proposed scheme, is shown in Fig. 15a-c, respectively. It may be noticed that the estimated SI frame with the proposed scheme is more similar to the original frame as compared to that in IST-TDWZ, and MLP-SI techniques. Moreover, similar findings have also been observed with other video sequences as well, but are not included here in view of space limitations.

6.2.3 Assessment of additional parity requests per SI frame

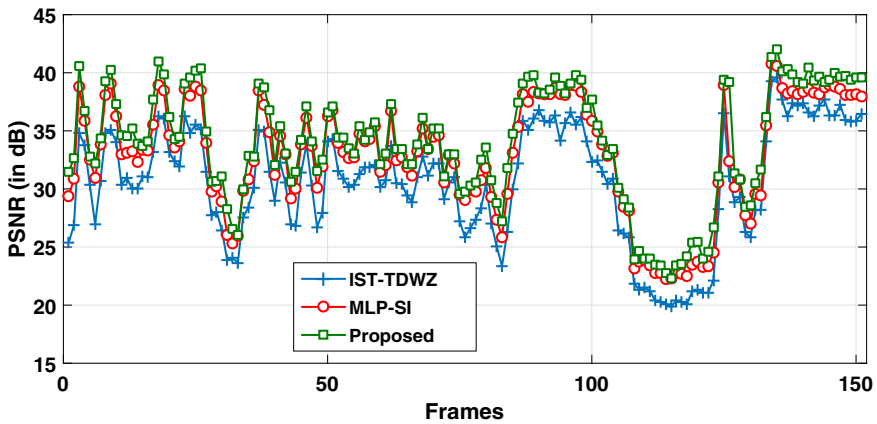
The efficacy of the decoder firmly depends on the number of parity request. Hence, it becomes utmost necessary to evaluate the number of additional parity bits requested by the decoder to correct the error that exists between the original WZ and the estimated SI frame. Figure 16a and b illustrate the additional requests initiated per SI frame with the IST-TDWZ, MLP-SI, and the proposed schemes, for the *Foreman*, and *Coastguard* sequences,



(a)



(b)



(c)

Fig. 11 PSNR (in dB) plot per estimated SI frame of: **a** *Carphone*, **b** *Hall-Monitor*, and **c** *Foreman*

Table 3 Average SSIM Values for different video sequences

Video Sequence	SSIM Values		
	IST-TDWZ	MLP-SI	Proposed
<i>Coastguard</i>	0.8744	0.9004	0.9250
<i>Foreman</i>	0.8361	0.8522	0.9415
<i>Hall-Monitor</i>	0.9324	0.9457	0.9582
<i>Carphone</i>	0.8856	0.9083	0.9154

The bold font signifies the improvement achieved with our proposed method over other schemes



Fig. 12 108th frame of *Carphone*: **a** Original, **b** IST-TDWZ (SSIM = 0.9375), **c** MLP-SI (SSIM = 0.9573), and **d** Proposed (SSIM = 0.9801)



Fig. 13 Difference frame of *Carphone*: **a** Original & IST-TDWZ, **b** Original & MLP-SI, **c** Original & Proposed

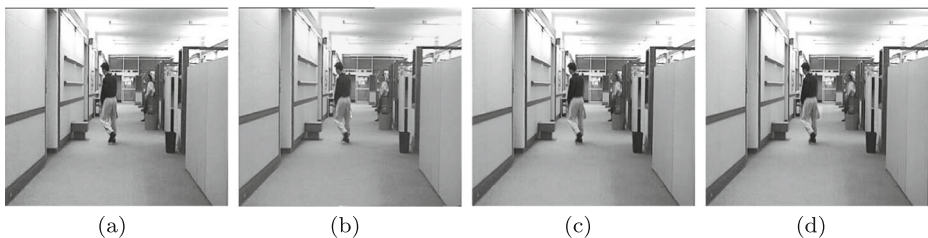


Fig. 14 84th frame of *Hall-Monitor*: **a** Original, **b** IST-TDWZ (SSIM = 0.9381), **c** MLP-SI (SSIM = 0.9535), and **d** Proposed (SSIM = 0.9767)

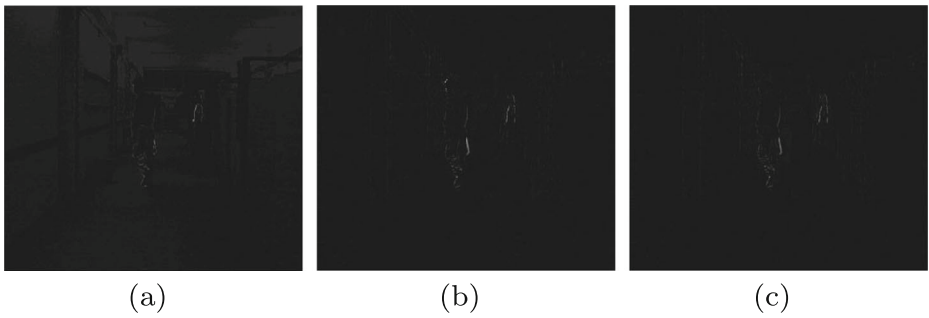
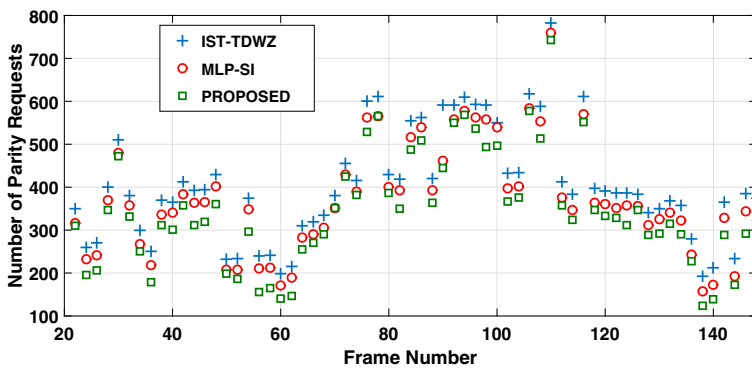


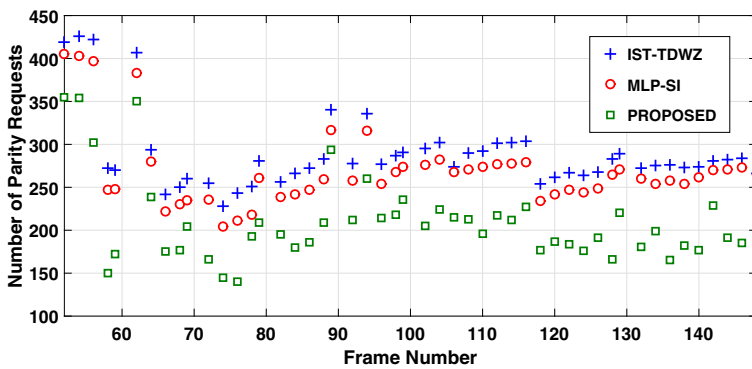
Fig. 15 Difference frame of *Hall-Monitor*: **a** Original & IST-TDWZ, **b** Original & MLP-SI, **c** Original & Proposed

respectively. To perform this experiment, a noiseless channel is adopted for transmission of parity bits.

From the experimental result, it is noticed that a maximum of 742 requests is made with the proposed scheme for the 110th frame of the *Foreman* sequence, whereas a maximum number of 760, and 782 requests are made with the MLP-SI, and IST-TDWZ schemes,



(a)



(b)

Fig. 16 Number of Parity Requests (at 15 fps) for: **a** *Foreman*, and **b** *Coastguard*

Table 4 Average PSNR values (in dB) obtained with the Proposed scheme over IST-TDWZ, Progressive-DVC (2 groups), Progressive-DVC (3 groups), Progressive-DVC (4 groups), and MLP-SI scheme for *Foreman* sequence with different bit-rates and GOP size 2, 4, and 8

Bit rate (in kbps)	IST-TDWZ	Progressive DVC (2 groups)	Progressive DVC (3 groups)	Progressive DVC (4 groups)	MLP-SI	Proposed
GOP = 2						
<i>0-100</i>	29.40	29.50	29.60	29.60	29.63	29.70
<i>101-200</i>	32.55	32.95	33.15	33.15	33.20	33.50
<i>201-300</i>	35.00	35.30	35.60	35.70	35.70	36.22
<i>301-400</i>	37.10	37.45	37.65	37.75	37.70	38.57
<i>401-500</i>	38.90	39.30	39.60	39.60	39.50	40.02
GOP=4						
<i>0-100</i>	28.50	28.90	29.10	29.17	29.15	29.25
<i>101-200</i>	31.00	31.50	32.50	32.57	32.29	32.99
<i>201-300</i>	33.80	34.80	35.40	35.47	35.68	36.19
<i>301-400</i>	36.10	36.75	38.10	38.30	38.38	39.20
<i>401-500</i>	38.00	38.50	39.40	39.50	39.28	40.20
GOP=8						
<i>0-100</i>	27.40	27.70	28.25	28.25	28.25	28.78
<i>101-200</i>	30.20	30.85	31.43	31.50	31.81	32.23
<i>201-300</i>	32.20	33.00	34.70	34.80	34.60	35.52
<i>301-400</i>	34.25	36.10	36.70	36.80	37.63	37.70
<i>401-500</i>	36.50	38.00	38.55	38.55	38.80	39.80

The bold font signifies the improvement achieved with our proposed method over other schemes

respectively. In general, similar improvements are observed with the proposed SI generation method for other video sequences as well.

6.2.4 Evaluation of comprehensive RD behavior

To evaluate the codec efficiency, this experiment assesses the overall RD performance of the proposed DVC codec. To compute the RD performance, only the *Y* (Luminance) component of the video frames is used. Additionally, both WZ and key frames are considered. The RD plots for *Foreman*, *Soccer*, *Hall-Monitor*, and *Coastguard* at 15 fps with $GOP = 2$ are shown in Fig. 17a-d, respectively. Similarly, the RD plots for $GOP = 4$, and $GOP = 8$, for the same video sequence, at 15 fps, are represented in Figs. 18a-d, and 19a-d, respectively.

Further, a brief comparison of the average PSNR values (in dB) obtained with the proposed approach over that of the IST-TDWZ, Progressive-DVC, and MLP-SI schemes at different bit rates (in kbps) and GOPs (2, 4 and 8), for the *Foreman*, *Soccer*, *Hall-Monitor*, and *Coastguard* sequences are presented in Tables 4, 5, 6 and 7, respectively. From an analysis of the results obtained, the following observations are made.

- (a) From Table 4, it is observed that for the *Foreman* sequence with $GOP = 2$, the proposed DVC codec has a maximum average PSNR gain of 1.47 dB and 0.87 dB over

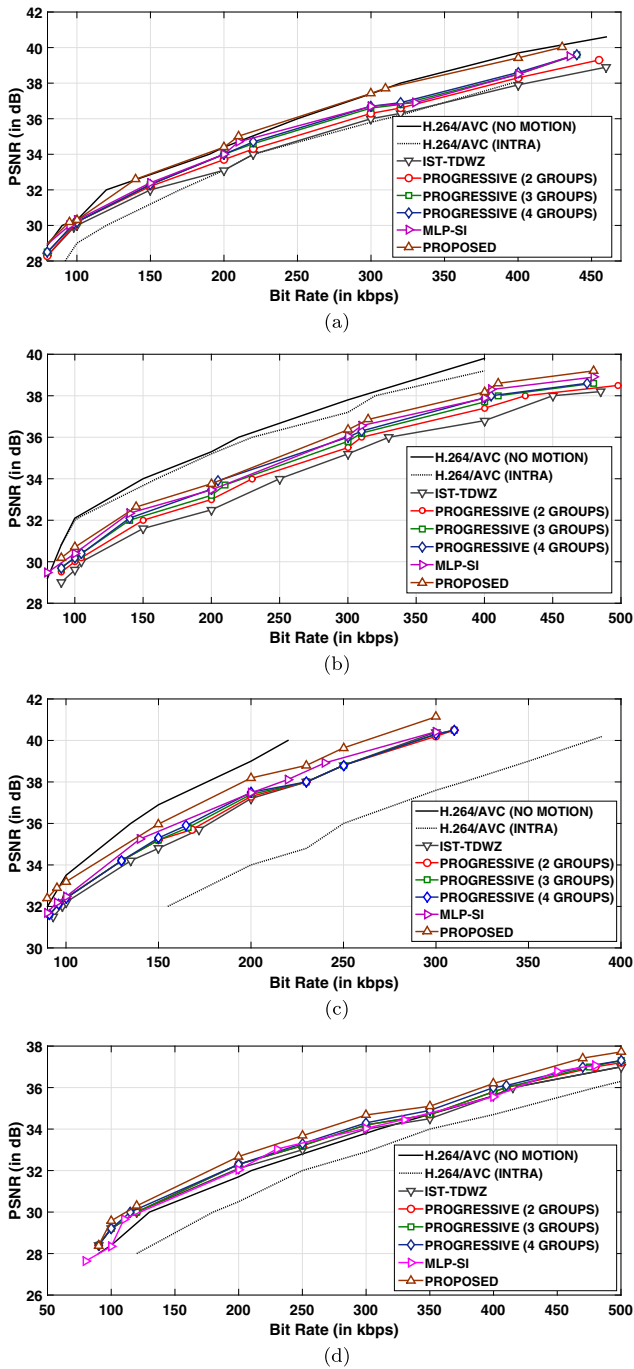


Fig. 17 Rate Distortion Plot with $GOP = 2$ at 15 fps for a Foreman, b Soccer, c Hall-Monitor, and d Coastguard

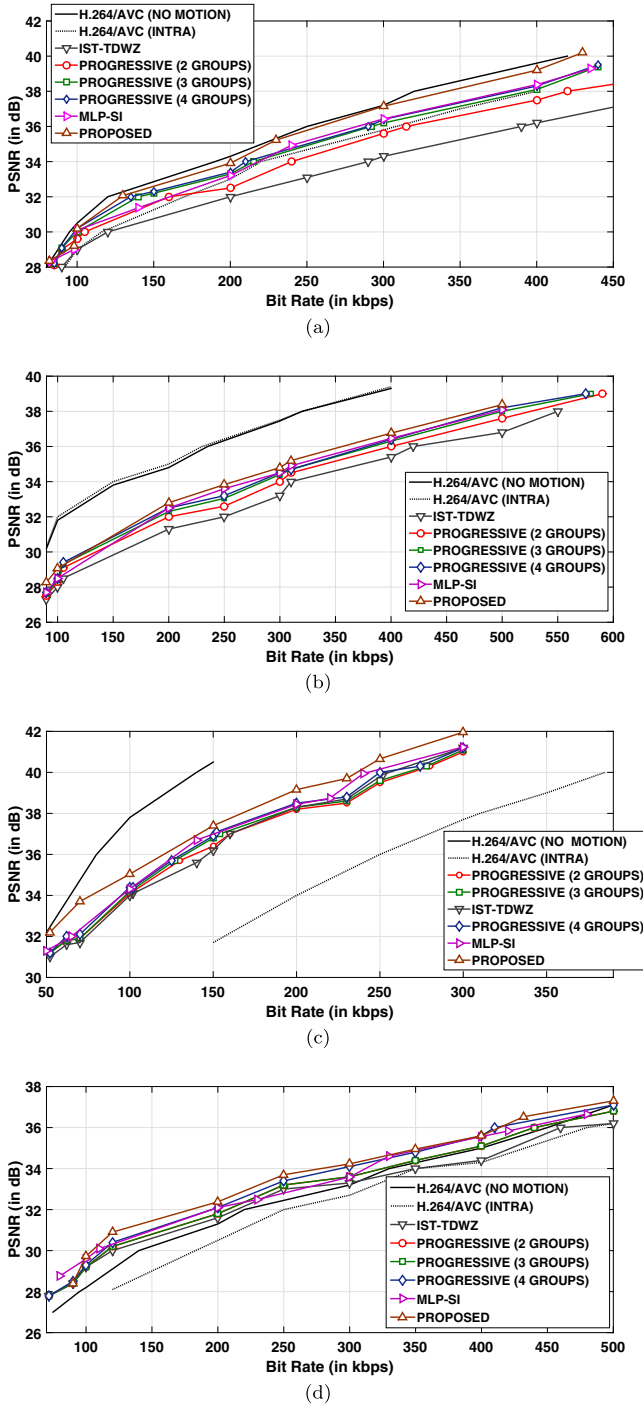


Fig. 18 Rate Distortion Plot with $GOP = 4$ at 15 fps for: **a** Foreman, **b** Soccer, **c** Hall-Monitor, and **d** Coastguard

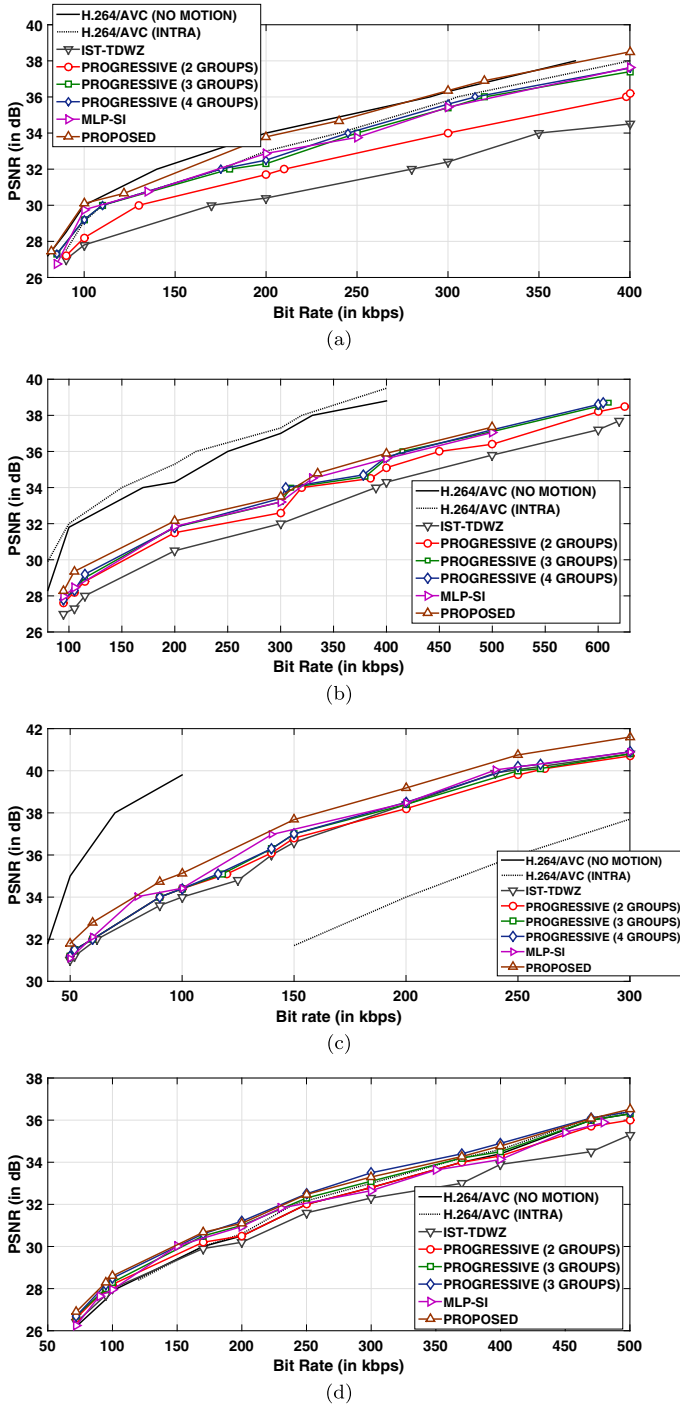


Fig. 19 Rate Distortion Plot with $GOP = 8$ at 15 fps for: **a** Foreman, **b** Soccer, **c** Hall-Monitor, and **d** Coastguard.

Table 5 Average PSNR values (in dB) obtained with the Proposed scheme over IST-TDWZ, Progressive-DVC (2 groups), Progressive-DVC (3 groups), Progressive-DVC (4 groups), and MLP-SI scheme for *Soccer* sequence with different bit-rates and GOP size 2, 4, and 8

Bit rate (in kbps)	IST-TDWZ	Progressive DVC (2 groups)	Progressive DVC (3 groups)	Progressive DVC (4 groups)	MLP-SI	Proposed
GOP = 2						
0-100	29.30	29.75	29.95	29.95	29.94	30.44
101-200	31.37	31.73	31.87	32.00	32.90	33.20
201-300	34.60	34.75	34.75	34.95	36.31	36.36
301-400	36.40	36.70	36.95	37.10	37.88	37.52
401-500	38.10	38.25	38.30	38.30	38.60	38.90
GOP=4						
0-100	27.65	27.90	28.00	28.10	28.12	28.67
101-200	29.90	30.55	30.80	30.95	32.50	32.80
201-300	32.60	33.30	33.80	34.05	34.05	34.32
301-400	34.70	35.25	35.50	35.55	35.68	35.98
401-500	36.40	37.60	38.00	38.20	38.08	38.38
GOP=8						
0-100	27.00	27.60	27.80	27.80	27.97	28.27
101-200	28.60	29.50	29.70	29.77	30.15	30.75
201-300	32.50	32.60	33.20	33.40	33.20	33.50
301-400	34.15	34.53	34.73	34.80	35.05	35.35
401-500	35.80	36.20	36.55	37.20	37.05	37.35

The bold font signifies the improvement achieved with our proposed method over other schemes

that of IST-TDWZ, and MLP-SI schemes, respectively. Similarly, with $GOP = 8$, a maximum average gain of 3.32 dB, 1.25 dB, and 1.00 dB is observed over that of IST-TDWZ, Progressive-DVC (3 Groups), and MLP-SI schemes, respectively.

- (b) From Table 5, it is noticed that for the *Soccer* sequence with $GOP = 2$, the proposed framework has a maximum average PSNR gain of 1.83 dB and 1.41 dB over that of IST-TDWZ, and Progressive-DVC (4 Groups) schemes, respectively. Similarly, with $GOP = 4$, a maximum average gain of 2.90 dB and 2.00 dB is observed over that of IST-TDWZ, and Progressive-DVC (3 Groups), respectively.
- (c) Similarly, it is seen that for the *Coastguard* sequence with $GOP = 2$ (see the Table 7), a maximum average PSNR gain of 1.07 dB and 0.99 dB is obtained with the proposed technique over that of IST-TDWZ, and MLP-SI scheme, respectively. Moreover, with $GOP = 8$, a maximum gain of 0.64 dB and 0.26 dB is noticed over that of IST-TDWZ, and Progressive-DVC (2 Groups) schemes, respectively.

6.2.5 Evaluation of decoding time

Generally, in WZ video codec, the decoder complexity is significantly higher than that of the encoder. Hence, in order to measure the decoder complexity, the average time taken

Table 6 Average PSNR values (in dB) obtained with the Proposed scheme over IST-TDWZ, Progressive-DVC (2 groups), Progressive-DVC (3 groups), Progressive-DVC (4 groups), and MLP-SI scheme for *Hall-Monitor* sequence with different bit-rates and GOP size 2, 4, and 8

Bit rate (in kbps)	IST-TDWZ	Progressive DVC (2 groups)	Progressive DVC (3 groups)	Progressive DVC (4 groups)	MLP-SI	Proposed
GOP = 2						
0-100	31.90	32.03	32.03	32.03	32.13	32.8
101-200	35.48	35.60	35.65	35.73	36.36	37.08
201-300	39.07	39.06	39.03	39.03	39.15	39.86
GOP=4						
0-100	32.08	32.25	32.28	32.43	32.54	33.65
101-200	36.24	36.30	36.42	36.52	37.57	38.28
201-300	40.03	39.83	39.90	40.08	39.97	40.77
GOP=8						
0-100	32.36	32.62	32.62	32.63	32.91	33.62
101-200	36.45	36.55	36.40	36.73	37.72	38.43
201-300	40.35	40.20	40.30	40.47	40.48	41.18

The bold font signifies the improvement achieved with our proposed method over other schemes

(in seconds) by the turbo decoder (TD) with the proposed and other benchmark schemes for different quantization matrices, namely, Q_1 , Q_4 , and Q_8 [6] for the *Foreman* sequence is reported in Table 8. Similarly, Table 9 depicts the average time taken (in seconds) by TD for the *Soccer* sequence. From the tabular data, the following observations can be inferred.

- From Table 8, it is observed that a maximum time reduction of 29.88%, and 12.82% is achieved with the proposed scheme over that of IST-TDWZ for $Q_i = 8$ with $GOP = 2$, and $GOP = 8$, respectively. Similarly, a maximum of 47.02%, and 36% is achieved with the proposed scheme over that of Progressive-DVC (4 Groups) for $Q_i = 1$ with $GOP = 2$, and $GOP = 8$, respectively.
- From Table 9, it is observed that a maximum time reduction of 33.95%, and 16.40% is achieved with the proposed scheme over that of IST-TDWZ for $Q_i = 8$ with $GOP = 2$, and $GOP = 8$, respectively. Similarly, a maximum of 35.95%, and 28.21% is achieved with the proposed scheme over that of Progressive-DVC (4 Groups) for $Q_i = 1$ with $GOP = 2$, and $GOP = 8$, respectively.

From these observations, it can be concluded that the proposed scheme requires considerably less decoding time as compared to that of the other competent schemes. Further, similar findings are also observed for other video sequences with different GOPs and quantization matrices as well.

6.2.6 Statistical analysis

Statistical analysis is a scientific method used to make judgments with a measurable confidence. Analysis of variance (ANOVA) is such a statistical approach used to validate whether

Table 7 Average PSNR values (in dB) obtained with the Proposed scheme over IST-TDWZ, Progressive-DVC (2 groups), Progressive-DVC (3 groups), Progressive-DVC (4 groups), and MLP-SI scheme for *Coastguard* sequence with different bit-rates and GOP size 2, 4, and 8

Bit rate (in kbps)	IST-TDWZ	Progressive DVC (2 groups)	Progressive DVC (3 groups)	Progressive DVC (4 groups)	MLP-SI	Proposed
GOP = 2						
<i>0-100</i>	28.80	28.80	28.80	28.80	27.99	28.98
<i>101-200</i>	31.05	31.15	31.15	31.15	30.85	31.49
<i>201-300</i>	33.50	33.70	33.70	33.80	33.54	34.18
<i>301-400</i>	35.05	35.25	35.25	35.45	35.01	35.65
<i>401-500</i>	36.50	36.73	36.77	36.80	36.93	37.57
GOP=4						
<i>0-100</i>	28.47	28.47	28.50	28.53	29.18	29.07
<i>101-200</i>	30.00	31.00	31.15	31.25	31.10	31.64
<i>201-300</i>	33.15	33.40	33.60	33.75	33.04	33.97
<i>301-400</i>	34.20	34.75	34.25	35.20	35.09	35.28
<i>401-500</i>	36.10	36.80	36.50	36.55	36.25	36.92
GOP=8						
<i>0-100</i>	27.60	27.67	27.80	27.37	27.28	27.93
<i>101-200</i>	30.25	30.75	30.90	30.05	30.49	30.89
<i>201-300</i>	32.40	32.70	33.00	31.95	32.24	32.86
<i>301-400</i>	34.15	34.35	34.65	33.45	33.88	34.52
<i>401-500</i>	35.85	36.15	36.25	34.90	35.66	36.30

The bold font signifies the improvement achieved with our proposed method over other schemes

the means of several groups are all equal. Typically, in ANOVA, a null and alternative hypothesis are defined. The null hypothesis states that there is no significant difference among the groups against the alternative hypothesis that there is a significant difference. The rejection or acceptance of the null hypothesis firmly relies on the resulting p-value of the ANOVA test. If $p \leq 0.05$ (considered significance level of 5%), the null hypothesis fails to be accepted. Further, for a better understanding, see the detailed explanation of ANOVA presented in [34].

Table 8 Comparative Analysis of Decoding Time (in secs) for *Foreman* sequence

Techniques	IST-TDWZ		Progressive DVC (4 Groups)		Proposed	
	GOP=2	GOP=8	GOP=2	GOP=8	GOP=2	GOP=8
1	393	919	621	1336	329	855
4	885	1964	1017	2202	850	1929
8	2912	6786	2898	5992	2042	5916

The bold font signifies the improvement achieved with our proposed method over other schemes

Table 9 Comparative Analysis of Decoding Time (in secs) for *Soccer* sequence

Techniques	IST-TDWZ		Progressive DVC (4 Groups)		Proposed	
	GOP=2	GOP=8	GOP=2	GOP=8	GOP=2	GOP=8
Qi						
1	545	1036	751	1354	481	972
4	1050	2040	1130	2150	925	1915
8	3411	7063	3163	5468	2253	5905

The bold font signifies the improvement achieved with our proposed method over other schemes

In the present work, ANOVA is used to validate that the proposed method produces statistically significant enhancement as compared to the benchmark schemes with respect to different parameters like PSNR (in dB), SSIM, and so on. For instance, the detailed analysis of the ANOVA test with respect to PSNR (in dB) for the *Hall-Monitor* sequence are reported in Table 10. It is noticed that the p-values obtained ($< .0001$ for *Hall-Monitor*) is considerably less than the set significance level of 5%. Similarly, the analysis with respect to SSIM for the *Carphone* sequence is shown in Table 11. The p-value of 0.0001 obtained is less than 5% significance level. Moreover, similar findings have been observed for other video sequences with different parameters as well. Hence, in general, it can be validated that the proposed technique produces statistically significant improvement as compared to the benchmark schemes.

Table 10 ANOVA test with respect to PSNR (in dB) for *Hall-Monitor*

Data Summary				
	Techniques			
	IST-TDWZ	MLP-SI	Proposed	Total
Sample Size	72	72	72	216
$\sum X$	2343.9947	2413.8077	2465.4001	7223.2025
Mean	32.5555	33.5251	34.2417	33.4408
$\sum X^2$	76595.317	81140.2266	84633.7102	242369.2538
Variance	4.0203	3.0183	3.0572	3.8136
Std. Deviation	2.0051	1.7373	1.7485	1.9528
Std. Error	0.2363	0.2047	0.2061	0.1329

Standard Weighted-Means Analysis				
ANOVA Summary		Independent Techniques (k=3)		
Source	SS	df	MS	
Treatments	103.1246	2	51.5623	F = 15.32
(Between-Groups)				
Within-Groups	716.8017	213	3.3653	p-value < 0.0001
Total	819.9263	215		

The bold font signifies the improvement achieved with our proposed method over other schemes

Table 11 ANOVA test with respect to SSIM for *Carphone*

Data Summary				
	Techniques			
	IST-TDWZ	MLP-SI	Proposed	Total
Sample Size	180	180	180	540
$\sum X$	159.415972	163.500716	164.768413	487.685102
Mean	0.885644	0.908337	0.91538	0.903121
$\sum X^2$	141.941608	149.203226	151.329241	442.474075
Variance	0.004222	0.003852	0.002813	0.003777
Std. Deviation	0.064978	0.062061	0.053037	0.061455
Std. Error	0.004843	0.004626	0.003953	0.002645
Standard Weighted-Means Analysis				
ANOVA Summary		Independent Techniques (k=3)		
Source	SS	df	MS	
Treatments	0.086927	2	0.043464	F = 11.98
(Between-Groups)				
Within-Groups	1.948706	537	0.003629	p-value < 0.0001
Total	2.035633	539		

The bold font signifies the improvement achieved with our proposed method over other schemes

7 Closing remarks

In this study, a hybrid approach utilizing the principles of discrete wavelet transform (DWT) and extreme learning machine (ELM) is proposed to estimate the side information (SI) in a distributed video coding (DVC) framework. The proposed scheme estimates the SI for the current WZ frame using two adjacent, neighboring, and formerly decoded key-key or key-WZ frames, as input. Initially, a level-3 Haar wavelet transform (HWT) is applied on the input frames to extract the low-low (LL_3) approximation coefficients. Similarly, a level-3 HWT is also applied on to the current WZ frame to obtain the (LL_3) approximation coefficients. Using the (LL_3) approximation coefficients obtained of both key and current WZ frames, the training pattern (input, target) is created. Next, the training patterns so created are used to train the ELM network. Once the network is trained, it is capable of generating the LL_3 coefficients for the estimated SI frame. Using the generated LL_3 coefficients and the previously retained detailed coefficients, an IDWT process is employed, level-by-level, to obtain the eventual estimated SI frame for the rest of the incoming WZ frames of a video sequence, in a real-time scenario. To exemplify the efficacy of the proposed technique, it is integrated over the Stanford-based transform-domain DVC (TDWZ) codec.

Comparisons have been made with respect to the existing contemporary video codecs. From the exhaustive simulations and analysis of the results obtained, it has been observed that the proposed SI generation scheme results in an improved in terms of both qualitative and quantitative measures. Additionally, to validate the inferred observations, a statistical test, namely, analysis of variance (ANOVA), has been utilized. From the ANOVA test, considering a significance level of 5%, it has been noticed that the proposed method and the other benchmark techniques are significantly different from one another. Further, considering the experimental results as well as the statistical (ANOVA test), it can be concluded

that the proposed SI estimation scheme is capable of achieving a significant enhancement in performance over that of the benchmark techniques. Moreover, it has also been shown that the proposed scheme minimizes the estimation error between the generated SI and the corresponding WZ frames.

The orthogonal wavelets represent the image feature information along different directions, namely, the horizontal, vertical, and diagonal directions. Therefore, it may not be possible to uphold sufficient information using these directions. Hence, advanced transformation techniques like curvelet transform may be exploited. Curvelet transforms are capable of analyzing images at various angles, scales, and locations. Similarly, another transformation technique, namely, contourlet transform shows a greater degree of directionality and anisotropy. It has multi-scale and time-frequency localization property which overcomes the drawback of wavelets. Further, advanced machine learning techniques, namely, convolutional neural network, recurrent neural network, extreme learning variants, and so on can be exploited to generate a better quality of SI. Furthermore, at present, some investigations have been made to develop an efficient parallel framework for intra-codec framework like high efficiency video coding (HEVC) [49, 50]. However, parallelization in other modules of HEVC are still some of the research direction to be investigated.

References

1. Aaron A, Rane SD, Setton E, Girod B et al (2004) Transform-domain wyner-ziv codec for video. In: Proceedings of SPIE, vol 5308, pp 520–528
2. Abou-Elailah A, Dufaux F, Farah J, Cagnazzo M, Pesquet-Popescu B (2013) Fusion of global and local motion estimation for distributed video coding. *IEEE Trans Circuits Syst Video Technol* 23(1):158–172
3. Artigas X, Ascenso J, Dalai M, Klomp S, Kubasov D, Oualet M (2007) The discover codec: architecture, techniques and evaluation. In: Picture Coding Symposium (PCS' 07), MMSPL-CONF-2009-014
4. Ascenso J, Brites C, Pereira F (2006) Content adaptive wyner-ziv video coding driven by motion activity. In: Image processing, 2006 IEEE International Conference on, IEEE, pp605–608
5. Ascenso J, Brites C, Pereira F (2010) A flexible side information generation framework for distributed video coding. *Multimedia Tools and Applications* 48(3):381–409
6. Brites C, Ascenso J, Pedro JQ, Pereira F (2008) Evaluating a feedback channel based transform domain wyner-ziv video codec. *Signal Process Image Commun* 23(4):269–297
7. Ciuti G, Mencias A, Dario P (2011) Capsule endoscopy: from current achievements to open challenges. *IEEE Rev Biomed Eng* 4:59–72
8. Dash B, Rup S, Mohapatra A, Majhi B, Swamy M (2017) Decoder driven side information generation using ensemble of mlp networks for distributed video coding. *Multimedia Tools and Applications* pp1–30
9. Deligiannis N, Verbist F, Slowack J, Rvd Walle, Schelkens P, Munteanu A (2014) Progressively refined wyner-ziv video coding for visual sensors. *ACM Transactions on Sensor Networks (TOSN)* 10(2):21
10. DISCOVER-Project ((accessed May 11, 2017)) Discover project page. <http://www.img.lx.it.pt/discover/home.html>
11. Dufaux F, Gao W, Tubaro S, Vetro A (2010) Distributed video coding: trends and perspectives. *EURASIP Journal on Image and Video Processing* 2009(1):508,167
12. El-Dahshan ESA, Hosny T, Salem ABM (2010) Hybrid intelligent techniques for mri brain images classification. *Digital Signal Process* 20(2):433–441
13. Girod B, Aaron AM, Rane S, Rebollo-Monedero D (2005) Distributed video coding. *Proc IEEE* 93(1):71–83
14. Gurav P, Patil G (2016) Full-reference video quality assessment using structural similarity (SSIM) index. *J Electr Commun Sys* 1(2)
15. Huang GB (2003) Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Trans Neural Netw* 14(2):274–281
16. Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. In: Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, IEEE, vol 2, pp 985–990

17. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. *Neurocomputing* 70(1):489–501
18. Huang GB, Wang DH, Lan Y (2011) Extreme learning machines: a survey. *Int J Mach Learn Cybern* 2(2):107–122
19. Huang X, Rakêt LL, Van Luong H, Nielsen M, Lauze F et al (2011) Multi-hypothesis transform domain wyner-ziv video coding including optical flow. In: *Multimedia Signal Processing (MMSP), 2011 IEEE 13th International Workshop on*, IEEE, pp 1–6
20. Jia Y, Wang Y, Song R, Li J (2015) Decoder side information generation techniques in wyner-ziv video coding: a review. *Multimedia Tools and Applications* 74(6):1777–1803
21. Kubasov D, Nayak J, Guillemot C (2007) Optimal reconstruction in wyner-ziv video coding with multiple side information. In: *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, IEEE, pp 183–186
22. Li R, Liu H, Chen J, Gan Z (2016) Wavelet pyramid based multi-resolution bilateral motion estimation for frame rate up-conversion. *IEICE Trans Info Sys* 99(1):208–218
23. Liu W, Dong L, Zeng W (2010) Motion refinement based progressive side-information estimation for wyner-ziv video coding. *IEEE Trans Circuits Syst Video Technol* 20(12):1863–1875
24. Mallat S, Hwang WL (1992) Singularity detection and processing with wavelets. *IEEE Trans Inf Theory* 38(2):617–643
25. Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans Pattern Anal Mach Intell* 11(7):674–693
26. Mallat S (1989) Multifrequency channel decomposition of image and wavelet modals. *IEEE Trans, Acoust, Speech, Signal Process* 37:2091–2110
27. Martins R, Brites C, Ascenso J, Pereira F (2009) Refining side information for improved transform domain wyner-ziv video coding. *IEEE Trans Circuits Syst Video Technol* 19(9):1327–1341
28. Martins R, Brites C, Ascenso J, Pereira F (2010) Statistical motion learning for improved transform domain wyner-ziv video coding. *IET image processing* 4(1):28–41
29. Ortega JM (1987) *Matrix theory. the university series in mathematics*
30. Pereira F, Brites C, Ascenso J (2009) Distributed video coding: basics, codecs and performance. *Distributed Source Coding* pp 189–245
31. Petrazzuoli G, Cagnazzo M, Pesquet-Popescu B (2010) High order motion interpolation for side information improvement in dvc. In: *Acoustics speech and signal processing (ICASSP), 2010 IEEE International Conference on*, IEEE, pp 2342–2345
32. Puri R, Majumdar A, Ramchandran K (2007) Prism: a video coding paradigm with motion estimation at the decoder. *IEEE Trans. Image Process.* 16(10):2436–2448
33. Qing L, Zeng W (2014) Context-adaptive modeling for wavelet-domain distributed video coding. *IEEE MultiMedia* 21(4):84–93
34. Rencher AC (2003) *Methods of multivariate analysis, vol 492*. John Wiley & Sons
35. Rup S, Majhi B (2013) A mixed framework for transform domain wyner-ziv video coding. *Optik-International Journal for Light and Electron Optics* 124(21):4929–4938
36. Rup S, Majhi B, Padhy S (2014) An improved side information generation for distributed video coding. *AEU-International Journal of Electronics and Communications* 68(3):201–209
37. Said A, Pearlman WA (1996) A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans Circuits Syst Video Technol* 6(3):243–250
38. Shapiro JM (1993) Embedded image coding using zerotrees of wavelet coefficients. *IEEE Trans Signal Process* 41(12):3445–3462
39. Slepian D, Wolf J (1973) Noiseless coding of correlated information sources. *IEEE Trans Inf Theory* 19(4):471–480
40. Tagliasacchi M, Tubaro S, Sarti A (2006) On the modeling of motion in wyner-ziv video coding. In: *Image processing, 2006 IEEE International Conference on*, IEEE, pp 593–596
41. Taieb MH, Chouinard JY, Wang D (2013) Spatial correlation-based side information refinement for distributed video coding. *EURASIP J Adv Signal Process* 2013(1):168
42. Thao NTH, Tien VH, Van Xiem H, Duong DT et al (2016) Side information creation using adaptive block size for distributed video coding. In: *Advanced technologies for communications (ATC), 2016 International Conference on*, IEEE, pp 339–343
43. Van Luong H, Raket LL, Huang X, Forchhammer S (2012) Side information and noise learning for distributed video coding using optical flow and clustering. *IEEE Trans Image Process* 21(12):4782–4796
44. Van Luong H, Raket LL, Forchhammer S (2014) Re-estimation of motion and reconstruction for distributed video coding. *IEEE Trans Image Process* 23(7):2804–2819
45. Varodayan D, Chen D, Flierl M, Girod B (2008) Wyner-ziv coding of video with unsupervised motion vector learning. *Signal Process Image Commun* 23(5):369–378

46. Vetterli M, Herley C (1992) Wavelets and filter banks: Theory and design. *IEEE Trans Signal Process* 40(9):2207–2232
47. Wiegand T, Sullivan GJ, Bjontegaard G, Luthra A (2003) Overview of the H.264/AVC video coding standard. *IEEE Trans Circuits Syst Video Technol* 13(7):560–576
48. Wyner A, Ziv J (1976) The rate-distortion function for source coding with side information at the decoder. *IEEE Trans Inf Theory* 22(1):1–10
49. Yan C, Zhang Y, Xu J, Dai F, Li L, Dai Q, Wu F (2014a) A highly parallel framework for HEVC coding unit partitioning tree decision on many-core processors. *IEEE Signal Process Lett* 21(5):573–576
50. Yan C, Zhang Y, Xu J, Dai F, Zhang J, Dai Q, Wu F (2014b) Efficient parallel framework for HEVC motion estimation on many-core processors. *IEEE Trans Circuits Syst Video Technol* 24(12):2077–2089
51. Zhang Y, Zhao D, Liu H, Li Y, Ma S, Gao W (2012) Side information generation with auto regressive model for low-delay distributed video coding. *J Vis Commun Image Represent* 23(1):229–236



Bodhisattva Dash received his M.Tech degree in Electronics and Telecommunication Engineering from Silicon Institute of Technology (SIT), Bhubaneswar, Odisha, India. He is continuing his Ph.D. degree in Computer Science and Engineering from International Institute of Information Technology, Bhubaneswar (IIIT-Bh), Odisha, India. His research interest includes Image and Video Processing, Distributed video coding, Scalable Video Coding, Image Compression, and Image Restoration.



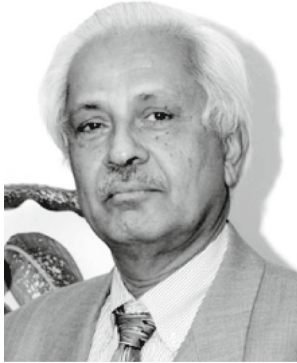
Suvendu Rup received his M.Tech degree in Computer Science and Engineering, from Jadavpur University, Kolkata, India. He received his Ph.D. degree in Computer Science and Engineering from National Institute of Technology (NIT), Rourkela, Odisha, India. Since 2010, he is with the Department of Computer Science and Engineering, International Institute of Information Technology Bhubaneswar (IIIT-Bh), India, and currently serving as an Assistant Professor. His research interest includes Medical Imaging, Distributed video coding, Image Compression, Computer Vision, Video Synopsis, and Video Object Detection and Tracking.



Anjali Mohapatra received her M.Tech and Ph.D. degree in Computer Science from Utkal University, Bhubaneswar, Odisha, India. In 2008, she joined the Department of Computer Science and Engineering, International Institute of Information Technology Bhubaneswar (IIIT-Bh), India, and currently serving as an Assistant Professor. Her research areas include novel application areas of computer science such as Molecular Biology, Soft Computing, Image Processing and Algorithms.



Banshidhar Majhi received his M.Tech degree and Ph.D. in Computer Science and Engineering in the year 1998 and 2003, respectively, from National Institute of Technology (NIT), Rourkela, Odisha, India. Since 1991, he is with the Department of Computer Science and Engineering, NIT, Rourkela, and currently serving as the Professor and Dean Academics. His research interest includes Image and Video Processing, Data Compression, Soft Computing, Bio-metrics and Network Security.



M. N. S. Swamy received the B.Sc. (Hons.) degree in mathematics from Mysore University, India, in 1954, the Diploma degree in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1957 and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Saskatchewan, Saskatoon, Canada, in 1960 and 1963, respectively. He was conferred in 2009 the title of Honorary Professor at National Chiao Tung University in Taiwan. He is presently a Research Professor and holds the Concordia Chair (Tier I) in Signal Processing in the Department of Electrical and Computer Engineering at Concordia University, Montreal, QC, Canada, where he served as the Founding Chair of the Department of Electrical Engineering from 1970 to 1977, and Dean of Engineering and Computer Science from 1977 to 1993. During that time, he developed the Faculty into a research-oriented one from what was primarily an undergraduate Faculty. He has also taught in the Electrical Engineering Department of the Technical University of Nova Scotia, Halifax, and the University of Calgary, Calgary, as well as in the Department of Mathematics at the University of Saskatchewan. He has published extensively in the areas of number theory, circuits, systems and signal processing, and holds five patents. He is the coauthor of nine books and several book chapters. He was a founding member of Micronet from its inception in 1990 as a Canadian Network of Centers of Excellence until its expiration in 2004, and also its coordinator for Concordia University. Dr. Swamy is a Fellow of the Institute of Electrical and Electronics Engineers, Fellow of the Institute of Electrical Engineers (United Kingdom), the Engineering Institute of Canada, the Institution of Engineers (India), and the Institution of Electronic and Telecommunication Engineers (India). He was inducted in 2009 to the Provosts Circle of Distinction for career achievements. He has served the IEEE in various capacities such as the President-Elect in 2003, President in 2004, Past-President in 2005, Vice President (Publications) during 2001-2002, Vice-President in 1976, Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I from June 1999 to December 2001, Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS during June 1985 to May 1987, Program Chair for the 1973 IEEE CAS Symposium, General Chair for the 1984 IEEE CAS Symposium, Vice-Chair for the 1999 IEEE Circuits and Systems (CAS) Symposium, and a member of the Board of Governors of the CAS Society. He is the recipient of many IEEE-CAS Society awards, including the Education Award in 2000, Golden Jubilee Medal in 2000, and the 1986 Guillemin-Cauer Best Paper Award. He is the Editor-in-Chief of the journal Circuits, Systems and Signal Processing (CSSP) since 1999. Recently CSSP has instituted a best paper award in his name.