Dictionary Learning Based Two-stage Near-lossless Video Compression

Zuhai Zhang[†], Luheng Jia[†]*, Li Song[‡], Shuyuan Zhu[§], Yuanfang Guo[¶] and Kebin Jia[†]
[†] Beijing University Of Technology, Beijing
[‡] Shanghai Jiao Tong University, Shanghai
[§] University of Electronic Science and Technology of China, Sichuan
[¶] Beihang University, Beijing
Corresponding email: luhengjia@bjut.edu.cn

***Abstract*—Traditional hybrid video coding framework using block based predictive coding and transform coding, such as the High Efficiency Video Coding (HEVC), cannot further dig out the redundancy remained in quantized transformed residual, causing extra bits consumption. Measured by rate-distortion (R-D) performance, the problem of higher bits consuming can be solved reversely by video quality enhancing. In this work, we proposed a video coding scheme that solve the problem by enhancing the reconstructed video quality using supplementary information from further compressed quantization error. Aiming at better R-D performance for near-lossless video coding, we propose a novel video coding scheme using a two-stage framework that extracts quantization error as complementary information which is compressed using dictionary learning and sparse representation. The employed over-complete dictionary is learned through K-SVD with orthogonal matching pursuit (OMP) for sparse representation. Statistically redundancy is further removed by a modified context-adaptive binary arithmetic coding (CABAC) with adaptive context models. This approach not only retains the advantages of the traditionally encoder for lossy compression but also exploits the redundancy in the quantization error to achieve high-quality near-lossless compression. Experimental results demonstrate that our method significantly outperforms traditional HEVC lossy encoder with over -20% BD-BR on average at high bitrate range for near-lossless coding, while the method is also proved to be efficient at low bitrate range achieving over -50% BD-BR on average with average PSNR over 41dB, which retains near-lossless performance.**

I. Introduction

Nowadays, the market share of the HEVC [1] is still rising with its ability to efficiently compress high-quality videos including 4K and 8K ultra-high-definition (UHD) videos. Besides the pursuits for higher resolution, lossless and near-lossless compression that provides higher data security and better quality have become a trend in recent years. Near-lossless video compression aims to preserve a significant amount of detail such that the loss of information is imperceptible to the human eye. Near-lossless compression offers a balanced trade-off between video bitrate and quality than fully lossless encoding and common lossy encoding in many applications such as medical video compression and surveillance video compression. However, most of the coding tools in HEVC are geared towards lossy compression. Merely using lower quantization parameters to achieve near-lossless compression often fails to fully leverage the encoder's potential for effective compression. There is still substantial room for improvement in optimizing near-lossless compression.

To ensure compression efficiency while preserving image details as much as possible, near-lossless compression was introduced and applied early on in JPEG-LS [2]. Subsequently, various near-lossless compression algorithms such as [3], [4] are incorporated into video encoding. Both approaches aim at video compression at high bitrate range. The former uses an adaptive strategy to select the optimal prediction algorithm to achieve smaller residuals, while the latter optimizes the quantization scheme at high bitrate range to achieve better compression results. While these methods attempt to modify traditional video encoders, we propose an optimization approach using a two-stage framework. This encoding process framework aims to enhance the encoder's efficiency while preserving its inherent advantages, allowing for flexible adjustments in compression rate. Moreover, the context models in context-adaptive binary arithmetic coding (CABAC) adopted in standard video encoder are not designed for lossless or near-lossless compressed data, causing reduced coding efficiency.

In addition to direct improvements to the encoder, another approach to lossless and near-lossless video compression adopts two-stage frameworks [5]–[8]. The two-stage framework decomposes video encoding process into two parts including a lossy compression stage and an subsequent error components compression stage. Heindel et al. [8] modified the two-stage encoding framework for near-lossless video compression. Bai et al. [9] introduced a learning-based coding strategy into the two-stage framework for near-lossless encoding. Both approaches have demonstrated the efficiency of the two-stage framework in near-lossless compression. Additionally, scalable video coding framework [7] with base-layer and enhancement-layer is considered as a two-stage coding framework. However, these approaches have not employed entropy encoder specifically designed according to the statistic

This work was supported by National Natural Science Foundation of China under Grant 61901012 and Science and Technology Commission of Shanghai Municipality under Grant 22DZ2229005. (*Corresponding author: Luheng Jia.*)

characteristics of the error components in the second encoding stage, and hence statistical redundancy has not been efficiently removed.

To address the aforementioned drawbacks of current methods, we propose a video coding scheme operates within a two-stage framework, employing dictionary learning and sparse representation to compress the residuals, followed by an specifically designed entropy encoder adapting the characteristics of sparse represented residual. The proposed modified two-stage framework retains efficiency of the standard lossy video encoder in the first stage, while the second stage leverages difference between original frame and reconstruction frames as complementary information to enhance the quality of reconstruction frame consuming small extra bits. And hence, better rate-distortion (R-D) performance is achieved. Additionally, the complementary information in the second stage is sparse represented using an patch-size adaptive dictionary learning method to characterize various distribution features of residuals at different bitrate. Finally, the sparse coefficients are entropy encoded using a modified CABAC-based entropy encoder with specificlly designed context models. The proposed scheme achieves significant coding efficiency gain for near-lossless coding scenario at high bitrate range, while the scheme outperforms standard lossy video encoder at lower birate range.

The remainder of the paper is organized as follows. Section II introduces the dictionary learning and sparse representation methods used. Section III describes the proposed framework and entropy encoding strategy. Section IV presents the experimental results. Section V concludes the paper.

II. Problem Formulation of R-D Optimized Two-Stage Near-lossless Video Coding

The proposed two-stage encoding framework is shown in Fig.1. We denote the original current frame as x. And the reconstructed version of x from the first stage is denoted as $\widehat{x}$. The difference between x and $\widehat{x}$ is denoted as r. And we have

$$r = x - \widehat{x} \tag{1}$$

which is adopted as complementary information to enhance the quality of final reconstructed frame $\widehat{x}'$. To reduce the extra bit consumption, r is sparse represented and entropy coded in the second stage. The reconstructed complementary information from the second stage is denoted as $\widehat{r}$. The final reconstructed frame $\widehat{x}'$ is obtained as follows.

$$\widehat{x}' = \widehat{x} + \widehat{r} \tag{2}$$

With equation (1) and (2), the final reconstruction error e is obtained as follows.

$$e = x - \widehat{x}' = r - \widehat{r} \tag{3}$$

Without using quantization as traditional lossy compression, the compression method adopted in the second stage controls R-D trade-off by sparsity setting, which can introduce very limited distortion. Therefore, $|e|$ is much smaller than $|r|$ leading to better reconstruction quality and near-lossless compression performance. To be noticed, r is the reconstruction error of traditional video encoder. The encoding distortion D of traditional video encoder is the variance of r, which is denoted as σ_r^2. While the encoding distortion D' of the proposed two-stage encoding framework is the variance of e, which is denoted as σ_e^2. The distortion difference can be considered as quality improvement. We assume e and r are of zero mean. We have

$$\begin{aligned} \Delta D &= D - D' = \sigma_r^2 - \sigma_e^2 \\ &= 2E(r \cdot \widehat{r}) - E(\widehat{r}^2) \\ &= 2 \cdot E(r^2) - 2 \cdot E(r \cdot e) - E(r^2) - E(e^2) + 2 \cdot E(r \cdot e) \\ &= E(r^2) - E(e^2) \end{aligned} \tag{4}$$

Under near-lossless coding scenario, $E(e^2)$ as the final reconstruction error of the two-stage coding scheme is small. By carefully adjusting the sparsity of sparse represented coefficients, we can always obtain $E(e^2)$ smaller than $E(r^2)$ and hence $\Delta D \geq 0$.

By employing the widely used logarithmic R-D model $R = \alpha \cdot log_2 \frac{\sigma^2}{D}$ [10], the D-R model for the traditional video encoder is as follows.

$$D = \frac{\sigma^2}{2^{\frac{R}{\alpha}}} \tag{5}$$

in which σ^2 is the quantized transformed coefficients of predictive residual before entropy coding. For easy comparison, we assuming the first stage of the proposed scheme using the same coding parameters set as standard video encoder, and hence the first stage has the same R-D characteristic as in (5). And the R-D characteristics of second stage in the proposed two-stage encoding framework is modeled as follows.

$$D' = \frac{\sigma'^2}{2^{\frac{R'}{\alpha}}} \tag{6}$$

where σ'^2 is the variance of the processed input difference from first stage before entropy coding. R' is the extra bits consumed in the second stage. We formulate the two-stage video coding as maximize the quality difference with given extra bits budget using the two-stage framework compared with the traditional standard encoder.

$$\max_{\theta} \ \Delta D \ \ s.t. \ \Delta R \leq R_T \tag{7}$$

where ΔD is the improved quality by introducing the complimentary information in the second stage as in (4). θ is the coding parameter sets. And R_T is the target extra bit budget. With the first stage encoder the same as the comparative standard encoder and D is considered as a constant, maximizing the distortion difference is equivalent to minimizing the final reconstruction distortion D'. Additionally, ΔR is the extra bits cost in the two-stage framework, which is R' if we use the same encoder in the first stage as the standard encoder. From (6), we can find that minimizing D' is equivalent to minimizing σ'^2. Finally, the problem is simplified as

$$\min_{\theta} \ \sigma'^2 \ \ s.t. \ R' \leq R_T \tag{8}$$

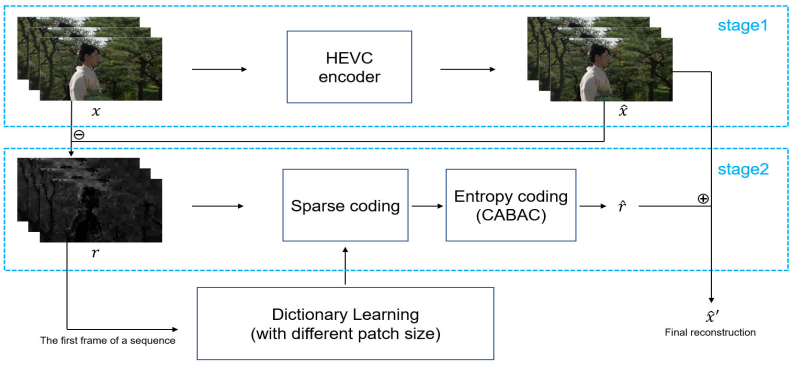

Fig. 1. The framework of proposed dictionary learning based two-stage video compress scheme

Since the input data of the second stage encoder r in (1) is the different between the original frame and reconstructed frame of the first stage, which is the quantization error. Under video coding scenario, quantization error is not fully decorrelated and contains structural information of original frame as shown in Fig.2, however, the difference frame r contains mainly high frequency noise. Therefore, traditional predictive coding and transform coding fail to compress the difference frame. To reduce the variance of the r and leading to small σ'^2, the dictionary learning based sparse representation method is an efficient approach, which is widely used in image demonising. The sparse representation not only removes high frequency noise in r, but also efficiently protects the remained structure information as patterns in learned dictionary. Finally, leading to small σ'^2 value and better overall R-D performance.

Moreover, to reduce the extra bits consumption, specifically designed CABAC encoder according to the statistic characteristic of sparse represented components is designed. The detailed algorithm design is introduced in the following section.

III. The Proposed Framework And Entropy Encoding Strategy

A. Dictionary Learning Based Two-stage Video Compression

Our proposed two-stage framework shown in Fig.1 consists of two main stages. The first stage is responsible for initially compressing the video to obtain a compressed lossy video bitstream. To ensure that the compressed video retains more details, it is necessary to restore the information lost in the first stage encoding. Therefore, in the second stage, we compress the different between the original video and the reconstructed video of first stage to restore high-quality near-lossless video. As is discussed in the first section, we adopt dictionary learning based sparse representation to compress the difference frame as complementary information to enhance reconstruction quality.

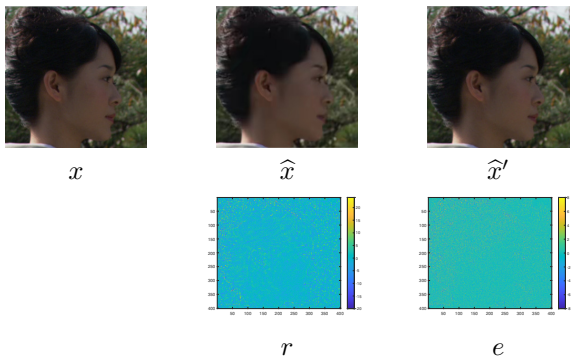

Fig. 2. Illustration of a segment of frames and residuals: x (original), $\widehat{x}$ (reconstructed), their residual r, and $\widehat{x}'$ (final reconstructed), and the final reconstruction error e.

In compressing the difference frame in the second stage, we use an over-complete dictionary, pre-trained with a large amount of data using the K-SVD algorithm, for sparse representation. The difference frame is divided into patches, the n_{th} patch $\mathbf{r_n}$ is decompose into a dictionary matrix Φ and a sparse coefficient matrix $\mathbf{s_n}$. The dictionary Φ stores the features of the original samples $\mathbf{r_n}$, and the sparse coefficient matrix s_n can represent r_n using fewer non-zero coefficient through the dictionary matrix Φ, as shown in the following equation.

$$\mathbf{r_n} = \Phi \cdot \mathbf{s_n} \tag{9}$$

where each column of $\mathbf{r_n}$ represents a $p \times p$ image patch, and each column of Φ represents a dictionary atom. We can represent an image block $\mathbf{r_{n,i}}$ in $\mathbf{r_n}$ using the sparse coefficients $\mathbf{s_{n,i}}$ corresponding to a column in $\mathbf{s_n}$ and the dictionary Φ, i.e., $\mathbf{r_{n,i}} = \Phi \cdot \mathbf{s_{n,i}}$.

To solve for Φ and $\mathbf{s_n}$ in order to represent the residual matrix R, an optimization problem needs to be addressed, as shown in the following equation:

$$\min_{\mathbf{\Phi},\mathbf{s_n}} ||\mathbf{r_n} - \mathbf{\Phi} \cdot \mathbf{s_n}||_F^2 \quad s.t.\ \forall i, ||\mathbf{s_{n,i}}||_0 \leq T_0 \tag{10}$$

where T_0 represents the number of nonzero elements. The notation $||\cdot||_F$ stands for the Frobenius norm and $||\cdot||_0$ stands for the zero norm. Using K-SVD, we first compute the sparse coefficient matrix $\mathbf{s_n}$ through sparse representation. Then, during the dictionary learning process, we sequentially update each atom of the dictionary $\mathbf{\Phi}$. This involves K iterations where each iteration requires performing singular value decomposition(SVD) to update the dictionary.

Furthermore, in our proposed DL-based two-stage framework, sparse representation is essential both during the learning of the K-SVD dictionary and when using the learned over-complete dictionary to represent image samples. This involves

computing a sparse solution s_i under the residual r and dictionary Φ, as shown in the following equation:

$$\min_{\mathbf{s_{n,i}}} ||\mathbf{s_{n,i}}||_0 \quad s.t. \ ||\mathbf{r_{n,i}} - \mathbf{\Phi} \cdot \mathbf{s_{n,i}}||_2^2 \leq \epsilon \tag{11}$$

ϵ represents a small allowable error. We use orthogonal matching pursuit (OMP) [11] to select the dictionary's basis vector that best matches the current difference patch $\mathbf{r_{n,i}}$ at each iteration. It updates the residual until the predetermined sparsity condition is met. The resulting sparse matrix is then entropy coded to obtain the a complementary bitstream.

Specifically, different patch sizes adopted in the dictionary learning and sparse representation show varying compression performance under different bit rates. We propose using two different patch sizes to handle video residuals across different bit rate ranges. When processing high bit rate videos using small quantization parameters (QP), the difference frame r to be compressed are relatively sparse. In this case, we use a small patch size of 8×8 for sparse encoding. Conversely, when handling low bit rate videos with large QP, where difference frame r contains more structural information, as shown in Figure 2, we use a larger patch size of 16×16.

To be noticed, the proposed scheme is different from traditional two-stage encoding methods that adjust compression ratio only by QP. In contrast, the method proposed in this paper allows for dynamic adjustment of multiple parameters including QP in the first stage compression, sparsity level and patch size during sparse representation based compression in the second stage. This multi-parameter controlled the compression scheme offers greater scalability compared to traditional methods, leading to better R-D performance. Through this framework, we can not only fully utilize the original encoder's algorithm for lossy compression but also supplement the lost information to flexibly enhance the video reconstruction quality and achieve a near-lossless performance.

B. Entropy Coding Designed for Sparse Represented Data

In this work, we modify the CABAC for compression and set reasonable context index to achieve the maximum compression efficiency. CABAC mainly consists of binarization, context modeling, and binary arithmetic coding. Our modified CABAC firstly iterates through each column of sparse coefficient matrix $\mathbf{s_n}$. For each column, we record the number of non-zero coefficients num, the values of the coefficients val, the sign of the coefficients $sign$, and the run lengths between coefficients run.

These three variables are then binarized in sequence. The number of non-zero coefficients (num) and the sign of non-zero coefficients ($sign$) are encoded using fixed-length coding, which ensures a straightforward and efficient representation of the count of significant values in each column. The coefficient values (val) and the run lengths (run) are encoded using exponential Golomb coding. Exponential Golomb coding is chosen for its efficiency in encoding a wide range of values with fewer bits, especially when dealing with sparse data. This method helps to minimize the bit rate by taking advantage of the distribution patterns in the sparse coefficient matrix, leading to more effective compression.

After binarization, we perform arithmetic coding on the four variables using different contexts. Unlike HEVC's CABAC, which establishes a table for coefficient probability changes and converts them by lookup, we simplify this step. We use traditional binary arithmetic coding, where the context probability is dynamically updated based on the previous symbol. However, for each variable to be encoded, we set different context models and mechanisms for selecting these models. This allows us to optimally encode according to the corresponding probability distribution of each variable. For example, for val, we use the preceding symbol to select different context models to achieve the optimal compression rate.

By utilizing context models specifically designed to each variable, we can more accurately predict the probability distribution of the symbols, thereby enhancing the efficiency of the arithmetic coding process. For num, a context model based on the distribution of previous counts can be employed. For val, the context is chosen based on the preceding coefficient, allowing the model to adapt to local variations in the data.

This context-adaptive approach enables the CABAC to better exploit the statistical properties of the sparse coefficient matrix, leading to a higher compression ratio and more efficient encoding.

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

A. Implementation

The method proposed in this paper consists of two stages. The first stage involves a basic encoder using HM16.20 for initial lossy compression of the video. In the second stage, difference frames are extracted between the first frames of the reconstructed and original videos. These frames undergo dictionary learning for feature extraction, and following frames are sparsely represented based on the shared dictionary. The resulting sparse coefficient matrix is entropy encoded to generate the bitstream for the second part. The experiments use the official HEVC datasets, with Classes A through E representing different resolutions. At the decoder, residuals are reconstructed into near-lossless form using the sparse coefficient matrix and the dictionary.

B. Experimental Results

To demonstrate the high coding efficiency of the proposed dictionary learning based two-stage compression method in near-lossless video coding in high bitrate range, we use small QP values of 1, 3, 5 and 7 in the first stage encoding. For the second stage coding, we employ a fixed dictionary learning sparsity L of 6, a patch size $p \times p$ of 8×8. Standard test sequence of different resolution and characteristics are employed encoded under Low-Delay Main configuration. The standard HEVC test model HM16.20 with same coding configuration is used as the comparative method. The Bjontegaard Delta Bitrate (BD-BR) [10] is used to measure the R-D performance

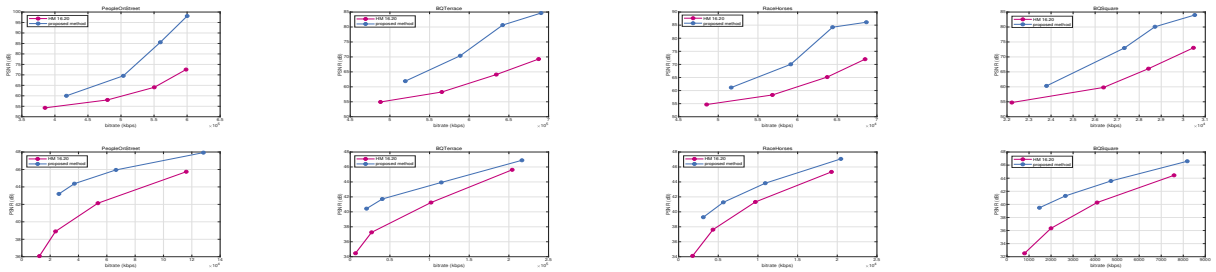

Fig. 3. TOP: High-bitrate R-D curves. Bottom: Low-bitrate R-D curves. From left to right: PeopleOnStreet (Class A), BQTerrace (Class B), RaceHorses (Class C), BQSquare (Class D).

TABLE I
BD-BR FOR HIGH-BITRATE CODING

Class	Name	QP	Bitrate(kbps)	PSNR(dB)	Bitrate(kbps)	PSNR(dB)	BD-BR
ClassA	Traffic	1	577236.76	76.04	668166.68	93.89	-14.20%
		3	535838.38	66.12	541805.17	93.50	
		5	483697.06	59.42	503812.08	73.68	
		7	372053.75	54.21	404999.16	60.34	
	PeopleOnStreet	1	598408.27	72.54	600148.11	98.11	-15.70%
		3	550358.25	64.06	559510.79	85.57	
		5	479540.71	58.03	503510.03	69.54	
		7	385080.43	54.26	417423.74	60.00	
ClassB	Cactus	1	649417.20	81.45	650635.68	95.68	-19.01%
		3	606358.76	69.81	609813.81	93.64	
		5	564251.79	60.07	580899.03	76.26	
		7	490673.41	55.20	517644.89	62.48	
	BQTerrace	1	687526.79	69.28	690565.55	84.66	-32.38%
		3	634135.23	64.10	642196.10	80.64	
		5	565464.34	58.27	588692.95	70.39	
		7	488142.68	54.94	519441.66	61.90	
ClassC	BQMall	1	130057.35	78.13	130143.79	88.72	-16.30%
		3	120625.53	67.56	121541.88	87.06	
		5	112504.52	60.02	116095.21	75.99	
		7	94904.86	54.92	101148.16	61.32	
	Racehorse	1	68508.28	71.99	68670.81	86.11	-25.95%
		3	63731.26	65.16	64413.56	84.20	
		5	56823.99	58.33	59134.56	70.03	
		7	48533.16	54.67	51630.90	61.16	
ClassD	BQSquare	1	30462.74	73.04	30521.32	83.51	-14.75%
		3	28421.90	66.07	28716.51	80.08	
		5	26381.74	59.80	27325.90	72.97	
		7	22218.88	54.75	23793.62	60.31	
	BlowingBublles	1	28431.05	76.06	28456.62	82.66	-9.70%
		3	26403.46	66.62	26625.27	80.42	
		5	24463.63	59.44	25293.60	73.30	
		7	21062.52	54.84	22371.51	60.56	
ClassE	Fourpeople	1	233402.96	74.89	233873.51	89.89	-31.33%
		3	217671.19	66.33	220293.33	88.22	
		5	192975.68	59.47	201753.80	72.45	
		7	135327.89	53.94	150377.98	59.31	
	Johnny	1	228459.63	74.70	190761.44	88.35	-43.60%
		3	213700.47	66.49	180062.38	86.49	
		5	189210.13	59.55	164890.23	71.54	
		7	130501.36	53.97	121242.64	59.47	
Average		1	323191.10	74.81	333009.58	89.16	-22.29%
		3	299724.44	66.23	303099.13	85.98	
		5	269531.36	59.24	280438.54	72.62	
		7	218849.89	54.57	235432.28	60.69	

TABLE II
BD-BR FOR LOW-BITRATE CODING

Class	Name	QP	Bitrate(kbps)	PSNR(dB)	Bitrate(kbps)	PSNR(dB)	BD-BR
ClassA	Traffic	17	70426.46	44.93	82743.45	46.76	-33.00%
		22	22852.28	41.73	35770.09	44.35	
		27	7813.48	38.95	21205.44	42.58	
		32	3166.67	36.39	16968.12	41.38	
	PeopleOnStreet	17	116064.08	45.74	128182.72	47.93	-56.40%
		22	53705.14	42.14	66462.02	45.95	
		27	23783.96	38.9	37111.35	44.37	
		32	12409.74	36.07	26159.39	43.20	
ClassB	Cactus	17	171672.82	44.45	182143.01	45.95	-59.10%
		22	45857.47	39.81	56948.36	42.67	
		27	10022.62	37.36	21482.60	40.99	
		32	4322.14	35.33	16065.08	40.01	
	BQTerrace	17	204200.25	45.62	216479.38	46.89	-61.81%
		22	101856.72	41.25	114826.91	43.93	
		27	26916.82	37.26	40449.17	41.72	
		32	6668.23	34.46	20574.82	40.43	
ClassC	BQMall	17	22682.86	44.32	25083.17	46.45	-36.43%
		22	7590.78	40.85	10093.14	44.04	
		27	3547.32	37.94	6147.50	42.26	
		32	1712.65	35.01	4405.24	40.75	
	Racehorse	17	19270.23	45.34	20453.38	47.08	-46.10%
		22	9677.00	41.32	10926.00	43.83	
		27	4330.79	37.62	5637.29	41.27	
		32	1763.41	34.11	3113.69	39.28	
ClassD	BQSquare	17	7581.41	44.44	8181.15	46.59	-68.10%
		22	4089.19	40.28	4716.88	43.57	
		27	1997.11	36.34	2650.36	41.29	
		32	796.38	32.52	1471.22	39.48	
	BlowingBublles	17	7091.50	44.17	7465.26	45.92	-42.35%
		22	3136.84	39.75	3613.39	42.56	
		27	1424.34	36.14	1981.21	39.89	
		32	646.98	33.03	1223.16	37.99	
ClassE	Fourpeople	17	17977.44	45.83	23432.35	49.25	-79.17%
		22	3654.89	43.25	9297.70	48.06	
		27	1228.06	40.93	7036.98	47.19	
		32	602.90	38.40	6574.01	46.46	
	Johnny	17	18119.69	46.22	23540.17	49.78	-32.70%
		22	4352.52	43.81	9932.94	48.67	
		27	1015.57	41.70	6704.56	47.84	
		32	389.91	39.64	6169.47	47.20	
Average		17	65508.67	45.11	71770.40	47.26	-51.52%
		22	25677.28	41.42	32258.74	44.76	
		27	8208.01	38.31	15040.65	42.94	
		32	3247.90	35.50	10272.42	41.62	

difference. Experimental results in Table 1 demonstrates that for near-lossless coding scenario in high bitrate range, our proposed scheme outperforms traditional HEVC encoder by -22.29% BD-BR on average. It is interesting to notice that in near-lossless video coding, the proposed method maintains similar bitrate while achieves significant quality improvement, for example, PSNR increase of 16.54 dB is achieved with only 1.27% bitrate increase for sequence BQTerrace encoded using QP of 3. The small increase of bitrate is due to that with small QP, the difference frame from the first stage is sparse. With small extra bits as complementary information, the lost details are well restored and the final reconstruction quality is significantly improved.

The proposed scheme also shows better R-D performance than traditional encoder in low bitrate coding scenario, and maintains near-lossless performance. The evaluation is conducted with a fixed dictionary learning sparsity L of 6, a patch size a patch size $p \times p$ of 16×16. The Low-Delay P Main configuration is used with QP values of 17, 22, 27, and 32 that follows the common test conditions. The standard HEVC test model HM16.20 with same coding configuration is used as the comparative method. The BD-BR is used to measure the R-D performance difference. As shown in Table 2, the proposed dictionary learning based two-stage compression method significantly outperforms standard HEVC lossy encoding in the low bit rate range that an average BD-BR of -51.52% is achieved. In video coding at low bitrate range, the difference frame generated from the first stage contains abundant noise and also structural information, which is sparse represented and lossy compressed in the second stage controlled by the sparsity level. Since the lost information complementary and quality restore in the second stage, the PSNR is always higher than that of standard encoder. It hence more reasonable to compare the bitrate reduction between the proposed scheme with higher QP and standard encoder with lower QP. For example, when encoding sequence Cactus, the proposed scheme using QP of 27 achieves PSNR of 40.99 dB, which is of 1.18dB higher than the standard encoder using QP of 22. However, the bitrate

TABLE III
R-D PERFORMANCE COMPARISON BETWEEN THE PROPOSED METHOD AND ELC [7]

	BD-BR (%)	
	ELC [7]	Our method
ClassA	-10.4%	-17.0%
ClassB	- 7.8%	-28.1%
ClassC	-15.8%	-24.1%
ClassD	-21.2%	-14.2%
ClassE	- 7.4%	-32.5%
Average	-12.5%	-23.2%

reduction is 59.10%. This demonstrates the proposed schemes always works in common bitrate range. Moreover, we can find from Table 2 that the average PSNR values using QP from 22, 27, 32 and 37 are all larger than 41 dB. Commonly, a PSNR measure of 40 dB, or above, typically constitutes visually lossless coding, considered as near-lossless [12]. We also depict the R-D curves in Fig. 3 to clearly shows the R-D performance at different bitrates range. It can be observed that our proposed method achieves significant improvements in both high and low bitrate video coding scenarios.

Moreover, comparison is made between our proposed method and a multi-layer coding framework with enhancement layer (ELC) proposed in [7]. The experimental results in Table 3 demonstrate that with the same coding parameters, our proposed scheme outperforms the ELC approach by -10.6% BD-BR on average. To be noticed, our proposed method shows better R-D performance in high resolution test sequences including Class A, B, C and E, indicating that the dictionary learning based compression is more efficient for high resolution videos. While for Class D of low resolution sequences, the comprasion method ELC shows better R-D performance.

V. CONCLUSIONS

In this work, we propose a dictionary learning based two-stage video encoding framework to compress and leverage the reconstruction errors as complementary information to enhance final reconstruction quality. CABAC is modified to be adaptive to the sparse represented reconstructed error, leading to high coding efficiency. Experimental results demonstrate that the proposed method achieves over 20% BD-BR reduction dealing with near-lossless video coding at high bitrate range. Additionally, the proposed framework is efficient under normal bitrate range, leading to over 50% BD-BR reduction and retaining near-lossless coding with PSNR values above 41 dB.

The proposed method can be implemented in video coding systems beyond HEVC. During training stage of dictionary learning, apart from adjusting the QP in the first lossy compression stage with standard encoder, we can also fine-tune training parameters such as sparsity L and patch size p during the second encoding stage with sparse coding tools to find the optimal coding parameter set. Moreover, rate control employing bit allocation algorithm between two encoding stages can further improve the overall R-D performance. In our future work, well designed inter-stage rate control method and optimal coding parameters determination will yield even greater benefits.

REFERENCES

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS," *IEEE Transactions on Image processing*, vol. 9, no. 8, pp. 1309–1324, 2000.

[3] A. Antony and S. Ganapathy, "Highly efficient near lossless video compression using selective intra prediction for HEVC lossless mode," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 11, pp. 1650–1658, 2015.

[4] N. Nazari, R. Shams, M. Mohrekesh, and S. Samavi, "Near-lossless compression for high frame rate videos," in *2013 21st Iranian Conference on Electrical Engineering (ICEE)*, IEEE, 2013, pp. 1–6.

[5] X. Cai and Q. Gu, "Improved HEVC lossless compression using two-stage coding with sub-frame level optimal quantization values," in *2014 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2014, pp. 5651–5655.

[6] S. Takamura and Y. Yashima, "Lossless scalable video coding with h. 264 compliant base layer," in *IEEE International Conference on Image Processing 2005*, IEEE, vol. 2, 2005, pp. II–754.

[7] A. Heindel, E. Wige, and A. Kaup, "Low-complexity enhancement layer compression for scalable lossless video coding based on HEVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1749–1760, 2016.

[8] A. Heindel and A. Kaup, "Scalable near-lossless video compression based on HEVC," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, IEEE, 2017, pp. 1–4.

[9] Y. Bai, X. Liu, W. Zuo, Y. Wang, and X. Ji, "Learning scalable ly=-constrained near-lossless image compression via joint lossy image and residual compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 946–11 955.

[10] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU SG16 Doc. VCEG-M33*, 2001.

[11] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[12] E. Palma and I. Tabus, "Near-lossless coding of plenoptic camera sensor images for archiving light field array of views," in *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, 2022, pp. 1–6.