HHS Public Access

Author manuscript

IEEE Trans Image Process. Author manuscript; available in PMC 2018 October 12.

Published in final edited form as:

IEEE Trans Image Process. 2017 June ; 26(6): 2972–2987. doi:10.1109/TIP.2017.2692882.

Progressive Dictionary Learning With Hierarchical Predictive Structure for Low Bit-Rate Scalable Video Coding

Wenrui Dai [Member, IEEE],
Department of Biomedical Informatics, University of California at San Diego, La Jolla, CA 92093 USA

Yangmei Shen,
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Hongkai Xiong [Senior Member, IEEE],
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Xiaoqian Jiang [Member, IEEE],
Department of Biomedical Informatics, University of California at San Diego, La Jolla, CA 92093 USA

Junni Zou [Member, IEEE], and
Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China

David Taubman [Fellow, IEEE]
School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia

Abstract

Dictionary learning has emerged as a promising alternative to the conventional hybrid coding framework. However, the rigid structure of sequential training and prediction degrades its performance in scalable video coding. This paper proposes a progressive dictionary learning framework with hierarchical predictive structure for scalable video coding, especially in low bitrate region. For pyramidal layers, sparse representation based on spatio-temporal dictionary is adopted to improve the coding efficiency of enhancement layers with a guarantee of reconstruction performance. The overcomplete dictionary is trained to adaptively capture local structures along motion trajectories as well as exploit the correlations between the neighboring layers of resolutions. Furthermore, progressive dictionary learning is developed to enable the scalability in temporal domain and restrict the error propagation in a closed-loop predictor. Under the hierarchical predictive structure, online learning is leveraged to guarantee the training and prediction performance with an improved convergence rate. To accommodate with the state-of-the-art scalable extension of H.264/AVC and latest High Efficiency Video Coding (HEVC), standardized codec cores are utilized to encode the base and enhancement layers. Experimental results show that the proposed method outperforms the latest scalable extension of HEVC and HEVC simulcast over extensive test sequences with various resolutions.

Correspondence to: Hongkai Xiong.

Index Terms

Scalable video coding; progressive dictionary learning; sparse representation

I. Introduction

With the rising demands of video transmission over heterogeneous networks and clients, scalable video coding (SVC) has been widely considered to adapt to varying network conditions and terminal capabilities [1], [2]. In the past two decades, state-of-the-art schemes, e.g. H.264/AVC scalable extension [3] and Barbell-lifting wavelet-based SVC [4], have been studied and standardized to achieve this goal. Recently, SHVC [5], the scalable extension of High Efficiency Video Coding (HEVC [6]) standard, has been finalized to minimize coding complexity and bridge the gap between non-scalable and scalable implementations for practical deployment. It enables multiple single-layer HEVC codec cores with inter-layer reference picture processing modules to efficiently achieve scalable functionality.

In general, temporal and spatial scalability are fundamental topics for scalable video coding. Inspired by wavelet-based schemes for image coding (i.e., EZW [8], SPIHT [9] and EBCOT [10]), subband/wavelet methods have been extensively adopted for scalable video coding. For video sequences, three dimensional wavelet decomposition can be achieved by extending 2-D spatial wavelet transform along the block-based motion information [11]–[13]. However, separable DWT in these methods fails to match motion trajectories, as perfect reconstruction cannot be guaranteed for block-based motion at subpixel accuracy. Thus, motion compensated temporal filtering (MCTF) was developed to leverage lifting structure of wavelet transform for motion alignment. In MCTF, lifting-based temporal transform can be directly applied to the original frames before spatial transform for desirable approximation performance [14]–[17] or performed on the subbands after wavelet decomposition to address specific objectives related to spatial scalability [18]–[21]. Furthermore, interlayer correlations were exploited with in-scale motion compensation in the spatial domain [22] to jointly consider the low-pass signals from lower-resolution layer and compensated high-pass signals within the same resolution layer for reconstruction. However, these methods still need to transmit block-level motion information in addition to the wavelet coefficients.

On the other hand, scalable extension of conventional hybrid motion compensation-discrete cosine transform (MC-DCT) framework has been widely studied for scalable video coding. Pyramidal layered methods and hierarchical B pictures are the state-of-the-art techniques to enable spatial and temporal scalability, respectively. Hierarchical B pictures [23] were first developed for H.264/AVC and its scalable extension, which adopted close-loop control to progressively refine frames at the finer layer of scalability without requiring motion-compensated update. In comparison to MCTF, it enables flexible reference structures with restricted propagation of errors and improved compression efficiency in the context of block-based motion [24]. To support multidirectional motion estimation, fully scalable motion model (SMM [25]) was developed to incorporate with the hierarchical B frame and

rate-distortion optimization. Inheriting the properties of H.264/AVC and its scalable extension, HEVC naturally leverages hierarchical B pictures to support the temporal scalability. To achieve spatial scalability, pyramid methods were considered to reconstruct higher resolution layers from the lower resolution one with upsampled texture and encoded motion information [26]. They are typically redundant in comparison to the original video sequences due to the overcomplete pyramid decomposition. To improve reconstruction, upsampling performance was improved with 2-D Wiener interpolation filter [27] and inverse filtering of down-sampled signals optimized for least-square errors [28]. To integrate with HEVC, Shi *et al.* [29] designed quadtree-based and learning-based interlayer prediction mode to support single-loop and multi-loop solutions for spatial scalability. Furthermore, Kang *et al.* [30] considered a cascaded two-layer representation of prediction residuals in HEVC, where sparse representation is adopted to capture regular patterns of residuals derived by DCT in a sequential manner. To optimize inter-layer prediction, Han *et al.* [31] formulated an estimation-theoretic framework that optimized the conditional distribution for relevant information from base and enhancement layers. In [32], empirical rate-distortion optimization coupling quantization-rate and quantization-distortion models was developed for spatially scalable video coding. Though optimized for upsampling performance, spatial-temporal correlations are not sufficiently exploited with conventional rigid motion compensation schemes.

Recently, dictionary learning has emerged as an alternative solution to the upsampling problem in the hybrid framework, where trained overcomplete dictionary improves the reconstruction from the low-quality visual data based on spatial-temporal correlations. Inspired by sparse representation based on the patch-based overcomplete dictionary [33], example-based approach [34] was developed for super-resolution based reconstruction for video coding. In [35], enhanced skip and direct modes were advanced to integrate spatial super-resolution and interpolation with H.264 and HEVC standard. Incorporating primal sketch priors [36], Xiong *et al.* [37] proposed a sparse spatio-temporal representation for prediction and reconstruction of frames. Furthermore, online learning [38] has been adopted to improve the convergence rate of spatio-temporal dictionary learning. Although dictionary learning based methods have been demonstrated competitive with conventional methods in terms of rate-distortion performance and visual quality, its sequential training and prediction structure is too rigid for scalable video coding.

This paper proposes a novel framework based on progressive dictionary learning for scalable video coding, especially in low bitrate regions. For pyramidal layers, sparse representation based on spatio-temporal dictionary is adopted to improve the coding efficiency of enhancement layers (ELs) with a guarantee of reconstruction performance. The overcomplete dictionary is trained to adaptively capture local structures along motion trajectories as well as exploit the correlations between neighboring layers of resolutions. Furthermore, progressive dictionary learning is developed to enable the scalability in temporal domain and restrict the error propagation in the close-loop predictor. Under the hierarchical predictive structure, online learning is leveraged to guarantee the training and reconstruction performance with an improved convergence rate.

To be concrete, spatio-temporal dictionary learning is adopted for inter-layer prediction, which exploits the spatio-temporal consistency along motion trajectory and inter-layer correlations between base and enhancement layer. To enable full scalability for video coding, two hierarchical predictive structures for progressive dictionary learning are proposed to consider local and long-term motion. In each temporal layer, non-reference frames in spatial EL are reconstructed based on the spatio-temporal dictionary progressively updated with the reference frames and reconstructed non-reference frames. This hierarchical structure can adaptively capture varying structures along the motion trajectory in reconstruction. In comparison to the previous work [37], the proposed method contributes to progressive dictionary learning for scalable video coding, where generation of training set with motion compensation, training process and dictionary-based prediction are improved and optimized to enable flexible learning and predictive structures for spatial and temporal scalability.

To validate the efficacy of this work, the proposed framework is integrated with standardized codec cores to accommodate with the state-of-the-art scalable schemes like SHVC and scalable extension of H.264/AVC. Experimental results show that the proposed scheme outperforms SHVC and HEVC simulcast over extensive test sequences with various resolutions and is competitive with HEVC in the low bit-rate region.

The rest of this paper is organized as follows. Section III presents the proposed framework for scalable video coding with spatio-temporal dictionary learning. In Section IV, progressive dictionary learning is formulated to enable spatial and temporal scalability. Experimental results are shown in Section V for validation. Finally, we conclude the contributions and discuss future work in Section VI.

II. Preliminaries

Dictionary learning has been develop to improve the upsampling performance in the hybrid framework. Consider that each GOP is divided into the first L reference frames (RFs) $\mathbf{X}_R$ and remaining non-reference frames (NRFs) $\mathbf{Z}_{NR}$ for separate coding. Here, $\boldsymbol{X}_R$ are encoded at their original resolution and utilized to train the spatio-temporal overcomplete dictionary to reconstruct the high-frequency details from downsampled version of $\mathbf{Z}_{NR}$. For video sequences, adaptive regularized dictionary learning [37] was developed to make spatial-temporal sparse representation based on dictionary pairs for spatial structures along motion trajectory.

$$\min_{\mathbf{D}_L,\mathbf{D}_H,\alpha_n} \sum_{n=1}^{N}\sum_{t=1}^{T}\left[\frac{1}{2}\|Z_l^n-\mathbf{D}_L\alpha_n^{(t)}\|_2^2+\lambda_i\|\alpha_n^{(t)}\|_0+\|\mathbf{D}_H\alpha_n^{(t)}-R_n^{(t)}X_h\|_2^2\right], \quad (1)$$

where $\{\mathbf{D}_L, \mathbf{D}_H\}$ is the trained dictionary pair for low- and high-frequency components, $\alpha_n^{(t)}$ is the sparse representation vector and R_n^t is the operator indicating block-based motion compensation. To maintain the spatio-temporal consistency along the motion trajectory,

dictionary pairs are trained based on 2-D patches and 3-D volumes, respectively. Since spatial components like edges are crucial for reconstruction performance based on dictionary learning, video sequences are decomposed into primitive and non-primitive regions. Consequently, primitive patches are extracted to represent multi-scale geometry of single frames, i.e. edges, ridge and corners. Introducing primal sketch priors, 2-D dictionary pairs are developed to capture local geometry of single frames based on the primitive patches with low dimensionality. On the other hand, non-primitive volumes are generated to maintain local spatio-temporal consistency. 3-D dictionary pair is trained for each GOP based on the 3-D volumes from reference frames. For local consistency, 3-D volumes are obtained by pairing a patch and its motion compensated version from the reference frames. Motion compensation frame interpolation is adopted with overlapped block motion compensation to obtain estimated reference frames for volume generation. For training and reconstruction, K-SVD is incorporated to iteratively optimize the dictionary and sparse representation vector, respectively.

In adaptive regularized dictionary learning, dictionary is trained from the first $L = 3$ RFs on a GOP basis and NRFs are reconstructed in a sequential manner based on the trained dictionary. This fact implies that it cannot naturally support spatial and temporal scalability. Recognizing its restriction in scalable video coding, this paper proposes progressive dictionary learning with hierarchical predictive structures in spatial and temporal dimensions. Flexible hierarchical predictive structures are developed for 3-D volume generation, dictionary learning and inter-layer prediction.

III. Scalable Video Coding With Hierarchical Predictive Structures

This section provides the proposed framework with hierarchical predictive structures for scalable video coding. Dictionary-based sparse representation is adopted to improve the prediction and reconstruction in the proposed framework. For spatial scalability, inter-layer prediction is achieved based on spatio-temporal dictionary learning to exploit motion information. Hierarchical predictive structures considering local and long-term motion are developed to enable temporal scalability with a guaranteed reconstruction performance.

A. The Proposed Framework

Fig. 1 depicts the proposed framework for scalable video coding, where sparse representation based on spatio-temporal dictionary learning is progressively leveraged for reconstructing the enhancement layer (EL). In comparison to conventional pyramidal layered methods, progressive dictionary learning is proposed to improve EL reconstruction by reducing required bits for HF details. For each GOP, the first L frames are selected as reference frames (RFs) $\mathbf{X}_R$, which are encoded at their original resolution and utilized to train the spatio-temporal overcomplete dictionary for EL reconstruction in the deocder side. In this paper, L is set to 2 to facilitate dyadic decomposition of each GOP. The base layer (BL) is generated by downsampling the group of pictures with a set of Gaussian filters. Here, we define $\mathbf{Z}_{BL}$ as the downsampled non-reference frames in BL. Standard codec cores are separately employed for $\mathbf{X}_R$ and $\mathbf{Z}_{BL}$ to produce a standardized bitstream for network

transmission, where motion information within and across GOPs is exploited to represent structures of different scales along motion trajectories in BL and EL.

In the decoder side, the non-reference frames (NRFs) are reconstructed from the decoded BL and EL. Let us denote $\hat{\mathbf{X}}_{BL}$ and $\hat{\mathbf{Z}}_{BL}$ the downsampled decoded RFs and NRFs in BL, respectively. Due to the Gaussian filter, reconstruction of NRFs in EL can be considered as a combination of signals for the low-frequency (LF) and high-frequency (HF) components. In the proposed framework, the LF component $\tilde{\mathbf{Z}}_{BL}$ can be easily obtained by scaling up $\hat{\mathbf{Z}}_{BL}$, while the HF details $\tilde{\mathbf{Z}}_{EL}$ are inferred by inter-layer prediction based on the trained dictionary pairs over the scale-up RFs $\tilde{\mathbf{X}}_{BL}$ from BL and corresponding $\hat{\mathbf{X}}_{EL}$ in EL. To maintain spatio-temporal consistency, 2-D patches from primal sketches are extracted for edges and textures, and 3-D volumes are obtained by concatenating patches from non-primitive regions to represent local motion. Specifically, concatenation for 3-D volumes involves patches from current frame for prediction and the estimation of its succeeding frame. The sparse representation of $\tilde{\mathbf{Z}}_{EL}$ is made from $\tilde{\mathbf{Z}}_{BL}$ based on the trained dictionary pair ($\mathbf{D}_{BL}$, $\mathbf{D}_{EL}$).

Furthermore, progressive dictionary update is proposed to achieve scalable functionality in the temporal domain. For each temporal layer, its spatio-temporal dictionary is progressively updated based on the inter-layer prediction from previous temporal layers for improved reconstruction performance. Inspired by hierarchical B-pictures in hybrid framework, two hierarchical predictive structures are developed to guarantee the causality and availability of RFs across multiple temporal layers. To balance the performance and efficiency of dictionary learning, online learning is adopted to update the dictionaries for each temporal layer with enhanced convergence rate.

The proposed framework can benefit scalable video coding in two aspects. First, inter-layer prediction can be improved with the spatio-temporal dictionary learning by reducing bits for HF details required in the conventional schemes. Second, it enables a flexible framework with competitive performance in comparison to the dictionary learning based video coding schemes. In Section III-B and III-C, we elaborate inter-layer prediction and hierarchical predictive structure, respectively. Section IV formulates progressive dictionary learning for the proposed framework.

B. Inter-Layer Prediction for Spatial Scalability

In this section, we develop inter-layer prediction based on spatio-temporal dictionary learning for the two-layer scalable architecture. Inter-layer prediction exploits the correlations between BL and EL to reconstruct the NRFs in EL, where 2-D primitive patches and 3-D non-primitive volumes are adopted to capture local geometry and maintain spatio-temporal consistency along motion trajectory, respectively. We assume that each patch/volume is obtained by linearly combining a small subset of patches/volumes with coefficients $\alpha_n \in \mathbb{R}^M$ for the n-th one. Here, reconstruction is based on overlapped patches/volumes to improve reconstruction performance with an acceptable efficiency.

1) 2-D Primitive Patches for Local Geometry—For primal sketches, 2-D patches are represented based on the instinctive features of block-based structures classified by K

Gaussian filters. To capture local geometry, a set of K 2-D dictionary pairs $\{\mathbf{D}_{bl}^k, \mathbf{D}_{el}^k\}$, $1 \leq k \leq K$ are generated by clustering the 2-D primitive patches extracted along the primal sketches. Given the scale-up NRFs $\tilde{\mathbf{Z}}_{EL}$ from BL and trained dictionary pair $\{\mathbf{D}_{bl}^k, \mathbf{D}_{el}^k\}$, the lost HF details for EL are recovered by

$$\{\tilde{\alpha}_n, \tilde{\mathbf{z}}_n^h\} = \arg \min_{\{\alpha_n^k, \mathbf{z}_n^h\}} \sum_{n=1}^{N} \left[\frac{1}{2}\|\mathbf{z}_n^l - \mathbf{D}_{bl}^k \alpha_n^k\|_2^2 + \lambda\|\alpha_n^k\|_0 + \|\mathbf{D}_{el}^k \alpha_n^k - \mathbf{z}_n^h\|_2^2\right], \quad (2)$$

where $\mathbf{z}_n^l$ is the n-th patch extracted from $\tilde{\mathbf{Z}}_{BL}$, $\mathbf{z}_n^h$ is the corresponding prediction for $\tilde{\mathbf{Z}}_{EL}$, and λ is a regularization parameter. In Eq. (2), the first and third terms evaluate the approximation and reconstruction error for the scale-up (LF) and HF frame, respectively, and the second term restricts the number of coefficients to maintain sparsity. $\tilde{\mathbf{Z}}_{EL}$ is reconstructed by collecting $\tilde{\mathbf{z}}_n^h$ that are obtained with joint optimization over patches from the scale-up (LF) BL and EL.

2) 3-D Non-Primitive Volumes for Spatio-Temporal Consistency—In non-primitive regions, 3-D volumes are predicted by block-matching based motion estimation to maintain the consistency of the motion trajectory based on incomplete visual patterns. Given trained 3-D dictionary pair $\{\mathbf{D}_{BL}, \mathbf{D}_{EL}\}$, the minimization function in Eq. (2) can be further extended to video sequences by considering the spatio-temporal consistency along motion trajectory.

$$f(\alpha_n, \hat{\mathbf{z}}_n^h, \mathbf{D}_{EL}, \mathbf{D}_{BL}) = \min_{\hat{\mathbf{z}}_n^h, \alpha_n} \sum_{i,j} \sum_{t} \left[\frac{1}{2}\|\mathcal{R}_n \tilde{\mathbf{z}}_n^l - \mathbf{D}_{BL}\alpha_n\|_2^2 + \lambda\|\alpha_n\|_0 + \|\mathbf{D}_{EL}\alpha_n - \mathcal{R}_n \hat{\mathbf{z}}_n^h\|_2^2\right] (3)$$

where $\alpha_n(i, j, t)$ denotes the identical sparse coefficients for dictionaries, $\mathcal{R}_n(i, j, t)$ is a projection matrix extracting a patch at time t and location (i, j). Under the assumption of sparse priors, sparse representation α_n is derived based on the 3-D volumes from scale-up frame $\tilde{\mathbf{Z}}_{BL}$ to reconstruct the HF details $\hat{\mathbf{z}}_n^h$, as shown in Fig. 2. Considering motion trajectory, 3-D volumes are obtained by concatenating the patches from the non-primitive $\tilde{\mathbf{Z}}_{BL}$ and its motion compensated estimation from RFs and reconstructed NRFs.

3) Motion Compensation for Volume Generation—To accommodate standardized codecs like SHVC and scalable extension of H.264, bidirectional motion estimation and compensation is adopted to generate 3-D volumes for spatio-temporal consistency of local structures. Given current frame to be reconstructed, its own patches $\tilde{\mathbf{z}}_n^l$ and corresponding motion-compensated estimation $\bar{\mathbf{z}}_n^l$ are concatenated to construct 3-D volumes. Since the spatio-temporal consistency tends to be maintained in both BL and EL, the motion

estimation and compensation in EL is performed based on the motion vector derived in BL. For each NRF $\hat{\mathbf{Z}}_{BL}$ in BL motion compensated frame interpolation (MCFI) is leveraged to obtain its estimation $\bar{\mathbf{Z}}_{BL}$ from its preceding and succeeding reconstructed frames $\mathcal{F}'_{BL}$ and $\mathcal{F}''_{BL}$. Here, $\mathcal{F}_{BL}$ could be RFs or reconstructed NRFs complying with decoding order.

Given arbitrary matching block $\mathbf{z}_{i,j}$ locating at (i, j), its candidate motion vector $\mathbf{v}$ is evaluated for motion estimation in terms of SAD. Without loss of generality, we first consider forward motion vector. To accommodate with HEVC, the optimal forward MV is recursively searched in a hierarchical manner for $\mathbf{z}_{i,j}$ and all its sub-blocks with sizes ranging from 64 × 64 to 4 × 4. Here, search window is adaptive with the size of (sub-)blocks to balance performance and complexity. Similarly, the backward MVs could also be acquired.

When MVs are obtained, $\bar{\mathbf{Z}}_{BL}$ is estimated with overlapped block motion compensation (OBMC [40]) to reduce unnatural artifacts. OBMC utilizes the MVs to determine a group of neighboring blocks for weighted estimation. The pixel $I_n(i', j')$ of $\mathbf{z}_{i,j}$ in the n-th frame can be esitmated by weighting the matched blocks ($\mathbf{z}_{i,j}$ and four adjacent blocks) with derived MVs in the n'-th frame.

$$I_n(i', j') = \sum_{p=-1}^{1} \sum_{q=-1}^{1} w_{p,q}(i', j')\, I_{n'}(i'', j''), \quad (4)$$

where $w_{p,q}(\mathbf{s})$ is the normalized weights for blocks and $(i'', j'') = (i', j') + \mathbf{v}(i+p, j+q)$ is determined by the MV $\mathbf{v}(i+p, j+q)$ for the block $\mathbf{z}_{i+p,j+q}$. Furthermore, the weights $w_{p,q}$ can be adjusted based on the reliability of MVs in compensating current block.

$$\hat{w}_{p,q}(i', j') = \frac{\Phi_{i,j}(\mathbf{v}_{i+p,j+q}) w_{p,q}(i', j')}{\sum_{s=-1}^{1} \sum_{t=-1}^{1} \Phi_{i,j}(\mathbf{v}_{i+s,j+t}) w_{s,t}(i', j')} \quad (5)$$

Here, $\Phi_{i,j}(v_{i+p,j+q})$ is the reliability of MV for $\mathbf{z}_{i+p,j+q}$ in proportion to $\mathbf{z}_{i,j}$, which is defined as the ratio of the prediction error $\varepsilon_{i,j}$ of current block $\mathbf{z}_{i,j}$ using current MV $\mathbf{v}_{i,j}$ against the one using neighboring MV $\mathbf{v}_{i+p,j+q}$.

$$\Phi_{i,j}(\mathbf{v}_{i+p,j+q}) = \frac{\varepsilon_{i,j}(\mathbf{v}_{i,j})}{\varepsilon_{i,j}(\mathbf{v}_{i+p,j+q})}$$

Using OBMC, $\bar{\mathbf{z}}_{i,j}$ in $\bar{\mathbf{Z}}_{BL}$ is estimated by comparing the weighted estimates derived by forward and backward MVs in the sense of SAD.

It is worth mentioning that the selection of preceding and succeeding frames for motion estimation and compensation could be adaptive with the decoding (reconstruction) order of

frames within or across GOPs. In Section III-C, we extend the motion estimation and compensation scheme to accommodate the hierarchical predictive structures for temporal scalability.

4) Equivalency to Inter-Layer Motion—Contrary to conventional scalable video coding framework, the proposed method does not explicitly perform inter-layer motion. In progressive dictionary learning, the pairs of 3-D trained dictionaries actually maintain a projection from BL to EL compensated with local motion. Let us denote $\mathcal{P}_{i,j,t}$ the projection operator for a patch at time t and location (i, j). Given arbitrary patch or volume, its motion vector $\mathbf{v}_{i,j,t}$ for BL can be determined based on RFs to minimize metrics like SAD. When $\mathbf{v}_{i,j,t}$ is employed in EL based on the assumption of sparse priors, the equivalent motion vector in EL can be derived as $\mathcal{P}_{i,j,t}\mathbf{v}_{i,j,t}$.

The proposed progressive dictionary learning method implicitly makes motion compensated prediction in EL reconstruction for NRFs. For each GOP, dictionary pairs are initially trained based on the patches and volumes from RFs. Consequently, reconstruction of each volume in EL for NRFs is based on the representation vector α derived by approximating the corresponding volume in BL for NRFs with a combination of volumes in BL for RFs. Thus, the presentation vector explicitly represents the motion for the specific patch from RFs. Since the dictionary pair is updated progressively, current frame tends to be estimated from the decoded frames. This fact implies that motion compensated prediction is implicitly performed in the EL construction based on the trained dictionary pair. Given the BL for RFs and NRFs, The projection operator $\mathcal{P}_{i,j,t}$ explicitly maps the base layer motion with the dictionary pair. In the proposed method, $\mathcal{P}_{i,j,t}$ is adaptively trained based on the training set.

C. Hierarchical Predictive Structure for Temporal Scalability

In previous dictionary learning based video coding schemes, NRFs are reconstructed in a sequential manner based on the dictionary trained from RFs. However, hierarchical B-pictures for hybrid framework cannot be directly employed, as 3-D volumes from each NRF are generated with MCFI based on previously reconstructed NRFs. In this section, we develop hierarchical predictive structures (HPS) that enable temporal scalability for progressive dictionary learning considering local and long-term motion information.

As shown in Fig. 1, it is a dominant problem to reconstruct NRFs $\hat{\mathbf{Z}}_{EL}$ using their sparse representation priors along the temporal dimension in addition to local structures between adjacent frames. The coding performance of ELs can be improved with the trained dictionaries that capture regular features to maintain spatio-temporal consistency along motion trajectories. To allow temporal scalability, we propose hierarchical prediction of NRFs based on the dictionary derived from the RFs and the reconstructed NRFs in the previous temporal layers. Under the proposed HPS, NRFs in each temporal layer can be independently reconstructed from RFs and NRFs in previous temporal layers. Furthermore, sequential learning can be leveraged for each temporal layer to update dictionaries based on previously reconstructed NRFs, so that the reconstruction performance of NRFs, especially those are distant from RFs in the GOP, can be improved.

1) Hierarchical Predictive Structure for Local Motion—Couples of frames are generated by concatenating decoded RFs and NRFs for spatio-temporal dictionary learning in a dyadic case. To consider local motion, we propose HPS-LO to extract 3-D volumes from two adjacent frames by concatenating the most matched patches in two adjacent RFs along the motion trajectory. RFs have the highest priority in encoding and decoding, which constitute the temporal BL T_0 and represent the coarsest supported temporal resolution. Subsequently, it can be refined by inserting one or more temporal EL for finer temporal levels. Each couple of adjacent frames is served as one B-picture to support refined prediction and dictionary update. Given the GOP with a size of 2^{Λ}, it can be divided into Λ temporal layers: one BL and $\Lambda-1$ ELs. Denote $T_0, \cdots, T_{\Lambda-1}$ the Λ temporal layers. For arbitrary $0 \leq \kappa < \Lambda$, T_{κ} produces an independently decodable bitstream with a ratio $(2^{\kappa-\Lambda+1})$ of the full frame rate for the GOP. To allow temporal scalability, motion-compensated prediction is restricted to reconstructed frames with a temporal layer identifier $\kappa' \leq \kappa$. Fig. 3(a) and (b) show the hierarchical predictive structures for a GOP with 16 frames under the hybrid MC-DCT and proposed framework.

2) Hierarchical Predictive Structure for Long-Term Motion—Besides HPS-LO, we propose HPS-LT to consider long-term motion across multiple GOPs. In HPS-LT, RFs and NRFs with corresponding positions in two adjacent GOPs are coupled for spatio-temporal dictionary learning. Fig. 3(c) provides an example for two GOPs with size 16. Given the GOP size 2^{Λ}, the trained dictionary tends to represent the motion with a gap of 2^{Λ} frames. The two adjacent GOPs can be reconstructed simultaneously with the hierarchical predictive structure for dictionary learning, where the required number of RFs can be reduced to one. Analogous to HPS-LO, the two GOPs can be divided into $\Lambda+1$ temporal layers for scalability. For temporal ELs T_{κ}, its 2^{κ} frames can be independently reconstructed to update dictionaries for $T_{\kappa+1}$.

In temporal layer T_{κ}, HPS-LO and HPS-LT perform motion estimation and compensation based on the RFs and reconstructed NRFs from $T_0, \cdots, T_{\kappa-1}$. In comparison to previous dictionary learning methods, each temporal layer can be progressively decoded and reconstructed for scalable frame rate. The training set is enlarged for each temporal layer by extracting volumes from couples of reconstructed NRFs. The atoms of the 3-D spatio-temporal dictionary are progressively updated from T_0 to $T_{\Lambda-1}$ (for HPS-LO) or T_{Λ} (for HPS-LT). Both alternatives are desirable for high-definition applications, as enriched prior knowledge can facilitate NRF reconstruction to achieve high-fidelity visual quality and improved rate-distortion performance. For each temporal layer, spatial scalable functionality is preserved using inter-layer prediction, as spatio-temporal dictionary learning is only performed on the reconstructed NRFs from previous temporal layers.

In summary, HPS-LO tends to achieve better reconstruction performance by considering the local motion between two adjacent frames. However, HPS-LO cannot provide arbitrary frame rate as the traditional hierarchical B-pictures, as it requires at least two RFs for each GOP and an increment proportional to two NRFs. By contrast, HPS-LT would degrade the reconstruction of ELs by only considering long-term motion between GOPs. In temporal domain, it is fully scalable, as only one RF is necessary for each GOP.

IV. Progressive Dictionary Learning With Hierarchical Prediction Structure

In this section, we elaborate the progressive dictionary learning with HPS for spatial and temporal scalability, including spatio-temporal dictionary learning for inter-layer prediction and progressive learning for temporal dictionary update. Integrating these methods, practical algorithm for progressive dictionary learning is developed for scalable video coding.

A. Spatio-Temporal Dictionary Learning for Inter-Layer Prediction

Fig. 4 illustrates spatio-temporal dictionary learning for inter-layer prediction. To exploit correlations along the motion trajectory, each frame in BL is segmented into a primitive layer, a non-primitive coarse layer, and a non-primitive smooth layer [37]. 2-D patches and 3-D volumes are collected from the primitive and non-primitive layers for local geometry and spatio-temporal consistency in the video sequences, respectively. In the decoder side, $\hat{\mathbf{X}}_R$ are used to learn a set of K 2-D sub-dictionary pairs ($\mathbf{D}_{bl}^k, \mathbf{D}_{el}^k$), $k = 1, \cdots, K$, and a 3-D dictionary pair ($\mathbf{D}_{BL}$, $\mathbf{D}_{EL}$) for EL reconstruction based on correlations across different spatial layers.

2-D sub-dictionary pairs are trained to capture low-dimensional local geometry like edge segments and textures in a multi-scale manner. The training set is generated by extracting 2-D patches at the same positions from corresponding frames in BL and EL. In training, 2-D primitive patches are clustered with the edge orientation in a scale space to specify such local geometry. A series of K Gaussian filters with various scales and orientations are introduced to record the maximum filter response and label the primitive patches along the primal sketch. For the k-th cluster $\mathcal{S}_{bl}^k$ and $\mathcal{S}_{el}^k$, ($\mathbf{D}_{bl}^k, \mathbf{D}_{el}^k$) is constructed with $\ell_{2,1}$ optimization.

$$\min_{\mathbf{D}_{bl}^k, \mathbf{D}_{el}^k, a^k} \|\mathcal{S}_{bl}^k - \mathbf{D}_{bl}^k \alpha^k\|_2^2 + \|\mathcal{S}_{el}^k - \mathbf{D}_{el}^k \alpha^k\|_2^2 + \|\alpha^k\|_1, \quad (6)$$

Here, α^k is the sparse representation vector for the k-th cluster.

To further exploit temporal correlations in inter-layer prediction, spatio-temporal dictionary pair ($\mathbf{D}_{BL}$, $\mathbf{D}_{EL}$) is trained to represent observed visual patterns in non-primitive regions. To maintain the spatio-temporal consistency, 3-D non-primitive volumes are extracted along the motion trajectory. Similar to inter-layer prediction, motion-compensated frame interpolation is adopted to generate 3-D volumes by concatenating decoded patches and its matched version with bidirectional motion estimation from RFs. $\mathbf{D}_{BL}$ and $\mathbf{D}_{EL}$ are jointly optimized over the training set of N volumes.

$$\min_{\mathbf{D}_{BL}, \mathbf{D}_{EL}, \alpha_n} \sum_{n=1}^{N} \left[\frac{1}{2} \|\mathbf{x}_n^{BL} - \mathbf{D}_{BL} \alpha_n\|_2^2 + \lambda \|\alpha_n\|_0 + \|\mathbf{D}_{EL} \alpha_n - \mathcal{R}_n \hat{\mathbf{x}}_n^{EL}\|_2^2 \right] \quad (7)$$

Here, $\mathbf{x}_n^{BL}$ and $\mathbf{x}_n^{EL}$ are volumes from the RFs and reconstructed NRFs. In optimization, iterative batch algorithms like K-SVD [41] would suffer from a high computational complexity and degraded coding efficiency due to minimizing the empirical cost over the entire training set. To achieve an enhanced convergence rate with a guarantee of approximation error, stochastic approximation [42] is leveraged to directly minimize the expected cost with dynamical update based on the randomly selected training data.

$$\min_{\mathbf{D}_{BL},\mathbf{D}_{EL},\alpha_n^t} \sum_{n=1}^{N}\sum_{t=1}^{\tau}\left[\frac{1}{2}\|\mathbf{x}_n^{BL}-\mathbf{D}_{BL}\alpha_n^t\|_2^2+\lambda\|\alpha_n^t\|_0+\|\mathbf{D}_{EL}\alpha_n^t-\mathscr{R}_n^t\mathbf{x}_n^{EL}\|_2^2\right] \quad (8)$$

where α_n^t is the representation coefficients at the t-th iteration.

Furthermore, we demonstrate the consistency for inter-layer prediction based on the trained dictionary. Under the two-layer architecture, the reconstructed NRFs with HF details from EL can be represented based on the trained dictionary pair.

$$\widehat{\mathbf{Z}}_{EL}=\widetilde{\mathbf{Z}}_{BL}+\widetilde{\mathbf{Z}}_{EL}=(\mathbf{D}_{BL}+\mathbf{D}_{EL})\alpha \quad (9)$$

In Proposition 1, we show that the approximation error $\|\hat{\mathbf{Z}}_{EL}-(\mathbf{D}_{BL}+\mathbf{D}_{EL})\alpha\|_2$ is related to $\|\tilde{\mathbf{Z}}_{BL}-\mathbf{D}_{BL}\alpha\|_2$ for BL.

Proposition 1—*Given the trained dictionary $\mathbf{D}_{BL}$ and $\mathbf{D}_{EL}$ for the base and enhancement layer, with sufficient sampling, the approximation error for the reconstructed NRFs with EL can be upper-bounded by the one for BL.*

Proof: *Please refer to* Appendix A.

Proposition 1 demonstrates that the consistency for inter-layer prediction can be maintained with the proposed dictionary learning method. It implies that the enhancement layer can be reconstructed based on the sparse representation coefficients derived from the trained dictionary $\mathbf{D}_{BL}$ for base layer. Since the scaled-up base layer $\tilde{\mathbf{Z}}_{BL}$ and $\mathbf{D}_{BL}$ are known in the inter-layer prediction, the reconstruction error could be minimized based on the sparse representation.

B. Progressive Learning for Temporal Dictionary Update

Fig. 5 depicts the progressive dictionary learning procedures for HPS. Decoded RFs $\hat{\mathbf{X}}_R$ are used to train the initial dictionary pair ($\mathbf{D}_{BL}^0, \mathbf{D}_{EL}^0$). Thus, spatio-temporal sparse representation can be made for NRFs $\widehat{\mathbf{Z}}_{BL}^0$ in BL based on the initial 3-D dictionaries $\mathbf{D}_{BL}^0$ and $\mathbf{D}_{EL}^0$. Subsequently, the reconstructed NRFs $\widehat{\mathbf{Z}}_{EL}^0$ in EL are decomposed to generate 3-D non-primitive volumes and the dictionary pair is updated for predicting the enhancement

layer T_1. The reconstructed $\widehat{\mathbf{Z}}_{BL}^0$ and $\widehat{\mathbf{Z}}_{EL}^0$ are taken as reference frames to make motion estimation and compensation. MCFI is performed within the frames (RFs and NRFs) obtained from T_0 and T_1 to generate the training set of 3-D volumes by concatenating the motion-compensated frames $\overline{\mathbf{Z}}_{BL}^0$ and $\overline{\mathbf{Z}}_{EL}^0$. As a consequence, the reconstructed NRFs in a coarser temporal layer T_κ are utilized as reference frames to sequentially update the dictionary and predict non-reference pictures in a finer layer $T_{\kappa+1}$. To guarantee the efficiency of dictionary learning, online dictionary learning is leveraged to recursively derive the dictionaries $\mathbf{D}_{BL}^{\kappa+1}$ and $\mathbf{D}_{EL}^{\kappa+1}$ with newly-added volumes in T_κ.

$$\mathbf{D}_{BL}^{\kappa+1} = \prod_{\mathscr{C}} [\mathbf{D}_{BL}^{\kappa} - \delta_t \nabla_{\mathbf{D}} l(\mathbf{z}^{\kappa}, \mathbf{D}^{\kappa})], \quad (10)$$

where $\mathscr{C}$ is a convex constraint set to prevent $\mathbf{D}$ from being arbitrarily large by restraining the ℓ_2-norm of dictionary atoms $\mathbf{d}_m$, $1 \le j \le M$ not greater than one and $l(\mathbf{z}^{\kappa}, \mathbf{D}^{\kappa}) = \min_{\alpha} \frac{1}{2}\|\mathbf{z}^{\kappa} - \mathbf{D}^{\kappa}\alpha\|_2^2 + \lambda\|\alpha\|_1$ is the regularized approximation error for the κ-th temporal layer. When $\mathbf{D}_{BL}^{\kappa+1}$ is obtained, the corresponding dictionary $\mathbf{D}_{EL}^{\kappa+1}$ for EL is derived by minimizing the representation error $\min_{\mathbf{D}_{EL}} \{\|\mathcal{S}_{EL} - \mathbf{D}_{EL}\alpha\|_F^2\}$ with a row full-rank coefficient matrix α.

$$\mathbf{D}_{EL} = \mathcal{S}_{EL}\alpha^{+} = \mathcal{S}_{EL}\alpha^{T}(\alpha\alpha^{T})^{-1} \quad (11)$$

Eq. (11) implies that the approximation error led by the progressive dictionary learning tends to depend on the trained dictionary for base layer. For simplicity, we use $\mathbf{D}^{\kappa}$ to represent the updated dictionary $\mathbf{D}_{BL}^{\kappa}$ for BL based on the volumes from the κ-th reconstructed EL.

The progressive dictionary update can improve the reconstruction performance with the hierarchical predictive structure. For each temporal layer, volumes from reconstructed NRFs would enrich the training set for progressive update of dictionary basis. This fact implies that the sparse priors tend to approach the actual statistics of the GOP. Moreover, online dictionary learning can aggregate the structures in current temporal layer and propagate the accumulated results from all the previous layers. Thus, reconstruction of NRFs in higher temporal layer can be refined with a warm restart to reduce the required number of iterations for convergence. As a support, we demonstrate in Proposition 2 that the approximate cost $f_t(\mathbf{D}^{\kappa}) = \frac{1}{t}\sum_{n=1}^{t} l(\mathbf{x}_n^{(\kappa)}, \mathbf{D}_n^{\kappa})$ for the progressive dictionary learning asymptotically converges to the expected cost $f(\mathbf{D}^*) = \int_{\mathbf{x}} l(\mathbf{x}, \mathbf{D}^*)\, \mathrm{d}p(\mathbf{x})$ w.r.t. the actual distribution $p(\mathbf{x})$ of the GOP consisting of $\mathbf{X}_R$ and $\mathbf{Z}_{NR}$.

Algorithm 1

Progressive Dictionary Learning Algorithm

Input: Decoded reference pictures $\hat{\mathbf{X}}_R$ in BL and EL, decoded NRFs in BL, and regularization parameter λ.

Extract i.i.d samples $\mathbf{x}^0 \in \mathbb{R}^{M \times N}$ of distribution $p(\mathbf{x})$ from $\hat{\mathbf{X}}_R$.

Initialization: $\mathbf{D}^0$.

Inter-layer reconstruction: Recover the frames $\widehat{\mathbf{Z}}_{EL}^{(1)}$ in T_0 from $\hat{\mathbf{X}}_R$ using $\mathbf{D}^0$.

for $\kappa = 1$ to $\Lambda - 1$ **do**

  Progressive dictionary learning:

  $\mathbf{D}_0^k = \mathbf{D}^{k-1}$.

  for $t = 1$ to τ **do**

    Draw $\mathbf{x}_t^{(\kappa)}$ from the pictures $\widehat{\mathbf{Z}}_{EL}^{(\kappa)}$ in the κ-the temporal layer in EL T_κ.

    Sparse coding: LARS-Lasso algorithm to solve Eq. (14).

    Dictionary update: Block-coordinate descent with warm restart $\mathbf{D}_{t-1}^{\kappa}$ to optimize Eq. (13).

  end for

  $\mathbf{D}^{\kappa} = \mathbf{D}_{\tau}^{\kappa}$.

  Inter-layer reconstruction: Recover the frames $\widehat{\mathbf{Z}}_{EL}^{(\kappa+1)}$ in T_κ from $\widehat{\mathbf{Z}}_{BL}^{(\kappa+1)}$ using $\mathbf{D}^{\kappa}$.

End for

Proposition 2—*In progressive update for dictionary* $\mathbf{D}_{BL}$*, given volumes* $\mathbf{z}$ *randomly selected from reconstructed frames, the divergence between the approximate cost* $f_k(\mathbf{D}_k)$ *and expected cost* $f(\mathbf{D})$ *converges almost surely to zero with sufficient sampling.*

Proof: *Please refer to* Appendix B.

Proposition 2 implies that the progressive dictionary learning can asymptotically achieve the equivalent approximation performance in comparison to the methods over the entire GOP. Since the dictionary is updated for the reconstructed frames in each temporal layer in an online manner, the efficiency of dictionary learning can be improved for video sequences with large-scale time-varying structures.

C. Main Algorithm

Algorithm 1 elaborates the progressive dictionary learning. Denote $\{\mathbf{x}_1^{(\kappa)}, \cdots, \mathbf{x}_{N_\kappa}^{(\kappa)}\}$ the training set of 3-D volumes for the temporal layer T_κ, $0 \le \kappa \le \lambda - 1$. From BL T_0 to the highest EL $T_{\lambda-1}$ in temporal domain, 3-D volumes are randomly and independently sampled from the underlying distribution $p(\mathbf{x})$ to constitute the training set. Frames located in the highest temporal layer T_λ are always coded as non-reference frames. The dictionary $\mathbf{D}^0$ is

initialized with the samples extracted from RFs $\hat{\mathbf{X}}_R$ in BL. Given arbitrary temporal layer T_κ, the progressive dictionary learning is iteratively performed to address the $\ell_{2,1}$ joint optimization problem defined in Eq. (12), where recursive optimization is formulated to make sparse coding for frames in current layer based on the trained dictionary $\mathbf{D}^\kappa$ and update the dictionary atoms to fit the newly-added training data.

$$\hat{f}_t(\mathbf{D}_t) = \min_{\mathbf{D} \in C, \alpha_n^t \in \mathbb{R}^M} \frac{1}{t} \sum_{n=1}^{t} \left[\frac{1}{2} \|\mathbf{x}_n - \mathbf{D}\alpha_n^t\|_2^2 + \lambda \|\alpha_n^t\|_1 \right] \quad (12)$$

At each iteration t of the inner loop (Step 7–11) in Algorithm 1, one training example $\mathbf{x}_t^{(\kappa)}$ is randomly picked from $\widehat{\mathbf{X}}_R^\kappa$ for current temporal layer, which is enlarged with reconstructed NRFs at the outer loop (Step 12–13) of previous iteration. Given the trained dictionary $\mathbf{D}_{t-1}^\kappa \in \mathbb{R}^{N \times M}$, the sparse representation vector $\alpha_t \in \mathbb{R}^M$ of $\mathbf{x}_t^{(\kappa)}$ is optimized for the ℓ_1-regularized linear least-squares problem in Eq. (14). Here, the optimization problem is solved by a Cholesky-based LARS-Lasso algorithm for high accuracy and fast implementation [43], [44].

Subsequently, the ℓ_2-constrained least-squares problem with the projected constraint set $\mathscr{C}$ is solved through the block-coordinate descent with a warm restart $\mathbf{D}_{t-1}^\kappa$. The atoms of $\mathbf{D}_t^\kappa$ are iteratively updated with the parameter-free estimation in a column-by-column manner.

$$\mathbf{D}_t^\kappa = \arg \min_{\mathbf{D} \in \mathscr{C}} \frac{1}{t} \sum_{n=1}^{t} \frac{1}{2} \|\mathbf{x}_n^{(\kappa)} - \mathbf{D}\alpha_n\|_2^2 + \lambda \|\alpha_n\|_1. \quad (13)$$

$$\alpha_t = \arg \min_{\alpha \in \mathbb{R}^M} \frac{1}{2} \|\mathbf{x}_t^{(\kappa)} - \mathbf{D}_{t-1}^\kappa \alpha\|_2^2 + \lambda \|\alpha\|_1, \quad (14)$$

In Fig. 5, we provide an example for the progressive dictionary learning with 4 dyadic hierarchical levels. A group of pictures (GOP) is decomposed into two reference frames (RFs) and the other non-reference frames (NRFs). NRFs are progressively reconstructed from the RFs of the previous and current GOP with hierarchical predictive structure HPS-LO. In each temporal layer, the 3-D dictionary pair is updated based on the reconstructed NRFs from previous layer.

V. Experimental Results

A. Implementation

For validation, we integrate the proposed framework with the latest SHVC Test Model 7 (SHM 7 [46]), where progressive dictionary learning is proposed to reconstruct NRFs in EL from BL based on RFs. The proposed scheme is evaluated over extensive test sequences with the YUV 4:2:0 format and various resolutions including CIF (352 × 288), WQVGA (416 × 240), standard definition (832 × 480), full high definition (1920×1080) and WQXGA (2500×1600). In evaluations, each GOP is composed of 16 successive pictures, where the first two adjacent frames are selected as reference frames. The base layer is generated via spatial down-sampling with a factor of 2, but the scheme is also applicable to other downsampling factors. The default low delay B (LDB) and random access (RA) configurations for main profile are adopted for HEVC and SHVC, respectively. The GOP size for HEVC and SHVC is set to 16. For the proposed method, RFs and NRFs are separately encoded by the HEVC codec cores in the order of "$IBBB\cdots$" with an intra-period of 12. Thus, the HPS-LO and HPS-LT require 3 and 4 temporal layers for each GOP, where RFs in the current and next GOPs are decoded as BL to train spatio-temporal dictionary pairs. For BL and EL, the maximum deltaQP is set to 4. In training, block sizes for motion estimation and compensation can vary from 64 × 64 to 4 × 4 to accommodate HEVC. For each temporal layer, the training set is generated by collecting 1024 volumes from each frame. The volume size is set to 13 × 13 × 2. The regularization parameter λ is 0.15.

B. Rate-Distortion Performance

Fig. 6–9 show the rate-distortion curves of various test sequences obtained by the proposed scheme, SHVC, HEVC simulcast and HEVC under low-delay B (LDB) and random access (RA) configurations, respectively. The proposed scheme noticeably outperforms SHVC and HEVC simulcast in most cases. In the low bit-rate region, it is demonstrated to be competitive with HEVC, which does not support spatial scalability. For a complete validation, Table II and III provide the rate-distortion performance in terms of BD-PSNR gain and BD-rate change [47] for the proposed scheme, SHVC, and HEVC simulcast over HEVC. For HEVC simulcast, the rate-distortion performance is obtained by assigning same QP value to both resolutions. In comparison to SHVC, the proposed scheme can achieve BD-PSNR gain and BD-rate reduction by up to 3.0 dB and 65%, respectively. It is worth mentioning that the coding performance obtained by SHVC is degraded for CIF test sequences due to overheads for CU semantic elements and motion vectors. By contrast, the proposed scheme can maintain the coding performance for test sequences with a wide range of resolutions.

Moreover, we evaluate the two alternatives for hierarchical predictive structures HPS-LO and HPS-LT for progressive dictionary learning. Fig. 10 provides rate-distortion curves for video sequences with various resolutions under random access configuration. It shows that HPS-LO and HPS-LT can achieve a gain in rate-distortion performance in comparison to SHVC and HEVC simulcast in a wide range of bit rates. In most cases, HPS-LO outperforms HPS-LT by focusing on local motion for prediction and reconstruction. This

fact implies that HPS-LT tends to enable fully scalability in temporal domain at the cost of reconstruction performance. It is worth mentioning that the rate-distortion performance is not significantly degraded in HPS-LT, which is rooted from the fact that spatial structures are mainly reconstructed based on the set of 2-D sub-dictionary pairs.

C. Visual Quality

Fig. 11 shows the reconstructed frames for sequences *News*, *BlowingBubbles*, *BQMall* and *ParkScene* obtained by the proposed scheme, SHVC, and HEVC simulcast, respectively. The proposed scheme is shown to obtain better visual quality in comparison to SHVC and HEVC simulcast, especially in the regions of texture and edges. To make further evaluation, SSIM [48] is introduced for quantitive measurement, where higher SSIM score represents better visual quality. Fig. 12 and 13 show the curves of SSIM vs bit-rates for the four sequences under the low-delay B and random access configuration, respectively. Under both configurations, the proposed scheme is shown to outperform SHVC and HEVC simulcast in a noticeable gap in terms of SSIM.

D. Computational Complexity

In this section, we discuss the the computational complexity for the proposed scheme, SHVC and HEVC simulcast. As mentioned in Section III and IV, the complexity of the proposed scheme mainly comes from the dictionary learning and reconstruction process. In learning phase, the proposed scheme commits a fast convergence speed to update the dictionary with the enlarged training data in a progressive manner. Decoding performance depends on the reconstruction of NRFs with inter-layer prediction. Since the proposed scheme directly optimizes the expected cost for progressive dictionary learning, it obtains sparser coefficients with a reduced number for iterative training in comparison to the batch-based dictionary learning methods like K-SVD.

To keep fairness, the encoder and decoder are realized with the same HEVC codec core. The experiments for dictionary learning and prediction are implemented with Matlab on a PC with 3.0 GHz dual-core CPU and 8G RAM. Table IV provides the computational complexity for the proposed scheme, SHVC, and HEVC simulcast. It should be noted that the complexity for the proposed scheme are evaluated in terms of learning speed (sec/GOP) and reconstruction speed (sec/frame). Table IV shows that progressive dictionary learning is efficient using online learning, which is not necessarily related with the resolutions of sequences. However, the proposed scheme requires about 20–100 times the computational complexity for reconstruction under various configurations in comparison to standard decoding process adopted by SHVC and HEVC simulcast. In fact, it is led by solving LASSO problem for overlapped volumes in reconstruction, which can be improved by adopting fast algorithm like selective coordinate descent [49] or introducing parallel algorithms for block-based reconstruction. It is also possible to adjust the number of overlapping pixels in each block to balance reconstruction performance and computational complexity, as overlapping pixels for each block is set to 1 in this paper for optimal reconstruction performance. Finally, it is possible to improve its running speed by transplanting into C implementation or optimizing the Matlab implementation.

VI. Conclusion

In this paper, we propose a progressive dictionary learning framework with hierarchical predictive structure for scalable video coding. Sparse representation based on spatio-temporal dictionary is leveraged for the inter-layer prediction of pyramidal spatial layers. Prediction and reconstruction of enhancement layers are improved by considering the motion trajectory under the invariance of sparse representation across various layers. Inspired by the hierarchical B-pictures adopted in conventional hybrid framework, progressive dictionary learning is developed to enable the scalability in temporal domain and restrict the error propagation in a close-loop predictor. Hierarchical predictive structure is adopted for temporal scalable layers, where the overcomplete dictionary for prediction is updated with the online learning for a guaranteed reconstruction error. For validation, the proposed progressive dictionary learning is integrated with the latest scalable extension of HEVC to accommodate the standardized bitstreams for video transmission. Experimental results show that the proposed method outperforms the latest SHVC and HEVC simulcast over extensive test sequences with various resolutions.

Acknowledgments

This work was supported in part by NSFC under Grant 61501294, Grant 61622112, Grant 61529101, Grant 61472234, and Grant 61425011, in part by the China Postdoctoral Science Foundation under Grant 2015M581617, and in part by the Program of Shanghai Academic Research Leader under Grant 17XD1401900. The work of X. Jiang was supported in part by the National Institute of Health under Award R00LM011392. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Vladimir Stankovic.

References

1. Schierl T, Stockhammer T, Wiegand T. Mobile video transmission using scalable video coding. IEEE Trans Circuits Syst Video Technol. Sep; 2007 17(9):1204–1217.
2. Wien M, Cazoulat R, Graffunder A, Hutter A, Amon P. Real-time system for adaptive video streaming based on SVC. IEEE Trans Circuits Syst Video Technol. Sep; 2007 17(9):1227–1237.
3. Schwarz H, Marpe D, Wiegand T. Overview of the scalable video coding extension of the H. 264/AVC standard. IEEE Trans Circuits Syst Video Technol. Sep; 2007 17(9):1103–1120.
4. Xiong R, Xu J, Wu F, Li S. Barbell-lifting based 3-D wavelet coding scheme. IEEE Trans Circuits Syst Video Technol. Sep; 2007 17(9):1256–1269.
5. Boyce JM, Ye Y, Chen J, Ramasubramonian AK. Overview of SHVC: Scalable extensions of the high efficiency video coding standard. IEEE Trans Circuits Syst Video Technol. Jan; 2016 26(1): 20–34.
6. Ugur K, et al. High performance, low complexity video coding and the emerging HEVC standard. IEEE Trans Circuits Syst Video Technol. Dec; 2010 20(12):1688–1697.
7. Wiegand T, Sullivan GJ, Bjøntegaard G, Luthra A. Overview of the H.264/AVC video coding standard. IEEE Trans Circuits Syst Video Technol. Jul; 2003 13(7):560–576.
8. Shapiro J. Embedded image coding using zerotrees of wavelet coefficients. IEEE Trans Signal Process. Dec; 1993 41(12):3445–3462.
9. Said A, Pearlman WA. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans Circuits Syst Video Technol. Jun; 1996 6(3):243–250.
10. Taubman D. High performance scalable image compression with EBCOT. IEEE Trans Image Process. Jul; 2000 9(7):1158–1170. [PubMed: 18262955]
11. Ohm JR. Three-dimensional subband coding with motion compensation. IEEE Trans Image Process. Sep; 1994 3(5):559–571. [PubMed: 18291952]

12. Choi SJ, Woods JW. Motion-compensated 3-D subband coding of video. IEEE Trans Image Process. Feb; 1999 8(2):155–167. [PubMed: 18267464]
13. Kim BJ, Xiong Z, Pearlman WA. Low bit-rate scalable video coding with 3-D set partitioning in hierarchical tree (3-D SPIHT). IEEE Trans Circuits Syst Video Technol. Dec; 2000 10(8):1374–1387.
14. Karlsson G, Vetterli M. Three dimensional subband coding of video. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process; New York, NY, USA. Apr. 1988; 1100–1103.
15. Xu J, Xiong Z, Li S, Zhang YQ. Three-Dimensional Embedded Subband Coding with Optimized Truncation (3-D ESCOT). Appl Comput Harmon Anal. May; 2001 10(3):290–315.
16. Secker A, Taubman D. Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression. IEEE Trans Image Process. Dec; 2003 12(12):1530–1542. [PubMed: 18244708]
17. Chen P, Woods JW. Bidirectional MC-EZBC with lifting implementation. IEEE Trans Circuits Syst Video Technol. Oct; 2004 14(10):1183–1194.
18. Park HW, Kim HS. Motion estimation using low-band-shift method for wavelet-based moving-picture coding. IEEE Trans Image Process. Apr; 2000 9(4):577–587. [PubMed: 18255431]
19. Andreopoulos Y, et al. In-band motion compensated temporal filtering. Signal Process Image Commun. Aug; 2004 19(7):653–673.
20. Li X. Scalable video compression via overcomplete motion compensated wavelet coding. Signal Process Image Commun. Aug; 2004 19(7):637–651.
21. Andreopoulos Y, Munteanu A, Auwera GVD, Cornelis JPH, Schelkens P. Complete-to-overcomplete discrete wavelet transforms: Theory and applications. IEEE Trans Signal Process. Apr; 2005 53(4):1398–1412.
22. Xiong R, Xu J, Wu F. In-scale motion compensation for spatially scalable video coding. IEEE Trans Circuits Syst Video Technol. Feb; 2008 18(2):145–158.
23. Schwarz H, Marpe D, Wiegand T. Hierarchical B Pictures. Joint Video Team; Jul, 2005 document JVT-P014
24. Schwarz H, Marpe D, Wiegand T. Analysis of hierarchical B pictures and MCTF. Proc. IEEE Int. Conf. Multimedia Expo; Toronto, ON, Canada. Jul. 2006; 1929–1932.
25. Chen H, Kao MP, Nguyen TQ. Bidirectional scalable motion for scalable video coding. IEEE Trans Image Process. Nov; 2010 19(11):3059–3064. [PubMed: 20494850]
26. Flierl M, Vandergheynst P. An improved pyramid for spatially scalable video coding. Proc. IEEE Int. Conf. Image Process; Sep. 2005; 878–881.
27. Zhang W, Men A, Chen P. Adaptive inter-layer intra prediction in scalable video coding. Proc. IEEE Int. Symp. Circuits Syst. (ISCAS); Taipei, Taiwan. May 2009; 876–879.
28. Wu X, Shao M, Zhang X. Improvement of H.264 SVC by model-based adaptive resolution upconversion. Proc. Int. Conf. Image Process. (ICIP); Hong Kong. Sep. 2010; 4205–4208.
29. Shi Z, Sun X, Wu F. Spatially scalable video coding for HEVC. IEEE Trans Circuits Syst Video Technol. Dec; 2012 22(12):1813–1826.
30. Kang JW, Gabbouj M, Kuo CCJ. Sparse/DCT (S/DCT) two-layered representation of prediction residuals for video coding. IEEE Trans Image Process. Jul; 2013 22(7):2711–2722. [PubMed: 23568505]
31. Han J, Melkote V, Rose K. An estimation-theoretic framework for spatially scalable video coding. IEEE Trans Image Process. Aug; 2014 23(8):3684–3697. [PubMed: 24956371]
32. Wang RJ, Huang CW, Chang PC. Adaptive downsampling video coding with spatially scalable rate-distortion modeling. IEEE Trans Circuits Syst Video Technol. Nov; 2014 24(11):1957–1968.
33. Yang J, Wright J, Huang TS, Ma Y. Image super-resolution via sparse representation. IEEE Trans Image Process. Nov; 2010 19(11):2861–2873. [PubMed: 20483687]
34. Shen M, Xue P, Wang C. Down-sampling based video coding using super-resolution technique. IEEE Trans Circuits Syst Video Technol. Jun; 2011 21(6):755–765.
35. Ates HF. Decoder-side super-resolution and frame interpolation for improved H.264 video coding. Proc. IEEE Data Compress. Conf; Snowbird, UT, USA. Mar. 2013; 83–92.

36. Sun J, Zheng N-N, Tao H, Shum H-Y. Image hallucination with primal sketch priors. Proc. IEEE Conf. Comput. Vis. Pattern Recognit; Nice, France. Oct. 2003; 729–736.
37. Xiong H, Pan Z, Ye X, Chen CW. Sparse spatio-temporal representation with adaptive regularized dictionary learning for low bit-rate video coding. IEEE Trans Circuits Syst Video Technol. Apr; 2013 23(4):710–728.
38. Dai W, et al. Sparse representation with spatio-temporal online dictionary learning for promising video coding. IEEE Trans Image Process. Oct; 2016 25(10):4580–4595. [PubMed: 27479966]
39. Segall CA, Sullivan GJ. Spatial scalability within the H.264/AVC scalable video coding extension. IEEE Trans Circuits Syst Video Technol. Sep; 2007 17(9):1121–1135.
40. Orchard MT, Sullivan GJ. Overlapped block motion compensation: An estimation-theoretic approach. IEEE Trans Image Process. Sep; 1994 3(5):693–699. [PubMed: 18291963]
41. Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans Signal Process. Nov; 2006 54(11):4311–4322.
42. Mairal J, Bach F, Ponce J, Sapiro G. Online learning for matrix factorization and sparse coding. J Mach Learn Res. Mar.2010 11:19–60.
43. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Statist. 2004; 32(2):407–499.
44. Osborne MR, Presnell B, Turlach BA. A new approach to variable selection in least squares problems. IMA J Numer Anal. Jul; 2000 20(3):389–404.
45. McCann K, , et al. High Efficiency Video Coding (HEVC) Test Model 16 (HM16) Improved Encoder Description. Sapporo, Japan: Jun, 2014 document JCTVC-R1002
46. Seregin V, He Y. Common SHM Test Conditions and Software Reference Configurations. Valencia, Spain: Mar, 2014 document JCTVC-Q1009
47. Bjontegaard G. Calculation of Average PSNR Differences Between RD-Curves. Austin, TX, USA: Apr, 2001 document VCEG-M33, ITU-T SG16/Q6
48. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process. Apr; 2004 13(4):600–612. [PubMed: 15376593]
49. Fujiwara Y, Ida Y, Shiokawa H, Iwamura S. Fast Lasso algorithm via selective coordinate descent. Proc. AAAI Conf. Artif. Intell; Phoenix, AZ, USA. Feb. 2016; 1561–1567.

Biographies

Wenrui Dai (M’15) received the B.S., M.S., and Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2005, 2008, and 2014, respectively, all in electronic engineering. He is currently a Post-Doctoral Scholar with the Department of Biomedical Informatics, University of California at San Diego. His current research interests include learning-based image/video coding, image/signal processing, and predictive modeling.

Yangmei Shen received the B.S. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2014. She is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. Her current research interests include dictionary learning-based signal reconstruction, and sparse representation.

Hongkai Xiong (M'01–SM'10) received the Ph.D. degree in communication and information system from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 2003.

From 2007 to 2008, he was with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, as a Research Scholar. From 2011 to 2012, he was a Scientist with the Division of Biomedical Informatics, University of California at San Diego, La Jolla, CA, USA. Since then, he has been with the Department of Electronic Engineering, SJTU, where he is currently a Full Professor. He has authored over 180 refereed journal/conference papers. His current research interests include source coding/network information theory, signal processing, computer vision, and machine learning.

Dr. Xiong has been a member of the Innovative Research Groups of the National Natural Science, since 2012. He served as TPC members for prestigious conferences, such as ACM Multimedia, ICIP, ICME, and ISCAS. He was a recipient of the Top 10% Paper Award at the 2016 IEEE Visual Communication and Image Processing (IEEE VCIP'16), the Best Student Paper Award at the 2014 IEEE Visual Communication and Image Processing (IEEE VCIP'14), the Best Paper Award at the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (IEEE BMSB'13), and the Top 10% Paper Award at the 2011 IEEE International Workshop on Multimedia Signal Processing (IEEE MMSP'11). In 2016, he was granted the Yangtze River Scholar Distinguished Professor from the Ministry of Education of China, and the Youth Science and Technology Innovation Leader from the Ministry of Science and Technology of China. He was awarded Shanghai Academic Research Leader. In 2014, he was granted National Science Fund for Distinguished Young Scholar and Shanghai Youth Science and Technology Talent. In 2013, he was a recipient of Shanghai Shu Guang Scholar. In 2011, he obtained the First Prize of

the Shanghai Technological Innovation Award for Network-Oriented Video Processing and Dissemination: Theory and Technology. In 2010 and 2013, he obtained the SMC-A Excellent Young Faculty Award of Shanghai Jiao Tong University. In 2009, he was awarded a recipient of New Century Excellent Talents in University, Ministry of Education of China.

Xiaoqian Jiang (S'06–M'10) received the Ph.D. degree in computer science from Carnegie Mellon University. He is currently an Assistant Professor with the Department of Biomedical Informatics, University of California at San Diego. He is an Associate Editor of *BMC Medical Informatics and Decision Making* and serves as an Editorial Board Member of the *Journal of American Medical Informatics Association*. He was involved primarily in health data privacy and predictive models in biomedicine. He was a recipient of the Distinguished Paper Award from American Medical Informatics Association, Clinical Research Informatics Summit in 2012 and 2013, respectively. His team won the best student paper award in 2016 AMIA Summit on Translational Bioinformatics.

Junni Zou (M'07) received the M.S. degree and the Ph.D. degree in communication and information system from Shanghai University, Shanghai, China, in 2004 and 2006, respectively.

From 2006 to 2016, she was with the School of Communication and Information Engineering, Shanghai University, and became a Full Professor in 2014. From 2011 to 2012, she was with the Department of Electrical and Computer Engineering, University of California at San Diego, as a Visiting Professor. She is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. She has authored over 80 IEEE journal/conference papers, and two book chapters, including 15 IEEE Transactions journal papers. She holds 12 patents and has 10+ under reviewing patents. Her current research interests include multimedia communication, network resource optimization, wireless communication, and network information theory.

Dr. Zou was granted National Science Fund for Outstanding Young Scholar in 2016. She was a recipient of Shanghai Yong Rising Star Scientist Award in 2011. She obtained the

First Prize of the Shanghai Technological Innovation Award in 2011, and the First Prize of the Shanghai Science and Technology Advancement Award in 2008. Also, she has served on some technical program committees for the IEEE and other international conferences.

David Taubman (F'15) received the B.S. degree and the B.E. degree in electrical from The University of Sydney, in 1986 and 1988, respectively, and the M.S. and Ph.D. degrees from the University of California at Berkeley, in 1992 and 1994, respectively. From 1994 to 1998, he was with the Hewlett- Packard's Research Laboratories, Palo Alto, CA, USA, joining the University of New South Wales in 1998, where he is currently a Professor with the School of Electrical Engineering and Telecommunications. He has authored the book with M. Marcellin *JPEG2000: Image Compression Fundamentals, Standards and Practice*. His current research interests include highly scalable image and video compression, motion estimation and modeling, inverse problems in imaging, perceptual modeling, and multimedia distribution systems. He was awarded the University Medal from The University of Sydney. He has received two best paper awards from the IEEE Circuits and Systems Society for the 1996 paper, A Common Framework for Rate and Distortion Based Scaling of Highly Scalable Compressed Video; and from the IEEE Signal Processing Society for the 2000 paper, High Performance Scalable Image Compression with EBCOT.

Appendix A. Proof of Proposition 1

Given that the sparse representation vector α, there exists a constant $\varepsilon > 0$ for $\tilde{\mathbf{X}}_{BL}$ and $\hat{\mathbf{X}}_{EL}$,

$$\|\tilde{\mathbf{X}}_{BL} - \mathbf{D}_{BL}\alpha\|_2^2 \leq \varepsilon$$
$$\|\hat{\mathbf{X}}_{EL} - \mathbf{D}_{EL}\alpha\|_2^2 \leq \varepsilon$$

The scaled-up BL $\tilde{\mathbf{X}}_{BL}$ is obtained by down- and up-sampling $\hat{\mathbf{X}}_R$ using operators $\mathcal{D}$ and $\mathcal{U}$ (i.e., bicubic).

$$\tilde{\mathbf{X}}_{BL} = \mathcal{U}\hat{\mathbf{X}}_{BL} = \mathcal{U}\mathcal{D}\hat{\mathbf{X}}_R = \mathcal{Q}\hat{\mathbf{X}}_R$$

Here, the global operator $\mathcal{Q}$ converts volumes in $\hat{\mathbf{X}}_R$ into corresponding ones in $\tilde{\mathbf{X}}_{BL}$. Thus, we can obtain that

$$\|(\mathcal{U}\mathcal{D}(\mathbf{D}_{BL}+\mathbf{D}_{EL})-\mathbf{D}_{BL})\alpha\|_2^2 \le \|\mathcal{U}\mathcal{D}(\widetilde{\mathbf{X}}_{BL}-\mathbf{D}_{BL}\alpha)\|_2^2+\|\mathcal{U}\mathcal{D}(\widehat{\mathbf{X}}_{EL}-\mathbf{D}_{EL}\alpha)\|_2^2+\|\mathcal{U}\mathcal{D}\widehat{\mathbf{X}}_R-\mathbf{D}_{BL}\alpha\|_2^2 \le C_{\mathcal{U},\mathcal{D}}\cdot\varepsilon$$

Here, $C_{\mathcal{U},\mathcal{D}}$ is related to $\|\mathcal{U}\mathcal{D}\|_2^2$. Under sufficient sampling, for the i-th volume $\hat{\mathbf{z}}_i$ in $\hat{\mathbf{Z}}_{NR}$ and the corresponding $\hat{\mathbf{z}}_i^{BL}$ in $\tilde{\mathbf{Z}}_{BL}$, there exists α_i derived from $\|\hat{\mathbf{z}}_i^{BL}-\mathbf{D}_{BL}\alpha\|_2^2 \le \varepsilon$ satisfying that

$$\|(\mathcal{U}\mathcal{D}(\mathbf{D}_{BL}+\mathbf{D}_{EL})-\mathbf{D}_{BL})\alpha\|_2^2 \le \mathcal{C}_{\mathcal{U},\mathcal{D}}\cdot\varepsilon.$$

Thus, we can find that $\|\hat{\mathbf{z}}_i-(\mathbf{D}_{BL}+\mathbf{D}_{EL})\,\alpha_i\| \le \varepsilon$. As a result, $\|\hat{\mathbf{Z}}_{NR}-(\mathbf{D}_{BL}+\mathbf{D}_{EL})\,\alpha\|$ is bounded by $\|\tilde{\mathbf{Z}}_{BL}-\mathbf{D}_{BL}\alpha\|$.

Appendix B. Proof of Proposition 2

Given arbitrary temporal layer T_κ, the divergence between the approximate and expected cost can be decomposed by

$$f(\mathbf{D}_*^\kappa)-f_t(\mathbf{D}_t^\kappa)=[f(\mathbf{D}_*^\kappa)-f_\tau(\mathbf{D}_*^\kappa)]+[f_\tau(\mathbf{D}_*^\kappa)-f_\tau(\mathbf{D}_\tau^\kappa)]+[f_\tau^\kappa(\mathbf{D}_\tau^\kappa)-f_t(\mathbf{D}_t^\kappa)], \quad (15)$$

where $f_\tau(\mathbf{D}_*^\kappa)$ is the empirical cost based on $\mathbf{D}_*^\kappa$. Given randomly selected samples, the divergence between the expected and empirical cost is upper-bounded by Vapnik-Chervonenkis (VC) bound $C\sqrt{(t/\tau)\log(\tau/t)}$ with positive constant C.

$$\mathbb{E}[\sup|f(\mathbf{D}_*^\kappa)-f_\tau(\mathbf{D}_*^\kappa)|] \le C\sqrt{\frac{t}{\tau}} \to 0,\ \tau\to\infty\ a.s.,$$

For the second term in Eq. (15), when $f_\tau(\mathbf{D})-f_\tau(\mathbf{D}_\tau)$ is Lipschitz, it is shown in [42] that

$$f_\tau(\mathbf{D}_*^\kappa)-f_\tau(\mathbf{D}_\tau^\kappa)\to 0,\ \tau\to\infty\ a.s. \quad (16)$$

For the approximate cost $f_t(\mathbf{D}_t^\kappa)$, we find that

$$f_{t+1}(\mathbf{D}_{t+1}^\kappa)-f_t(\mathbf{D}_t^\kappa)=f_{t+1}(\mathbf{D}_{t+1}^\kappa)-f_{t+1}(\mathbf{D}_t^\kappa)+\left[\frac{l(\mathbf{x}_{t+1}^\kappa,\mathbf{D}_t^\kappa)-f_\tau(\mathbf{D}_t^\kappa)}{t+1}+\frac{f_\tau(\mathbf{D}_t^\kappa)-f_t(\mathbf{D}_t^\kappa)}{t+1}\right]\le 0.$$

Here, $\mathbf{x}_{t+1}^{\kappa}$ is the newly-added samples at the $t+1$-th iteration for $T_{\kappa+1}$. Given the set of dictionaries in all the previous layers $\mathcal{D}_t^{\kappa}$, the upper bound of the conditional approximate cost for sparse representation can be developed.

$$\mathbb{E}\left[f_{t+1}(\mathbf{D}_{t+1}^{\kappa}) - f_t(\mathbf{D}_t^{\kappa}) \mid \mathcal{D}_t^{\kappa}\right] \leq \frac{\mathbb{E}[l(\mathbf{x}_{t+1}^{\kappa}, \mathbf{D}_t) \mid \mathcal{D}_t^{\kappa}] - f_{\tau}(\mathbf{D}_t^{\kappa})}{t+1}$$
$$\leq \frac{\|f(\mathbf{D}_t^{\kappa}) - f_{\tau}(\mathbf{D}_t^{\kappa})\|}{t+1} \leq \frac{C}{t+1} \cdot \sqrt{\frac{\tau}{t}}$$

When $t \rightarrow \tau$, $f_t(\mathbf{D}_t^{\kappa})$ approaches $f_{\tau}(\mathbf{D}_{\tau}^{\kappa})$, as $\mathbb{E}[\|f(\mathbf{D}_*^{\kappa}) - f_{\tau}(\mathbf{D}_*^{\kappa})\|_{\infty}] \sim o(\frac{1}{\tau}) \rightarrow 0$, which means that $f_t(\mathbf{D}_t^{\kappa})$ converges almost surely for T_{κ} with the growth of τ. As a result, with sufficient sampling, the divergence between approximate and expect cost $\sum_{\kappa=0}^{\Lambda-1} | f_t(\mathbf{D}_t^{\kappa}) - f(\mathbf{D}_*^{\kappa}) |$ asymptotically varnishes for progressive dictionary learning almost surely with $t \rightarrow \tau$.

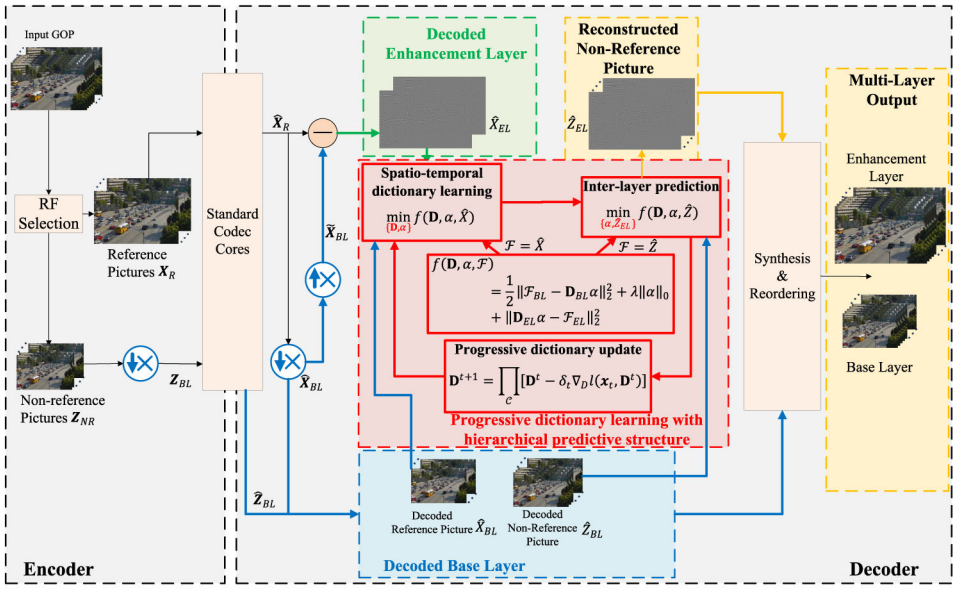


Fig. 1.

The proposed framework for scalable video coding based on progressive dictionary learning. Given groups of pictures (GOP) to be encoded, the first L consecutive frames are selected as reference frames (RFs) $\mathbf{X}_R$. The remaining non-reference frames (NRFs) $\mathbf{Z}_{NR}$ are downsampled at a predetermined ratio, e.g. 2× and 1.5×. In the decoder side, the base layer (BL) consists of the down-sampled decoded RFs $\hat{\mathbf{X}}_{BL}$ and NRFs $\hat{\mathbf{Z}}_{BL}$. The enhancement layer (EL) $\hat{\mathbf{X}}_{EL}$ collects the residual of RFs by substracting the corresponding scaled-up ones $\tilde{\mathbf{X}}_{BL}$ from decoded RFs $\hat{\mathbf{X}}_R$ Overcomplete dictionary is trained based on the pair of blocks extracted from reference pictures from BL and EL, respectively. NRFs are reconstructed with the inter-layer prediction based on the trained dictionary. Here, $\mathcal{F}$ represents $\hat{\mathbf{X}}$ for training and $\hat{\mathbf{Z}}$ for reconstruction. To enable temporal scalability, the update of dictionary and reconstruction are performed based on the proposed hierarchical structure.

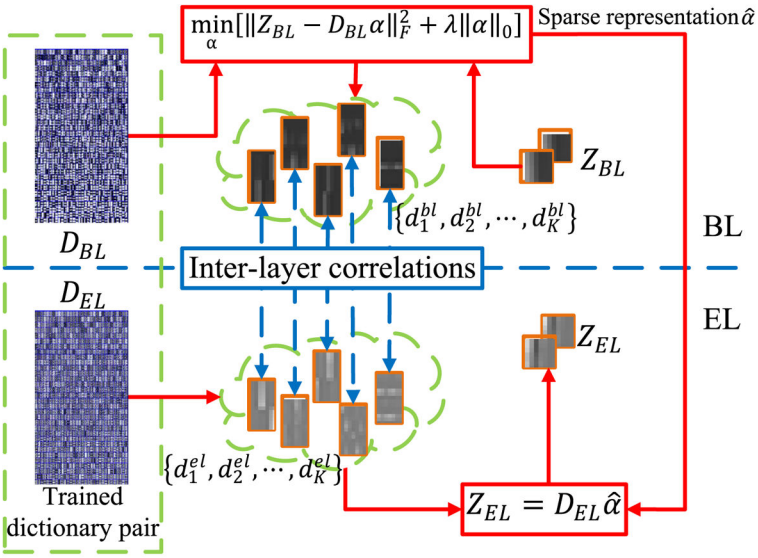


Fig. 2.

Inter-layer prediction for the two-layer architecture. Patches from EL are reconstructed based on the trained dictionary pair $\mathbf{D}_{BL}$ and $\mathbf{D}_{EL}$ for BL and EL. Sparse representation α is estimated with $\mathbf{D}_{BL}$ and non-reference frames $\tilde{\mathbf{Z}}_{BL}$ in BL. Non-reference frames $\tilde{\mathbf{Z}}_{EL}$ are reconstructed by exploiting inter-layer correlations for atoms between BL and EL.

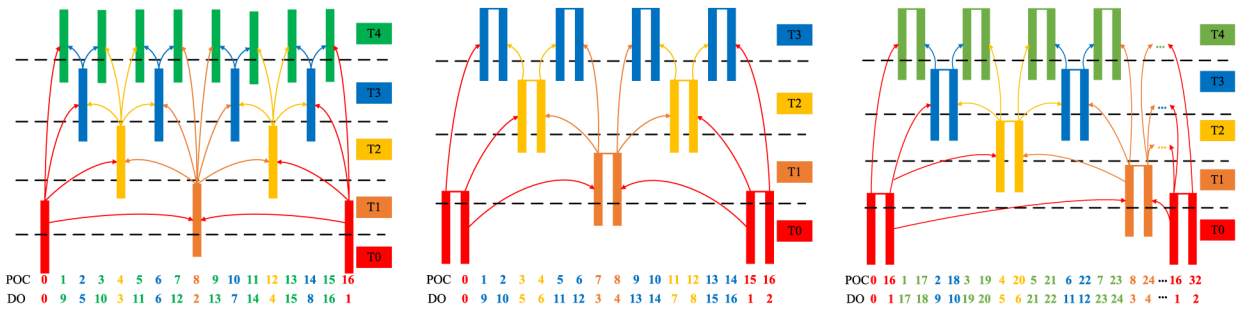


Fig. 3.

Hierarchical predictive structure for temporal scalability within the GOP with a size of 16 frames. Picture order count (POC) and decoding order (DO) are shown for each frame. (a) Hierarchical B-pictures for hybird MC-DCT framework; (b) Hierarchical predictive structure with local motion (HPS-LO) for progressive spatio-temporal dictionary learning; (c) Hierarchical predictive structure with long-term motion (HPS-LT) for progressive spatio-temporal dictionary learning.

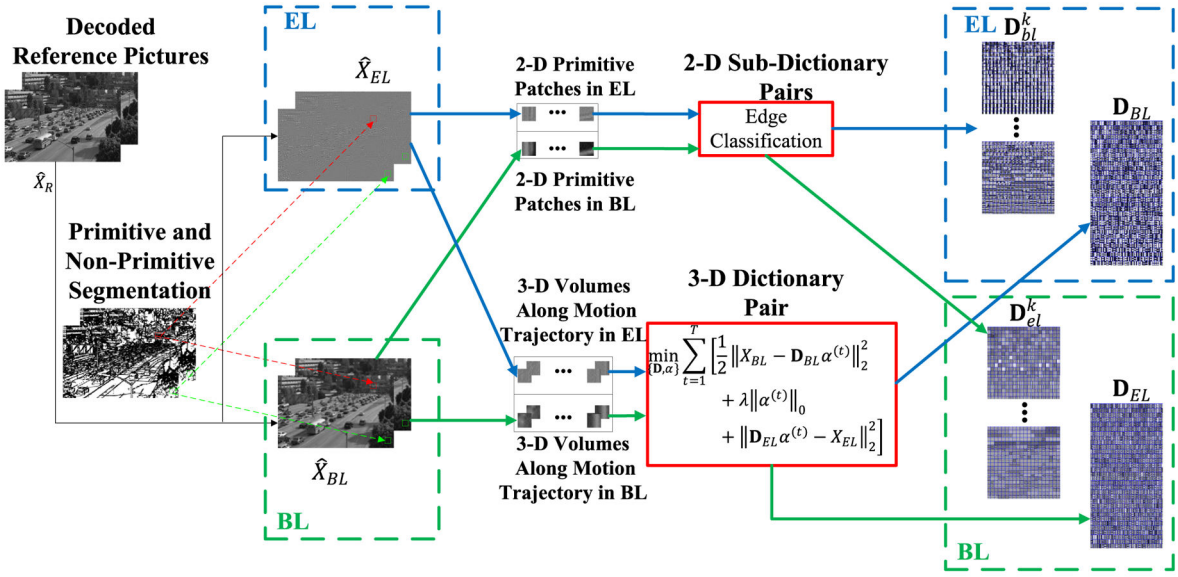


Fig. 4.

Spatio-temporal dictionary learning for inter-layer prediction. 2-D sub-dictionary and 3-D dictionary pairs are trained for the base and enhancement layer. Given reference frames, 2-D primitive patches and 3-D non-primitive volumes are extracted for dictionary learning. The 2-D sub-dictionary paris are generated by matching 2-D patches with edge classification. Joint optimization over 3-D volumes are made for 3-D dictionary pair to exploit spatial and temporal correlations along motion trajectory.

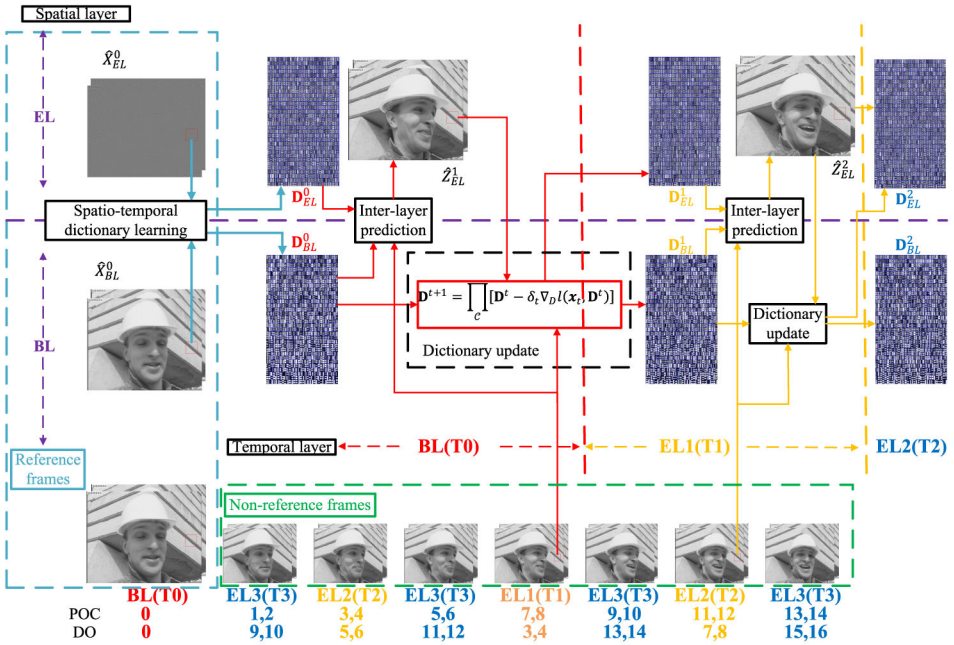
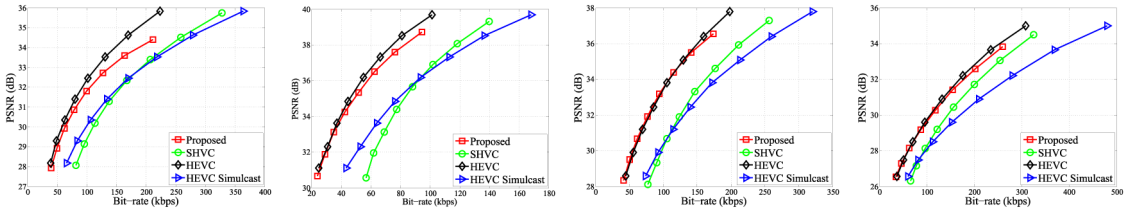


Fig. 5.

Progressive dictionary learning with hierarchical predictive structure. The GOP with 16 frames are progressively trained and predicted based on the hierarchical predictive structure. Firstly, spatio-temporal dictionary is trained based on the reference frames. Consequently, for each (base or enhancement) layer, non-reference frames are predicted based on the updated dictionary from previous layer and updated with stochastic gradient descent based on the reconstructed frames in current layer.



(a) (b) (c) (d)

Fig. 6.

Rate-distortion curves for CIF test sequences by the proposed scheme, SHVC, HEVC, and HEVC simulcast under low-delay B (LDB) configuration. (a) *Foreman* (352 × 288). (b) *Akiyo* (352 × 288). (c) *News* (352 × 288). (d) *Waterfall* (352 × 288).

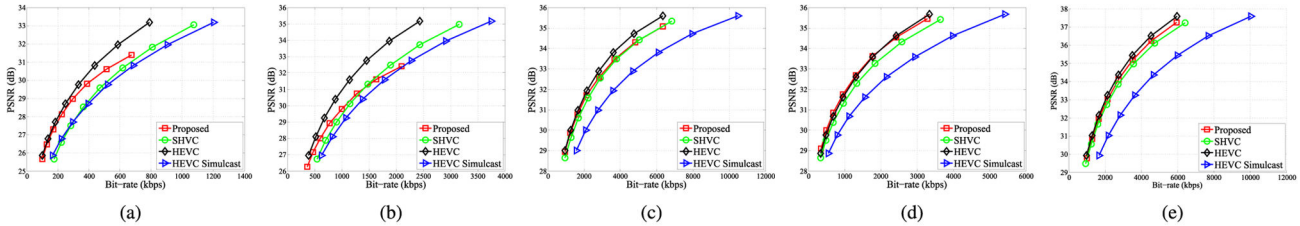


Fig. 7.

Rate-distortion curves for test sequences with various resolutions obtained by the proposed scheme, SHVC, HEVC, and HEVC simulcast under the low-delay B (LDB) configuration. (a) *BlowingBubbles* (416 × 240). (b) *BQMall* (832 × 480). (c) *Cactus* (1920 × 1080). (d) *ParkScene* (1920 × 1080). (e) *Traffic* (2500 × 1600).

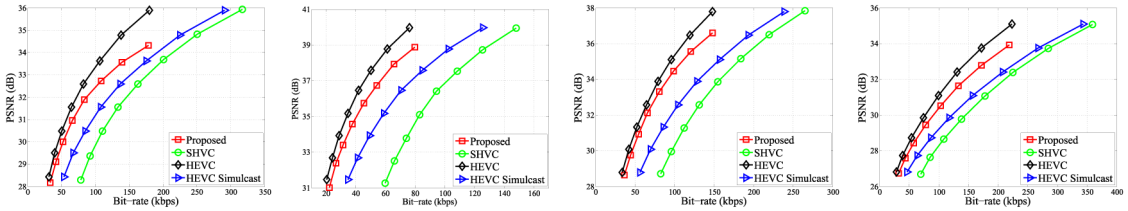


Fig. 8.

Rate-distortion curves for CIF test sequences by the proposed scheme, SHVC, HEVC, and HEVC simulcast under random access (RA) configuration. (a) *Foreman* (352 × 288). (b) *Akiyo* (352 × 288). (c) *News* (352 × 288). (d) *Waterfall* (352 × 288).

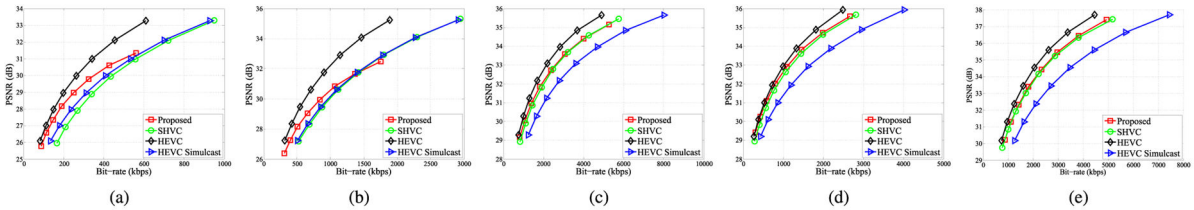
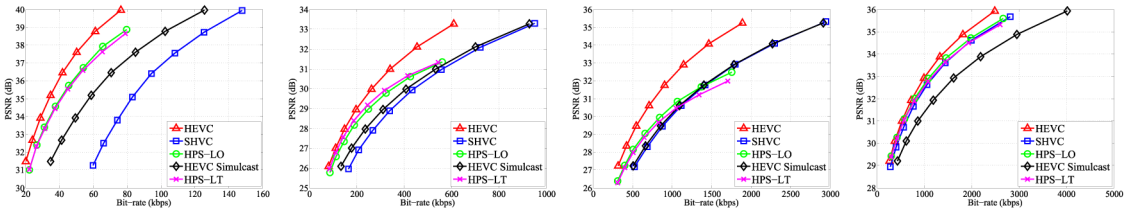


Fig. 9.

Rate-distortion curves for test sequences with various resolutions obtained by the proposed scheme, SHVC, HEVC, and HEVC simulcast under the random access (RA) configuration. (a) *BlowingBubbles* (416 × 240). (b) *BQMall* (832 × 480). (c) *Cactus* (1920 × 1080). (d) *ParkScene* (1920 × 1080). (e) *Traffic* (2500 × 1600).



(a) (b) (c) (d)

Fig. 10.

Rate-distortion performance (dB) for the proposed method with hierachical predictive structures considering local (HPS-LO) and long-term (HPS-LT) motion under random access (RA) configuration. (a) *Akiyo* (352×288). (b) *BlowingBubbles* (416×240). (c) *BQMall* (832×480). (d) *ParkScene* (1920×1080).

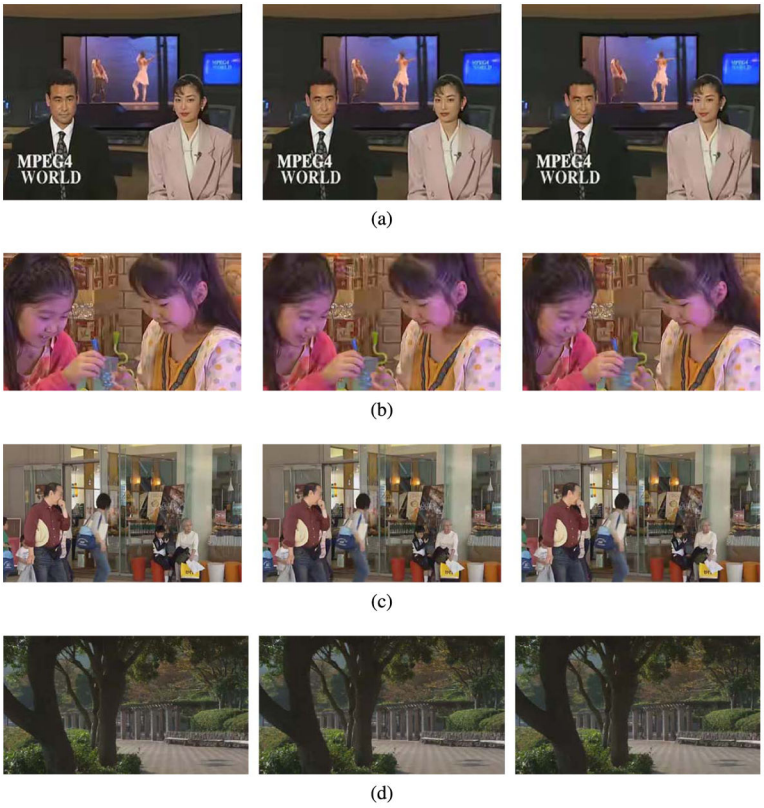


Fig. 11.

Visual results of reconstructed sequences. From top to bottom and from left to right, there are reconstructed frames in the enhancement layer of sequences *News*, *BlowingBubbles*, *BQMall*, and *ParkScene* obtained by the proposed scheme, SHVC, and HEVC simulcast, respectively. (a) Visual performance of *News* sequence. From left to right, PSNR of the reconstructed frames are 36.26 dB, 33.05 dB, and 32.38 dB, respectively. (b) Visual performance of *BlowingBubbles* sequence. From left to right, PSNR of the reconstructed frames are 30.23 dB, 27.49 dB, and 27.65 dB, respectively. (c) Visual performance of *BQMall* sequence. From left to right, PSNR of the reconstructed frames are 31.60 dB, 29.24 dB, and 29.33 dB, respectively. (d) Visual performance of *ParkScene* sequence. From left to right, PSNR of the reconstructed frames are 31.48 dB, 30.50 dB, and 29.92 dB, respectively.

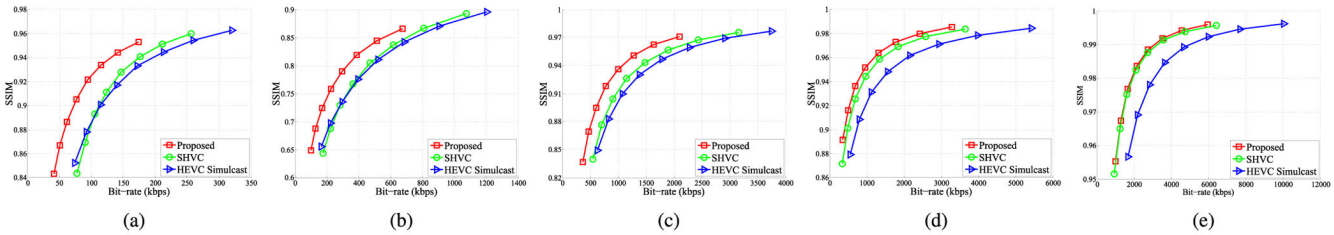


Fig. 12.

SSIM performance for test sequences *News*, *BlowingBubbles*, *BQMall*, *ParkScene*, and *Traffic* obtained by the proposed scheme, SHVC, and HEVC simulcast under low-delay B (LDB) configuration. (a) *News* (352 × 288). (b) *BlowingBubbles* (416 × 240). (c) *BQMall* (832 × 480). (d) *ParkScene* (1920 × 1080). (e) *Traffic* (2500 × 1600).

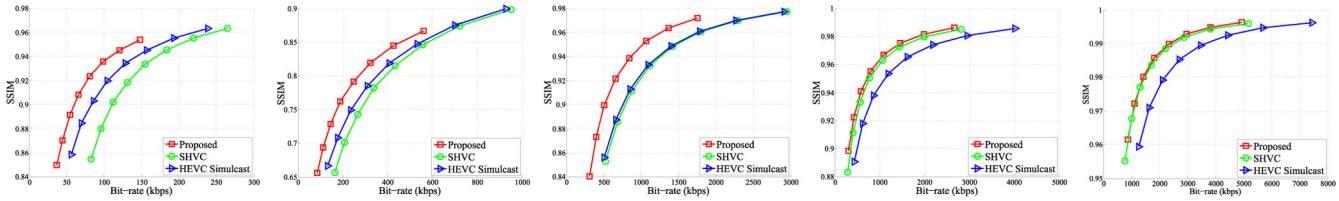


Fig. 13.

SSIM performance for test sequences *News*, *BlowingBubbles*, *BQMall*, *ParkScene*, and *Traffic* obtained by the proposed scheme, SHVC, and HEVC simulcast under random access (RA) configuration. (a) *News* (352 × 288). (b) *BlowingBubbles* (416 × 240). (c) *BQMall* (832 × 480). (d) *ParkScene* (1920 × 1080). (e) *Traffic* (2500 × 1600).

TABLE I

Summary of Notations in Progressive Dictionary Learning

Notation	Description
$\mathbf{X}_R$, $\mathbf{Z}_{NR}$	Reference frame (RF) and non-reference frame (NRF)
$\mathbf{Z}_{BL}$	Downsampled NRF at the encoder side
$\hat{\mathbf{X}}_R$	Decoded RF
$\hat{\mathbf{X}}_{BL}$, $\hat{\mathbf{Z}}_{BL}$	Decoded BL for RF and NRF
$\tilde{\mathbf{X}}_{BL}$, $\tilde{\mathbf{Z}}^l_{EL}$	Scaled-up frame for RF and NRF from $\hat{\mathbf{X}}_{BL}$ and $\hat{\mathbf{Z}}_{BL}$
$\hat{\mathbf{X}}_{EL}$	Enhancement layer (EL) for RF obtained by subtracting scaled-up $\tilde{\mathbf{X}}_{BL}$ from $\hat{\mathbf{X}}_R$ at the decoder side
$\tilde{\mathbf{Z}}_{EL}$	Recovered high-frequency details from $\tilde{\mathbf{Z}}^l_{EL}$
$\hat{\mathbf{Z}}_{EL}$	Reconstructed EL for NRF
$\mathbf{x}^{BL}_n$, $\mathbf{x}^{EL}_n$	Patch or volume extracted from BL and EL for RF
$\mathbf{z}^l_n$, $\mathbf{z}^h_n$	Patch or volume extracted from $\tilde{\mathbf{Z}}^l_{EL}$ and $\tilde{\mathbf{Z}}^h_{EL}$
$\{\mathbf{D}^k_{bl}, \mathbf{D}^k_{el}\}$	The k-th 2-D sub-dictionary pair for the BL and EL
$\{\mathbf{D}_{BL}, \mathbf{D}_{EL}\}$	The 3-D sub-dictionary pair for the BL and EL
α	Sparse representation vector
$\vec{\mathcal{F}}$	Generalized notation for frames ($\hat{\mathbf{X}}$ for RF and $\hat{\mathbf{Z}}$ for NRF)
T_κ	The κ-the temporal layer

TABLE II

BD-PSNR (dB) in Comparison to HEVC Obtained by the Proposed Method, SHVC, and HEVC Simulcast Under Low-Delay B and Random Access Configuration, Respectively

Sequence	Resolution	Low-delay B			Random access		
		Proposed	SHVC	HEVC Simulcast	Proposed	SHVC	HEVC Simulcast
Akiyo	352 × 288	−0.31	−5.84	−3.31	−1.16	−7.70	−3.39
Foreman		−0.61	−2.82	−2.28	−0.86	−3.58	−2.20
News		0.02	−3.04	−3.05	−0.77	−5.01	−3.01
Waterfall		−0.12	−1.38	−1.90	−0.69	−2.85	−1.90
BlowingBubbles	416 × 240	−0.44	−1.63	−1.66	−0.87	−2.14	−1.65
Harbor	704 × 576	−0.43	−1.39	−1.82	−0.79	−1.73	−1.74
BQMall	832 × 480	−1.16	−1.52	−2.06	−1.47	−2.13	−2.02
Cactus	1920 × 1080	−0.28	−0.47	−1.77	−0.58	−0.73	−1.69
ParkScene		−0.07	−0.38	−1.49	−0.29	−0.51	−1.49
Traffic	2500 × 1600	−0.25	−0.45	−2.27	−0.64	−0.70	−2.20
Average		−0.35	−1.89	−2.16	−0.81	−2.71	−2.13

IEEE Trans Image Process. Author manuscript; available in PMC 2018 October 12.

TABLE III

BD-Rate Change (%) in Comparison to HEVC Obtained by the Proposed Method, SHVC, and HEVC Simulcast Under Low-Delay B and Random Access Configuration, Respectively

Sequence	Resolution	Low-delay B			Random access		
		Proposed	SHVC	HEVC Simulcast	Proposed	SHVC	HEVC Simulcast
Akiyo	352 × 288	5.68	81.88	70.9	20.47	140.74	69.66
Foreman		16.02	76.76	67.39	23.85	105.27	65.89
News		−0.44	57.04	64.23	13.91	104.97	63.64
Waterfall		2.49	41.57	59.50	18.60	91.37	58.95
BlowingBubbles	416 × 240	13.77	55.30	58.32	29.18	75.43	57.66
Harbor	704 × 576	13.49	49.21	67.90	27.40	65.40	67.26
BQMall	832 × 480	33.79	39.40	57.13	45.17	59.28	56.24
Cactus	1920 × 1080	8.89	14.73	66.58	19.33	24.32	65.22
ParkScene		−2.70	13.95	64.51	9.95	18.67	63.95
Traffic	2500 × 1600	6.22	11.94	70.34	16.79	18.84	69.50
Average		9.72	44.18	64.63	22.47	70.43	63.79

TABLE IV

Computational Complexity for the Decoders of the Proposed Scheme, SHVC, and HEVC Simulcast, Respectively. For the Proposed Scheme, Both the Dictionary Learning Speed (sec/GOP) and Reconstruction Speed (sec/Frame) are Provided

Sequence	Low-delay B				Random access			
	Proposed		SHVC	HEVC Simulcast	Proposed		SHVC	HEVC Simulcast
	Learning	Reconstruction	Decoding	Decoding	Learning	Reconstruction	Decoding	Decoding
Akiyo	111.58	69.12	1.08	0.77	118.33	32.52	1.28	0.91
Foreman	226.84	63.14	1.78	1.59	228.38	54.05	1.79	1.41
News	157.74	31.14	1.32	1.05	169.67	32.13	1.39	1.09
Waterfall	245.94	36.06	1.49	1.21	271.05	26.17	1.49	1.12
BlowingBubbles	299.50	36.21	1.81	1.67	307.41	31.24	1.72	1.54
Harbor	289.39	449.71	7.37	6.81	369.33	329.73	6.55	5.88
BQMall	190.28	222.52	5.46	5.12	204.53	176.06	5.59	5.17
Cactus	150.99	1891.05	24.83	24.02	172.62	1750.46	24.94	2.76
ParkScene	166.55	2551.93	25.86	26.01	182.18	2408.28	25.48	22.39
Traffic	160.55	4456.55	46.34	40.56	169.85	3920.23	38.79	38.61