Sparse Representation With Spatio-Temporal Online Dictionary Learning for Promising Video Coding

Wenrui Dai, *Member, IEEE*, Yangmei Shen, Xin Tang, Junni Zou, *Member, IEEE*, Hongkai Xiong, *Senior Member, IEEE*, and Chang Wen Chen, *Fellow, IEEE*

***Abstract*— Classical dictionary learning methods for video coding suffer from high computational complexity and interfered coding efficiency by disregarding its underlying distribution. This paper proposes a spatio-temporal online dictionary learning (STOL) algorithm to speed up the convergence rate of dictionary learning with a guarantee of approximation error. The proposed algorithm incorporates stochastic gradient descents to form a dictionary of pairs of 3D low-frequency and high-frequency spatio-temporal volumes. In each iteration of the learning process, it randomly selects one sample volume and updates the atoms of dictionary by minimizing the expected cost, rather than optimizes empirical cost over the complete training data, such as batch learning methods, e.g., K-SVD. Since the selected volumes are supposed to be independent identically distributed samples from the underlying distribution, decomposition coefficients attained from the trained dictionary are desirable for sparse representation. Theoretically, it is proved that the proposed STOL could achieve better approximation for sparse representation than K-SVD and maintain both structured sparsity and hierarchical sparsity. It is shown to outperform batch gradient descent methods (K-SVD) in the sense of convergence speed and computational complexity, and its upper bound for prediction error is asymptotically equal to the training error. With lower computational complexity, extensive experiments validate that the STOL-based coding scheme achieves performance improvements than H.264/AVC or High Efficiency Video Coding as well as existing super-resolution-based methods in rate-distortion performance and visual quality.**

***Index Terms*— Online dictionary learning, sparse representation, video coding, stochastic gradient descent, K-SVD.**

Manuscript received October 13, 2014; revised September 26, 2015 and May 18, 2016; accepted July 17, 2016. Date of publication July 27, 2016; date of current version August 9, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61425011, Grant U1201255, Grant 61271218, Grant 61501294, Grant 61529101, Grant 61472234, and Grant 61271211, in part by the China Postdoctoral Science Foundation under Grant 2015M581617, and in part by the Shu Guang Project under Grant 13SG13. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Anuj Srivastava.

W. Dai is with the Department of Biomedical Informatics, University of California at San Diego, La Jolla, CA 92093 USA, and also with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wed004@ucsd.edu).

Y. Shen, X. Tang, and H. Xiong are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shenyangmei0214@sjtu.edu.cn; xint14@sjtu.edu.cn; xionghongkai@sjtu.edu.cn).

J. Zou is with the Key Laboratory of Special Fiber Optics and Optical Access Networks, Shanghai University, Shanghai 200072, China (e-mail: zoujn@shu.edu.cn).

C. W. Chen is with the Department of Computer Science and Engineering, The State University of New York at Buffalo, Buffalo, NY 14260 USA (e-mail: chencw@bu_alo.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2594490

I. Introduction

THE state-of-the-art video coding schemes, e.g. H.264/AVC [1] and its successor High Efficiency Video Coding (HEVC) standard [2]), have achieved a vital efficiency by exploring redundancies among pixels through intra- and inter-prediction. For substantial improvement, a perspective of disruptive techniques has arisen from the hybrid framework. One promising attempt for video coding is to reconstruct high-resolution (HR) video contents with their sampled low-resolution (LR) versions. For example, scalable video coding [3] maintained the spatial capability through down-sampling and inter-layer prediction with up-sampling, but suffered from a heavy coding burden of the encoder. To relieve, distributed video coding [4] shifted the complexity of intensive prediction to the decoder for applications with constrained encoders. Inspired by Wyner-Ziv coding scheme, it approaches the rate of joint entropy by separate modeling of correlated sources. However, its practical performance is degraded by the estimation of correlated side information.

As an alternative, texture synthesis and hallucination for up-sampling based reconstruction were introduced. Dumitras and Haskel [5] developed a texture analysis-synthesis scheme to reduce the entropy of source information, where the homogeneous area is clustered into a small patch and represented with the epitome contents of associated regions. It can be regarded as an infant stage to apply sparse dictionary learning to video coding. Since these patches are close to uniform, they can be handled under the framework of Markov random fields (MRFs) with iterative optimization, e.g. belief propagation. Various side information has been considered to restore the missing information. For example, spatial correlations could be inferred based on edges [6], static textures [7], and assistant parameters [8]. To be consistent with the motion trajectory, spatio-temporal structures have been developed for global optimization [9], [10]. Although these methods claimed to achieve a good perceptual quality, they failed to ensure pixel-wise fidelity.

To guarantee both evaluations, state-of-the-art super-resolution methods have been widely considered for video reconstruction. They estimated correlations between high-frequency (HF) contents and their sampled sparse low-resolution versions in a nonparametric sense. According to the assumptions and methodologies for exploiting such correlations, they are classified into three categories: interpolation-based, reconstruction-based, and learning-based. Interpolation-based methods, such as bilinear and bicubic methods, exploit the strong correlations within adjacent pixels,

but blur the discontinuities and edges. Reconstruction-based methods [11] regularized the super-resolution image with reconstruction constraints derived from prior knowledge.

Recently, learning-base approaches are prevailing in super-resolution task, which incorporate dictionary learning to train an over-complete dictionary for reconstruction with sparse representation. Pioneered in [12], lost HF information was inferred based on learned co-occurrence priors from low-level vision [13] with MRFs in a patch manner. Its core idea is that pairs of LR-HR patches can be similarly represented with a corresponding pair of learned dictionaries under the assumption of sparse representation invariance. Under such assumption, fully-coupled learning methods [14] utilized bi-level optimization to train the dictionary pair. However, these methods are restricted by the supposed sparse representation invariance. To relax, [15] utilized data clustering to learn a set of linear mappings between the non-zero representation coefficients of LR and HR patches. Peleg and Elad [16] designed a multi-level scale-up scheme to make MMSE estimation for reconstruction in the sense of feed-forward neural network. Meanwhile, a Bayesian non-parametric approach [17] adopted beta-Bernoulli process to learn the dictionary.

In the line of video coding, motion trajectory is considered as well as spatial correlations. An example-based method [18] presented a super-resolution approach for video coding, where a set of pairs of non-adaptive LR-HR patches are trained to enhance the reconstruction of high-frequency details. Ates [19] introduced enhanced skip and direct modes to integrated spatial super-resolution and frame interpolation with H.264 and HEVC standard. Considering that primal sketch priors could enhance blurred edges, ridges and corners [20], a sparse spatio-temporal representation [21] was developed for bit-rate video coding. It facilitated reconstruction by learning an adaptive regularized dictionary of 2-D patches and 3-D volumes and outperformed H.264/AVC in terms of both objective and subjective comparisons. Since batch learning algorithms like K-SVD [23] are adopted to solve the optimized inverse problem, these methods are prohibitive for image and video signals. In comparison to iterative batch procedures [23]–[25], online learning has been recognized to be capable of significantly reducing the computational complexity and memory consumption for training [22]. Furthermore, it could achieve more sparse representation based on the trained dictionary than improved batch learning algorithms, e.g. fast alternatives using combination of analytic and adaptive learned dictionaries [26], analysis-based sparse representation [27], [28], submodular dictionary selection [29], [30]. Currently, there exist a challenge to balance the sparse representation for approximated signals and overhead of trained dictionary for coding [31]–[33]. However, these general frameworks do not optimize the sparse representation for video sequences with inherent structured sparsity and hierarchical sparsity. Along with the insight, it stimulates us to investigate an efficient learning algorithm for training dynamic time-varying signals.

In this paper, we propose spatio-temporal online dictionary learning (STOL) for sparse representation with the application to video coding. It trains a 3-D spatio-temporal dictionary by iteratively updating the atoms for asymptotically optimal representation and fast convergence rate. Unlike the classical batch gradient descents, it formulates optimized stochastic approximations by exploiting the structure of sparse coding. For the training set of i.i.d. samples drawn from the underlying distribution, STOL sequentially predicts the decomposition coefficients for each sample over the trained dictionary and updates the dictionary with stochastic gradient descent algorithm to minimize the expected cost. Hence, the proposed method could obtain sparse representation for dynamic time-varying signals with the trained dictionary. Theoretically, it is proved that the proposed STOL could maintain both structured sparsity and hierarchical sparsity with better approximation than K-SVD. It is shown to outperform batch gradient descent methods (K-SVD) in the sense of convergence speed and computational complexity for large-scale optimization problems. Furthermore, its upper bound for prediction error is proven to be asymptotically equal to the training error.

For dictionary learning, online learning incorporates stochastic approximations to exploit the temporal and spatial correlations by sequentially adapting the small patches from training data. With respect to classical dictionary learning based video coding schemes, the design of STOL provides two advantages. On the one hand, STOL behaves faster than iterative batch alternatives, e.g. K-SVD and improved dictionary learning schemes. It depends on lower computational complexity and memory consumption without explicit learning rate tuning. On the other hand, STOL updates the dictionary by minimizing expect cost over a convex set of constraints instead of empirical cost in iterative batch learning. It achieves sparser representation for practical coding than the improved dictionary learning schemes for efficiency. For validation, we apply the proposed STOL algorithm into synthetic data and video coding. With a sharp reduction of computational complexity, sufficient experiments show that the STOL-based scheme achieves both objective and visual quality improvements than H.264/AVC or HEVC as well as current super-resolution based methods.

The remainder of this paper is organized as follows. All the abbreviations are summarized in Table I.

TABLE I
ABBREVIATION TABLE

SUMMARY OF ABBREVIATIONS AND MODEL PARAMETERS	
HF	High-frequency
LF	Low-frequency
HR	High-resolution
LR	Low-resolution
KF	Key frame
NKF	Non-key frame
$X_h, \hat{X}_h$	HR KF, and $\hat{X}_h$ is a reconstructed version
X^r_h	The reference frame for learning 3-D dictionary
$\hat{X}^L_{LF}$, $\hat{X}^L_{HF}$	Scaled-up LF and HF frame from $\hat{X}_h$
$Z_l, \hat{Z}_l$	LR NKF, and $\hat{Z}_l$ is a reconstructed version
$\hat{Z}_{LF}$	Scaled-up HR frame from $\hat{Z}_l$
$\hat{Z}_{HF}$	Recovered HF frame from $\hat{Z}_{LF}$ by super-resolution
$\hat{Z}_h$	Recovered HR NKF by super-resolution
$\{\mathbf{D}^i_l, \mathbf{D}^i_h\}$	The i-th 2-D sub-dictionary pair for LR and HR frames
$\{\mathbf{D}_L, \mathbf{D}_H\}$	The 3-D dictionary pair for LR and HR frames

Section II provides the motivation and definition of STOL, including the proposed framework, the update properties, and analysis on spatio-temporal and cross-band consistency. The main learning algorithm is described in Section III with the analysis of the convergence rate and upper bound of prediction error. For validation, Section IV compares STOL, K-SVD and MOD over the synthetic data. Applied into video coding, Section V provides extensive experimental results to evaluate in terms of rate-distortion performance and visual quality. Finally, the conclusion is drawn in Section VI.

II. Sparse Representation With Spatio-Temporal Online Dictionary Learning

A. Preliminary

Given arbitrary signal $\mathbf{x} \in \mathbb{R}^m$ generated by a source with the underlying distribution $p(\mathbf{x})$, sparse coding is the process of computing its representation coefficients $\alpha \in \mathbb{R}^k$ based on the overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{m\times k}$. It admits a sparse approximation over $\mathbf{D}$ with the k columns referred to as atoms, when a linear combination of "few" atoms is "close" to $\mathbf{x}$.

To evaluate the sparse representation of $\mathbf{x}$ with $\mathbf{D}$, a loss function $l(\mathbf{x}, \mathbf{D})$ is to measure the divergence between the actual signal $\mathbf{x}$ and its optimal reconstruction with $\mathbf{D}$. Typically, it is formulated as an ℓ_1-sparse coding problem.

$$l(\mathbf{x}, \mathbf{D}) \triangleq \min_{\alpha\in\mathbb{R}^k} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\alpha\|_2^2 + \lambda\|\alpha\|_1, \tag{1}$$

where λ is a regularization parameter and α is solved under the ℓ_1 penalty. When $l(\mathbf{x}, \mathbf{D})$ is minimized, $\mathbf{D}$ is a "good" sparse representation for the signal $\mathbf{x}$. Thus, the expected cost $f(\mathbf{D})$ is minimized for arbitrary $\mathbf{x}$ generated by $p(\mathbf{x})$, which evaluates $l(\mathbf{x}, \mathbf{D})$ over the entire support of $\mathbf{x}$.

Definition 1 (Expected Cost): The expected cost $f(\mathbf{D})$ for the signal $\mathbf{x}$ generated by the underlying distribution $p(\mathbf{x})$ is

$$f(\mathbf{D}) \triangleq \mathbb{E}_{\mathbf{x}} l(\mathbf{x}, \mathbf{D}) = \int_{\mathbf{x}} l(\mathbf{x}, \mathbf{D})\, dp(\mathbf{x}), \tag{2}$$

where $\mathbf{D}$ is the trained overcomplete dictionary.

To avoid arbitrarily small values of α, the columns $(\mathbf{d}_j)_{j=1}^k$ of $\mathbf{D}$ are commonly enforced to have an ℓ_2 norm less than or equal to one. Consequently, we define $\mathcal{C}$ as the convex set of matrices verifying the constraint.

$$\mathcal{C} \triangleq \{\mathbf{D} \in \mathbb{R}^{m\times k} s.t. \forall j = 1, \ldots, k,\ \mathbf{d}_j^T\mathbf{d}_j \leqslant 1\} \tag{3}$$

Thus, $f(\mathbf{D})$ is a joint optimization problem with respect to the dictionary $\mathbf{D}$ and the coefficients $\alpha = [\alpha_1, \ldots, \alpha_n]$ of the sparse decomposition. However, the occurrence of $\mathbf{x}$ is modeled as the random independent sampling from the underlying distribution $p(\mathbf{x})$, which leads to inability of computing the cost function $f(\mathbf{D})$ directly. As an alternative, empirical cost is defined as an approximation of $f(\mathbf{D})$. Given a finite training set of interdependent signals $\mathcal{S} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, it minimizes the loss function $l(\mathbf{x}, \mathbf{D})$ over $\mathcal{S}$ to design optimal dictionary $\mathbf{D}$.

Definition 2 (Empirical Cost): The empirical cost $f_n(\mathbf{D})$ for the training set $\mathcal{S}$ with n samples is

$$f_n(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n} l(\mathbf{x}_i, \mathbf{D}), \tag{4}$$

where $\mathbf{D}$ is the trained overcomplete dictionary.

Eq. (4) implies that the overcomplete dictionary $\mathbf{D}$ is generated as a response to fit the mini-batch sample $\mathbf{x}_i$ in the training set $\mathcal{S}$ within an acceptable margin of error. Generally, $f_n(\mathbf{D})$ provides a good estimate of the expected cost $f(\mathbf{D})$ when the training set is large enough.

B. Definition and Motivation

Video sequences have similar structures and textures within one frame or among various frames, which tends to grow along the contours. When the support set for learning is consistent with the underlying structures and textures, coefficients for representation can be reduced by capturing them over the feature set. A simple example is the contourlet transform [34]. It approximates contour with a multi-resolution directional tight frame to make a sparser representation than wavelet. However, analytic dictionaries suffer from the curse of generality, which cannot adapt to the varying structures and textures with implicit and parametric mathematical models [35].

To improve generality, trained dictionaries are adopted to explicitly make sparse representation for specific video contents. Their atoms are iteratively optimized over the training set. At each iteration, These atoms are updated by minimizing the empirical cost $f_n(\mathbf{D})$ with batch gradient descents. Therefore, the trained dictionary $\mathbf{D}_{k+1}$ at iteration $k+1$ is obtained in a successive form.

$$\mathbf{D}_{k+1} = \mathbf{D}_k - \Phi_k \nabla_{\mathbf{D}} f_n(\mathbf{D}_k) = \mathbf{D}_k - \Phi_k \frac{1}{n}\sum_{i=1}^{n} \nabla_{\mathbf{D}} l(\mathbf{x}_i, \mathbf{D}_k), \tag{5}$$

where Φ_k is an appropriately selected positive definite symmetric matrix, which can make the convergence super-linear or quadratic in favorable cases. Nonetheless, batch gradient descents involve a burdening computation of averaging gradients of the loss function $\nabla_{\mathbf{D}} l(\mathbf{x}_i, \mathbf{D}_k)$ over the entire training set and massive memory consumption required to store the training set, which makes it impractical for video coding.

Online dictionary learning incorporates stochastic gradient descents [36] to relieve the high computational complexity and memory consumption. Instead of averaging over the entire training set, it randomly selects one sample $\mathbf{x}_t$ at iteration t and updates the dictionary $\mathbf{D}_t$ with gradient descent.

$$\mathbf{D}_{t+1} = \mathbf{D}_t - \eta_t \nabla_{\mathbf{D}} l(\mathbf{x}_t, \mathbf{D}_t), \tag{6}$$

where η_t is the learning rate. Averaging the update over all possible states of the samples $\mathbf{x}_t$ would restore the batch gradient descent algorithm. Given the training set composed of i.i.d. samples with distribution $p(\mathbf{x}) = 1/t$, online dictionary learning iteratively minimizes the approximate cost to alternate classical sparse coding steps for sparse representation.

Definition 3 (Approximate Cost): Given the training set $\mathcal{S}$ composed of i.i.d. samples, the approximate cost for $\mathcal{S}$ is

$$\hat{f}_t(\mathbf{D}) \triangleq \frac{1}{t}\sum_{i=1}^{t} l(\mathbf{x}_i, \mathbf{D}), \tag{7}$$

where $\mathbf{D}$ is the trained overcomplete dictionary.

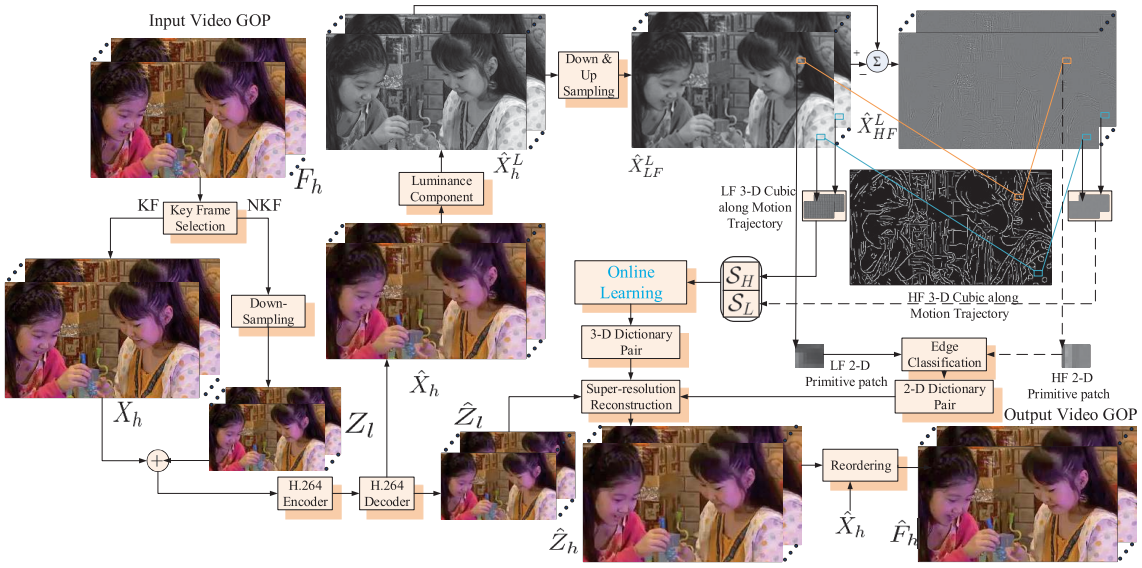


Fig. 1. The framework of sparse representation with spatio-temporal online dictionary learning for video coding.

Furthermore, online dictionary learning directly optimizes the expected cost with stochastic gradient descent rather than the empirical cost over the entire training set. The fact implies that its convergence rate can be enhanced by minimizing the approximate cost, as a gradient descent method with variable step $1/t$ is applied to the iterative optimization of dictionaries.

C. The Proposed Framework for STOL

Fig. 1 depicts the proposed framework of sparse representation with STOL for video coding. As in [21], a video sequence F_h with group of pictures (GOP) is decomposed into the selected high-resolution (HR) key frames (KFs) X_h and the down-sampled low-resolution (LR) non-key frames (NKFs) Z_l. X_h and Z_l are encoded and decoded by a standard H.264 codec. Denote $\hat{X}_h$ and $\hat{Z}_l$ the corresponding reconstructed KFs and NKFs, respectively. In the decoder, the low-frequency (LF) band is obtained by the down- and up-sampling operators, and the high-frequency (HF) band is obtained by subtracting LF band from HR band. The goal of super-resolution is to reconstruct the missing HF band for the decoded LR NKFs $\hat{Z}_l$. Consequently, the HR version $\hat{Z}_h$ of $\hat{Z}_l$ is recovered from $\hat{Z}_l$ by the learning-based super-resolution reconstruction.

In the learning phase, the LF frames are classified into a primitive layer, a non-primitive coarse layer, and a non-primitive smoothness layer. Hence, training data are collected to learn two corresponding kinds of dictionaries in alignment with adaptive reconstruction of the HF frames. The sparse representation of 2-D patches and 3-D volumes is optimized by adaptive regularized dictionary learning: a set of 2-D subdictionary pairs $(\mathbf{D}_l^i, \mathbf{D}_h^i), i = 1, 2, \ldots, K$ trained from primitive patches and a 3-D dictionary $(\mathbf{D}_L, \mathbf{D}_H)$ from non-primitive volumes. In detail, 2-D dictionary pairs are dedicated to spatial components, e.g. edge segments, bars, blobs, and terminations. While the 3-D dictionary pair focuses on blocks with high-frequency energy. The sparse representation of non-primitive volumes is constrained by the consistency along the motion trajectory by a trained 3-D spatio-temporal dictionary.

Considering that batch learning algorithms for training such dictionary suffer from a high computational complexity and a degraded coding efficiency for minimizing empirical cost $f_n(\mathbf{D})$ [36], it is desirable to directly optimize the expected cost $f(\mathbf{D})$ for an improved convergence rate with a tolerance of approximation error. Denote $\mathcal{S}_L$ the LF training set in $\mathbb{R}^{m\times n}$ and $\mathbf{D}_L$ the corresponding LF dictionary in $\mathbb{R}^{m\times k}$. The spatio-temporal dictionary pair $(\mathbf{D}_L, \mathbf{D}_H)$ is obtained from training sets $(\mathcal{S}_L, \mathcal{S}_H)$, where the 3-D LF and HF volumes are extracted and normalized from a set of frames along motion trajectory for consistency. For each reference frame X_h^r from KFs, its estimation $\tilde{X}_h$ is attained by motion-compensated frame interpolation approach to consider scene changes from its preceding and succeeding frames, as shown in Fig. 1. Consequently, an over-complete dictionary $\mathbf{D}_L$ with size $m \times k$ $(n \gg k > m)$ is trained to represent video sequences in a sparse manner.

Assuming that each patch for prediction is a linear combination of a small subset of patches with a coefficient matrix $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]$, where the ith column $\alpha_i \in \mathbb{R}^k$ is called the sparse solution to the i-th patch. Under such assumption, the super-resolution task is an energy minimization procedure.

$$f_{Video}(\alpha_i^{(t)}, X_h) = \min \sum_{i=1}^{n} \sum_{t=1}^{T} \left[\frac{1}{2} \| Z_l^i - \mathbf{D}_L \alpha_i^{(t)} \|_2^2 + \lambda \| \alpha_i^{(t)} \|_0 + \| \mathbf{D}_H \alpha_i^{(t)} - R_i^{(t)} X_h \|_2^2 \right], \quad (8)$$

where Z_l^i is the i-th patch extracted from the blurred and down-sampled version of X_h, $R_i^{(t)}$ is a projection matrix that selects the i-th patch from X_h at time t, and λ is a regularization parameter. At time t, $\alpha_i^{(t)}$ is derived based on $\mathbf{D}_L$ for sparse solution. In Eq. (8), the first and third terms evaluate the approximation and reconstruction error, respectively. While the second term restricts the number of coefficients to

maintain sparsity. For video sequences, 3-D volumes are obtained by concatenating the patches from the non-primitive coarse layer of X_h^r and those of $\tilde{X}_h$ in the same location.

$$f_{Video}^r(\alpha_i^r, X_h^r, F_h^r) = \min \sum_i \left(\frac{1}{2} \| Z_l^i - \mathbf{D}_L \alpha_i^r \|_2^2 + \lambda \| \alpha_i^r \|_0 + \| \mathbf{D}_H \alpha_i^r - R_i^r X_h \|_2^2 \right), \tag{9}$$

Cost function in Eq. (9) shows that dictionary pair $(\mathbf{D}_L, \mathbf{D}_H)$ can be learned from the training sets $(\mathcal{S}_L, \mathcal{S}_H)$.

The proposed STOL algorithm is developed to maintain the spatio-temporal consistency as structured sparsity, and consistency between the HF and LF band as hierarchical sparsity. While 2-D patches are represented based on the instinctive features for block-based structures by iteratively updating atoms of the dictionary, 3-D volumes are predicted by block-matching based motion estimation to maintain the consistency of the motion trajectory based on incomplete visual patterns. Furthermore, we relate the HF and LF band with the linear mapping for the missing details by scaling-up a down-sampled version of HF band.

D. Spatio-Temporal Consistency for STOL

In this section, we demonstrate the asymptotic equivalence among approximation cost, empirical cost and expected cost with the growth of sample size. The fact implies that the proposed STOL algorithm (based on approximate cost) and K-SVD (based on empirical cost) have the equivalent convergence rate for each iteration.

Since $f_n(\mathbf{D})$ provides a good estimate of $f(\mathbf{D})$ and $\hat{f}_t(\mathbf{D})$ is an approximation of $f_n(\mathbf{D})$, the divergence between $f(\mathbf{D})$ and $\hat{f}_t(\mathbf{D})$ is

$$\begin{aligned} &\mathbb{E}[f(\mathbf{D}^*) - \hat{f}_t(\mathbf{D}_t^*)] \\ &\quad = \mathbb{E}[|f_n(\mathbf{D}_n^*) - f(\mathbf{D}^*)|] + \mathbb{E}[|\hat{f}_t(\mathbf{D}_t^*) - f_n(\mathbf{D}_n^*)|], \end{aligned} \tag{10}$$

where $\mathbf{D}_n^* = \arg\min_{\mathbf{D}} f_n(\mathbf{D})$ and $\mathbf{D}_t^* = \arg\min_{\mathbf{D}} f_t(\mathbf{D})$ are the optimal solutions to the empirical and approximate cost. Here, we assume that $\mathbf{D}^*$ is a stationary point of the dictionary learning problem. The first term of Eq. (10) measures the accuracy of minimizing the empirical cost instead of the expected cost. The second term measures the approximation error for the proposed algorithm based on optimizing empirical cost. In Proposition 1, we demonstrate that the divergence between $f(\mathbf{D})$ and $\hat{f}_t(\mathbf{D})$ vanishes as t increases.

Proposition 1: Given training volume $\mathbf{x}$ *from video sequence and dictionary* $\mathbf{D}$ *in constraint* $\mathcal{C}$*, the approximate cost* $\hat{f}_t(\mathbf{D}_t)$ *converges almost surely to the expected cost* $f(\mathbf{D})$ *and their divergence converges almost surely to 0 as* $t \to \infty$.

Proof: Please refer to Appendix A. ∎

Proposition 1 shows that the approximate cost $\hat{f}_t(\mathbf{D}_t)$ approaches the expected cost $f(\mathbf{D})$ with sufficient sampling. Consequently, it also approximates the empirical cost $f_n(\mathbf{D}_n)$. Therefore, $\hat{f}_t(\mathbf{D}_t)$ is naturally minimized instead of $f_n(\mathbf{D})$ as an approximation of $f(\mathbf{D})$. By minimizing approximate cost to the expected cost, STOL has better ability to exploit the structured sparsity than K-SVD. Actually, experimental results also show that STOL achieves better approximation for sparse representation than K-SVD and its improved versions IDL.

E. Cross-Band Consistency for STOL

Besides spatio-temporal consistency, STOL also maintains the consistency between HF and LF bands. We construct a hierarchical structure like tree with 2 level and n disjoint branches corresponding to the number of sub-dictionary pairs, where a node and its parent should be selected at the same time. At the decoder, the HF band is reconstructed based on its recovered LF sparse representation coefficients over the proposed LF-HF dictionary pair.

$$\hat{Z}_h = \hat{Z}_{LF} + \hat{X}_{HF} = (\mathbf{D}_L + \mathbf{D}_H)\alpha = \sum_i \alpha_i (\mathbf{d}_{li} + \mathbf{d}_{hi}). \tag{11}$$

Using an interpolation operator $\mathcal{U}$, e.g. bicubic or bilinear, to fill in the missing rows and columns, the LF band $\hat{Z}_{LF}$ is obtained by scaling-up a down-sampled version Z_l of the reconstructed HF band $\hat{X}_h$.

$$\hat{Z}_{LF} = \mathcal{U}\hat{Z}_l = \mathcal{U}\mathcal{D}\hat{X}_h, \tag{12}$$

where $\mathcal{D}$ is the down-sampling operator. Denote $R_{i,j}$ the projection that selects the j-th pixel of i-th volume, $\hat{x}_{i,j} = R_{i,j}\hat{X}_h$. Analogically, the corresponding pixel $\hat{z}_{i,j}$ in the LR volume is obtained by $\hat{z}_{i,j} = R_{i,j}\hat{Z}_{LF} = R_{i,j}\mathcal{U}\mathcal{D}\hat{X}_h$. Therefore, we assume that $\hat{z}_{i,j} = (\mathcal{U}\mathcal{D})'\hat{x}_{i,j} + \hat{v}_{ij}$, where $(\mathcal{U}\mathcal{D})'$ is a local operator of $\mathcal{U}\mathcal{D}$ and $\hat{v}_{i,j}$ is the additive noise. Since $\mathbf{D}_H$ is the over-complete HF dictionary of k bases ($k > m$), $\hat{x}_{i,j}$ can be represented as $\hat{x}_{i,j} = \mathbf{D}_H\alpha_{i,j}$ with $\|\alpha_{i,j}\|_0 \ll n$.

$$\hat{z}_{i,j} = (\mathcal{U}\mathcal{D})'\mathbf{D}_H\alpha_{i,j} + \hat{v}_{i,j}. \tag{13}$$

Eq. (13) implies that $\hat{x}_{i,j}$ can be reconstructed with $\alpha_{i,j}$ over HR dictionary only when $\|\hat{z}_{i,j} - (\mathcal{U}\mathcal{D})'\mathbf{D}_H\alpha_{i,j}\|_2^2 \le \epsilon$ is satisfied for $z_{i,j}$ within a limited error ϵ. Under such condition, the consistency between the LF and HF bands is maintained, so that the missing HF details can be recovered from the its LR sparse representation coefficients derived from the corresponding LF dictionary $\mathbf{D}_L$. Therefore, reconstruction error shown as the third term in Eq. (9) can be omitted, as it approaches the limited error ϵ for sparse representation with LR frames.

III. Main Algorithm

A. Online Dictionary Learning Algorithm

As defined in Section II, the online dictionary learning technique aims to optimize the expected cost function over i.i.d. samples drawn from the underlying distribution $p(\mathbf{x}) = 1/t$.

$$f_t(\mathbf{D}) = \min_{\mathbf{D}\in\mathcal{C}, \alpha\in\mathbb{R}^{k\times n}} \frac{1}{t} \sum_{i=1}^{t} \left[\frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right] \tag{14}$$

The dictionary $\mathbf{D}$ is updated by the first-order projected stochastic gradient descent.

$$\mathbf{D}_{t+1} = \prod_{\mathcal{C}} [\mathbf{D}_t - \delta_t \nabla_{\mathbf{D}} l(\mathbf{x}_t, \mathbf{D}_t)], \tag{15}$$

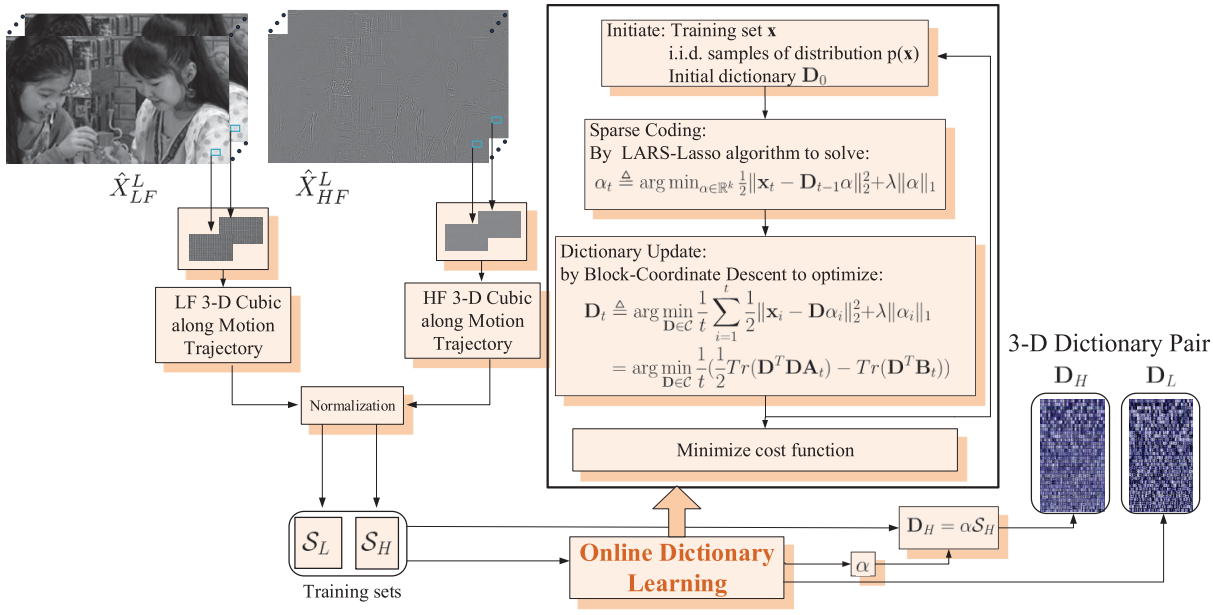


Fig. 2. The diagram of online dictionary learning for 3-D spatio-temporal dictionary. The 3-D dictionary pair $(\mathbf{D}_H, \mathbf{D}_L)$ is constructed by minimizing the expected cost over the training set $(\mathcal{S}_L, \mathcal{S}_H)$ for consistency along the motion trajectory.

where $\prod_{\mathcal{C}}$ is the orthogonal projector onto $\mathcal{C}$. In practice, $\mathbf{x}_t$ is obtained by cycling on a randomly permuted training set. The step for gradient descent is determined by $\delta_t \triangleq a/(t+b)$, where a and b are selected according to the statistics of signals.

For large-scale training data, online learning has competitive efficiency as iterative batch alternatives. In the proposed STOL algorithm for video coding, cost function Eq. (16) is minimized, which omits the reconstruction error in Eq. (9).

$$f(\mathbf{D}_L) = \min_{\mathbf{D}_L\in\mathcal{C},\alpha} \mathbf{E}_{\hat{Z}_l^i}\left(\frac{1}{2}\|\hat{Z}_l^i - \mathbf{D}_L\alpha_i\|_2^2 + \lambda\|\alpha_i\|_0\right) \quad (16)$$

Since ℓ_0-optimization is an NP-hard problem, a feasible strategy is to substitute ℓ_0-norm with ℓ_1-norm to deduce the optimal convex approximation.

$$\hat{f}_t(\mathbf{D}_L) = \min_{\mathbf{D}_L\in\mathcal{C},\alpha} \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}\|\hat{Z}_l^i - \mathbf{D}_L\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1\right) \quad (17)$$

Fig. 2 shows the diagram for the proposed STOL algorithm, where 3-D volumes are extracted from both LF and HF frames along the motion trajectory to maintain the spatio-temporal consistency. At arbitrary t-th iteration, one sample $\mathbf{x}_t$ is drawn from $p(\mathbf{x})$ at random. The minimization of cost function Eq. (17) is an ℓ_1-regularized least-squares problem, which can be solved by Cholesky decomposition for the classical LARS-Lasso algorithm. At iteration t, the coefficient α_t for $\mathbf{x}_t$ is obtained based on the trained dictionary $\mathbf{D}_{t-1}$ at iteration $t-1$, and each column of the dictionary D_t is updated under the convex constraint $\mathcal{C}$.

$$\mathbf{u}_j \leftarrow \frac{1}{\mathbf{A}_{jj}}\left(\mathbf{b}_j - \mathbf{D}\mathbf{a}_j\right) + \mathbf{d}_j, \ \ \mathbf{d}_j \leftarrow \frac{1}{\max\left(\|\mathbf{u}_j\|_2, 1\right)}\mathbf{u}_j$$

Here, the auxiliary matrices $\mathbf{A}, \mathbf{B}$ are updated by $\mathbf{A}_t \leftarrow \mathbf{A}_{t-1} + \alpha_i\alpha_i^T$ and $\mathbf{B}_t \leftarrow \mathbf{B}_{t-1} + \mathbf{x}_i\alpha_i^T$.

The dictionary $\mathbf{D}_t$ is updated by block-coordinate descent with warm restarts, which minimizes the expected cost Eq. (17) without tuning learning rate. For global optimum, $\mathbf{D}_t$ is updated column by column for convergence under $\mathcal{C}$. In STOL, it incorporates stochastic gradient descent to keep faster convergence than K-SVD with a guarantee of approximation error. Actually, the optimal dictionary to minimize the expected cost is served as the LF dictionary $\mathbf{D}_L$. While HF dictionary $\mathbf{D}_H$ is derived by relating projection matrix R, coefficients α, and the LF training set $\mathcal{S}_H$, in detail $\mathbf{D}_H = R\mathcal{S}_H\alpha^T(\alpha\alpha^T)^{-1}$.

B. Computational Cost for STOL

Video sequences are large-scale signals with high redundancy characterized by high-dimensional and low-rank visual structures. Consequently, prohibitive computational complexity is a fatal limitation for super-resolution based video coding schemes with batch learning algorithms. In this section, we show that the proposed learning algorithm outperforms batch gradient descent methods such as K-SVD in the sense of convergence speed and computational complexity for large-scale optimization problems. Inspired by Bottou and Bousquet [36], the proposed algorithm makes an estimation-approximation tradeoff.

$$\min_{\rho,n} \ \mathbb{E}\left[|f_n\left(\mathbf{D}_n^*\right) - f\left(\mathbf{D}^*\right)|\right] + \mathbb{E}\left[|\hat{f}_t\left(\mathbf{D}_t^*\right) - f_n(\mathbf{D}_n^*)|\right]$$
$$\text{s.t. } n \le n_{max}, T(\rho, n) \le T_{max} \quad (18)$$

According to Eq. (18), large-scale optimization problems are constrained by the maximal computing time T_{max} that enables processing more training samples to achieve better generalization, rather than the number of samples n_{max}.

For sufficient sampling under limited computing resources, STOL aims to enhance the convergence speed to rapidly derive the approximate solution $\mathbf{D}_t^*$ within a tolerable margin ρ of approximation error, $\hat{f}_t(\mathbf{D}_t^*) < f_n(\mathbf{D}_n^*) + \rho$. Thus, STOL can sufficiently exploit high spatial and temporal redundancies

within video sequences. Given optimal dictionary $\mathbf{D}^*$ for NKF reconstruction, the Hessian matrix $\mathbf{H}$ and gradient covariance matrix $\mathbf{G}$ are adopted for evaluation.

$$\mathbf{H} \triangleq \mathbb{E}\left[\nabla^2_{\mathbf{D}} l(\mathbf{x}, \mathbf{D}^*)\right], \quad \mathbf{G} \triangleq \mathbb{E}\left[\nabla_{\mathbf{D}} l(\mathbf{x}, \mathbf{D}^*) \nabla_{\mathbf{D}} l(\mathbf{x}, \mathbf{D}^*)^T\right], \tag{19}$$

Given arbitrary $\eta > 0$, with the probability at least $1 - \eta$, it satisfies under sufficient sampling

$$tr\left(\mathbf{G}\mathbf{H}^{-1}\right) \leq v \text{ and } EigenSpectrum\,(\mathbf{H}) \subset [\lambda_{min}, \lambda_{max}], \tag{20}$$

Denote $\kappa = \lambda_{min}/\lambda_{max}$ the ratio of eigenvalues. In Proposition 2, we show that STOL converges much faster than batch learning algorithms like K-SVD.

Proposition 2: Given the same computing resources, the time needed to reach accuracy ρ by STOL and batch learning algorithm are $\mathcal{O}(kn\kappa \log(1/\rho))$ and $\mathcal{O}(kv\kappa^2/\rho)$, respectively.

Proof: Please refer to Appendix B. ∎

Proposition 2 implies that STOL achieves better generalization performance for large-scale dynamic signals, as it performs more efficiently under limited computing resources. The accuracy for STOL and batch learning algorithms within same time t can be evaluated.

$$\left(\mathbf{D}_t - \mathbf{D}^*\right)^2 \sim \frac{1}{kn\kappa \log N} \ll \frac{1}{kv\kappa N} \sim (\mathbf{D}^*_N - \mathbf{D}^*)^2. \tag{21}$$

Remarkably, the convergence speed of batch learning algorithm decreases dramatically when the number of training samples grows. Its high computational cost is due to averaging cost over the entire training set and allocating memory for storage.

C. Update for Sequential Coding

In this section, the upper bound for prediction error is shown to be asymptotically equal to the training error. As shown in Eq. (6), $\mathbf{D}_t$ is updated based on the previous dictionary $\mathbf{D}_{t-1}$ and the selected sample $\mathbf{x}_t$. Thus, the generalization error between training and prediction shall be evaluated for sequential coding of NKFs.

To assure the properness of optimal $\mathbf{D}_t$ at iteration t, a relaxed approximate cost is adopted for evaluation, $f_t^{\gamma}(\mathbf{D}_t) = \sup_{\|\mathbf{D}-\mathbf{D}_t\|<2\gamma} f_t(\mathbf{D})$. In Proposition 3, we demonstrate the consistency between training and prediction.

Proposition 3: Given the trained dictionary $\mathbf{D}_t$ and arbitrary constant $\eta > 0$, with sufficient sampling, there exists $\varepsilon(l, \gamma, n, \eta) \to 0$ satisfying

$$Pr\left[\sup\left[\mathbb{E}_{\mathbf{x}} f(\mathbf{D}_t) - \mathbb{E}_{\mathcal{S}} f_t^{\gamma}(\mathbf{D}_t)\right] \leq \varepsilon\right] > 1 - \eta, \tag{22}$$

where $\mathbb{E}_{\mathbf{x}} f(\mathbf{D}_t)$ and $\mathbb{E}_{\mathcal{S}} f_t^{\gamma}(\mathbf{D}_t)$ are the estimated expectation for prediction and relaxed training error, respectively.

Proof: Please refer to Appendix C. ∎

In view of generalization error, Proposition 3 is translated as: given $\mathbf{D}_t$ and arbitrary constant $\eta > 0$, with probability of at least $1 - \eta$, the prediction error is bounded by

$$\mathbb{E}_{\mathbf{x}} f(\mathbf{D}_t) \leq \mathbb{E}_{\mathcal{S}} f_t^{\gamma}(\mathbf{D}_t) + \varepsilon(l, \gamma, n, \eta). \tag{23}$$

In Eq. (23), the second term is the excess error conditioned on the complexity of training model, which vanishes with sufficient sampling. The fact implies that the average prediction error $\mathbb{E}_{\mathbf{x}} f(\mathbf{D}_t)$ is asymptotically equivalent to the well-tuned average error in training.

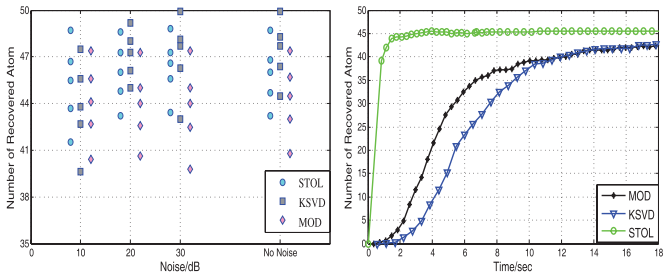
Proposition 3 ensures the predictive performance by relating the upper bound for prediction error to the tunable traing error. With sufficient sampling, STOL can asymptotically minimize the expected cost over the training set. In each GOP, since dictionary learning is based on the patches extracted from the key frames (KFs), the reconstruction of NKFs based on the trained dictionary would be affected by a deviation in statistics due to motion of objects. For the patch-based dictionary learning, such deviation in statistics is typically based on three facts. First, it is often difficult to match the motion trajectory in video sequence with a linear representation of patches, especially in low-level vision tasks related to intensity components [37]. The uncertainy of motion trajectory still remains unsolved, though [38] implies that it would be better to fit it under the implicit manifold constraints. Second, the training set would be incomplete for representing those patches for reconstruction, as distortions like deformation and occlusion would change the statistics of patches in a nonlinear manner [39]. Third, reconstruction of HF patches from LF ones based on the trained dictionary is an ill-posed problem [12]. Thus, Proposition 3 provides a theoretical validation to guarantee the reconstruction performance based on the dictionary trained by STOL, especially for those patches cannot be exactly inferred from the matched atoms in the training set.

IV. Application into Synthetic Signals

As in [23], we employ STOL on synthetic signals to evaluate its ability to recover the underlying dictionary that generates these signals. STOL is compared with the batch learning methods K-SVD and MOD [24]. The generating dictionary $\mathbf{D} \in \mathbb{R}^{20\times 50}$ was obtained based on i.i.d. samples drawn from uniform distribution with each of its columns normalized to a unit ℓ^2-norm. The training set $\{\mathbf{x}_i \in \mathbb{R}^{20}\}_{i=1}^{1500}$ was collected based on a linear combination of atoms of the three generating dictionary in random and independent locations. White Gaussian noise with variable SNR was added.

In training, the number of iteration was 50. 50 trials were conducted for noise levels of 10, 20, 30 dB and noiseless case, respectively. The number of recovered dictionary atom was estimated by measuring the error $1 - |d_i^T \tilde{d}_i|$. A dictionary atom was supposed to be recovered, when its approximation error was less than a threshold $\Delta = 0.01$.

Fig. 3 shows synthetic results for the three algorithms and the Y-axis is the mean number of recovered dictionary atoms beyond 50. Each point in Fig. 3(a) means the average result for ten experiments. It shows that STOL has competitive results with K-SVD and is more effective than MOD for all tested noise levels. Remarkably, STOL is more efficient in the sense of time consumption as shown in Fig. 3(b), where each trial was conducted 10 times for noise level of 10dB. The fact means that STOL updates more atoms under the same time.



(a) Recovery performance vs noise. (b) Recovery performance vs time.

Fig. 3. Recovery performance of synthetic signals for STOL, KSVD, and MOD, respectively. The number of recovered atoms are evaluated: (a) under free noise and noise levels 10, 20, and 30 dB; (b) with the growth of time.

V. Application into Super-Resolution Based Video Coding

A. Implementation

The proposed STOL algorithm is applied into super-resolution based video coding, where a GOP is composed of 16 successive frames, or 3 KFs and 13 NKFs. For each GOP, the first three frames are selected as key frames (KFs) and the remaining 13 frames are down-sampled with 1/4 size to serve as NKFs. Both KFs and NKFs are encoded by the H.264/AVC coding engine in the form of "$IPPP\cdots$." The KFs in both current GOP and next GOP are used to learn dictionary pairs for the consistency along motion trajectory. With the decoded low-resolution NKFs and its corresponding dictionary of 2-D patches and 3-D volumes, the HF details of NKFs are recovered.

In STOL, a volume size is set to $7 \times 7 \times 2$ for optimal coding with 3-D sparse representation. The training set collects 1024 volumes from each frame, which is a 98×1024 matrix. The numbers of iteration for STOL, K-SVD, and IDL are 150, 15 and 15, respectively. The regularization parameter λ is 0.15.

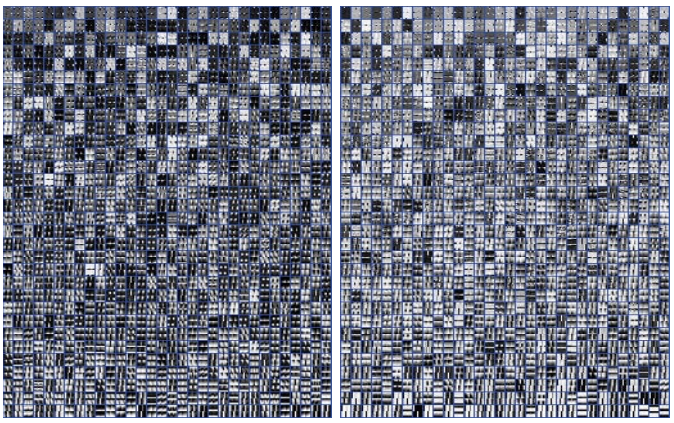
In experiments, we evaluate the performance over test sequences with the YUV 4:2:0 format, 30Hz frame rate, and various resolutions including CIF (352×288), WQVGA (416×240), and DVD (720×480). Benchmarks adopted for validation include the state-of-the-arts H.264/AVC, HEVC, adaptive regularized dictionary (ARD) scheme [21] with K-SVD, the scheme with IDL [27], and the fast batch alternative with Sparse K-SVD (S-KSVD) [26]. It is noted that IDL is the latest improved dictionary learning method based on K-SVD.

B. Dictionary Learning Performance

Initially, we compared learning performance of STOL, K-SVD, and IDL for video coding with fixed 100 iterations. Fig. 4 presents the spatio-temporal dictionaries learned by KSVD and STOL. For evaluation, we define the accuracy f_{na}, sparsity f_{ns} based on cost function $f(\mathbf{x}, \mathbf{D})$.

$$f_{na}(\mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2$$

$$f_{ns} \triangleq \frac{1}{n}\sum_{i=1}^{n}\|\alpha_i\|_1 f(\mathbf{x}, \mathbf{D}) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda\|\alpha_i\|_1\right)$$



(a) (b)

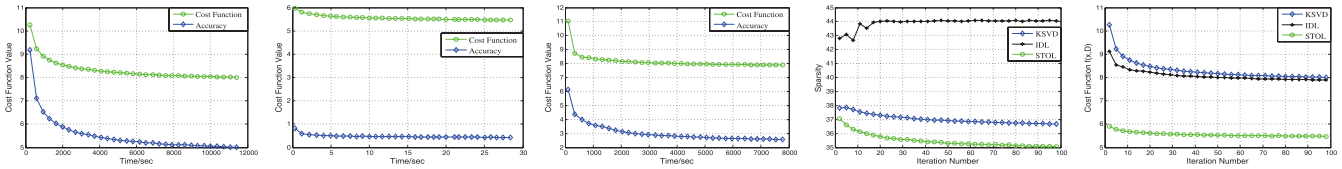
Fig. 4. The trained dictionary with K-SVD and STOL, respectively. (a) K-SVD. (b) STOL.

They indicate the ability of learning-based dictionary in sparse representation for video frames. Fig. 5 displays the variation of accuracy f_{na}, sparsity f_{ns}, and cost function $f(\mathbf{x}, \mathbf{D})$ with the growth of time and iteration number, respectively. Comparing Fig. 5 (a)-(c), it shows that the proposed scheme significantly reduces the computational complexity with an improvement of prediction performance. In comparison to K-SVD and IDL, STOL requires less time for training each sample with a rapid convergence in large-scale learning of dynamic signals. Furthermore, STOL is the most trivial in the sense of cost function, which means that it is desirable to directly optimize the expected cost for video coding.

C. Rate-Distortion Performance

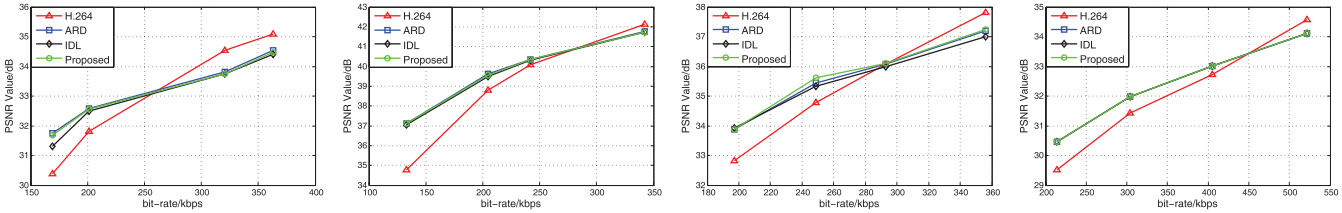
Fig. 6 provides rate-distortion curves of various video sequences obtained by the proposed scheme, H.264/AVC, ARD and IDL. Within the low bit-rate region, the proposed scheme is competitive with ARD and H.264/AVC and outperforms IDL in the rate-distortion performance. For a complete validation, Table II provides BD-PSNR gain and BD-rate reduction [42] for the proposed scheme, ARD, and IDL over H.264/AVC. In comparison to H.264/AVC, BD-PSNR gain and BD-rate reduction are up to 0.75 dB and 7.5%, respectively. In summary, the proposed STOL is comparable to ARD and outperforms H.264/AVC and IDL, especially for video sequences with complex motion. Moreover, the proposed scheme is compared with HEVC (HM 8.0) over *Foreman*, *Akiyo*, and *Waterfall*. In Fig. 7, it can be seen that the proposed scheme keeps comparative performance with HEVC, especially for video sequences with low bit-rates and subtle motion, e.g. *Akiyo*. Furthermore, we compare the proposed scheme with the fast batch alternative Sparse K-SVD. Fig. 8 shows the proposed scheme outperforms Sparse-K-SVD in a noticeable margin.

To further validate the proposed method, we evaluate STOL and H.264/AVC with "full power" over extensive test sequences with various resolutions. Here, the GOP structure is set to "IBPB" with up to five reference frames and quater-pixel interpolation for H.264/AVC. Table III shows the



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

Fig. 5. Results for convergence rate of K-SVD, STOL, and IDL in terms of accuracy, sparsity and cost function with iteration number 100. (a) K-SVD. (b) STOL. (c) IDL. (d) Sparsity comparison. (e) Cost comparison.



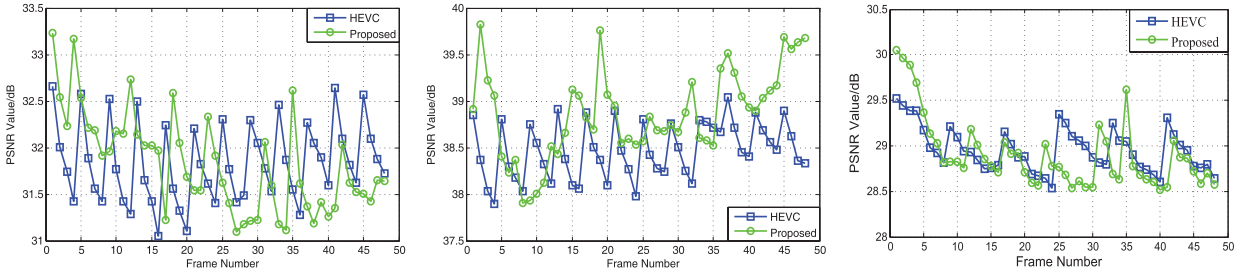
(a) (b) (c) (d)

Fig. 6. Rate-distortion curve for performance comparison. The proposed model is compared with H.264/AVC, ARD, and IDL, respectively. (a) *Foreman*. (b) *Akiyo*. (c) *News*. (d) *Waterfall*.

TABLE II

PSNR (dB) Performance for the Proposed Scheme, H.264/AVC, ARD, and IDL in the Low Bit-Rate Region, Respectively. The Proposed Scheme Is Compared With ARD and IDL in Terms of BD-PSNR (dB) Gain and BD-Rate (%) Reduction Over H.264/AVC

Sequence	Bit-rate (kbps)	H.264	Proposed			ARD			IDL		
		PSNR	PSNR	BD-PSNR	BD-rate	PSNR	BD-PSNR	BD-rate	PSNR	BD-PSNR	BD-rate
Foreman	169.4	30.393	31.698	0.011	-3.823	31.739	0.047	-3.350	31.313	-0.020	-2.152
	201.5	31.815	32.552			32.578			32.505		
	320.3	34.542	33.753			33.803			33.737		
	363.0	35.096	34.478			34.544			34.411		
Akiyo	132.9	34.791	37.109	0.752	7.409	37.106	0.777	7.952	37.054	0.698	6.430
	204.5	38.784	39.569			39.607			39.490		
	241.7	40.078	40.340			40.358			40.318		
	342.4	42.125	41.744			41.764			41.734		
News	197.0	32.822	33.879	0.437	7.507	33.830	0.356	4.742	33.886	0.242	3.361
	248.8	34.773	35.446			35.328			35.610		
	293.0	36.086	36.083			35.983			36.088		
	356.2	37.812	37.193			36.994			37.252		
Waterfall	213.9	29.516	30.471	0.405	7.189	30.461	0.405	7.182	30.472	0.405	7.219
	303.8	31.434	31.968			31.971			31.968		
	405.2	32.728	33.014			33.016			33.015		
	521.2	34.567	34.106			34.101			34.103		



(a) (b) (c)

Fig. 7. PSNR (dB) performance of the reconstructed frames for the proposed scheme and HEVC, respectively. (a) *Foreman*: 89.3 kbps. (b) *Akiyo*: 88.3 kbps. (c) *Waterfall*: 57.6 kbps.

coding performance for the proposed STOL and H.264/AVC in terms of PSNR and SSIM, respectively. In overall, the average BD-PSNR gain and BD-rate reduction in comparison to H.264/AVC with GOP structure "IBPB" are about 0.10 dB and 1%, respectively. Moreover, Fig. 9 provides the rate-distortion curve for test sequences with various resolutions. We can find that the proposed method still outperforms H.264/AVC in most cases, especially in the low and moderate bit-rate region, even though better motion estimation and motion compensation would be achieved by using B-frames.

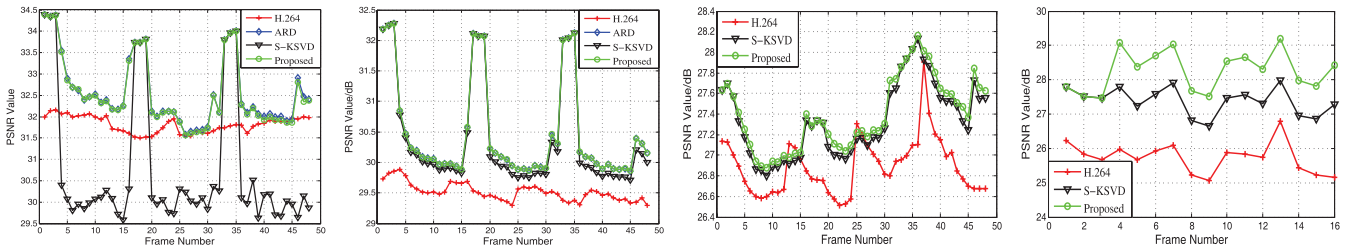


Fig. 8. PSNR (dB) performance of the reconstructed frames from the proposed scheme, H.264/AVC, ARD, and Sparse K-SVD, respectively. (a) *Foreman*: 201.5 kbps. (b) *Waterfall*: 213.9 kbps. (c) *BlowingBubbles*: 185.7 kbps. (d) *Driving*: 758.1 kbps.

TABLE III

PSNR (dB) AND SSIM PERFORMANCE FOR TEST SEQUENCES WITH VARIOUS RESOLUTION OBTAINED BY THE PROPOSED METHOD AND H.264/AVC WITH GOP STRUCTURE "IBPB", RESPECTIVELY. HERE, BD-PSNR AND BD-RATE REFER TO BD-PSNR GAIN (dB) AND BD-RATE REDUCTION (%)

Sequence	Resolution	Bit-rate (kbps)	H.264		Proposed			
			SSIM	PSNR	SSIM	PSNR	BD-PSNR	BD-rate
Foreman	352 × 288	142.2	0.8494	30.63	0.8683	31.63	0.12	0.67
		172.6	0.8672	31.71	0.8683	32.37		
		261.2	0.9010	33.97	0.9018	33.66		
		332.1	0.9149	35.08	0.9126	34.29		
Akiyo		136.2	0.9548	37.45	0.9611	38.28	0.04	0.15
		178.8	0.9647	39.22	0.9684	39.57		
		241.6	0.9722	41.02	0.9737	40.69		
		302.9	0.9769	42.28	0.9768	41.36		
News		176.7	0.9297	33.51	0.9397	34.60	0.15	-0.86
		212.9	0.9427	34.91	0.9476	35.33		
		258.9	0.9520	36.21	0.9542	35.96		
		309.6	0.9606	37.51	0.9605	37.08		
Waterfall		204.7	0.7691	30.28	0.8052	30.97	0.33	6.68
		266.9	0.8205	31.52	0.8463	31.99		
		349.8	0.8658	32.84	0.8824	33.16		
		445.5	0.9033	34.26	0.8992	33.81		
BlowingBubbles	416 × 240	270.1	0.6965	27.07	0.7264	27.55	0.10	1.32
		353.2	0.7408	28.12	0.7613	28.38		
		461.8	0.7808	29.19	0.7910	29.15		
		603.6	0.8165	30.31	0.8182	29.94		
BQMall	832 × 480	1121.6	0.7988	28.31	0.8238	28.73	-0.08	-1.99
		1388.7	0.8331	29.51	0.8490	29.75		
		1723.2	0.8618	30.69	0.8647	30.32		
		2144.6	0.8869	31.92	0.8833	31.29		
ParkScene	1920 × 1080	1074.1	0.8981	29.75	0.9101	30.22	0.07	0.42
		1447.5	0.9258	30.79	0.9298	30.92		
		1936.8	0.9456	31.82	0.9486	31.78		
		2557.2	0.9618	32.90	0.9594	32.67		
Average		-	-	-	-	-	0.10	0.91

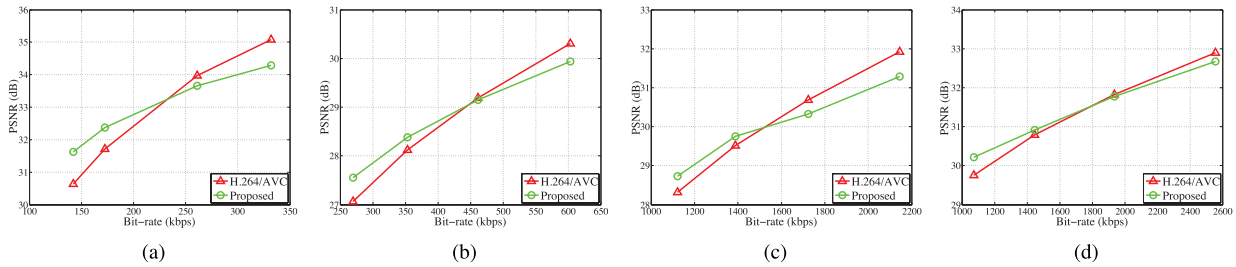


Fig. 9. Rate-distortion curves for the proposed method and H.264/AVC with GOP structure "IBPB", respectively. (a) *Foreman* (352 × 288). (b) *BlowingBubbles* (416 × 240). (c) *BQMall* (832 × 480). (d) *ParkScene* (1920 × 1080).

Remarkably, the proposed method can achieve up to 0.33 dB gain in PSNR and 6.68% reduction in bit rate.

It is worth mentioning that rate-distortion performance for the proposed method would vary with the statistics of the video sequences. This fact implies that training and reconstruction with various patch sizes would fit various statistics within the video sequences. Thus, the proposed method would be improved by considering a hierarchical tree to adaptively determine the patch size for training and reconstruction. In future work, it would be promising to adopt fast decision algorithm

Fig. 10. The visual comparison between the proposed scheme and H.264/AVC. From left to right and from top to bottom: original and reconstructed video frames of *Foreman*, *Waterfall*, *Driving*, and *BlowingBubbles* obtained by H.264/AVC, and the propose scheme, respectively.

TABLE IV

SUBJECTIVE QUALITY IN TERMS OF SSIM AND DMOS FOR THE PROPOSED ALGORITHM, H.264/AVC, ARD, AND IDL, RESPECTIVELY

Sequence	Bit-rate (kbps)	Average SSIM				Average DMOS			
		H.264	ARD	IDL	Proposed	H.264	ARD	IDL	Proposed
Foreman	169.4	0.8447	0.8694	0.868	0.8691	40.716	36.833	36.994	36.883
	201.5	0.8680	0.8831	0.8828	0.8828	37.054	34.680	34.730	34.725
	320.3	0.9079	0.9003	0.8999	0.9000	30.893	32.034	32.090	32.067
	363.0	0.9146	0.9106	0.9100	0.9104	29.926	30.513	30.588	30.532
Akiyo	132.9	0.9339	0.9544	0.9538	0.9543	27.317	24.862	24.935	24.881
	204.5	0.9628	0.9682	0.9673	0.9680	23.957	23.408	23.503	23.437
	241.7	0.9690	0.9713	0.9711	0.9712	23.334	23.109	23.125	23.122
	342.4	0.9764	0.9773	0.9771	0.9772	22.629	22.545	22.556	22.555
News	197.0	0.9221	0.9355	0.9349	0.9359	28.879	27.113	27.193	27.064
	248.8	0.9418	0.9483	0.9475	0.9490	26.332	25.557	25.654	25.473
	293.0	0.9516	0.9526	0.9517	0.9528	25.177	25.069	25.164	25.044
	356.2	0.9629	0.9597	0.9586	0.9601	23.948	24.285	24.285	24.242
Waterfall	213.9	0.7329	0.7794	0.7793	0.7794	53.623	49.461	49.473	49.461
	303.8	0.8165	0.8406	0.8406	0.8406	44.869	41.344	41.340	41.346
	405.2	0.8621	0.8734	0.8734	0.8734	38.001	36.206	36.203	36.207
	521.2	0.9113	0.9022	0.9022	0.9022	30.402	31.742	31.746	31.748

to determine current patch size from the reconstructed patches with convolutional neural network (CNN) or probabilistic models with joint optimization for adjacent patches.

D. Visual Quality

Fig. 10 shows the proposed scheme achieves better visual quality than H.264/AVC, especially in the texture regions like "tree" region in *Waterfall* and *Driving*. To further compare the visual effects, SSIM [40] and DMOS [41] are introduced, where higher SSIM or smaller DMOS scores represent better visual quality. Obviously, Table IV shows the proposed scheme achieves best visual quality. To be concrete, it is competitive with ARD and IDL and obviously outperforms H.264/AVC. Similarly, Fig. 11 and 12 compare SSIM performance with HEVC and Sparse K-SVD, respectively. It is consistent with the rate-distortion comparison in Section V-C.

E. Computational Complexity

Without loss of generality, we discuss the the computational complexity involving with learning and decoding process. In learning phase, the proposed scheme commits a fast convergence speed to update with the training data

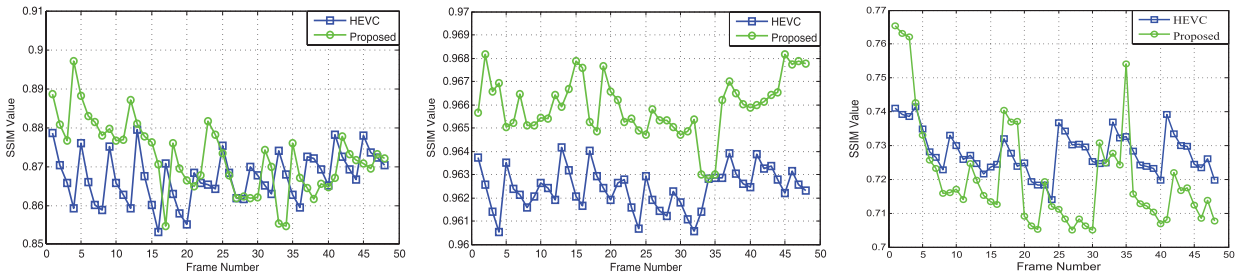


Fig. 11. SSIM performance of the reconstructed frames for the proposed scheme and HEVC, respectively. (a) *Foreman*: 89.3 kbps. (b) *Akiyo*: 88.3 kbps. (c) *Waterfall*: 57.6 kbps.

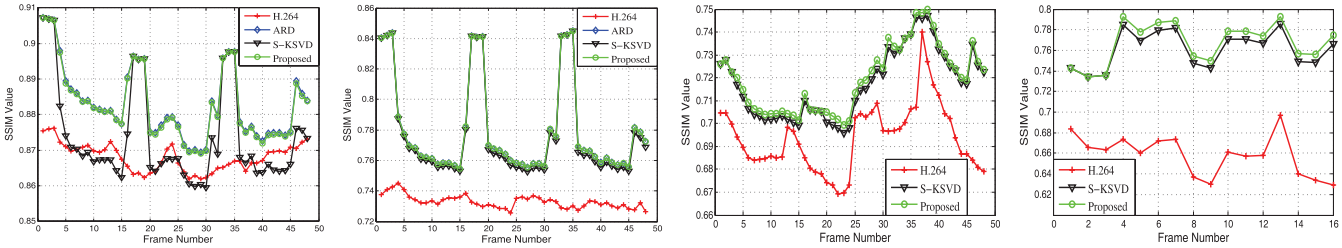


Fig. 12. SSIM performance of the reconstructed frames obtained by the proposed scheme, H.264/AVC, ARD, and Sparse K-SVD, respectively. (a) *Foreman*: 201.5 kbps. (b) *Waterfall*: 213.9 kbps. (c) *BlowingBubbles*:185.7 kbps. (d) *Driving*:758.1 kbps.

TABLE V

COMPUTATIONAL COMPLEXITY IN TERMS OF LEARNING SPEED (SEC/GOP) AND DECODING SPEED (SEC/FRAME) FOR THE PROPOSED SCHEME, ARD, AND IDL, RESPECTIVELY. RUN-TIME RATIOS (%) FOR THE COMPARATIVE METHODS ARE ASSESSED AS: RT-RATIO $= t_{COMP}/t_{PROP} \times 100\%$.

Sequence	Learning complexity					Decoding complexity				
	Proposed	ARD		IDL		Proposed	ARD		IDL	
	Speed	Speed	RT-RATIO	Speed	RT-RATIO	Speed	Speed	RT-RATIO	Speed	RT-RATIO
Foreman	115.45	1556.19	1347.93	1314.86	1138.90	437.20	527.25	120.60	506.54	115.86
Akiyo	83.28	1567.17	1881.81	1341.20	1610.47	301.73	394.42	130.72	380.29	126.04
News	81.46	1433.20	1759.39	1245.91	1529.47	368.40	446.76	121.27	434.43	117.92
Waterfall	100.69	351.32	348.91	332.09	329.81	392.19	407.86	104.00	405.80	103.47

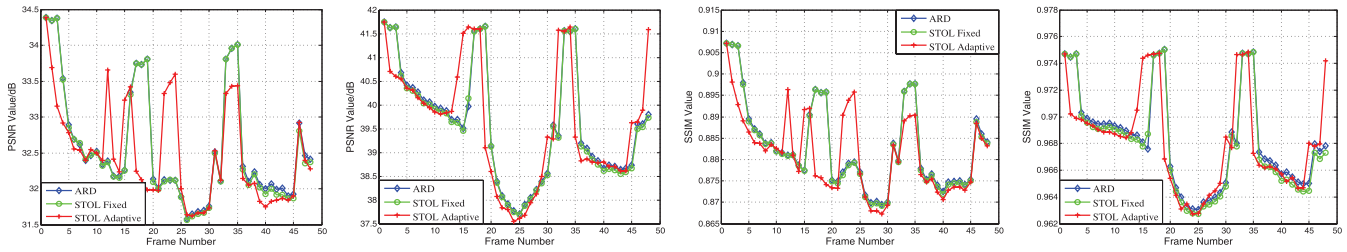
TABLE VI

COMPUTATIONAL COMPLEXITY FOR THE DECODERS OF THE PROPOSED SCHEME, H.264/AVC, AND HEVC, RESPECTIVELY. FOR THE PROPOSED SCHEME, BOTH THE DICTIONARY LEARNING SPEED (SEC/GOP) AND RECONSTRUCTION SPEED (SEC/FRAME) ARE PROVIDED

Sequence	H.264		Proposed		HEVC		Proposed	
	Encoding	Decoding	Learning	Reconstruction	Encoding	Decoding	Learning	Reconstruction
Foreman	2942.78	0.41	187.47	205.84	1245.48	1.14	177.52	214.36
Akiyo	2464.43	0.24	117.13	168.24	985.46	0.55	99.13	119.61
News	2939.22	0.27	128.44	235.15	766.27	0.75	107.33	216.38
Waterfall	2463.28	0.36	291.39	658.23	649.51	0.88	223.47	469.57
BlowingBubbles	2203.40	0.44	421.17	358.70	1319.58	1.22	247.84	365.57
BQMall	11710.52	1.24	170.92	4567.19	4473.75	3.77	125.88	2522.26
ParkScene	94500.23	8.72	334.63	43784.60	22553.45	19.72	375.44	42871.34

in a refined manner. It is worth mentioning that the proposed scheme iterates ten-fold more times than K-SVD in comparable elapsed time. Decoding performance depends on the reconstruction of NKFs by combining atoms with sparse representation coefficients. Since the proposed scheme directly optimizes the expected cost, it obtains sparser coefficients for less computation. To keep fairness, all the experiments are implemented with Matlab on a PC with 3.0 GHz dual-core CPU and 8G RAM. Table V provides the computational complexity for the proposed scheme, ARD, and IDL in terms of learning and decoding speed, where the run-time ratio is used for evaluation. For ARD and IDL, their run-time ratios for learning range from 351% to 1567% and 392% to 1610%, respectively. While there is an approximately 20% reduction in decoding complexity. The facts mean that the proposed scheme is efficient for video coding schemes with learning-based super-resolution, which relieves the prohibitive complexity for batch learning algorithms like K-SVD and IDL.

In addition, Table VI provides the complexity comparison with H.264/AVC and HEVC. For clarity, the complexity for learning and reconstruction are both provided. Here, the learning complexity is evaluated in terms of seconds per GOP, as we make dictionary learning for each GOP. In general, the computational complexity is about 1.5-10 times and 3-20 times



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

Fig. 13. PSNR and SSIM performance for fixed and adaptive selection of key frames in the proposed scheme, respectively. (a) PSNR for *Foreman*: 201.5*kbps*. (b) PSNR for *Akiyo*: 204.5*kbps*. (c) SSIM for *Foreman*: 201.5*kbps*. (d) SSIM for *Akiyo*: 204.5*kbps*.

in comparison to H.264/AVC and HEVC, respectively. The computational complexity for learning has been significantly reduced to about 0.5 ms for each pixel. Considering that the proposed method is implemented with Matlab, it is possible to further reduce the computational complexity for practical purpose by transplanting the proposed algorithm to C/C++. Remarkably, it would be better to develop a flexible and hierarchical structure to determine patch sizes. Thus, fast decision algorithm can be adopted to reduce the computational complexity for dictionary-based reconstruction with a guarantee of rate-distortion performance.

F. Key Frames Selection

Naturally, the selection of key frame would affect reconstruction performance. The optimal point of PSNR and SSIM value is generally located at the third key frame in a GOP. The proposed scheme utilizes two couples of first three successive frames for dictionary learning, and it is shown to be close to the adaptive scheme in the sense of coding performance as follows.

The adaptive scheme searches uncorrelated key frames to remove temporal redundancies based on the trained dictionary pairs. To measure the difference on structure characteristics for each frame, primitive frames are extracted by orientation energy through a set of Gaussian derivative filters.

$$OE_{\sigma,\theta} = (I * f_{\sigma,\theta}^{odd})^2 + (I * f_{\sigma,\theta}^{even})^2, \tag{24}$$

where $f_{\sigma,\theta}^{odd}$ and $f_{\sigma,\theta}^{even}$ are the first and second Gaussian derivative filters at scale σ and orientation θ, respectively. The key frames in the GOP G_n are selected by maximizing their mutual difference $D(i, j)$.

$$i = \arg \max_{i,j \in G_n, i \neq j} D(i, j) = arg \max_{i,j \in G_n, i \neq j} \sum_{j=1}^{L-1} (OE_i - OE_j) \tag{25}$$

Here, L is the size of G_n, and OE_i and OE_j are primitive areas in i-th and j-th frame, respectively. We select three frames with large differences $D(i, j)$ as KFs.

Fig. 13 compares the PSNR and SSIM performance with fixed and adaptive KF selection scheme for *Foreman* and *Akiyo*. The adaptive scheme does not show advantages over the fixed one, as large motion vectors between KFs require more bit-rates for coding. For example, the adaptive scheme selects the 33-th to 35-th frames as key frames and reconstructs them with a lower PSNR than the fixed scheme. Moreover, the adaptive scheme would be trivial for video sequences with subtle motion, as similar structures in one GOP narrows the gap between the largest and smallest value of $D(i, j)$.

VI. Conclusion

In this paper, a spatio-temporal online dictionary learning (STOL) algorithm is proposed for sparse representation in efficient video coding. By randomly selecting one i.i.d. sample from the underlying distribution to update the dictionary at each iteration, it trains the 3-D low-frequency and high-frequency dictionary pair for asymptotically optimal representation and fast convergence rate. Moreover, it directly optimizes the expected cost rather than the empirical cost to maintain the spatio-temporal and cross-band consistency for video volumes. It requires lower memory consumption and computational cost without explicitly tuning the learning rate. For large-scale optimization problems, it has been shown that the STOL scheme could maintain both structured sparsity and hierarchical sparsity with better approximation and convergence speed than batch gradient descent algorithms. Applied into large-scale dynamic video signals, STOL is integrated into the framework of super-resolution based video coding. Experimental results show that the proposed scheme achieves better coding performance with a significantly reduced computational complexity in comparison to K-SVD and IDL.

Appendix A
Proof of Proposition 1

Firstly, we verify that the proposed cost function $f(\mathbf{D})$ satisfies the four assumptions in [22].

i) According to Eq. (14), $f(\mathbf{D})$ is three-order differentiable with continuous derivatives. Its lower bound is 0.

ii) For the learning rates $\eta_t = 1/t$, we have $\sum_{i=1}^{\infty} \eta_t = \infty$ and $\sum_{i=1}^{\infty} \eta_t^2 = \frac{\pi^2}{6} < \infty$.

iii) Since eigenvalues of Hessian matrix $\mathbf{H}$ are constrained in $[\lambda_1, \lambda_2]$, it holds for non-negative constants A_k and B_k.

$$\mathbb{E}_{\mathbf{x}}(\|\nabla_{\mathbf{D}} l(\mathbf{x}, \mathbf{D})\|^k) \leq A_k + B_k \|\mathbf{D}\|^k, \ k = 2, 3, 4 \tag{26}$$

iv) Columns $\{\mathbf{d}_i\}$ of $\mathbf{D}$ are constrained by $\mathcal{C}$ in Eq. (3).

Hence, we consider Eq. (10) for proof. When randomly selecting samples for updating $\{\mathbf{d}_i\}$, the divergence

between the expected and empirical cost is upper-bounded by Vapnik-Chervonenkis (VC) bound $c\sqrt{(k/n)\log(n/k)}$. To simplify, its logarithmic term can be eliminated with chain techniques.

$$\mathbb{E}[\sup|f(\mathbf{D}^*) - f_n(\mathbf{D}^*)|] \le c\sqrt{\frac{k}{n}} \xrightarrow[n\to\infty]{} 0 \quad a.s.,$$

where c is arbitrary positive constant. Thus, we obtain

$$f(\mathbf{D}^*) - f_n(\mathbf{D}^*) \xrightarrow[n\to\infty]{} 0 \quad a.s. \tag{27}$$

For the divergence between the approximate and empirical cost, we show their convergence at first. Since $f_n(\mathbf{D}) - f_n(\mathbf{D}_n)$ is Lipschitz, [31, Proposition 4] shows that the divergence between $\mathbf{D}_n^*$ and $\mathbf{D}^*$ converges almost surely to 0 when $n \to \infty$.

$$f_n(\mathbf{D}^*) - f_n(\mathbf{D}_n^*) \xrightarrow[n\to\infty]{} 0 \quad a.s. \tag{28}$$

For the approximate cost, $\hat{f}_{t+1}(\mathbf{D}_{t+1}) - \hat{f}_t(\mathbf{D}_t)$ is considered.

$$\begin{aligned}
&\hat{f}_{t+1}(\mathbf{D}_{t+1}) - \hat{f}_t(\mathbf{D}_t) \\
&\quad = \hat{f}_{t+1}(\mathbf{D}_{t+1}) - \hat{f}_{t+1}(\mathbf{D}_t) \\
&\qquad + \left[\frac{l(\mathbf{x}_{t+1}, \mathbf{D}_t) - f_n(\mathbf{D}_t)}{t+1} + \frac{f_n(\mathbf{D}_t) - \hat{f}_t(\mathbf{D}_t)}{t+1}\right].
\end{aligned}$$

Here, $\hat{f}_{t+1}(\mathbf{D}_{t+1}) - \hat{f}_{t+1}(\mathbf{D}_t) \le 0$, as $\mathbf{D}_t$ is updated by $\mathbf{D}_{t+1}$. Since f_t is the lower bound of its approximation $\hat{f}_t$, it holds $\hat{f}_{t+1}(\mathbf{D}_{t+1}) - \hat{f}_{t+1}(\mathbf{D}_t) \le 0$. Thus, the upper bound conditioned on previous information $\mathcal{F}_t$ for dictionaries and sparse representation coefficients are developed.

$$\begin{aligned}
&\mathbb{E}\left[\hat{f}_{t+1}(\mathbf{D}_{t+1}) - \hat{f}_t(\mathbf{D}_t)\,|\mathcal{F}_t\right] \\
&\le \frac{\mathbb{E}[l(\mathbf{x}_{t+1}, \mathbf{D}_t)|\mathcal{F}_t] - f_n(\mathbf{D}_t)}{t+1} \le \frac{\|f(\mathbf{D}_t) - f_n(\mathbf{D}_t)\|}{t+1} \\
&\le \frac{c\sqrt{k}}{\sqrt{t}(t+1)}.
\end{aligned}$$

Here, $f_t(\mathbf{D}_t)$ approaches $f_n(\mathbf{D}_n)$ when $t \to n$. Since $\mathbb{E}[\|f(\mathbf{D}^*) - f_n(\mathbf{D}^*)\|_\infty] \le c\sqrt{(k/n)}$, approximate cost converges almost surely as t increases to ∞.

Finally, we prove the divergence ρ between the approximate and empirical cost vanishes when $t \to n$ and $n \to \infty$. Since $\hat{f}_t(\mathbf{D}_t)$ is non-negative quasi-martingale [22], its holds $\sum_{t=0}^{\infty}\mathbb{E}[\hat{f}_{t+1}(\mathbf{D}_{t+1}) - \hat{f}_t(\mathbf{D}_t)|\mathcal{F}_t] < +\infty\, a.s$ almost surely for the conditional expectation of its variations. Introducing the converging f_n, the cumulative divergence between the approximate and empirical cost is bounded.

$$\sum_{t=0}^{\infty}\mathbb{E}[\hat{f}_t(\mathbf{D}_t) - f_n(\mathbf{D}_t)|\mathcal{F}_t] < +\infty \quad a.s.$$

Therefore, the divergence ρ converges almost surely to 0, when $t < n \to \infty$.

$$\hat{f}_t(\mathbf{D}_t^*) - f_n(\mathbf{D}_t^*) \xrightarrow[t<n\to\infty]{} 0 \quad a.s. \tag{29}$$

Eq. (29) implies that ρ can be reduced with sufficient sampling. Combining Eq. (27)-(29), the upper bound for Eq. (10) is developed.

$$\begin{aligned}
&\mathbb{E}[|f(\mathbf{D}^*) - f_n(\mathbf{D}_n^*)|] + \mathbb{E}[|\hat{f}_t(\mathbf{D}_t^*) - f_n(\mathbf{D}_n^*)|] \\
&\quad \le c\sqrt{\frac{k}{n}} + \rho = \mathcal{O}\left(\rho + \sqrt{\frac{k}{n}}\right) \xrightarrow[n\to\infty]{} 0 \quad a.s.
\end{aligned}$$

As a result, the divergence between the expected and approximate cost vanishes almost surely with sufficient sampling.

Appendix B
Proof of Proposition 2

Murata [43] has proved that the learning curve of optimal 1/t-annealed online learning is asymptotically given by

$$\mathbb{E}\left[\hat{f}_t(\mathbf{D}_t)\right] = f(\mathbf{D}^*) + \frac{1}{2t}tr(\mathbf{G}_*\mathbf{H}^{-1}). \tag{30}$$

It coincides with batch learning algorithm in the order of $\mathcal{O}(1/t)$, which means STOL has the same convergence rate with K-SVD at iteration t. However, Eq. (5) and (6) show that STOL only needs k derivatives for one sample rather than all n samples required by K-SVD. Thus, their computational complexities for each iteration are $\mathcal{O}(nk)$ and $\mathcal{O}(k)$, respectively.

Batch learning algorithms require $\mathcal{O}(\kappa\log(1/\rho))$ for each derivative to reach accuracy ρ [44]. Similarly, we can estimate for STOL.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}\left[\hat{f}_t(\mathbf{D}_t) - f_n(\mathbf{D}_n)\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[tr\left(\mathbf{H}(\mathbf{D}_t - \mathbf{D}_n)(\mathbf{D}_t - \mathbf{D}_n)'\right)\right] + \mathcal{O}(1/t) \\
&= tr\left(\mathbf{H}\mathbb{E}_{\mathbf{x}}[(\mathbf{D}_t - \mathbf{D}_n)]\,\mathbb{E}_{\mathbf{x}}[(\mathbf{D}_t - \mathbf{D}_n)]' + \mathbf{H}\mathbf{Var}_s[\mathbf{D}_t]\right) \\
&\quad + \mathcal{O}(1/t) \\
&\le tr(\mathbf{GH})/t + \mathcal{O}(1/t) \le \nu\kappa^2/\rho + \mathcal{O}(1/t).
\end{aligned}$$

It means that STOL can reach accuracy ρ within $\nu\kappa^2/\rho + o(1/\rho)$ with convergence rate $\mathcal{O}(1/t)$. Thus, the time needed for K-SVD and STOL to asymptotically reach accuracy ρ is $\mathcal{O}(kn\kappa\log(1/\rho))$ and $\mathcal{O}(\frac{k\nu\kappa^2}{\rho})$, respectively.

Appendix C
Proof of Proposition 3

Since $f(\mathbf{D}_t)$ and $f_t(\mathbf{D}_t)$ are based on $l(\mathbf{D}_t, \alpha_t)$, it holds [45]

$$\begin{aligned}
&Pr\left(\sup\|\mathbb{E}_{\mathbf{x}}f(\mathbf{D}_t) - \mathbb{E}_{\mathcal{S}}f_t^{\gamma}(\mathbf{D}_t)\| > \varepsilon\right) \\
&\qquad \le 4\mathbb{E}\left[\mathcal{N}_\infty(l, \gamma, \mathcal{S})\exp\left(\frac{-n\varepsilon^2}{32}\right)\right],
\end{aligned} \tag{31}$$

where $\mathcal{N}_\infty(l, \gamma, \mathcal{S})$ is the covering number in ∞-norm for $\mathcal{S}$. Denote $\mathcal{N}_\infty(l, \gamma, n)$ its supremum for training set consisting of n samples. The upper bound of such supremum is derived [45].

$$\ln\mathcal{N}_\infty(l, \gamma, n) \le 36(p-1)\frac{a^2b^2}{\gamma^2}\ln\left(2\lceil\frac{4ab}{\gamma}\rceil n + 1\right), \tag{32}$$

where $\|\mathbf{x}\|_p \le b$ and $\|\alpha_t\|_q \le a$ with $1/p + 1/q = 1$. Let $\eta = 4\mathcal{N}_\infty(l, \gamma, n)\exp(-n\epsilon^2/32)$. These exists $\epsilon(l, \gamma, n, \eta)$:

$$\epsilon(l, \gamma, n, \eta) = \sqrt{\frac{32}{n}\left(\ln 4\mathcal{N}_\infty(l, \gamma, n) + \ln\frac{1}{\eta}\right)}. \tag{33}$$

With sufficient sampling, we compare Eq. (32) and (33).

$$\epsilon \leq \epsilon_0(\gamma)\sqrt{\ln n/n} \sim o(\ln n/n) \to 0.$$

Consequently, given arbitrary η, ϵ vanishes when $t \to n$ and $n \to \infty$. Since $\mathbb{E}[\mathcal{N}_\infty(l,\gamma,\mathcal{S})] \leq \sup_{\mathcal{S}} \mathcal{N}_\infty(l,\gamma,\mathcal{S}) = \mathcal{N}_\infty(l,\gamma,n)$, $\varepsilon \leq \epsilon_0 \to 0$ can be arbitrarily small. As a conclusion, Proposition 3 is drawn from Eq. (31).

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[3] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[4] B. Girod, A. M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[5] A. Dumitras and B. G. Haskell, "An encoder-decoder texture replacement method with application to content-based movie coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 6, pp. 825–840, Jun. 2004.

[6] D. Liu, X. Sun, F. Wu, S. Li, and Y.-Q. Zhang, "Image compression with edge-based inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 10, pp. 1273–1287, Oct. 2007.

[7] J. Ballé, A. Stojanovic, and J.-R. Ohm, "Models for static and dynamic texture synthesis in image and video compression," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 7, pp. 1353–1365, Nov. 2011.

[8] Z. Xiong, X. Sun, and F. Wu, "Block-based image compression with parameter-assistant inpainting," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1651–1657, Jun. 2010.

[9] Y. Wexler, E. Shechtman, and M. Irani, "Space-time completion of video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 3, pp. 463–476, Mar. 2007.

[10] H. Xiong, Y. Xu, Y. F. Zheng, and C. W. Chen, "Priority belief propagation-based inpainting prediction with tensor voting projected structure in video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 8, pp. 1115–1129, Aug. 2011.

[11] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 1838–1857, Jul. 2011.

[12] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[13] W. T. Freeman and E. C. Pasztor, "Learning low-level vision," in *Proc. IEEE Int. Conf. Comput. Vis.*, Corfu Island, Greece, Oct. 1999, pp. 1182–1189.

[14] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3467–3478, Aug. 2012.

[15] K. Jia, X. Wang, and X. Tang, "Image transformation based on learning dictionaries across image spaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 367–380, Feb. 2013.

[16] T. Peleg and M. Elad, "A statistical prediction model based on sparse representations for single image super-resolution," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2569–2582, Jun. 2014.

[17] G. Polatkan, M. Zhou, L. Carin, D. Blei, and I. Daubechies, "A Bayesian nonparametric approach to image super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 346–358, Feb. 2015.

[18] M. Shen, P. Xue, and C. Wang, "Down-sampling based video coding using super-resolution technique," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 6, pp. 755–765, Jun. 2011.

[19] H. F. Ates, "Decoder-side super-resolution and frame interpolation for improved H.264 video coding," in *Proc. IEEE Proc. Data Compress. Conf.*, Snowbird, UT, USA, Mar. 2013, pp. 83–92.

[20] J. Sun, N.-N. Zheng, H. Tao, and H.-Y. Shum, "Image hallucination with primal sketch priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. 729–736.

[21] H. Xiong, Z. Pan, X. Ye, and C. W. Chen, "Sparse spatio-temporal representation with adaptive regularized dictionary learning for low bit-rate video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 710–728, Apr. 2013.

[22] L. Bottou, "On-line learning and stochastic approximations," *On-Line Learn. Neural Netw.*, vol. 17, no. 9, pp. 9–42, 1998.

[23] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[24] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5. Mar. 1999, pp. 2443–2446.

[25] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 801–808.

[26] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: Learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, Mar. 2010.

[27] L. N. Smith and M. Elad, "Improving dictionary learning: Multiple dictionary updates and coefficient reuse," *IEEE Signal Process. Lett.*, vol. 20, no. 1, pp. 79–82, Jan. 2013.

[28] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Trans. Signal Process.*, vol. 61, no. 3, pp. 661–677, Feb. 2013.

[29] A. Krause and V. Cevher, "Submodular dictionary selection for sparse representation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, Jun. 2010, pp. 567–574.

[30] A. Krause and C. Guestrin, "Submodularity and its applications in optimized information gathering," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 4, Jul. 2011, Art. no. 32.

[31] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[32] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 791–804, Apr. 2012.

[33] J. Yang *et al.* (2014). *Video Compressive Sensing Using Gaussian Mixture Models*. [Online]. Available: http://people.ee.duke.edu/~lcarin/videoCS_GMM_v1.32.pdf

[34] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.

[35] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proc. IEEE*, vol. 98, no. 6, pp. 1045–1057, Jun. 2010.

[36] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2007, pp. 161–168.

[37] E. P. Simoncelli, *Distributed Representation and Analysis of Visual Motion*, Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. #209, Jan. 1993.

[38] X. Li and Y. Zheng, "Patch-based video processing: A variational Bayesian approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 1, pp. 27–40, Jan. 2009.

[39] B. Wang, H. Xiong, X. Jiang, and Y. F. Zheng, "Data-driven hierarchical structure kernel for multiscale part-based object recognition," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1765–1778, Apr. 2014.

[40] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[41] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[42] G. Bjontegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document VCEG-M33, ITU-T SG16/Q6, 13th VCEG Meeting, Austin, TX, USA, Apr. 2001.

[43] N. Murata, "A statistical study of on-line learning," in *On-Line Learning in Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[44] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Upper Saddle River, NJ, USA: Prentice-Hall, 1983.

[45] T. Zhang, "Covering number bounds of certain regularized linear function classes," *J. Mach. Learn. Res.*, vol. 2, pp. 527–550, Mar. 2002.

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, all in electronic engineering. He is currently a Post-Doctoral Scholar with the Department of Biomedical Informatics, University of California, San Diego. His 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.

Xin Tang received the B.S. and M.S. degrees in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2012 and 2015, respectively. Her research interest is learning-based video coding.

Junni Zou (M'07) received the M.S. and Ph.D. degrees in communication and information system from Shanghai University, Shanghai, China, in 2004 and 2006, respectively. Since then, she has been with the School of Communication and Information Engineering, Shanghai University, where she is a Full Professor. From 2011 to 2012, she was a Visiting Professor with the Department of Electrical and Computer Engineering, University of California, San Diego.

Her research interests include distributed resource allocation, multimedia networking and communications, and network information theory. She has published over 70 international journal/conference papers on these topics. She was a recipient of Shanghai Young Rising-Star Scientist in 2011.She also acts as a member of Technical Committee on Signal Processing of Shanghai Institute of Electronics.

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. Since then, he has been with the Department of Electronic Engineering, SJTU, where he is currently a Full Professor. He was a Research Scholar with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, from 2007 to 2008. He was a Scientist with the Division of Biomedical Informatics, University of California, San Diego, CA, USA, from 2011 to 2012.

His research interests include source coding/network information theory, signal processing, computer vision, and machine learning. He has published over 170 refereed journal/conference papers. He was a recipient of the Best Student Paper Award at the 2014 IEEE Visual Communication and Image Processing, the best paper award at the 2013 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, and the Top 10% Paper Award at the 2011 IEEE International Workshop on Multimedia Signal Processing.

Dr. Xiong was a recipient of the New Century Excellent Talents in University of the Ministry of Education of China in 2009. He received the SMC-A Excellent Young Faculty Award of Shanghai Jiao Tong University in 2010 and 2013. He also received the First Prize of the Shanghai Technological Innovation Award for Network-Oriented Video Processing and Dissemination: Theory and Technology, in 2011. He has been a member of Innovative Research Groups of the National Natural Science since 2012. He was also a recipient of the Shanghai Shu Guang Scholar in 2013. He also received the National Science Fund for Distinguished Young Scholar and Shanghai Youth Science and Technology Talent as well, in 2014. He served as a TPC Member for prestigious conferences, such as ACM Multimedia, ICIP, ICME, and ISCAS.

Chang Wen Chen (F'04) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1983, the M.S.E.E. degree from the University of Southern California, Los Angeles, in 1986, and the Ph.D. degree from the University of Illinois at Urbana–Champaign, Urbana, in 1992.

He was with the Faculty of Electrical and Computer Engineering, University of Missouri–Columbia, Columbia, from 1996 to 2003, and the University of Rochester, Rochester, NY, from 1992 to 1996. He served as the Head of the Interactive Media Group with David Sarnoff Research Laboratories, Princeton, NJ, from 2000 to 2002. He was an Allen S. Henry Distinguished Professor with the Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, from 2003 to 2007. He has been a Professor of Computer Science and Engineering with The State University of New York at Buffalo, Buffalo, since 2008. He has also consulted with Kodak Research Laboratories, Microsoft Research, Beijing, China, Mitsubishi Electric Research Laboratories, the NASA Goddard Space Flight Center, and the U.S. Air Force Rome Laboratories.

Dr. Chen was elected as a fellow of the SPIE for his contributions in electronic imaging and visual communications. He has been the Editor-in-Chief of the IEEE TRANSACTIONS ON MULTIMEDIA since 2014. He served as the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2006 to 2009. He has served as an Editor of the PROCEEDINGS OF THE IEEE, the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, the IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS, the *IEEE Multimedia Magazine*, the *Journal of Wireless Communication and Mobile Computing*, the *EURASIP Journal of Signal Processing: Image Communications*, and the *Journal of Visual Communication and Image Representation*. He has also chaired and served on numerous technical program committees for the IEEE and other international conferences.