Multiscale Online Dictionary Learning for Quality Scalable Video Coding

Xin Tang*†, Hongkai Xiong*, Xiaoqian Jiang†

*Dept. of Electronic Engineering
Shanghai Jiao Tong University
Shanghai 200240, China
{xint14,xionghongkai}@sjtu.edu.cn

†Division of Biomedical Informatics
University of California, San Diego
CA 92093, USA
x1jiang@ucsd.edu

This paper proposes a novel multiscale online dictionary learning algorithm with double sparsity structure for scalable video coding. Along hierarchical structures on the feature set by wavelet transform, the search space of online learning is optimized to sub-blocks for hierarchical sparsity. The group sparsity is exploited on lowest sub-band in the base layer to obtain the low-frequency sub-dictionary and sparse coefficient. We also designed cross-scale decomposition and reconstruction, for which the recovery error can be bounded. The dictionary is updated by stochastic gradient descent to optimize the expected cost. Hierarchical high-frequency information is predicted from a pre-learned corresponding sub-dictionary pairs for scalable coding. We demonstrated that the proposed algorithm can achieve scalable signal to noise ratio (SNR).

Given a video sequence F_h with group of pictures (GOP), it is decomposed into a selected HR key frames X_H and the down-sampled LR non-key frames Z_L. Through a standard codec, X_H and Z_L are coded and decoded. At the receiver, the multi-scale dictionary will be learned from HR key frames for each wavelet level. The HR version $\hat{X}_{Hl}$ would be recovered from $\hat{Z}_L$ by the scalable super-resolution reconstruction via sparse representation in a SNR-scalable manner.

The multiscale super-resolution reconstruction can be defined as an energy minimization:

$$f(\alpha_L^d, \mathbf{X}_H) = arg \min_{\mathbf{X}_{Hl}\alpha_L^d} \frac{1}{2}\sum_d \|\mathbf{W}_A\mathbf{Z}_L - \mathbf{D}_L^d\alpha_L^d\|_2^2 + \lambda\sum_d \|\alpha_L^d\|_0 + \sum_{l=2}^{k} \|\mathbf{D}_{Hl}^d\alpha_L^d - \mathbf{W}_A\mathbf{X}_H\|_2^2 \quad (1)$$

The HR key frames are decomposed into four sub-band via non-decimated 2D-DWT. For all the $3K$ sub-bands, overlapping patches with the same size are sketched as the training set in identical position according to each direction. Thus, we have a set of training pair $\mathbf{T}_L^i, \mathbf{T}_{Hl}^i, i = 1,2,3, l = 1 \ldots k-1$. For each lowest sub-band, the sub-dictionary $\mathbf{D}_L^i$ and the sparse representation over each sub-dictionary α_L^i are obtained.

Through the patches sketched from three level NDWT, it can be observed that the different level patches of each column share the same structure and texture. The supports of wavelet transform at different scales create a parent/child relationship between wavelet coefficients, and the multi-scale dictionary learning algorithm selects the set of multi-scale dictionary atom F as a connected sub-tree. Each direction has such set F which defines a subspace of signal that all representation coefficients outside F are zero. For the the sub-dictionary D as an $N \times K$ matrix , where $N \ll K$ and satisfies the (ε_F, α)-RAmP for model $\mathbb{T}$ with small enough signal residual ε_F, the signal estimate $\hat{\mathbf{x}}_i = \mathbf{D}\alpha$ has the upper-bound with noise n

$$\|\mathbf{x} - \hat{\mathbf{x}}_i\|_2 \leq 2^{-i}\|\mathbf{x}\|_2 + 35(\|n\|_2 + \delta(1 + \ln\lceil K/F\rceil)) \quad (2)$$

where δ is a parameter limited by signal set $\mathbb{T}_F$ and nonzero index F. Thus, we have a guarantee recovery for the proposed parent-child structure. Also, experiments of several CIF sequences with a variety of motion activities show that the proposed algorithm achieves the SNR-scalable video coding and behaves comparative performance with H.264/SVC.

Reference

[1] J. Huang, T. Zhang, D. Metaxas, "Learning with structured sparsity," *The Journal of Machine Learning Research*, vol. 12, pp. 3371-3412. Feb. 2011.

1068-0314/14 $31.00 © 2014 IEEE
DOI 10.1109/DCC.2014.30