MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Efficient Dictionary Based Video Coding with Reduced Side Information

Kang, J-W.; Kuo, C.C. J.; Cohen, R.; Vetro, A.

TR2011-026 May 2011

Abstract

In this paper, we propose a novel dictionary based video coding technique with adaptive construction of over complete dictionaries and advanced coding methods tailored to sparse signal representations. A set of dictionaries is trained off-line using inter or intra predicted residual samples and is applied for encoding. New coding tools are developed so that the encoder can more compactly represent the residual signal. The same set of dictionary elements can be reused for neighboring blocks, and the optimal number of dictionary elements can be decided using rate-distortion optimization. Experimental results demonstrate that the proposed algorithm yields both improved coding performance and improved perceptual quality at low bit rates.

IEEE International Symposium on Circuits and Systems (ISCAS)

EFFICIENT DICTIONARY BASED VIDEO CODING WITH REDUCED SIDE INFORMATION

Je-Won Kang and C.-C. Jay Kuo
Ming Hsieh Department of Electrical Engineering and
Signal and Image Processing Institute,
University of Southern California, Los Angeles,
CA 90089-2564, USA
Email: jewonkan@usc.edu and cckuo@sipi.usc.edu

Robert Cohen and Anthony Vetro
Mitsubishi Electric Resarch Laboratories
Cambridge, MA 02139, USA
Email: cohen@merl.com and avetro@merl.com

***Abstract*— In this paper, we propose a novel dictionary based video coding technique with adaptive construction of overcomplete dictionaries and advanced coding methods tailored to sparse signal representations. A set of dictionaries is trained off-line using inter or intra predicted residual samples and is applied for encoding. New coding tools are developed so that the encoder can more compactly represent the residual signal. The same set of dictionary elements can be reused for neighboring blocks, and the optimal number of dictionary elements can be decided using rate-distortion optimization. Experimental results demonstrate that the proposed algorithm yields both improved coding performance and improved perceptual quality at low bit rates.**

I. Introduction

The High Efficiency Video Coding (HEVC) standardization project recently has been launched by the Joint Collaborative Team on Video Coding (JCT-VC) [1]. New research efforts are being made to improve the coding performance over H.264/AVC for broader application and wider ranges of bit rate. So far, the high-level architecture of HEVC is not significantly different than that of current standards, which use block-based prediction techniques, 2-D Discrete Cosine Transforms, and context adaptive entropy coding. The new coding tools in HEVC are more advanced and flexible, but they generate significant increases in computational complexity.

For the past few decades, orthogonal and bi-orthogonal complete dictionaries such as the DCT or wavelets were the dominant transform-domain representation in image and video coding standards. Recently, sparse and redundant representations of signals over overcomplete dictionaries have been intensively studied and successfully applied to various applications such as image denoising. In this paper, owing to the advances in sparse representation theory, we propose an efficient adaptive dictionary-based video coding technique. In this algorithm, the conventional DCT transform can be replaced by a set of trained dictionaries. Experimental results demonstrate that the proposed algorithm provides gains in rate-distortion (R-D) performance and improvements in perceptual quality in low bit rates.

II. Review of Previous Work

Overcomplete video coding techniques have been found to achieve competitive coding gains at very low bit rates as compared to modern video coding standards. Basically, the block based 2-D DCT transform is replaced with with an expansion of larger and more suitable basis functions during overcomplete video coding. At lower bit rates, residual signals are represented with fewer nonzero DCT coefficients because of coarse quantization. Thus, only low frequency components of a macroblock are retained. In this scenario, an overcomplete dictionary set can provide a more flexible and faithful representation of residual signals, as compared to the complete dictionary set. Thus, the residual signal can be approximated better with fewer coefficients.

The overcomplete video coding technique, which was initially proposed by [2], constructs a dictionary set with modulated Gabor functions. Matching Pursuits (MP) [3] is employed to select the most appropriate dictionary elements in the representation. MP gives us a computationally tractable solution to the sparse signal representation. Dictionary sets can be varied by concatenating dictionaries generated by several analytic functions such as Wavelets, the DFT, Curvelets, and so on. However, these mathematical models can have problems capturing complex characteristics of natural images, and they occasionally introduce artifacts such as ringing.

Dictionary training is a more recent approach to dictionary based video coding. It has been shown that residual signals tend to have directional orientations after prediction [4]. Therefore, a good set of dictionaries could be designed by leveraging the residual signals' characteristics. In [5], a mode dependent directional transform was proposed for coding intra prediction residuals. A complete dictionary was trained using intra prediction residuals corresponding to the prediction direction. In [6], dictionary training was used for image coding applications. An adaptive dictionary was locally trained and applied for predicting neighboring blocks.

In this paper, we propose an efficient dictionary based video coding scheme integrated into the H.264/AVC framework. Basically, the conventional transform-domain representation is replaced by a sparse representation of residual signals using trained dictionaries. We perform adaptive dictionary learning and create a set of dictionaries which can efficiently represent the residual signals' characteristics, e.g., directional components. Advanced coding tools are also presented. The optimal size of dictionary elements can be determined via R-D optimization so that the signals are more compactly represented. We also introduce the *dictionary index copy method*, which reduces the size of the dictionary used to code a given block.

III. The Proposed Algorithm

A. Sparse representation of signals and context adaptive dictionary training

A signal $y \in \mathbb{R}^n$ can be represented as a sparse linear combination of elements in an overcomplete dictionary $D \in \mathbb{R}^{n \times m}$, where m is considerably larger than n and D is a full-rank matrix. We aim to approximate y with the smallest number of dictionary elements. Mathematically, the sparse representation can be found by solving

$$\min_{\mathbf{x}} \|x\|_0 \quad \text{s.t.} \quad \|y - Dx\|_2 \leq \delta, \tag{1}$$

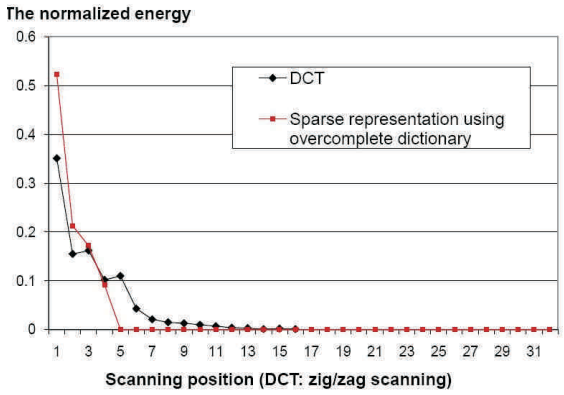


Fig. 1. The normalized energy distribution of DCT and overcomplete dictionary coefficients in QP32. The distribution is from several CIF sequences such as "Foreman," "City," and "Bus."

where $x \in \mathbb{R}^m$ is the vector which consists of the coefficients representing the signal y, and $\|.\|_p$ denotes the l_p norm of a signal.

We expect that a dictionary can be trained to better fit the sample data. Thus, the minimization problem in (1) can be converted to find the best dictionary in the given sparsity constraint C for the sparse representation of y as follows:

$$\min_{\mathbf{x},\mathbf{D}} \|y - Dx\|_2 \ \text{ s.t. } \ \|x\|_0 \leq C. \tag{2}$$

The dictionary is trained to provide a better representation of the actual signal when the number of nonzero coefficients are less than or equal to C. Fig. 1 shows the comparison of the energy compaction property between the trained dictionary used in our algorithm and 4×4 DCT transform in H.264/AVC. As shown in Fig. 1, the normalized energy distribution of the dictionary requires fewer coefficients than for the DCT, and the energy level quickly decreases. For computing this distribution, the sparsity constraint, i.e., the number of dictionary elements used to represent a data sample, was set to 4. In the proposed algorithm, we employ K-SVD [7] to train a dictionary. Also, MP is adopted to find the approximated sparse representation.

For video coding we perform either intra or inter prediction, and then we encode the prediction residual signal. Thus, the dictionary is trained ahead of time using prediction residual samples. However, if we use all prediction residual blocks from a set of sequences to train the dictionary, the resulting dictionary may not be large enough to represent all salient features of a residual signal. To tackle this problem, we classify the residual signals into contexts during the training process. Fig.2 shows the generic model of context based dictionary training with a classifier.

The classifier can operate based upon defined characteristics of the residual signal. For example the properties of intra-coded residuals in H.264/AVC are correlated with the intra prediction direction [5]. Therefore, the multiple sets of dictionaries are trained, where each set uses samples corresponding to the prediction direction used to generate the residual. During encoding, the dictionary corresponding to the intra prediction mode is used as the context for the classifier. Thus, no additional side information is required to inform the decoder which dictionary to use for a given block. For inter-coded residuals, the classifier computes the energy of the sample data and uses it for classification during the training process. The contexts are also determined by neighbor blocks during encoding. As shown in

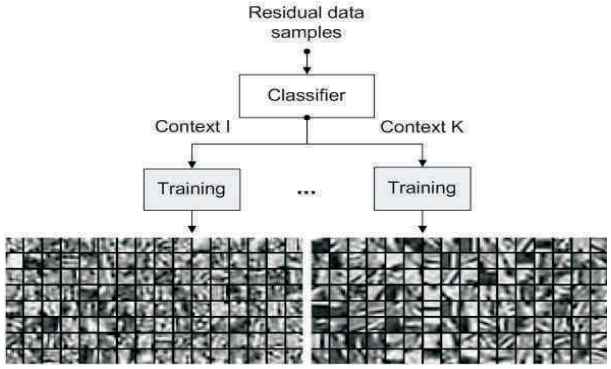


Fig. 2. A generic model of the context based dictionary training of residual data samples.

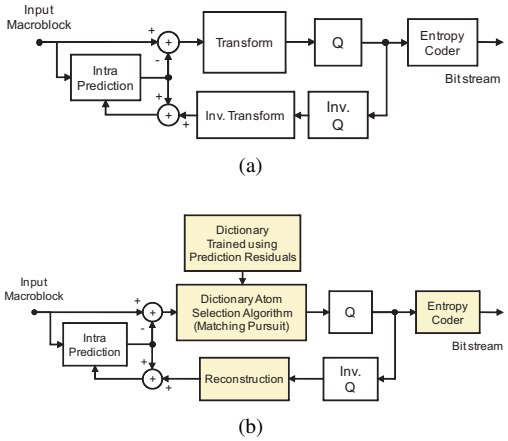


Fig. 3. The block diagram of (a) H.264/AVC encoder and (b) the dictionary based video coding.

Fig.2, the dictionary pictured for context K contains more diagonal components, which will be useful for representing oriented edges in inter-predicted residuals.

B. Elements of the dictionary based video coding

The dictionaries are trained with prediction residuals resulting from the encoding of a wide variety of sequences. These dictionaries are used for sparse representations of the input residual data during encoding. Fig. 3(b) shows a block diagram of the proposed algorithm as compared with the modern video coding standards in Fig. 3(a). Thus, the transform, reconstruction, and entropy coding portions of H.264/AVC are changed for the proposed algorithm.

MP is known to be a computationally tractable solution for sparse signal representation. We employ MP to choose the appropriate elements from dictionaries. In each iteration, MP forms a linear combination of dictionary elements by minimizing the residue of the reconstructed signal. Thus, we must transmit the coefficients values and indices of the selected dictionary elements to a decoder. The number of MP iterations may not exceed the sparsity constraint. However, it can be reduced by stopping iterations when the R-D costs are minimized, so that the signals can be optimally represented with fewer nonzero coefficients, less than or equal to a given constraint. We will address this criteria in the next subsection.

After the selection of dictionary elements, the coefficients are quantized and entropy copied as shown in Fig. 3(b). We found that

a Laplacian distribution approximates the coefficient distributions of the dictionary sets used in the proposed algorithm. Thus, an uniform quantizer is adopted to the proposed algorithm [8]. In an entropy coder, the coefficient values are binarized via a Huffman-coding table based on statistics of the coefficients. We also found that the statistical distribution of dictionary indices is mostly uniform, so any adaptive or fixed scanning order is not guaranteed to code the nonzero coefficients first. Thus, the indices are encoded with fixed length codes whose size are $\log_2 \lceil m \rceil$, where m is the number of elements in the dictionary. For reconstruction, which is the same as in the encoder, the indices specify which dictionary elements to use, and the quantized coefficients are used as weighting factors.

C. R-D optimal dictionary element selection

The dictionary elements are chosen by MP, and the coefficients are computed during each iteration up to a certain number, denoted by N . An important feature in modern video coding standards is R-D optimization. Instead of the fixed number of coefficients, an encoder can provide the best sparse approximation by minimizing the R-D costs defined by $D(N) + \lambda R(N)$, where $R(N)$ is the estimated number of bits, $D(N)$ is the MSE between the original and reconstructed signal, and $\lambda = 0.85 \times 2^{(QP-12)/3}$, which is the same Lagrangian multiplier as that used for the mode decision in H.264/AVC. Note that the number also tells when the encoder should stop the MP iterations. By simply dropping N, therefore, the optimal number of nonzero coefficients can be determined by,

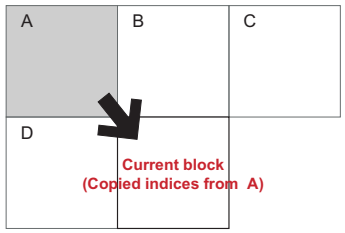
$$\begin{aligned} N^* &= \underset{N\in\{0,1,\ldots,C\}}{\arg\min} \{D(N) + \lambda R(N)\} \\ &= \underset{N\in\{0,1,\ldots,C\}}{\arg\min} \{D + \lambda(R_H + R_C + R_I)\}, \end{aligned} \quad (3)$$

where C is the given sparsity constraint, and R_H, R_C, and R_I is the required number of bits for header, MP coefficients, and index representations, respectively. The header information includes the number of nonzero coefficients, i.e., the number of iterations, so that the decoder knows the correct number of elements for reconstruction. In total, the encoder provides the best sparse approximation given the bit budget constraing.

D. An advanced coding technique: dictionary index copy method

In addition to the normal coding process using MP, a new coding tool, called the *dictionary index copy method*, is developed for the proposed algorithm. Generally, the current block has contexts similar to those of one or more of the four previously-coded neighboring blocks. Therefore, instead of sending both dictionary indices and coefficient values to the decoder, we can send coefficient values along with a two-bit flag indicating from which neighboring block the dictionary indices should be copied. During encoding, the R-D cost for using each of the four neighbors are computed. The two-bit flag indicates which neighbor yielded the lowest cost. Fig. 4(a) shows an example of the index copy method. An encoder may copy one of the index sets, e.g., block A, for coding the current block.

The index copy mode can be efficient for coding homogeneous regions in pictures. Fig. 4(b) shows the spatial distribution of blocks that use the index copy method. The light blocks indicate where the index copy method was used, and the dark blocks indicate where new indices have been generated. As shown in the figure, the index copy method tends to be selected in the smoother areas of the picture.



(a)

(b)

Fig. 4. (a) An example of index copy coding method and (b) Example picture showing where the index copy method was used, as indicated by light blocks.

TABLE I
PROPERTIES OF ENCODER PARAMETERS.

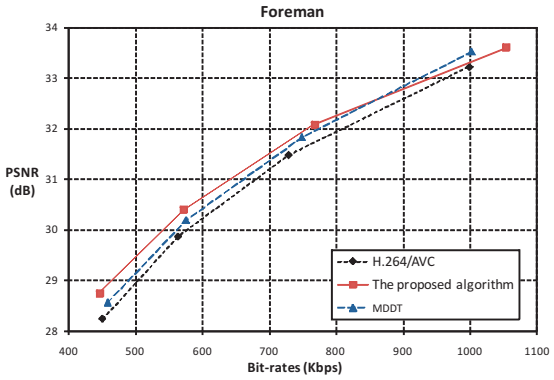
RDO	ON
Entropy Coding	CAVLC
QP	35, 38, 41, 43
Overcomplete dictionary size	64×128 and 16×32 for intra coding and 64×128 for inter coding

IV. EXPERIMENTAL RESULTS

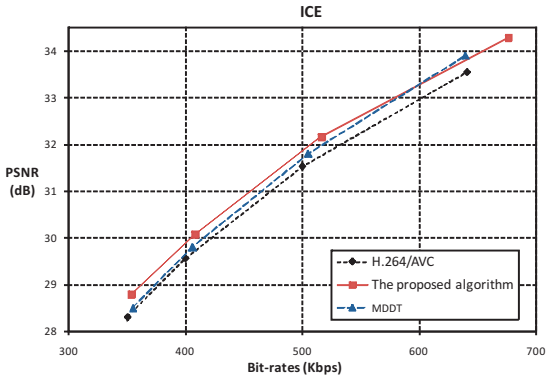
We present the R-D performance of the proposed algorithm and evaluate the subjective quality in this section. The experimental conditions are summarized in Table I. The proposed algorithm is implemented in JM11KTA2.6r1 [9]. The reference used for computing performance metrics is the unmodified JM11KTA2.6r1. In the experiments, universal overcomplete dictionaries are trained with several training sequences such as "Bus," "Coastguard," "Tempete," and others. Test sequences such as "Foreman" and "Ice" are not included in the training set. The overcomplete dictionaries are constructed as 16×32 and 64×128 matrices, where the columns represent the number of pixels in a block, and the rows represent the number of dictionary elements. This dictionary-based coding method was used for 4x4 and 8x8 blocks in intra pictures and for 8x8 only in inter pictures.

Fig. 5 compares the R-D performance of intra coding among the proposed algorithm, Mode-Dependent Directional Transforms (MDDT) [5], and H.264/AVC. MDDT uses complete KLT-based transforms which are trained using intra prediction residuals. At lower bit-rates, the proposed dictionary-based algorithm achieves better coding performance than both of the other methods. The proposed algorithm yields BD bit-rate reductions [10] of about 5% in "Foreman" and 3% in "Ice", as compared to H.264/AVC. In these comparisons, the proposed algorithm also outperforms MDDT up to 0.3 dB at very low bit rates in "Foreman".

Subjective differences can be seen in Fig. 6. We allowed the encoder to make an R-D optimized decision between the DCT and our dictionaries. In Fig. 6(a), the brighter blocks use the dictionary sets, while darker blocks use the DCT. Note that dictionary sets



(a)



(b)

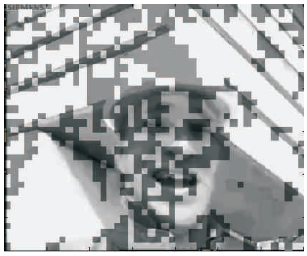
Fig. 5. The R-D performance comparison between the proposed algorithm, MDDT, and H.264/AVC in intra coding of (a) "Foreman" and (b) "Ice".

in the proposed algorithm include suitable diagonal components for residual signals, while the low frequency components of DCT are mainly oriented vertically or horizontally. Fig. 6(b) and (c) show the subjective differences between the proposed algorithm and DCT. The PSNR of these two frames are 29.7 dB and 29.8 dB, respectively. It is shown that DCT suffers from artifacts along straight edges due to the loss of high frequency components, while the proposed algorithm provides better representation of diagonal components along edges.

Fig. 7 gives an R-D performance comparison among the proposed algorithm, H.264/AVC, and the proposed algorithm without the classifier for inter coding. Here, the proposed algorithm is comparable with H.264/AVC, and a BD-rate loss of about 3% occurred when the classifier was not used. In the classification, standard deviation of residual signals are computed, and the contexts are decided by the values with empirically obtained thresholds from training sequences.

V. Conclusions

In this paper, we presented a dictionary based video coding technique based on a context adaptive overcomplete dictionary and novel coding techniques used for achieving bit-rate savings. We trained a dictionary using inter or intra prediction residual samples and used them to obtain sparse representations. We also developed several coding tools including the dictionary index copy method, and the decision of the optimal number of nonzero coefficients. Experimental results show that the proposed algorithm can provide both gains in R-D performance as well as perceptually improved reconstruction at low bit rates.



(a)

(b) (c)

Fig. 6. The perceptual quality evaluation between the proposed algorithm and H.264/AVC: (a) The block distribution of DCT and MP (brighter blocks), and the subjective evaluation between (b) the proposed algorithm and (c) DCT.

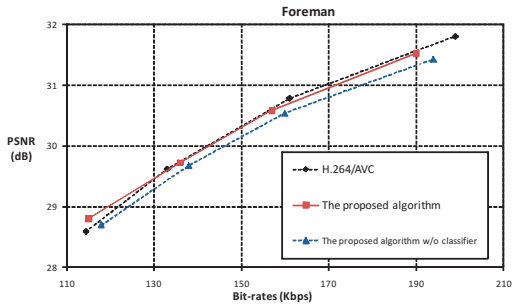


Fig. 7. R-D performance impact of the classifier used during inter coding.

References

[1] "Joint Call for Proposals on Video Compression Technology," ITU-T Q.6/SG16, Doc. VCEG-AM91, Jan. 2010.

[2] Ralph Neff, Avideh Zakhor, and Martin Vetterli, "Very low bit rate video coding using matching pursuit," in *Proc. SPIE Conference on Visual Communications and Image Processing*, Sep 1994, pp. 47–60.

[3] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, pp. 3397–3415, Dec. 1993.

[4] Bo Tao and Michael T. Orchard, "Gradient-based residual variance modeling and its applications to motion-compensated video coding," *IEEE Trans. Image Process.*, vol. 10, no. 1, pp. 24–35, Jan. 2001.

[5] Yan Ye and Marta Karczewicz, "Improved H.264 Intra Coding Based on Bi-directional Intra Prediction, Directional transform, and Adaptive Coefficient Scanning," in *Proc. ICIP*, Oct. 2008.

[6] Mehmet Turkan and Christine Guillemot, "Sparse Approximation with Adaptive Dictionary for Image Prediction," in *Proc. ICIP*, Oct. 2009.

[7] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. Signal Process.*, vol. 54, pp. 4311–4322, Nov. 2006.

[8] Gary J. Sullivan and Shijun Sun, "On Dead-Zone Plus Uniform Threshold Scalar Quantization," in *Proc. IEEE VCIP*, July 2005.

[9] "Key technology area (KTA) software of the ITU-T," Downloadbale: http://iphome.hhi.de/suehring/tml/.

[10] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," ITU-T Q.6/16, Doc. VCEG-M33, Mar. 2001.