

# CONTEXT MODELING AND ACCESSIBILITY FOR 3D SCALABLE COMPRESSION

Raymond Leung and David Taubman

The University of New South Wales, Sydney, Australia

## ABSTRACT

Highly scalable video compression based on invertible motion adaptive lifting transforms has emerged as a promising area in image processing research and an important component in interactive multimedia technology. However, within this feed-forward framework, the potential for coding efficiency improvement and its impact on random accessibility still has not been carefully assessed. In this paper, we compare the merits of several three-dimensional context coding strategies from an information-theoretic perspective. The variation in random access cost in response to coding parameter adjustments is analyzed, for a variety of spatial and temporal configurations.

## 1. INTRODUCTION

This paper is concerned with the feasibility of three-dimensional context-based adaptive arithmetic coding techniques in scalable volumetric and video compression. Researchers in the past have extended the principles of embedded zero trees [1] and layered zero coding to multiple dimensions for 3D volumes with isotropic distribution [2]. Boulgouris et.al.[3] devised a lexicographical scheme which uses category codes to encode the magnitude of each wavelet coefficient, using context information gathered from its vicinity and parent coefficients. It has been found in [4] that motion-compensated predictive coding is superior to statistical, auto-regressive modeling of 3D spatial dependency.

In this paper, we look at the random access properties of volumetric images encoded using spatial and 3D context schemes, to establish whether the benefits attributed to 3D context modeling outweigh the escalated cost associated with slice recovery during synthesis. The first motivation for this work is to quantify the effectiveness of exploiting context information from neighboring slices or video frames relative to the effectiveness of subband transforms applied along the slice/temporal direction. Our aim is to determine the best combination of these techniques, given the various costs involved. One of the key questions concerns how much information the context models are able to capture in the absence of any temporal transformation. A complementary question is the amount of residual information contained in the temporal subbands, after the source has been approximately decorrelated through  $L$  levels of axial wavelet decomposition.

The second motivation stems from coding efficiency and random access concerns. Our goal is to establish the best spatio-temporal code block dimensions under the EBCOT paradigm, striking the best balance between reconstruction quality and the degree of localization. The latter refers to the ability to recover an arbitrary slice from an embedded codestream at a reasonable cost. Thus, the important criterion for an effective context coding scheme is the degree of compression which can be achieved for a given level of quality, subject to constraints on the ease of random accessibility.

## 2. CONTEXT MODELING

In this section, we restrict our attention to bit plane coding and concentrate on intrascale magnitude clustering of wavelet coefficients. As recent studies have shown, exploitation of interscale dependencies tends to offer little benefit [5]. The context models which we investigate are constructed using a variety of primitive significance context labels. We write  $Nx$  for a spatial neighborhood context, similar to that specified in the JPEG2000 image compression standard. We use a prefix  $P$  to signify Markov-1 dependence on the previous slice (or frame, in the case of video). Specifically,  $Px$  is the context label constructed from the bit corresponding to  $x$  in the previous slice, while  $PNx$  is a context label constructed from the 8 bits in the immediate neighbourhood of  $x$ , but in the previous slice. We use  $(x,z)$  to denote an event corresponding to a current bit value of  $x \in \{0, 1\}$  and a composite spatiotemporal context label  $z = \{Nx, (Px, Nx), (Px, PNx, Nx)\}$ .

### 2.1. Conditional Mutual Information

Let  $P(x|z)$  be the conditional probability of the outcome  $X = x$  given a context label of  $Z = z$ . The conditional mutual information between random variables  $X$  and  $Y$ , given  $Z$ , is defined as  $I(X; Y|Z) \equiv E[\log_2 \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}] = H(X|Z) - H(X|Y, Z) \geq 0$ .  $I(X; Y|Z)$  provides a practical bound which indicates the compression advantage (bit rate saving) that might be expected, when aspects of  $Y$  are taken into consideration during coding, in addition to  $Z$ . Intuitively,  $Y$  should convey some information about  $X$ , in such a way that it reduces the statistical uncertainty of  $X$ , given  $Z$  is known; if not,  $X$  and  $Y$  are said to be conditionally independent. This helps us in measuring the effectiveness of a particular conditional coding scheme for exploiting inter-frame redundancy. Specifically, with the additional knowledge of the corresponding pixel  $Px$  from the previous slice,  $I(X; Px|Nx)$  indicates how many less bits one can expect to spend encoding the outcome of  $X$ , compared to the scenario where only  $Nx$  is known. Similarly,  $I(X; Px, PNx|Nx)$  indicates the extent to which the bit rate may be reduced if the context,  $Px, PNx$ , is exploited together with  $Nx$ .

### 2.2. Information and Probability Estimation

In order to assess the effectiveness of 3D context coding schemes, we need to be able to estimate  $P(x|z)$ . We choose to emulate the behavior of an arithmetic coder, by initializing it in an unbiased state (with all events equally probable) and accumulating the incremental bit rates  $\log_2 P_n(x_n|z_n)$  contributed by each successive symbol,  $x_n$ , as we go. The conditional probabilities are continuously augmented in accordance with  $P_n(x|z) = [C_{n-1}(x, z) + \Delta] / [C_{n-1}(0, z) + C_{n-1}(1, z) + 2\Delta]$ , where  $\Delta = 1$  and  $C_n(x, z)$  is the number of times  $x$  has been seen in context  $z$ . Context re-

duction techniques are used to minimize the impact of learning penalties as the model adapts from its initially unbiased state.

### 2.3. Localization versus Coding Efficiency

Initially, the probability adaptation process incurs a learning penalty. While the probability estimates are being driven toward the true statistics, we pay a price for inaccuracies in the estimates. Thus, a context coding scheme will be rendered inefficient, if the code blocks processed by the coding engine contain insufficient samples and have a relatively short span along the temporal (slice) axis. For a given spatial dimension, the adaptation cost may be leveraged by allowing the span of the code block to grow in the temporal dimension. This has the potential of reaping the full benefit of context extension, provided that the source is stationary.

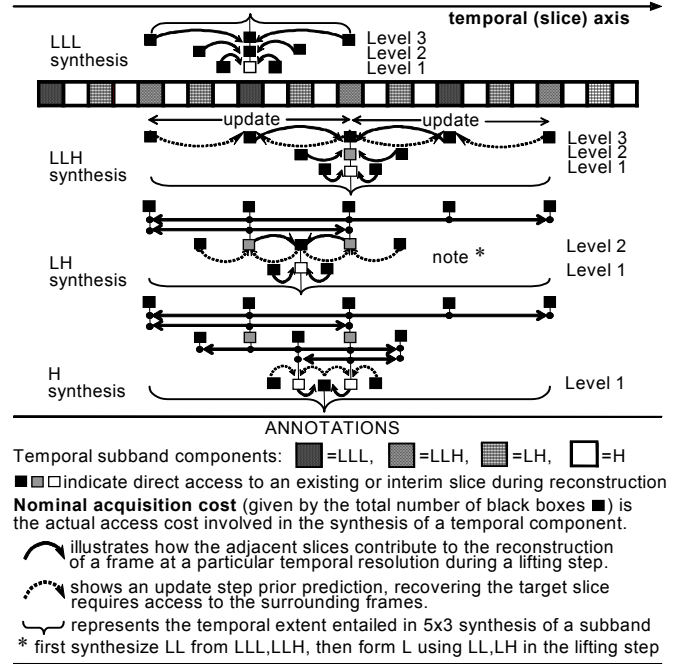
By contrast with video, “scene” changes are seldom encountered in medical image sequences. The gradual evolution of cross-section features and long-term similarity generally allow us to use accrued knowledge in our probability estimates (from the very first slice) and not be penalized for its decreasing sensitivity to recent innovation in the sequence. Unfortunately, the coding efficiency improvement achieved through the use of longer code blocks forces us to sacrifice random accessibility. The Markov-1 dependence seen in causal 3D context coding prevents us from gaining access to an image slice of interest without retrieval and decompression of all prior slices. To reduce access cost, we need to limit the length of these spatio-temporal codeblocks in the slice direction to eliminate long-term dependence. The idea is to permit some degree of random accessibility, bearing in mind that this also compromises coding efficiency.

### 2.4. Temporal Extent & Random Access Cost

The span of the temporal synthesis kernels and the number of levels of temporal wavelet decomposition represent fundamental constraints on the degree of localization that is possible for random access. When 3D context modeling is not employed, these contribute to a *nominal acquisition cost*,  $A_N$ , indicating the number of slices that the decoder needs to directly access to reconstruct an image slice from the subband domain. In the ensuing discussion, we will concentrate more on the concept of *temporal extent*,  $T_E$ , since this provides a useful upper bound on the actual access cost.

For the purpose of calculating temporal extent, each subband slice is assigned a location with respect to the original sequence. With one level of decomposition, low and high pass subband frames are assigned the even and odd sequence locations respectively. For further levels of decomposition, the same interleaving principle is applied recursively. The *temporal extent*,  $T_E$ , is the length of the smallest interval in this interleaved sequence of subband frames, such that the interval contains all subband slices (or frames) required to reconstruct a given slice (or frame) of interest.

The nominal acquisition cost may differ slightly from the temporal extent, due to the appearance of holes (frames which are not required) within the interleaved sequence of subband frames. If 3D context modeling is employed to code the subbands, the collection of subband frames required to reconstruct a slice (or frame) of interest generally increases, filling in these holes. In this case, the total access cost,  $A'_N$ , increases beyond the nominal acquisition cost ( $A_N$ ) associated with 2D spatial context coding alone. Similarly, the access cost upper bound,  $T'_E$ , is also elevated in the event of 3D context coding. We distinguish the cost associated with 3D context modeling using the prime notation, to highlight



**Fig. 1.** Slice random accessibility for 3 levels of temporal decomposition (without 3D context coding). Concept of temporal extent and nominal acquisition cost in the case of 5x3 wavelet synthesis.

the fact that both  $T'_E$  and  $A'_N$  depend on the code block temporal dimensions; while  $T_E$  and  $A_N$  are both independent of code block temporal dimensions.

As seen in Figure 1, the temporal extent,  $T_E$ , and indeed, the nominal acquisition cost,  $A_N$ , may differ depending on the location of the slice (or frame) to be reconstructed. When we average  $T_E$  over all possible slices (or frames) that one may wish to randomly access, we arrive at an expected value for the upper bound on the random access cost,  $E[T_E]$ .

In an  $L$ -level temporal decomposition,  $L + 1$  subbands are produced. Conceptually, a collection of interleaved temporal subbands may be assembled to fit nicely inside a virtual container of size  $2^L \times M$ , called a *frame slot*. Adopting the convention that each subband contributes one independently coded 3D code block to each frame slot, the code blocks at depth  $d$  in the temporal decomposition measure  $2^{L-d} \times M$  slices in length, where  $1 \leq d \leq L$ . Flushing of probability models is performed at the code block boundary. The first slice within each code block is always encoded using spatial contexts only, while an inter-frame coding policy is evoked on the remaining slices. The code block temporal dimensions and access cost may be adjusted by selecting different values of  $M$ , which we refer to as the *frame slot multiplier*.

## 3. SIMULATION RESULTS

Various three-dimensional context coding strategies are examined with and without the use of an  $L$  level temporal subband decomposition, based on the 5x3 wavelet kernels. In every instance, the source is subject to 5-levels of spatial subband decomposition, based on the 9x7 wavelet kernels. Quantization parameters

**Table 1.** Theoretical performance of 3D magnitude context models on a MR medical volumetric image of the human brain (with 10 significant bits, dimension=256x256x200)

Context Coding Mutual Information & Gain Factors				
Levels of transform	$L = 0$	$L = 1$	$L = 2$	$L = 3$
$I'(X_T; Nx_T)$	.2173	.1150	.1059	.1057
$I'(X_T; Px_T, Nx_T)$	.2937	.1292	.1120	.1101
$I'(X_T; PNx_T, Px_T, Nx_T)$	.3066	.1309	.1131	.1108
$I'(X_T; Px_T Nx_T)$	.0763	.0141	.0061	.0044
$I'(X_T; PNx_T, Px_T Nx_T)$	.0892	.0158	.0072	.0051
Temporal Gain $G_T^{(L)}$	-	+3.033	+3.321	+3.549
$G(X_T; Px_T Nx_T)$	+0.657	+0.202	+0.084	+0.060
$G(X_T; PNx_T, Px_T Nx_T)$	+0.770	+0.229	+0.099	+0.069

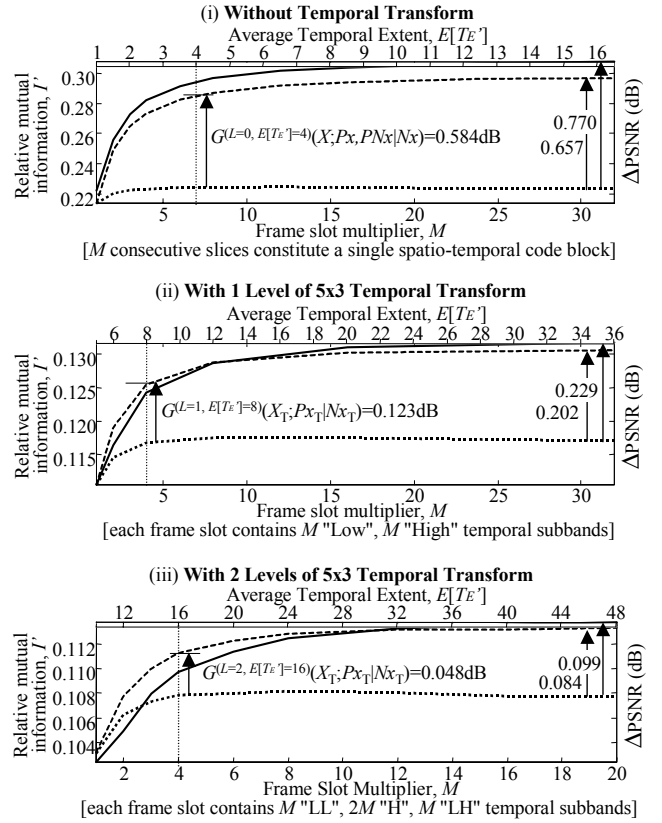
are selected so as to yield a target PSNR of around 40dB, using the JPEG2000 compression scheme to process the temporal subband slices jointly as separate image components. A quantization scaling factor is selected for each temporal subband component as  $\Delta_t = \Delta_o/\sqrt{W_t}$  where  $W_t$  is the energy weight associated with the 5x3 synthesis system. The code block dimensions are set to  $64 \times 64$ . Coding proceeds in a hierarchical fashion, walking through each resolution, subband, frame slot, code block, slice and bitplane in stripes. Samples are classified using the relevant context model.

The performance of the 3D context models is assessed in terms of (conditional) mutual information  $0 \leq I(X|Z) \leq 1$  and a coding gain,  $G$ . The coding gain represents the difference between the reconstructed PSNR obtained using the 3D context model and the PSNR obtained using the purely spatial context model of JPEG2000, at the same compressed bit-rate. The figures  $I'(X|Z)$  presented in Table 1, represent *relative mutual information*, which is obtained by normalizing  $I(X|Z)$  with respect to the bit-rate obtained by coding subband samples without any context modeling at all (i.e., a single context). We use the notation  $x_T$  in place of  $x$  to distinguish the case  $L > 0$  (i.e., using a temporal transform) from  $L = 0$  (without a temporal transform).

### 3.1. Performance Analysis

Temporal transformation generally yields significantly greater quality improvement compared to 3D context modeling. In Table 1,  $I'(X; Nx)_{L=0}$  shows that without temporal transform, spatial contexts alone achieve a rate reduction of 21% relative to the bit-rate obtained without any context modeling. When 3D context modeling is used, the previous bit,  $Px$ , and the previous neighborhood,  $PNx$ , together only manage to convey 9% more information about  $X$ , given  $Nx$  is known. The corresponding coding gain for this margin of rate reduction is 0.77dB. In contrast, temporal transformation yields an improvement of at least 3 dB, relative to pure intra-frame coding, while the best 3D composite context model considered in the absence of temporal transformation yields an improvement of less than 0.8 dB, relative to spatial context modeling alone. In view of the fact that  $G(X; PNx, Px|Nx) \ll G_T^{(L)}$ , 3D context coding is generally not a viable alternative to temporal transform in scalable compression; especially if a high degree of random access is desired.

When temporal transformation is used, the additional benefit of 3D context coding is typically very small. Due to the whitening property of the transform, information provided by the 3D contexts



**Fig. 2.** Coding efficiency and localization tradeoffs amongst various context modeling strategies. The axes represent: (*left*) relative mutual information, (*right*) PSNR gain from exploiting conditional mutual information conveyed by the 3D-contexts (dB), (*top*) average temporal extent  $E[T_E']$ , (this forms an upper bound on the expected random access cost), (*bottom*) frame slot multiplier,  $M$  (from which, the size of pyramidal code blocks are determined). Annotations:  $\cdots$  denotes  $X|Nx$ ;  $---$  denotes  $X|Px, Nx$ ;  $—$  denotes  $X|PNx, Px, Nx$ .

$I'(X_T; PNx_T, Px_T, Nx_T)$  surpasses  $I'(X_T; Nx_T)$  by no more than 2%. This suggests that the residual information contained in the wavelet transform decorrelated source is negligible.

Furthermore, the utility of 3D context information diminishes rapidly with decreasing temporal resolution. Measuring this utility as  $U = \max\{I'(X; Px, PNx|Nx), I'(X; Px|Nx)\}/I'(X; Nx)$ , we find that  $U = \{0.41, 0.14, 0.07\}$  for  $L = \{0, 1, 2\}$  respectively. These ratios indicate that any residual information present, can be sufficiently captured by spatial context alone when  $L \neq 0$ .

The same trend may be observed by considering the coding gains,  $G(X_T; PNx_T, Px_T|Nx_T)$ , reported in Table 1. Evidently, very little gain in coding efficiency may be expected from hybrid “slice transform+3D context” schemes. In fact, considering that 3D context modeling and temporal transformation both introduce a cost in terms of random accessibility, the total cost is generally minimized by avoiding 3D context modeling altogether.

To see this, observe firstly that the expected temporal extent  $E[T_E]$  (upper bound on the random access cost) associated

**Table 2.** Actual 3D context coding gain results obtained using a full coding system, for various numbers of temporal decomposition levels,  $L$ , and spatial code-block dimensions,  $w \times h$ . Gains in each row represent the improvement in PSNR (dB) resulting from the use of 3D context models in place of 2D models.

$L$	$w \times h$	Frame slot size, $2^L \times M$				
		$M=8$	$M=16$	$M=24$	$M=32$	$M=40$
0	16 x 16	0.3972	0.5337	0.6049	0.6444	0.6646
0	64 x 64	0.5464	0.6805	0.7055	0.7468	0.7748
1	16 x 16	0.0706	0.1193	0.1362	0.1443	0.1493
1	64 x 64	0.1140	0.1529	0.1816	0.1892	0.1891

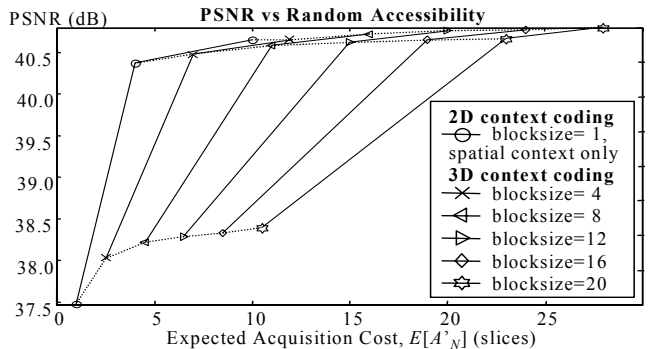
with an  $L$  level temporal transform, using the 5x3 subband filters and only spatial context modeling, takes values of 4, 8 and 16, for  $L = 1, 2$ , and 3, respectively. Now suppose that the number of temporal transform levels is reduced to  $L - 1$  and 3D context modeling is introduced, with a frame slot multiplier,  $M$ , set large enough so as to achieve the same temporal extent as the  $L$  level transform. Comparing the coding gains,  $G_T^{(L)}$ , and  $G_T^{(L-1)} + G^{(L-1, E[TE])}(X_T; Px_T, PNx_T | Nx_T)$ , for these two cases produces strong evidence in favor of our claims ( $G$  values appear in Table 1 and Figure 2). Our results also confirm that periodic flushing of the probability tables adversely affects coding efficiency, although this is necessary to achieve localization for random access purposes. (Figure 2 shows the possible tradeoffs).

#### 4. PRELIMINARY EXPERIENCE WITH A REAL CODING SYSTEM

We have also been able to verify the simulation results presented above, in the context of an experimental spatio-temporal coding system. The experimental coder is essentially an extension of the JPEG2000 coding system, incorporating motion-compensated temporal transformation and 3D context models, within a fractional bit plane coding framework. Results obtained using the best observed 3D context models are reported in Table 2. The coding gains presented in the table are expressed relative to the performance of the purely spatial context models used by JPEG2000.

A comparison between the theoretical gains in Table 1 and the observed 3D context coding gains in Table 2 confirms that the trends observed in Section 3.1 are indeed reflected in the performance of the experimental coder. Our experiments also suggest that code-block spatial dimensions as small as 16x16 are able to achieve similar compression efficiency to the larger code-block sizes such as 64x64 commonly used with JPEG2000. The reason for this is that code-blocks span multiple subband frames (within the frame slot). The 3D context coding gain also appears to be largely insensitive to the spatial dimensions of the 3D code blocks, at least for the sizes reported in Table 2. These observations may have important implications for spatial accessibility within compressed volumes and video sequences.

Inspecting the performance curves in Figure 3 provides further insight into the true value of 3D context coding. It demonstrates that the benefits of context modeling are significantly outweighed by the penalty it imposes on random accessibility. The actual random access cost,  $A'_N$ , dominates the context coding gain as the curve flattens. The results presented in Figure 3 are obtained using the MR medical volume; however, similar trends are observed with typical video sequences such as *mobile* and *flower garden*.



**Fig. 3.** PSNR versus the expected acquisition cost,  $E[A'_N]$ . Each solid line is parameterized by a pyramidal block size. Dotted lines (bottom-up) correspond to  $L=\{0,1,3\}$  levels of slice decomposition

#### 5. CONCLUSIONS

In this study, we have investigated the potential of 3D context modeling for scalable compression. Our experience in the context of volumetric imagery suggests that 3D context modeling is not a viable alternative to the use of 3D transforms. From both an information-theoretic and a practical viewpoint, the coding gain attributed to 3D conditional coding is at best modest and its utility is limited to high temporal resolutions. The goals of high coding efficiency and random accessibility present two conflicting requirements. Both spatio-temporal context modeling and spatio-temporal transforms introduce interframe dependencies which compromise the degree of localization, suggesting that a tradeoff exists between the degree to which each of these techniques is used. Our studies show, however, that this tradeoff always favours the use of temporal transform methods in place of 3D context modeling. Although the conditional coding gain is typically less than 1 dB, 3D context coding may still be a worthwhile technique for applications where random accessibility is of less interest.

#### 6. REFERENCES

- [1] M. Bénétière, V. Bottreau, A. Collet-Billon, and T. Deschamps, "Scalable compression of 3d medical datasets using a (2d+t) wavelet video coding scheme," *Proc. ISCAS*, vol. 2, pp. 537–540, August 2001.
- [2] G. Menegaz, L. Grewe, and J. Thiran, "Multi-rate coding of 3d medical data," *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, pp. 656–659, 2000.
- [3] N. Boulgouris, A. Leontaris, and M. Strintzis, "Wavelet compression of 3d medical images using conditional arithmetic coding," *IEEE ISCAS*, vol. 4, pp. 557–560, May 2000.
- [4] M. Orchard, A. Nosratinia, and R. Rajagopalan, "On interframe coding models for volumetric medical data," *Proc. IEEE Int. Conf. Image Proc.*, pp. 17–20, October 1995.
- [5] J. Liu and P. Moulin, "Information-theoretic analysis of inter-scale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Proc.*, vol. 10, pp. 1647–1658, November 2001.