An Embedded Wavelet Video Coder Using Three-Dimensional Set Partitioning in Hierarchical Trees (SPIHT)

Beong-Jo Kim and William A. Pearlman
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180

ABSTRACT

The SPIHT (set partitioning in hierarchical trees) algorithm by Said and Pearlman is known to have produced some of the best results in still image coding. It is a fully embeded wavelet coding algorithm with precise rate control and low complexity. In this paper is presented an application of the SPIHT algorithm to video sequences, using three-dimensional (3D) wavelet decompositions and 3D spatio-temporal dependence trees. A full 3D-SPIHT encoder/decoder is implemented in software and is compared against MPEG-2 in parallel simulations. Although there is no motion estimation or compensation in 3D SPIHT, it performs measurably and visually better than MPEG-2, which employs complicated means of motion estimation and compensation.

I. INTRODUCTION

Embedded zero-tree coding by Shapiro [[Sha92]] is a coding scheme which exploits inter-subband correlations/similarities. It uses self-similarity to efficiently transmit the significance map with a tree structure called a zero-tree which denotes a tree of zero symbols across subbands starting at a root being also zero. The zero-tree is based on the simple hypothesis that if a vector at a coarse scale is insignificant, then vectors in the same spatial orientation at finer scales are also likey to be insignificant. The resulting algorithm where zero-tree has been combined with bit plane coding in an elegant way is called embedded zero-tree wavelet(EZW) algorithm. Improved two-dimensional (2D) zero-tree coding (IEZW) by Said and Pearlman [[SP93]] has been extended to three dimensions (3D-IEZW) by Chen and Pearlman [[CP96]] and shows promise of an effective and computationally simple video coding system without any motion compensation, and obtained excellent numerical and visual results. Three dimensionsional zerotree coding through a modified EZW algorithm has also been

used with excellent results in compression of volumetric medical images [[LWCP96]]. Said and Pearlman [[SP96]] recently provided a new and more efficient implementation of EZW through the procedure of set partitioning in hierarchical trees or SPIHT algorithm. Here, we offer a SPIHT video coding system extended from two to three dimensions without motion compensation having the following SPIHT characteristics: (1) partial ordering by magnitude of the 3D subband/wavelet transformed video with a set partitioning algorithm, (2) ordered bit plane transmission of refinement bits, and (3) exploitation of self-similarity across different spatio-temporal scale(subbands). The compressed bit-stream is completely embedded, so that a single file for a video sequence can provide progressive video quality, that is, the algorithm can be stopped at any compressed file-size or let run until nearly lossless reconstruction is obtained, which is desirable in many applications including HDTV. Since the coding takes place on a three-dimensional (3D) wavelet transform, the video is also scalable both in spatial and temporal directions, allowing delivery of different size frames at different frame rates from a single compressed bit stream. Comparative simulations against a full implementation of MPEG2 with its complicated motion compensation show superiority of 3D-SPIHT video coding, which has no motion compensation.

II. THREE-DIMENSIONAL SUBBAND FRAMEWORK AND TREE STRUCTURE

Although a variety of subband structures may be obtained by combining separable spatial and temporal filtering operations, in this work, a classical two channel spatial decomposition hierarchy is adopted and is extended to 3D spatiotemporal decomposition by applying filtering operations first in the temporal domain and then in the spatial domain in a recursive fashion to obtain some desired pyramid levels for video compression. The complete subband structure of 2-level decomposition (for simplicity of illustration) is shown in Figure 1, where ' H_t ' and ' L_t ' represent temporal highpass and lowpass subbands respectively, and ' H_h ', ' L_h ', ' H_v ', and ' L_v ' represent horizontal highpass, horizontal lowpass, vertical highpass and vertical lowpass subbands in the spatial domain respectively. A total of 21 subbands results from the two-level spatiotemporal subband/wavelet decomposition. To obtain a tree structure similar to 2D SPIHT, we consider the 2D case first. In two dimensions, the tree is defined in such a way that each node has either no offspring (the leaves) or four offspring, which always form a group of 2x2 adjacent pixels. Figure 2 shows the parent-offspring relationship. The pixels in the highest level of the pyramid are tree roots and 2x2 adjacent pixels are also grouped into blocks. However, their offspring branching rule is different, and in each group, one of them(indicated by the star in Figure 2) has no decendendants. Hence, the parent-children linkage except at the highest and lowest pyramid levels is

$$O(i,j) = \{(2i,2j),(2i,2j+1),(2i+1,2j),(2i+1,2j+1)\}, \tag{1}$$

where $O(i,j)$ represents a set of coordinates of all the offspring of node (i,j).

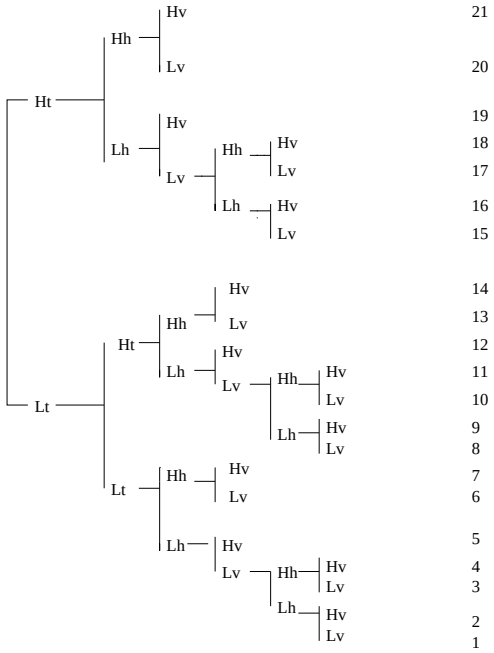

Figure 1: Spatio-temporal decomposition.

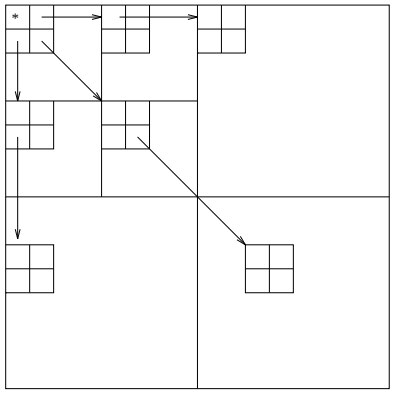

Figure 2: Parent-Offspring Dependency in 2D SPIHT

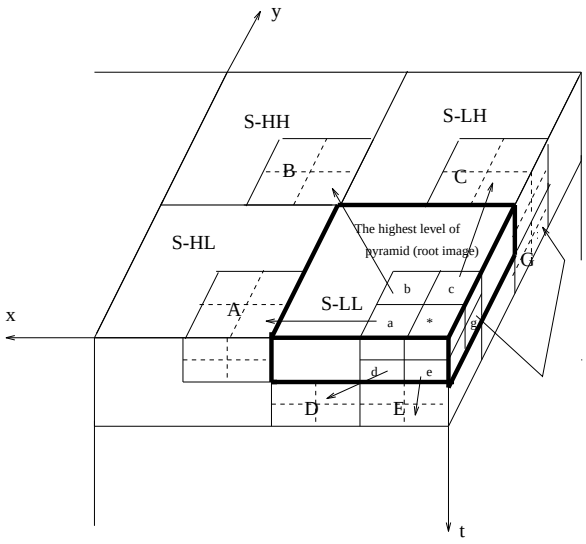

Figure 3: Parent-Offspring Dependency in 3D SPIHT at the highest level

For three dimensions, each node has either no offspring (the leaves) or eight offspring which is a group of 2x2x2 adjacent pixels. Hence, similar parent-offspring relationship can be established as shown in Figure 3. That is, a simple extension to a 3D hierarchical tree, except at the highest and lowest pyramid levels, is

$$\begin{aligned} O(i,j,k) \;=\; & \{(2i,2j,2k),\ (2i,2j+1,2k),\ (2i+1,2j,2k), \\ & (2i+1,2j+1,2k)(2i,2j,2k+1),\ (2i+1,2j,2k+1), \\ & (2i,2j+1,2k+1),\ (2i+1,2j+1,2k+1)\}. \end{aligned}$$

The pixels in the highest level of the pyramid are also grouped 2x2x2 adjacent pixels, and one of the pixels (*) in each group has no offspring, as in the 2D case. Figure 3 depicts the parent-offspring relationships in the highest level of the pyramid, assuming the highest level of pyramid has dimension of 4x4x2 for simplicity. There is a group of 8 pixels (*, a,b,c,d,e,f,g) in S-LL, where pixel ' f ' is hidden under pixel ' b '. Every arrow originated from a root pixel to a 2x2x2 block shows the parent-offspring relationship. Offspring block ' F ' of pixel ' f ' is hidden under block ' B ' in the figure.

III. 3D SPIHT VIDEO CODING SYSTEM AND IMPLEMENTATION DETAILS

In this section, we present a complete 3D SPIHT coding scheme. The basic procedure is that a segment of a video sequence to be coded is first subband transformed. The number of spatiotemporal subband decompositions directly depends on the number of frames to be processsed at a time. In this work where 16 frame segments are sequentially processed, three-level decomposition in both temporal and spatial domain is used. After subband/wavelet transformation, the 3D SPIHT algorithm is applied

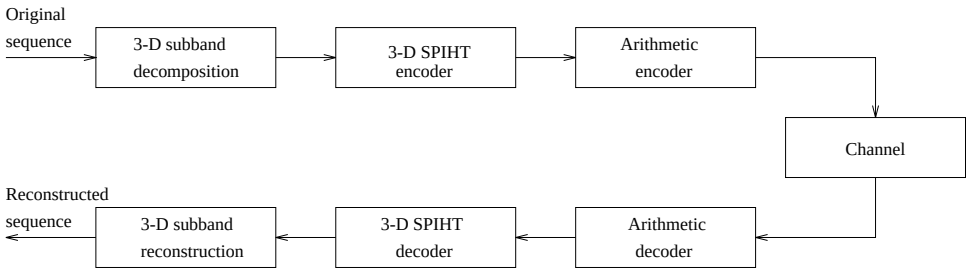

Figure 4: 3D SPIHT video coding system

to the resulting multiresolution pyramid. Then, the output bit stream is further compressed with an arithmetic encoder. To increase the coding efficiency, groups of 2x2x2 coordinates were kept together in the list, and their significance values are coded as a single symbol by the arithmetic encoder. Since the amount of information to be coded depends on the number of insignificant pixels m in that group, we use several different adaptive models, each with 2^m symbols,where $m \in \{1,2,3,4,5,6,7,8\}$, to code the information in a group of 8 pixels. By using different models for the different number of insignificant pixels, each adaptive model becomes a better estimate of the probability conditioned to the fact that a certain number of adjacent pixels are significant or insignificant. The decoder does exactly the opposite, that is, first arithmetic decoding, then 3D SPIHT decoding , and finally inverse subband/wavelet transformation. These encoding and decoding procedures are shown in Figure 4.

We used the 9/7 biorthogonal wavelet filters of [[ABMD92]] separably in all dimensions. The same filtering operation is performed in both temporal and spatial domain with reflection extensions both at each image boundary and at the boundary of each video segment of 16 frames. In the simulation tests, the memory required to process 16 frames as a unit is not a problem in a Pentium PC with 16 MB memory or in most workstations.

IV. SIMULATION RESULTS

Parallel simulations of 3D-SPIHT and MPEG-2 were run on two gray level (8 bits/pixel) SIF (352x240) sequence ' table tennis ', and ' football ' sampled at 30 frames per seconds at test bit rates of 760 kbps (0.3 bits/pixel) and 2.53 Mbps (1.0 bits/pixel). Like MPEG-2, the 3D-SPIHT coder is a fully implemented software encoder and decoder. It is important to note that we can obtain a reconstructed video sequence at any bit rate from just one compressed bit stream file with 3D-SPIHT. Quality of reconstruction is measured by peak signal to noise ratio (PSNR) defined by

$$PSNR = 10 \log_{10}(\frac{255^2}{MSE})\ dB, \tag{2}$$

where MSE denotes the mean squared-error between the original and reconstructed frame.

Sequence	Rate (bpp)	3D-SPIHT(dB)	3D-IEZW (dB)	MPEG-2 (dB)
tennis	0.3	31.0	30.7	30.3
tennis	1.0	37.2	36.7	36.4
football	0.3	27.9	27.3	26.9
football	1.0	34.2	33.5	33.0

Table I: Coding Results (Average PSNR in dB)

Eighty (80) and forty-eight (48) frames were coded for the 'table tennis' and 'football' sequences, respectively. Table I shows that average PSNR results with 3D-SPIHT are 0.3 – 0.7 dB better than 3D-IEZW and 0.6 – 1.2 dB better than MPEG-2. The trend is similar for visual quality. The visual effects the coding are shown in Figure 5, where appear 3D-SPIHT and MPEG-2 reconstructions of the same 'football' frame at the bit rate of 0.2 bpp averaged over 48 frames. Both 3D-SPIHT and EZW exhibit some blurring effects in small regions at low bit rate, while MPEG-2 suffers additionally from blocking effects. Figures 6 and 7 compare 3D-SPIHT and MPEG-2 in terms of PSNR versus frame number at 1.0 bpp and 0.3 bpp for the 'table tennis' and 'football' sequences. For both rates for the 'football' sequence, the PSNR of 3D-SPIHT is generally higher at every frame. (Frame 1 for MPEG-2 is intra-coded at a high rate, so will always show larger PSNR.) But for 'table tennis', which has much more localized motion, there are alternating epochs of PSNR superiority between the two coders. These different patterns of PSNR fluctuations stem from the different allocation of average rate to the individual frames. In SPIHT, the rate is exactly specified across 16 frames, while the fluctuation of bit rate in MPEG-2 follows that of the magnitude levels in the inter- and intra-coded frames. Lastly, 3D-SPIHT exhibits the same phenomenon as 3D-IEZW [[CP96]], that, at the beginning and end of each 16 frame segment, the PSNR's decrease somewhat abruptly, probably due to boundary effects from the subband/wavelet transformation.

V. CONCLUSION

In this paper, we presented a 3D SPIHT video coding scheme which is based on the subset partitioning algorithm in a 3D hierarchical tree. It features the same simplicity and high performance of the 2D SPIHT algorithm for still images. Even without motion compensation in its extension to 3D, it performs better measurably and visually than a full implementation of MPEG-2 with its complicated motion compensation. Finally, the fact that the bit stream is the output of a fully embedded wavelet coder renders it capable of delivering progressive buildup of fidelity and scalability in frame size and rate.

REFERENCES

[ABMD92] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding

using wavelet transformation. *IEEE Trans. Image Processing*, 1:205–220, April 1992.

[CP96] Y. Chen and W. Pearlman. Three-dimensional subband coding of video using the zero-tree method. *in Visual Communications and Image Processing '96, Proc. SPIE 2727*, pages 1302–1309, March 1996.

[LWCP96] J. Luo, X. Wang, C. W. Chen, and K. J. Parker. Volumetric medical image compression with three-dimensional wavelet transform and octave zerotree coding. *in Visual Communications and Image Processing'96, Proc. SPIE 2727*, pages 579–590, March 1996.

[Sha92] J. Shapiro. An embedded wavelet hierarchical image coder. *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, 4:657–660, March 1992.

[SP93] A. Said and W. Pearlman. Image compression using the spatial-orientation tree. *Proc. IEEE Intl. Symp. Circuits and Systems*, pages 279–282, May 1993.

[SP96] A. Said and W. A. Pearlman. A new fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. on Circuits and Systems for Video Technology*, 6:243–250, June 1996.

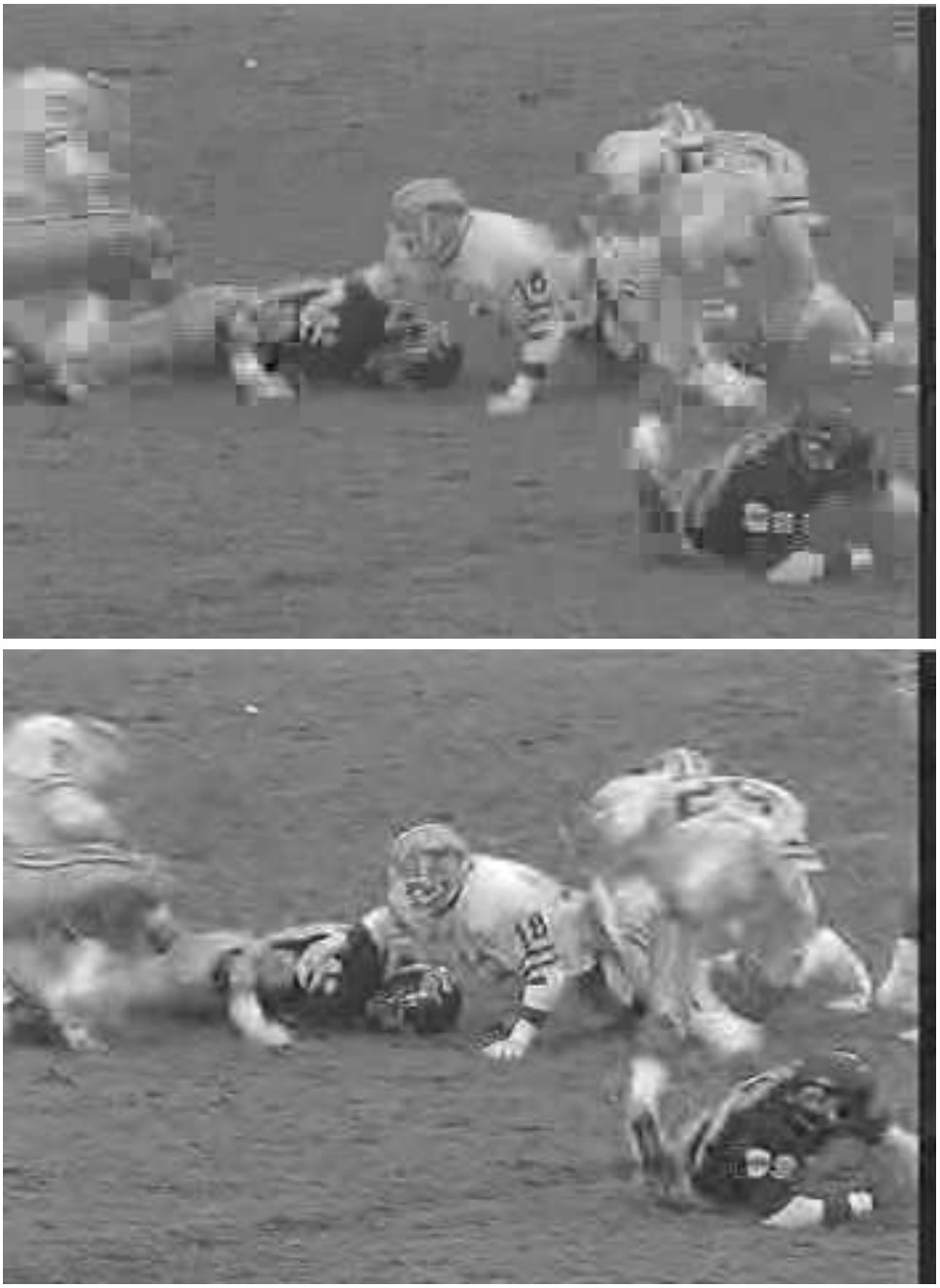

Figure 5: Same 'football' frame reconstructions at 0.2 bpp average rate for MPEG-2 (top) and 3D SPIHT (bottom).

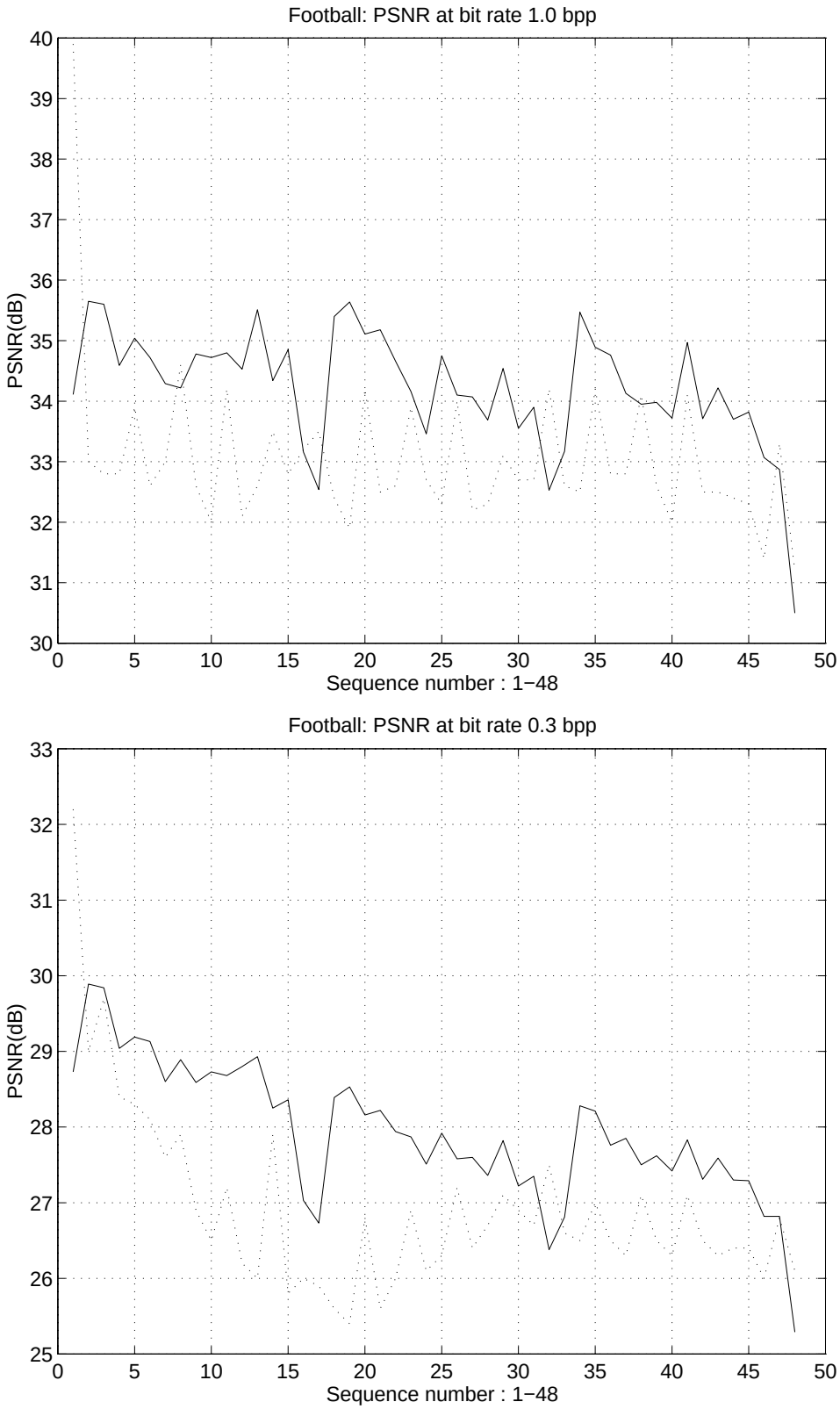

Figure 6: PSNR vs. frame number (1-48) for 'football' at 1.0 and 0.3 bpp. Solid line: 3D-SPIHT; broken line: MPEG-2.

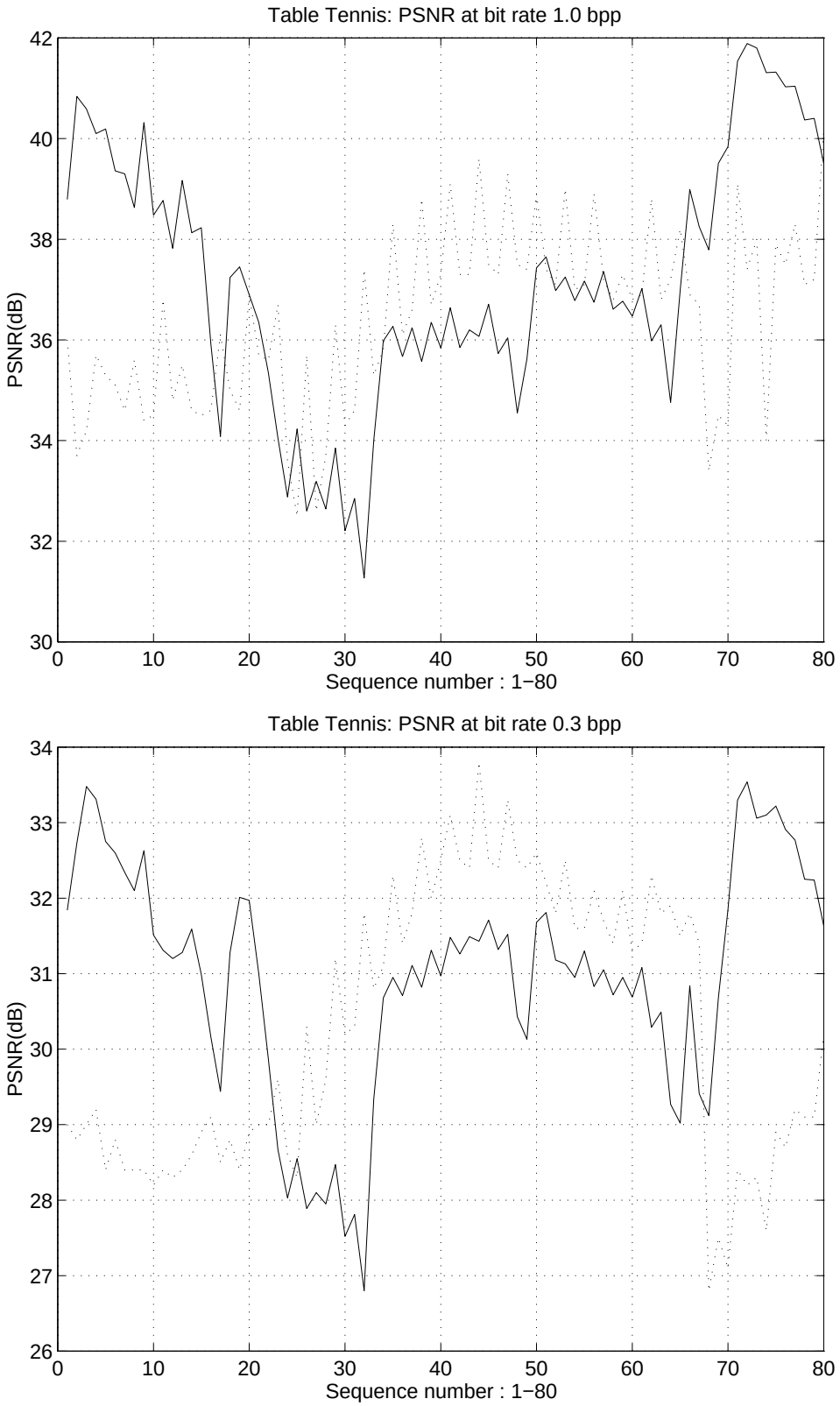

Figure 7: PSNR vs. frame number(1-80) for 'table tennis' at 1.0 bpp and 0.3 bpp
Solid line: 3D-SPIHT; broken line: MPEG-2.