Chapter 4

3-D Video Coding with WBTC

4.1 Introduction

In contrast to the traditional hybrid video coding [44], 3-D video coding systems use one-dimensional (1-D) temporal decomposition to exploit the temporal redundancy present in the video signal. The resulting temporal frames are then spatially decomposed using a two-dimensional (2-D) wavelet transform followed by coefficients encoding. An efficient spatio-temporal analysis and coefficient encoding is the key to achieve effective compression performances. However, in this chapter, the focus is on the coding algorithm to efficiently encode the wavelet coefficients generated by any 3-D wavelet transformations.

Motivated by the success of wavelet-based algorithms [14], [62]-[64] for image compression, they have been extended for 3-D video coding as well [16], [115]-[117]. Because of its simplicity, low complexity, and good rate-distortion performance with embedded bitstream, the set partitioning in hierarchical trees (SPIHT) algorithm [14] has also been successfully extended for 3-D video coding [16]. However, the 3-D SPIHT algorithm puts a restriction on the size of the lowest temporal subband. This is to keep the basic coefficient units of $2 \times 2 \times 2$ for arithmetic coding. By removing this restriction, a more efficient 3-D coefficient tree structure is developed in [142]. This tree structure is quite flexible and has no limitation on the number of wavelet decomposition levels along temporal and spatial directions. Recently, a decoupled 3-D tree structure is also proposed in which temporal and spatial trees are coded independently [143].

Moreover, in these codecs, the color video sequences are coded by treating wavelet transformed luminance-chrominance color planes independently, assuming that they
are mutually exclusive. However, experimental investigation on several sequences has shown that the probability of the chrominance coefficients having smaller magnitudes than the luminance ones in lowest resolution subband is very high. Thus, if a luminance coefficient has insignificant offspring at a given bitplane, the chrominance coefficients at the same location also have a high probability to have insignificant offspring. This interdependency of the transformed color planes has been exploited for efficient coding of color videos [107]-[109].

In this chapter, a 3-D video coding system employing a 3-D extension of the WBTC still image codec is proposed. The key idea is the use of a composite block-tree hierarchical structure to link blocks of wavelet coefficients in spatial, temporal and the color planes in such a way that the insignificant sets are coded together. The proposed 3-D WBTC algorithm integrates the set partitioning strategies of hierarchical trees and blocks into a single algorithm, resulting in improvement in the coding efficiency and reduction in the computational complexity.

4.2 System Overview

The basic structure of the 3-D video coding system is shown in Fig. 4.1. The encoder consists of spatio-temporal analysis filters to perform 3-D wavelet transform and 3-D WBTC encoder to encode the resulting wavelet coefficients. The decoder has structure symmetric to that of the encoder. Due to delay and buffer constraint, a video sequence is divided into a number of GOF and each GOF is coded separately and independently. Theoretically, a larger GOF length may give a higher compression ratio, but requires

![Figure 4.1 The basic structure of 3-D WBTC video coding system.](image-url)
more memory space and causes delay. Therefore, the length of GOF is chosen in order to trade-off the delay in reconstruction (important in real-time applications) and the efficiency of the subsequent coding algorithm.

In the spatio-temporal analysis, a GOF is first temporally decomposed with 1-D wavelet transform into a set of temporal frequency bands followed by spatial decomposition of each temporal frame with 2-D wavelet transform. The temporal decomposition helps in reducing the temporal correlations present in a video, whereas spatial decomposition reduces the spatial correlations present with in each frame. The first-level temporal decomposition is performed by taking the pixels of the same location in different frames of the GOF as 1-D data and applying 1-D wavelet transform. This results in a temporal low-pass and high-pass subbands. The temporal high-pass subband represents the residual signal with most of the coefficients having small values. Since, most of the energy is concentrated in the temporal low-pass subband, therefore it is recursively decomposed to the desired temporal level. Fig. 4.2 shows a 4-level temporal decomposition of a GOF of sixteen frames, resulting into five temporal frequency bands (t-L4, t-H4, t-H3, t-H2 and t-H1). After temporal analysis, each of the resulting frames separately undergoes a spatial dyadic 2-D wavelet decomposition to complete the 3-D decomposition as shown in Fig. 4.3. The resulting spatio-temporal wavelet coefficients are then quantized and coded by the proposed 3-D WBTC algorithm.

Figure 4.2 Temporal decomposition of a GOF of sixteen frames into five temporal frequency bands (t-L4, t-H4, t-H3, t-H2 and t-H1).
Figure 4.3 3-D wavelet decomposition of a GOF of sixteen frames (a) input GOF and (b) the resulting spatio-temporal subbands.

An important issue associated with 3-D wavelet transform is the choice of filters. Different filters in general show quite different signal characteristics in the transform domain in terms of energy compaction, and error signal in the high-frequency bands [148]. However, the investigation of optimum filter design is beyond of the scope of this thesis. In the present work, only the known filters which have shown good performance in wavelet coding systems will be employed.

The rate allocation is done for each GOF based on the overall bit rate, which is defined as

\[
b_g = n_g \times \frac{R_b}{f_r} \text{ bits}
\]

(4.1)

where

\begin{align*}
b_g &= \text{bit budget allocated to a GOF} \\
R_b &= \text{total bit rate (bits/sec.)} \\
n_g &= \text{number of frames in a GOF} \\
f_r &= \text{frame rate (frames/sec.)}
\end{align*}
With the 3-D WBTC, the current bit budget of a GOF will be allocated over each frame within the GOF automatically according to the distribution of actual wavelet coefficient magnitude. Therefore, no explicit rate control is required. However, it is possible to introduce a scheme for bit re-alignment by simply scaling one or more subbands to emphasize or de-emphasize the bands so as to artificially control the visual quality of the video [16].

4.3 3-D WBTC

This section introduces the extension of WBTC image coding (already discussed in Chapter 3) to 3-D video coding. Each GOF of the input video is first temporally decomposed into a set of temporal frequency bands according to Fig. 4.2. For color videos (YUV 4:2:0 format), each color plane is temporally decomposed separately to \( n_t \) levels. Then the resulting temporal frames are wavelet transformed using \( n_t \) levels of decomposition for the luminance (Y) plane and \((n_t-1)\) levels for each of the chrominance (U, V) planes. Since in the 4:2:0 color format, the resolutions of the chrominance planes are one-quarter to that of the luminance plane, wavelet decomposition of the chrominance planes by one level less than that of the luminance plane will result in the lowest resolution band of each transformed color planes of the same dimensions. This simplifies the child-parent relationship to link the three-color planes together.

After spatio-temporal analysis of a GOF, the wavelet coefficients in each color plane are divided into blocks of \( n \times n \) coefficients. To exploit the self-similarity and magnitude localization property in the spatio-temporal subbands, a spatial orientation block-tree is used. It is based on the hypothesis that if a 3-D wavelet transform coefficient at a coarser scale of spatio-temporal decomposition is insignificant with respect to a given threshold, all the 3-D wavelet transform coefficients in the same spatio-temporal location at a finer scale are most likely to be insignificant. Fig. 4.4 shows a typical spatio-temporal block-tree (3-D block-tree) structure. In order to maintain the clarity, only eight temporal frames with three levels of temporal decompositions are shown. However, this basic structure can be extended to \( n_t \) levels of temporal decomposition on a GOF size of \( 2^n \). Only the blocks in the lowest spatial frequency bands are the root blocks and each root block has spatial decedents in the
same frame up to the bottom of the block-tree. Additionally, to exploit the interdependency of the color planes, each block of the lowest spatial frequency band of the Y plane is associated with the blocks at the same location in the corresponding subband of the U and V planes as shown in Fig. 4.5. The child-parent relationship in the composite block-tree structure can be described as follows. Let

\[ B(i, j, t; X) : \text{an arbitrary block of size } n \times n \text{ with top-left corner at column } i, \text{row } j \text{ and frame } t \text{ in color plane } X, \text{where } X \in \{Y, U, V\} \]

\[ w : \text{width of the lowest spatial frequency bands} \]

\[ h : \text{height of the lowest spatial frequency bands} \]

\[ f : \text{number of frames in the lowest temporal frequency band} \]

\[ O(i, j, t; X) = \text{offsprings of a block } B(i, j, t; X) \]

\[ \bullet \text{ if } X = Y, \text{ then} \]

\[ \bullet \text{ if } i < w, j < h, t < f \]

\[ O(i, j, t; Y) = \begin{cases} \{B(i+w,j,t;Y),B(i,j+h,t;Y),B(i+w,j+h,t;Y)\} \\ B(i,j,t;U),B(i,j,t;V),B(i,j+t+f;Y) \end{cases} \]

\[ \bullet \text{ else if } i < w, j < h, \text{ in the highest temporal frequency band} \]

\[ O(i, j, t; Y) = \begin{cases} \{B(i+w,j,t;Y),B(i,j+h,t;Y),B(i+w,j+h,t;Y)\} \\ B(i,j,t;U),B(i,j,t;V) \end{cases} \]

\[ \bullet \text{ else if } i < w, j < h \]

\[ O(i, j, t; Y) = \begin{cases} \{B(i+w,j,t;Y),B(i,j+h,t;Y),B(i+w,j+h,t;Y)\} \\ B(i,j,t;U),B(i,j,t;V),B(i,j,2t;Y),B(i,j,2t+1;Y) \end{cases} \]

\[ \bullet \text{ else} \]

\[ O(i, j, t; Y) = \begin{cases} \{B(2i,2j,t;Y),B(2i+1,2j,t;Y)\} \\ B(2i,2j+1,t;Y),B(2i+1,2j+1,t;Y) \end{cases} \]

\[ \bullet \text{ else if } X \in \{U, V\} \]

\[ \bullet \text{ if } i < w, j < h \]
It is worth mentioning here that the proposed block-tree structure allows the linking of all the coefficients of the three color planes of a GOF through 3-D spatio-temporal block-trees having roots in the lowest spatial frequency band of the lowest temporal frequency band only.

Figure 4.4 Spatio-temporal block-tree structure used in 3-D WBTC (for $n_t = 3$ and GOF size = 8).
After constructing the 3-D spatio-temporal block-trees as defined, the next step is encoding of the coefficients into a bitstream. Essentially, it can be done by feeding the 3-D data structure to the 3-D WBTC coding algorithm. 3-D WBTC algorithm is very much similar to the WBTC algorithm described in Chapter 3 except that it has to process much more complex spatio-temporal block-tree structure. Just like the WBTC, 3-D WBTC also uses bitplane-based coding comprising of sorting and refinement passes. The only difference being it has to process 3-D rather than 2-D sets and three color planes. At the initialization step, only the root blocks from the lowest temporal frequency band of Y plane are added to LIB and LIBS. The LSP starts as an empty list. Also, initial threshold value corresponding to the maximum wavelet coefficients magnitude is computed. The 3-D WBTC will sort the data of a GOF according to the magnitude bitplane-by-bitplane. After the sorting is over, the refinement stage of 3-D WBTC will be exactly the same as WBTC. At the destination, the decoder will follow the same execution path of the encoder, which is conveyed by the received significance decision bits.

4.4 Simulation Results

The 3-D WBTC video coding system has been implemented in software and the performance is evaluated on different color test video sequences in YUV 4:2:0 format, namely; Hall-Monitor (QCIF), Mother-Daughter (QCIF), Salesman (QCIF), Akiyo (CIF) and Salesman (CIF). For temporal decomposition, S/3 biorthogonal filters [54]
are used due to its better overall performance [120]. Spatial decomposition is performed using 9/7 biorthogonal filters [21] as follows. For QCIF sequences number of spatial decomposition level \( n_s = 3 \) and for CIF sequences, \( n_s = 4 \) is used. All the tests were performed using 96 frames and frame rate of 30 frames per second for each sequence. The initial block size in 3-D WBTC is considered as 2x2. All results are without arithmetic coding and without motion compensation. The objective quality of the decoded frames is measured in terms of the peak signal-to-noise ratio (PSNR) of the three color planes. The implementations were done in ‘C’ programming language under LINUX operating system. Tests were performed using a PC having Pentium-4 processor with CPU speed of 3.06 GHz and 1 GB RAM.

4.4.1 Effect of GOF Length

To assess the effect of GOF length on the coding performance of 3-D WBTC, tests were performed with different GOF lengths. For different GOF lengths, the temporal wavelet transform is always applied up to the end (until only one frame is left in temporal low-low band), and the same levels of the spatial decompositions are used in each case. Fig. 4.6 shows the average luminance PSNR performance with GOF length of 8, 16, and 32 for Hall-Monitor (QCIF) and Salesman (CIF) sequences at a wide range of bit rates. Temporal decomposition level used is 3, 4, and 5 for GOF size of 8, 16, and 32, respectively. It shows that by increasing the GOF length, coding performance can be improved. The reason is with larger GOF lengths more temporal decompositions can be applied resulting in more clustering of energy and thereby improvement in coding performance. However, this improvement is more pronounced when moving from GOF length of 8 to 16. Similarly, GOF length of 32 offers marginally better compression performance than 16, but increases the memory requirement and delay significantly. Therefore, for subsequent experiments in this chapter a GOF length of 16 frames is used.

4.4.2 Effect of Temporal Decomposition

Here the aim is to assess the effect of number of temporal decomposition levels on the coding performance. Tests were performed with a GOF size of 16 frames using three and four levels of temporal decompositions. Fig. 4.7 shows the average luminance
PSNR (dB) for *Hall-Monitor* (QCIF) and *Salesman* (CIF) sequences at a range of bitrates. It is observed that coding performance improves with increase in number of temporal decomposition levels. One of the reasons is that higher temporal decompositions results in better energy clustering and thereby improving the coding performance. The other reason is that with increased temporal decomposition, the total number of trees to be coded is reduced and the number of elements per tree is also increased, which in turn improves the coding performance. Therefore, it is beneficial to always apply the temporal decomposition up to the end.

### 4.4.3 Coding Performance

The frame-by-frame luminance PSNR (dB) obtained with the proposed 3-D WBTC video coder for the QCIF sequences (*Hall-Monitor* and *Mother-Daughter*) coded at a bitrate of 10, 50, and 100 kbps and CIF sequences (*Salesman* and *Akiyo*) coded at 100, 500, and 1000 kbps are shown in Figs. 4.8 and 4.9 respectively.

It is observed that the luminance PSNR improves as the bitrate increases. The reason is that as the bitrate increases, more bits will be available to reconstruct a wavelet coefficient and hence improving the PSNR of the decoded frames. This also demonstrates the quality scalability feature of the 3-D WBTC codec in which the video is coded once at the highest bitrate and decoded many times at the required bitrate (quality) from the same embedded bitstream.

The coding performance of 3-D WBTC is next compared with the original 3-D SPIHT [16]. For 3-D wavelet transform, a GOF size of 16 frames is first temporally decomposed into four temporal subbands. The resulting frames are then spatially decomposed to three levels. The initial block size in the proposed algorithm is considered as 2×2 and no arithmetic coding is used. The results of 3-D SPIHT are obtained by running the executables available at [149]. The software uses 3-level decomposition along both temporal and spatial directions to keep 2×2×2 coefficient units for arithmetic coding. This software also uses arithmetic coding. All results are without motion compensation in both coders.
Figure 4.6 Rate-distortion performances with different GOF lengths for (a) Hall-Monitor (QCIF) and (b) Salesman (CIF) sequences.
Figure 4.7 Rate-distortion performances with 3- and 4-level temporal decompositions of a GOF of 16 frames for (a) Hall-Monitor (QCIF) and (b) Salesman (CIF) sequences.
Figure 4.8 Frame-by-frame luminance PSNR (dB) for QCIF sequences (a) Hall-Monitor and (b) Mother-Daughter at bitrates of 10, 50, and 100 kbps.
Figure 4.9 Frame-by-frame luminance PSNR (dB) for CIF sequences (a) *Salesman* and (b) *Akiyo* at bitrates of 100, 500 and 1000 kbps.
Figure 4.10 show the average luminance PSNR in the range of 20-100 kbps for the QCIF sequences Hall-Monitor and Salesman. A comparison reveals that 3-D WBTC always outperforms 3-D SPIHT. In particular, it brings a performance gain of about 1.4-2.1 dB with respect to 3-D SPIHT for both sequences. One of the reasons for the superior performance is the better aggregation of insignificant coefficients by using a block-tree, and hence increases the coding efficiency. The other reason for better performance is the partial exploitation of the intra-subband correlations in the form of zero-blocks.

Finally, the performance of 3-D WBTC is also compared with two improved versions of 3-D SPIHT, namely; 3-D SPIHT with optimum tree structure (3-D SPIHT-OT) [142] and an adapted version of the original 3-D SPIHT [16]. In these codecs, each color plane is spatially decomposed into four levels for CIF sequences and three levels for QCIF sequences. The adapted version of 3-D SPIHT is similar to the one given by Kim et al. [16] but with the following differences. The CIF and QCIF resolution images with four and three levels of decompositions respectively, result in odd dimensions of LL-band of chrominance planes and hence not suitable for SPIHT-type child-parent relationship. Kim et al. have solved this problem by extending the chrominance planes before the 2-D spatial transformation [16]. However, this increases the system complexity and also reduces the coding efficiency slightly, due to coding of artificial coefficients. In the present implementation, the last rows and columns of the LL-band of chrominance planes follow the EZW’s tree structure. 3-D SPIHT also uses separate coding of each color plane as in [16], but initialization structure of LIP and LIS is similar to that in [99]. However, the implementation of 3-D SPIHT-OT is exactly based on the algorithm given in [142].

Table 4.1 compares the average PSNR (dB) of the three color planes at three different bit budgets for all the test sequences. A comparison reveals that the 3-D WBTC always outperforms both 3-D SPIHT and 3-D SPIHT-OT in terms of luminance PSNR. The performance gain is more pronounced at lower bitrates and for some sequences. In particular, it brings a performance gain of about 0.2-1.5 dB for QCIF test sequences and up to 0.4 dB for CIF sequences when compared with 3-D SPIHT. It is also interesting to note that at lower bitrates, 3-D SPIHT is superior to 3-D SPIHT-OT
by up to 0.2 dB. This is due to the adaptation used in the present implementation as discussed above.

Figure 4.10 Average luminance PSNR (dB) comparison at various bitrates for QCIF sequences (a) Hall-Monitor and (b) Salesman.
However, the chrominance PSNR of 3-D WBTC is sometimes superior and some times inferior and even some times comparable to the 3-D SPIHT. Since the human eyes are more sensitive to the changes in the brightness, therefore the gain in luminance PSNR as obtained from the 3-D WBTC will yield visually better results. This is also evident from the subjective quality comparisons as given in Fig. 4.11. The reason for the superior performance is due to the better exploitation of both inter- and intra-subband correlations as well as the interdependency of the color planes.

**Table 4.1** Color plane wise average PSNR (in dB) comparisons of different 3-D video codecs

<table>
<thead>
<tr>
<th>Bitrate (kbps)</th>
<th>3-D SPIHT</th>
<th>3-D SPIHT-OT</th>
<th>3-D WBTC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>U</td>
<td>V</td>
</tr>
<tr>
<td><strong>Hall-Monitor (QCIF)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>21.9</td>
<td>31.8</td>
<td>36.3</td>
</tr>
<tr>
<td>50</td>
<td>29.5</td>
<td>36.1</td>
<td>38.5</td>
</tr>
<tr>
<td>100</td>
<td>34.0</td>
<td>37.8</td>
<td>40.2</td>
</tr>
<tr>
<td><strong>Mother-Daughter (QCIF)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>26.2</td>
<td>33.7</td>
<td>33.0</td>
</tr>
<tr>
<td>50</td>
<td>32.8</td>
<td>38.8</td>
<td>39.7</td>
</tr>
<tr>
<td>100</td>
<td>35.8</td>
<td>40.5</td>
<td>41.3</td>
</tr>
<tr>
<td><strong>Salesman (CIF)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>28.7</td>
<td>37.0</td>
<td>37.9</td>
</tr>
<tr>
<td>500</td>
<td>35.1</td>
<td>41.2</td>
<td>41.8</td>
</tr>
<tr>
<td>1000</td>
<td>37.1</td>
<td>42.8</td>
<td>43.6</td>
</tr>
<tr>
<td><strong>Akiyo (CIF)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>33.2</td>
<td>37.3</td>
<td>39.4</td>
</tr>
<tr>
<td>500</td>
<td>41.3</td>
<td>44.9</td>
<td>46.2</td>
</tr>
<tr>
<td>1000</td>
<td>44.9</td>
<td>48.5</td>
<td>49.5</td>
</tr>
</tbody>
</table>
Figure 4.11 Subjective quality comparisons for 13th frame of the QCIF sequence *Mother-Daughter* coded at 10 kbps (a) original frame (b) 3-D SPIHT (c) 3-D SPIHT-OT and (d) 3-D WBTC.
4.4.4 Computational Complexity

Computational complexity is assessed in terms of the run times of encoder and decoder. Table 4.2 summarizes average encoding and decoding timings (in milliseconds) for one GOF by the three codecs. It can be observed that the encoding and decoding complexity of both 3-D SPIHT and 3-D SPIHT-OT is almost of the same order. On the other hand 3-D WBTC encodes a set of wavelet coefficients about 1.5-2.2 times faster than 3-D SPIHT. The possible reason for this is that being a block-based encoder it has to process considerably smaller number of elements in its lists and hence reduces the encoding time. The decoder timings of 3-D WBTC are higher than 3-D SPIHT. This is because reconstruction of a significant coefficient in larger sets is more time consuming than the relatively smaller sets of 3-D SPIHT. However, in power limited devices such as mobile handsets, where encoder sets the computational complexity limit, the extra complexity of the decoder does not create any burden.

Table 4.2 Comparison of encoding and decoding timings of different 3-D video codecs.

<table>
<thead>
<tr>
<th>Bitrate (kbps)</th>
<th>Encoding time (ms)</th>
<th>Decoding time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3-D SPIHT</td>
<td>3-D SPIHT-OT</td>
</tr>
<tr>
<td>Hall-Monitor (QCIF)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>146</td>
<td>148</td>
</tr>
<tr>
<td>50</td>
<td>271</td>
<td>270</td>
</tr>
<tr>
<td>100</td>
<td>324</td>
<td>326</td>
</tr>
<tr>
<td>Salesman (CIF)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1017</td>
<td>926</td>
</tr>
<tr>
<td>500</td>
<td>1360</td>
<td>1314</td>
</tr>
<tr>
<td>1000</td>
<td>1524</td>
<td>1514</td>
</tr>
</tbody>
</table>

73
4.5 Summary

In this chapter, a 3-D video coding system employing a 3-D extension of the WBTC algorithm is proposed. The key idea is the use of a composite block-tree hierarchical structure to link blocks of wavelet coefficients in spatial, temporal and the color planes in such a way that insignificant sets are coded together. Simulation results show that the 3-D WBTC gives better coding efficiency and has lower encoder complexity as compared to 3-D SPIHT and 3-D SPIHT with optimum tree structure. The proposed 3-D WBTC poses features such as fidelity embedded bitstream for progressive transmission, precise rate control for constant bitrate traffic and low complexity. Therefore, it is quite attractive for video delivery over heterogeneous networks through hand held portable devices.