Scalable Video Coding

by

Susie J. Wee

S.M., Massachusetts Institute of Technology (1991)
S.B., Massachusetts Institute of Technology (1991)

Submitted to the
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the
Massachusetts Institute of Technology

June 1996

© Massachusetts Institute of Technology, 1996.
All rights reserved.

Author

Department of Electrical Engineering and Computer Science
May 24, 1996

Certified by

William F. Schreiber
Professor Emeritus of Electrical Engineering
Thesis Supervisor

Accepted by

Frederic R. Morgenthaler
Chairman, Committee on Graduate Students

JUL 16 1996
LIBRARIES
Scalable Video Coding

by

Susie J. Wee

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In terrestrial broadcasting, a centrally located transmitter delivers a signal to receivers distributed throughout the broadcast area. Receivers closest to the transmitter receive the highest signal power, and therefore have the capacity to receive the highest data rates. In a typical television broadcasting scenario, the capacity available in the central region can be 4 times greater than the capacity available at the fringe of reception.

An advanced television system has been designed to exploit the wide range of capacities available in the terrestrial broadcasting environment. Joint source/channel coding techniques allow the system to be efficient, where efficiency implies that the decoded video quality should depend on the available capacity at each receiver. This variation in video quality is achieved with scalable video coding and hybrid transmission.

Scalable video coding allows the video to be decoded at several predefined spatial resolutions, where the highest decodable resolution depends on the received signal strength. Layered coding methods are used to achieve natural-looking video at every resolution. In addition, a novel concept of conditioning is introduced to combat the artifact propagation problem inherent to layered systems. Within each spatial resolution, hybrid transmission allows the video quality to vary gracefully with received signal power. Hybrid transmission combines the advantages of analog and digital methods; it allows the system to use the powerful video compression techniques associated with digital coding while retaining the graceful degradation properties of conventional analog systems.

Thesis Supervisor: William F. Schreiber
Title: Professor Emeritus of Electrical Engineering
Acknowledgements

I would like to express my deepest appreciation and gratitude to Professor Schreiber. He taught me many valuable lessons during our little group meetings and influenced me in more ways than he knows. I always enjoyed watching him analyze pictures – his insights never ceased to amaze me. I am both proud and lucky to have the honor of being Professor Schreiber’s last student!

Mike Polley has been a wonderful friend and colleague throughout the years. I value the time we spent working together and the many cups of cappuccino that we shared. I would like to thank Mike and Raj Aggarwal for their Texan BBQs and all the great times – they have been true friends.

I would like to thank my thesis readers, Professors Adelson and Oppenheim. Their technical insights are truly inspiring. Professor Adelson first introduced me to image processing many years ago; Professor Oppenheim has been both a friend and a mentor. I would also like to thank Professor Lim for the use of his facilities, and the entire Advanced Television and Signal Processing group for the many friendships and good times. I give a special thanks to Cindy LeBlanc for always making me feel like a part of the group. I would also like to thank Marilyn Pierce and Monica Bell for taking care of me and the EECS department.

Many close friends have made my graduate years an enjoyable experience. Thanks to Diane Ho, Aradhana Narula, and Cynara Wu for their endless friendship throughout the years; thanks to Kathleen Wage for our hallway chats; and thanks to Patti Jankun and Kevin Yu for always having an ice hockey team for me to play on, even in the summertime! My involvement with the MIT women’s ice hockey team has enriched my graduate experience. I am thankful for having had the opportunity to participate both as a player and as a coach.

I give special thanks to John Apostolopoulos for being my best friend. John’s patience, encouragement, dedication, and honesty make him a true inspiration. He has filled my graduate years with laughter and happiness which we shall share for many years to come.

Finally, I thank my family for their never-ending love and support. My big brother Kee has protected and advised me throughout the years, and my little sister Sylvia has supported me, particularly as my roommate during this past year. I cannot express enough thanks to my parents, whose selflessness has made this day possible – it is to them that I dedicate this thesis.
Dedication

To Mom and Dad
# Contents

1 Introduction ................................................................. 19  
1.1 Motivation ............................................................... 20  
1.2 Thesis Organization .................................................... 22  
1.3 Contributions .......................................................... 23  

2 Video Coding ............................................................... 25  
2.1 Still-Image Compression .................................................. 25  
  2.1.1 Transform/Subband Representation ................................ 26  
  2.1.2 Quantization and Entropy Coding ................................ 27  
  2.1.3 Nature of Transform Coding Artifacts ............................. 28  
  2.1.4 Remarks ............................................................... 29  
2.2 Video Compression ....................................................... 30  
  2.2.1 Intraframe Methods .................................................. 30  
  2.2.2 Interframe Methods .................................................. 30  
  2.2.3 Remarks ............................................................... 32  
2.3 Summary ................................................................. 34  

3 Scalable Video Coding .................................................... 35  
3.1 Introduction ............................................................... 35  
3.2 Scalable Coding Methods ................................................. 36  
  3.2.1 Scalable Coding of Still Images .................................. 36  
  3.2.2 Scalable Coding of Video .......................................... 45  
3.3 Conventional Approaches ................................................. 47  
  3.3.1 Laplacian Pyramid .................................................. 47  
  3.3.2 Embedded Wavelet Coder .......................................... 47  
  3.3.3 MPEG-2 ............................................................... 47  
  3.3.4 Spatio-Temporal Pyramid .......................................... 48  
3.4 A Generalized Framework for SVC ...................................... 49
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.1 Overview</td>
<td>49</td>
</tr>
<tr>
<td>3.4.2 Generalized Framework</td>
<td>49</td>
</tr>
<tr>
<td>3.4.3 Conditioning</td>
<td>51</td>
</tr>
<tr>
<td>3.4.4 Example: Layered JPEG/MPEG</td>
<td>53</td>
</tr>
<tr>
<td>3.4.5 Summary</td>
<td>56</td>
</tr>
<tr>
<td>4 Hybrid Video Communication for Terrestrial Broadcasting</td>
<td>59</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>59</td>
</tr>
<tr>
<td>4.2 Fundamental Concepts</td>
<td>60</td>
</tr>
<tr>
<td>4.3 Hybrid Transmission</td>
<td>66</td>
</tr>
<tr>
<td>4.3.1 Description</td>
<td>67</td>
</tr>
<tr>
<td>4.3.2 Hybrid Channel Efficiency</td>
<td>68</td>
</tr>
<tr>
<td>4.3.3 Discussion</td>
<td>70</td>
</tr>
<tr>
<td>4.4 Hybrid Video Coding</td>
<td>70</td>
</tr>
<tr>
<td>4.4.1 Background</td>
<td>70</td>
</tr>
<tr>
<td>4.4.2 Relative Importance of Different Video Data</td>
<td>71</td>
</tr>
<tr>
<td>4.4.3 Hybrid Video Coder</td>
<td>72</td>
</tr>
<tr>
<td>4.4.4 Experimental Results</td>
<td>74</td>
</tr>
<tr>
<td>4.4.5 Remarks</td>
<td>75</td>
</tr>
<tr>
<td>4.5 Summary</td>
<td>78</td>
</tr>
<tr>
<td>5 An ATV Terrestrial Broadcasting System</td>
<td>79</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>79</td>
</tr>
<tr>
<td>5.2 Hybrid Transmission</td>
<td>80</td>
</tr>
<tr>
<td>5.3 Joint Source/Channel Coding</td>
<td>81</td>
</tr>
<tr>
<td>5.4 Channel Coder</td>
<td>81</td>
</tr>
<tr>
<td>5.5 Scalable Source Coder</td>
<td>82</td>
</tr>
<tr>
<td>5.5.1 Pyramid Filtering</td>
<td>84</td>
</tr>
<tr>
<td>5.5.2 Hybrid Video Coding</td>
<td>86</td>
</tr>
<tr>
<td>5.6 Results</td>
<td>89</td>
</tr>
<tr>
<td>6 Summary and Future Work</td>
<td>95</td>
</tr>
<tr>
<td>6.1 Future Work</td>
<td>96</td>
</tr>
<tr>
<td>A Motion-Compensated Prediction in a Hybrid-Transmission HDTV System</td>
<td>99</td>
</tr>
<tr>
<td>A.1 Background</td>
<td>99</td>
</tr>
<tr>
<td>A.2 Experiment</td>
<td>100</td>
</tr>
<tr>
<td>A.2.1 Digital Video Coding System</td>
<td>100</td>
</tr>
</tbody>
</table>
A.2.2 Modified Digital Video Coding System .......................... 101
A.3 Experimental Results .............................................. 102

B Adaptive Selection Coding .......................................... 103
B.1 Introduction ....................................................... 103
  B.1.1 Coefficient Selection .......................................... 103
  B.1.2 Correlation Across Orientations .............................. 105
B.2 Adaptive Selection Algorithm .................................... 105
B.3 Experimental Results .............................................. 109

C A Wiener Filter Approach to Reducing Transform Coding Artifacts 119
C.1 Introduction .......................................................... 119
C.2 Transform Coding Artifacts ....................................... 119
C.3 Reduction of Artifacts ............................................. 121
  C.3.1 Previous Approaches .......................................... 121
  C.3.2 Wiener-Filter Approach ....................................... 122
C.4 Remarks ............................................................. 123
C.5 Summary ............................................................. 124
List of Figures

1.1  *A video communication system.* The encoder codes the original video into a data stream that is transmitted across the channel. The received data stream is decoded into the video that will be viewed by the end user. .................................................. 19

1.2  *Terrestrial broadcast environment.* A centrally located transmitter delivers a signal to receivers distributed throughout the broadcast area. .................................................. 20

2.1  *Conventional transform coding.* Transform coding achieves compression by exploiting the spatial correlation inherent in typical imagery. .................................................. 26

2.2  *Basis functions.* Two-dimensional basis functions for an 8 × 8 block DCT and a three-level wavelet transform. .................................................. 27

2.3  *Conventional transform coding – a more detailed view.* In conventional transform coding, the transform coefficients are quantized. After quantization, the DC coefficients and the locations and amplitudes of the nonzero AC coefficients are encoded into digital data stream. .... 28

2.4  *Motion-compensated transform coding.* A high degree of compression is achieved by exploiting the temporal correlation inherent in typical video. .................................................. 31

2.5  *Successive frames of original video.* Each frame contains 256 × 512 pixels. .................................................. 32

2.6  *Motion-compensated prediction and residual.* The MC prediction and residual were computed for the two video frames shown in Figure 2.5. Standard block-matching techniques were used with 16 × 16 pixel non-overlapping windows. .................................................. 33

3.1  *Scalable video coding.* The video is encoded into data streams of prioritized importance. The base data stream can be decoded to reconstruct baseline-quality video; additional enhancement data streams can also be decoded to reconstruct increased-quality video. .................................................. 36

3.2  *Spatial scalability.* Sample frames of the “Girl” sequence are shown at low (1/2 × 1/2), medium (3/4 × 3/4), and high (1 × 1) resolutions. .................................................. 37

3.3  *Spatial scalability: Viewing full-size video.* Viewers often prefer to view the lower-resolution video at full size. Sample frames of video are shown at low (1/2 × 1/2), medium (3/4 × 3/4), and high (1 × 1) resolutions. .................................................. 38
3.4 One-dimensional subband coding. The analysis $F$ and synthesis $G$ filterbanks are typically designed to be orthogonal, maximize compression efficiency, and yield perfect reconstruction in the case of lossless coding. ................................................................. 39

3.5 Octave-band wavelet decomposition. The higher-frequency subbands are larger than the lower-frequency subbands. This results in finer spatial frequency resolution at lower frequencies, and finer spatial resolution at higher frequencies. ................................................................. 40

3.6 Partial reconstruction of wavelet subbands. Spatial scalability can be achieved by grouping the subbands by frequency and coding each group into the base and enhancement data streams. The images that result from partially reconstructing the subbands are shown in Figure 3.7. ................................................................. 41

3.7 Partially reconstructed images. Octave-band wavelet transform at $1/8 \times 1/8$, $1/4 \times 1/4$, and $1/2 \times 1/2$ resolutions. ................................................................. 42

3.8 Octave-band recursive pyramid. Burt and Adelson’s recursive pyramid produces natural-looking images at every resolution. The Resample 2:1 is a lowpass filter and subsample operation and the Resample 1:2 is an upsampling and lowpass filter operation. ................................................................. 43

3.9 Octave-band layered pyramid. This slight generalization of the recursive pyramid also produces natural-looking images at every resolution. This generalization provides added flexibility and further insight into the oversampled representation. ................................................................. 44

3.10 Octave-band pyramid representation. The smallest band contains the lowest-frequency components. Each larger band contains additional higher-frequency components as well as the lower-frequency components that were not coded in the lower levels. ................................................................. 45

3.11 Partial reconstruction of pyramid subbands. Spatial scalability can be achieved by coding each subband into a data stream. The lower-resolution images that result (using different relative bandwidths) are shown in Figure 3.3. ................................................................. 46

3.12 A generalized framework for scalable video coding. This framework encompasses current approaches and accommodates an improved class of layered scalable video coders. A novel aspect lies in the use of pre- and post-conditioning within each level of the coder. The decoder, post-conditioner, and adder in the highest level are shown for illustrative purpose; they are used only in the receiver and would not be needed in the actual encoder. ................................................................. 50

3.13 An improved generalized framework for scalable video coding. The fully decoded video can also be conditioned in each level of the pyramid. ................................................................. 52

3.14 Conditioning example: Two-layer MPEG, with and without low-level conditioning. ................................................................. 53

3.15 Conventional approaches vs. conditioning. A video sequence was coded with the conventional two-level spatially scalable coder shown in Figure 3.14. The PSNRs of the reconstructed high-resolution video sequences are shown for the conventionally coded video, the conditioned video, and the post-processed video. ................................................................. 56
3.16 *Performance comparison.* A video sequence was coded in a two-level spatially scalable video coder. An MPEG coder was used in each level. An enlarged view of the fourth frame of the reconstructed high-resolution video is shown for the conventional approach (37.22 dB, left), conventional approach with post-processing (37.33 dB, center), and proposed approach featuring conditioning (37.93 dB, right). Conditioning outperforms conventional coding and post-processing. .............................. 57

3.17 *Effectiveness of conditioning.* An experiment was performed to better understand the source of the improvement achieved by conditioning. Post-conditioning was applied in only the low level, in only the high level, and in both the low and the high levels. Results show that the majority of the gain is due to post-conditioning the low-level video. .............................. 58

4.1 *Communication system.* A communication system is composed of a transmitter, a communication channel, and a receiver. .............................. 60

4.2 *Capacity of a Gaussian channel.* The channel capacity is the maximum data rate that can be delivered reliably across a channel. .............................. 61

4.3 *Separate source/channel coding.* The separation theorem states that the source and channel coders can be designed separately without loss of optimality. A number of conditions must be satisfied for this to hold true. .............................. 62

4.4 *Joint source/channel coding.* The conditions for the separation theorem are not satisfied for all types of practical video communication systems. Therefore, joint design of the source and channel coders can be beneficial. .............................. 62

4.5 *Terrestrial broadcast environment.* A centrally located transmitter delivers a signal to receivers distributed throughout the broadcast area. Typically, the transmitter is located downtown in a densely populated area. This results in a high concentration of receivers near the transmitter. .............................. 63

4.6 *SNR vs. distance.* In terrestrial broadcasting, the received SNR depends on the receiver’s distance from the transmitter. (Data provided by Dr. Oded Bendov.) .............................. 63

4.7 *Decoded data rate in digital transmission.* In a single-quality all-digital system, the data rate is dictated by the channel capacity of the receivers operating under the worst conditions. Notice the capacity that is wasted at the higher SNRs. .............................. 64

4.8 *Multiple data streams.* Two lower-rate signals can be transmitted across the channel as long as $R_1 + R_2 \leq R$. .............................. 65

4.9 *Theoretical channel capacity of embedded data streams.* The various plots represent different power distributions between the two embedded components. .............................. 66

4.10 *A simple one-dimensional hybrid constellation.* The hybrid signal is composed of a digital discrete-amplitude component and an analog continuous-amplitude component. .............................. 67

4.11 *Hybrid transmission.* The hybrid signal can be characterized by the decoded digital data rate and the received analog SNR as a function of the channel SNR. .............................. 69
4.12 Video coders. Transform coding and motion-compensated transform coding result in different types of data. ................................................. 71

4.13 Digital transform coding. In conventional digital transform coding, the transform coefficients are quantized, and the location and amplitude information of the nonzero quantized coefficients are encoded into a digital data stream. ............................................. 73

4.14 Hybrid transform coding. In hybrid transform coding, transform coefficients are selected for coding. The DC coefficients are coded separately into a digital data stream. The locations of the selected AC coefficients are encoded into a digital data stream, and the amplitudes are transmitted in full precision. .............................................. 73

4.15 Comparison of digital and hybrid systems. Reconstructed images in digital receivers (top) and in hybrid receivers (bottom) located at the fringe of the broadcast area. ..................... 76

4.16 Hybrid system. The quality of the reconstructed video in a hybrid system varies gracefully with channel conditions. This is the image that would be reconstructed at a higher channel SNR. In a digital system, the reconstructed image would be identical to that image shown in Figure 4.15. ......................................................... 77

5.1 Joint source/channel coding. The source and channel coders were designed jointly with the goal of efficiently using the channel capacity available to each receiver in the terrestrial broadcast area. .................................................. 81

5.2 Source and channel coders. The source coder encodes the video into three data streams. The channel coder delivers these data streams across the broadcast area. ........................................... 81

5.3 Coverage area. A digital data stream is delivered to all the receivers in the broadcast area. Receivers with higher SNRs can also decode a first hybrid data stream; and receivers with even higher SNRs can also decode a second hybrid data stream. .............................. 82

5.4 Channel coder. The channel coder prepares the data stream for transmission across the channel. Forward error correction coding, spread-spectrum processing, and orthogonal frequency division multiplexing protect the system against typical analog channel impairments such as additive noise and multipath interference. ........................................ 83

5.5 Scalable source coder. The scalable source coder compresses video into three data streams (Digital, Hybrid-1, and Hybrid-2) using pyramid filtering, MPEG video coding, and hybrid video coding. Receivers closest to the transmitter can use all three data streams to decode high-resolution video; receivers farther from the transmitter can use the Digital and Hybrid-1 data streams to decode medium-resolution video; and receivers farthest from the transmitter can use the Digital data stream alone to reconstruct low-resolution video. .............. 84
5.6 *Intraframe hybrid video coder.* The intraframe hybrid video coder compresses enhancement video into a hybrid data stream made up of analog and digital components. The DC coefficients are coded separately into a digital data stream; the locations and amplitudes of selected AC coefficients are encoded into a digital data stream and an analog data stream.

5.7 *Interframe hybrid video coder.* The interframe hybrid video coder compresses enhancement video into a hybrid data stream made up of analog and digital components. The motion vectors and the locations of selected coefficients are encoded into the digital data stream; the amplitudes of selected coefficients are communicated in the analog data stream.

5.8 *High-resolution original video.* The fifth frame of the progressively scanned “Traffic-girl” sequence.

5.9 *Low-resolution reconstructed video.* The low-resolution video can be decoded throughout the broadcast area, which includes receivers with SNRs exceeding 6 dB.

5.10 *Medium-resolution reconstructed video.* The medium-resolution video can be decoded by receivers with SNRs exceeding 17 dB. This particular frame was decoded at a channel SNR of 19 dB.

5.11 *High-resolution reconstructed video.* The high-resolution video can be decoded by receivers with SNRs exceeding 29 dB. This particular frame was decoded at a channel SNR of 32 dB.

B.1 *Conventional transform coding.* In conventional transform coding, the nonzero quantized coefficients are coded into a digital data stream.

B.2 *Generalized transform coding.* In a more general approach to transform coding, coefficients are explicitly selected and coded into a data stream.

B.3 *Selection mask.* The traffic-girl sequence was coded with the ATV system described in Chapter 5. This is a part of the selection mask that results when selecting 5% of the highest-amplitude wavelet coefficients in the middle level.

B.4 *Adaptive selection algorithm.* This algorithm is used to select transform coefficients in a hybrid video coder.

B.5 *Naming convention for wavelet bands.*

B.6 *AND operation.* The AND operation exploits the inter-band correlation that exists in the classification mask.

B.7 *Adaptive selection mask.* The mask shown in Figure B.3 resulted from simply thresholding the coefficients. This is the final selection mask that results from the adaptive selection algorithm. The data rate required to code this mask is significantly lower than that required to code the initial mask.

C.1 *Original and DCT-coded images.* The coded Lena image is afflicted with blocking effects.
C.2  *Transform coding artifacts due to quantization.* The average noise distribution over blocks of the DCT-coded image is shown in the transform and spatial domains. Note the inverted pincushion distribution of the quantization noise in the spatial domain. .......................... 121

C.3  *Wiener filtering methods for removing transform coding artifacts.* The Wiener filter can be applied as an alternative set of synthesis functions in the decoder or as a post-processing operation following conventional decoding. ................................................................. 122

C.4  *Wiener filtered image and spatial-block noise distribution.* The Wiener filtered image has reduced blocking effects. This results in a flattened spatial-block noise distribution. ......................... 123
Chapter 1

Introduction

Video communication will play an important role in emerging technologies. Applications such as video conferencing, video telephony, and advanced television broadcasting all require the communication of video. The challenge in designing a video communication system lies in the fact that video has very large raw data rates, and this information must be conveyed across a communication channel with limited bandwidth. This imbalance requires the development of very efficient video communication techniques.

A video communication system is shown in Figure 1.1. The encoder codes the original video into a data stream that is transmitted across the channel. The received data stream is decoded into the video that will be viewed by the end user. The encoder is composed of a source and channel coder; the source coder encodes the video into data streams that the channel coder delivers across the communication channel.

Figure 1.1: A video communication system. The encoder codes the original video into a data stream that is transmitted across the channel. The received data stream is decoded into the video that will be viewed by the end user.
The communication channel imposes constraints on system design due to factors such as bandwidth limitations and channel impairments. These constraints can be different for each application. For example, an ATM network suffers from impairments such as packet loss, while a terrestrial broadcast channel suffers from additive noise, interference, and multipath. A high-performance video communication system can best be achieved by designing it to exploit the properties of the particular channel.

We have developed an advanced television system for terrestrial broadcasting under the premise that a video communication system should be designed for the particular application. The primary objective of this thesis was to design the source coder of the system. The channel coder was developed in a companion thesis by Polley [1]. These coders were designed jointly with the goal of achieving high performance by exploiting the wide range of channel capacities available in the broadcast channel [2]. In particular, the best video quality that any individual receiver can decode depends on the channel capacity available to that receiver.

1.1 Motivation

In terrestrial broadcasting, a centrally located transmitter delivers a signal to receivers distributed throughout the broadcast area, as shown in Figure 1.2. Typically, the transmitter is located downtown in a densely populated area. This results in a high concentration of receivers near the transmitter.

![Figure 1.2: Terrestrial broadcast environment. A centrally located transmitter delivers a signal to receivers distributed throughout the broadcast area.](image)

The receivers closest to the transmitter receive the highest signal power, while those farther away receive a lower signal power. For this reason, the receivers closest to the transmitter have the highest channel capacity, which implies that they capable of receiving the highest data rates. Meanwhile, the receivers located farther out have lower channel capacity and thus can only receive lower data rates. In a typical television broadcast scenario, the channel capacity available in
central regions of the broadcast area can be four times greater than the capacity available at the fringe area. An efficient use of this channel implies that the higher-capacity receivers should be able to decode higher-quality video.

Conventional television systems use analog transmission methods. In these systems, the wide range of received signal power directly translates to cleaner video for closer-in receivers and noisier video for farther-out receivers.\(^1\) This desirable property is called graceful degradation, and it naturally allows the decoded video quality to depend on the channel conditions. The shortcoming of this purely analog approach is that high compression rates cannot be achieved with purely analog methods.

The current proposals for digital television deliver a single-rate, single-quality video signal to all the receivers in the broadcast area. Powerful channel coding methods deliver the digital data with a very low probability of error at signal-to-noise ratios greater than the threshold. This enables the use of very powerful video compression algorithms such as motion-compensated prediction and entropy coding. However, only a single video quality is decodable by all the receivers. This does not exploit the extra channel capacity available in the central region of the broadcast area.

In order to simultaneously achieve the benefits of graceful degradation and high-performance video coding, Schreiber developed the concept of hybrid transmission [3]. In hybrid transmission, a discrete-time signal is composed of a digital discrete-amplitude component and an analog continuous-amplitude component. This hybrid signal is transmitted across the broadcast area. Channel coding methods ensure that the digital component is received with very low probability of error and that the analog component is received with an SNR proportional to the SNR in the channel. The reliable reception of the digital component allows the system to use the powerful video compression algorithms traditionally associated with digital methods. Meanwhile, the graceful reception of the analog component allows the quality of the decoded video to vary gracefully with channel conditions.

Schreiber also proposed the concept of multiresolution video transmission for terrestrial broadcasting [4]. When using multiresolution transmission of video, the video can be decoded at various spatial resolutions, depending on the channel conditions available to the individual receiver. The receivers farthest from the transmitter at the fringe of the broadcast area can only decode low-resolution video, the closer receivers can decode medium-resolution video, and the closest receivers can decode the highest-resolution video.

We have designed an advanced television system with the goal of efficiently using the wide range of channel capacities available in the terrestrial broadcast environment, where efficiency im-

---

\(^1\) It is interesting to note that the type of noise generated in conventional analog television receivers is primarily white noise. This is one of the most benign of the possible distortions. The white noise is spread across the entire screen, and does not cause unnatural, structured artifacts that are distracting to the viewer. However, the ghosting that results from multipath can be quite distracting.
pies that higher channel capacity translates to better video quality. This system is based on the multiresolution and hybrid transmission concepts described above. Thus, the improvement in video quality occurs in two ways: multiple spatial resolutions can be decoded where the highest decodable resolution is dictated by the available channel capacity; and within each resolution the video quality improves gracefully with improving channel conditions.

This thesis investigates the video coding and joint source/channel coding issues that arose in designing the system. Spatially scalable video coding techniques were designed to match the multiresolution transmission method. In particular, we focused on layered approaches to spatial scalability. The shortcoming of conventional layered coding schemes is error propagation, i.e. lower-level coding errors propagate to higher levels, thereby decreasing the efficiency of the overall system. A novel method of conditioning is developed to control the error propagation problem.

Hybrid video coding methods were developed to match the hybrid transmission scheme. These methods were based on understanding the importance of the different types of video data that result from conventional video compression algorithms. It is shown that the high compression efficiency achievable with digital video compression algorithms are also achievable with hybrid transmission. Furthermore, in terrestrial broadcasting, improved performance can be achieved by exploiting the extra capacity available in central regions of the broadcast area.

1.2 Thesis Organization

In Chapter 2, basic video compression techniques are reviewed. The chapter focuses on transform coding and motion-compensated transform coding, because they are the most popular methods of compressing still images and video. This chapter also includes a brief discussion on some issues that arise when using these techniques in a video communication system.

The scalable video coding (SVC) problem is discussed in Chapter 3. In SVC, the video is encoded into multiple data streams of prioritized importance. This chapter concentrates on spatially scalable video coders in which the video is coded at multiple spatial resolutions. In particular, it considers layered approaches to spatially scalable video coding. A generalized framework is presented for layered coding, and a novel concept of conditioning is presented as a solution to the error propagation problem that afflicts layered coders. It is shown that the performance of layered MPEG and JPEG coders can be improved with simple conditioning techniques.

Chapter 4 examines the problem of video communication for terrestrial broadcasting. The idea of providing improved video quality to receivers with greater received signal strength is quantified with the concept of channel capacity. It is shown that a hybrid of analog and digital methods can be useful for video communication in the terrestrial broadcast environment. This chapter investigates how practical video coding techniques can be used in the context of hybrid transmission. In
particular, the performance of a hybrid transmission system is compared with the performance of a conventional all-digital transmission system with comparable complexity. It is shown that the hybrid system achieves comparable performance at the fringes of the broadcast area, and improved performance in the central regions of the broadcast area.

Chapter 5 describes the advanced television system developed for terrestrial broadcasting. It provides a general overview of the system, and focuses on the novel video coding algorithm that was developed based on the concepts introduced in the earlier chapters. The system is scalable in spatial resolution, and within each resolution it achieves graceful degradation by means of hybrid video coding. A complete computer simulation is used to demonstrate the feasibility of the system.

Concluding remarks and directions for future work are discussed in Chapter 6. A series of appendices discusses various topics that were addressed during the completion of this work.

1.3 Contributions

The major contributions of this thesis are summarized below.

Spatially scalable video coding.
- A generalized framework was developed for layered approaches to scalable video coding.
- A novel concept of conditioning was developed to control the error propagation problem inherent to layered systems.
- A solution was provided to the problem of layering JPEG and MPEG algorithms at non-octave subsampling ratios.

Hybrid transmission.
- Further insight was provided into the hybrid transmission problem.
- A practical hybrid video coding algorithm was developed for hybrid transmission.
- The feasibility of using motion-compensated prediction in hybrid transmission was demonstrated.
- A single-resolution hybrid system was compared to a single-resolution digital system with comparable complexity.

Advanced television system.
- A video coding algorithm was designed for the hybrid, multiresolution transmission system.
- An adaptive selection algorithm was designed to efficiently represent the location information of the selected subband coefficients.
- The feasibility of the system was demonstrated with a complete software simulation.
Chapter 2

Video Coding

Due to modern-day bandwidth constraints, efficient communication of video requires a high degree of compression. The basic video compression problem has been examined in depth for decades, and is a fairly well-understood topic [5, 6, 7, 8, 9, 10]. Currently, the most successful image and video compression algorithms are based on transform coding and motion-compensated transform coding, respectively. A large number of tutorials can be found on these subjects [11, 12, 13, 14]. For this reason, the topic is reviewed only briefly, and only in the context that will be useful for later chapters of this dissertation. For more in-depth coverage, the reader is referred to the references.

2.1 Still-Image Compression

Image coding is based on the principle that neighboring image pixels are very similar in nature and therefore compression can be achieved by exploiting this similarity. Of course, the degree of spatial similarity varies from image to image and depends on many factors such as spatial resolution. However, typical imagery generally contains objects that extend over many pixels, so it is natural for a good amount of spatial correlation to exist.

Transform coding is a widely used image compression technique that achieves compression by reducing the spatial redundancy inherent in typical imagery. A block diagram of a conventional transform coder is shown in Figure 2.1. The objective of the coder is to encode the image into a compressed digital data stream. First, an orthogonal transform is used to compress the signal energy into a small fraction of the transform coefficients. The coefficient amplitudes are then quantized into fewer discrete levels. In this process, many of the low-amplitude coefficients are set to zero. Finally, the locations and amplitudes of the remaining nonzero quantized coefficients are encoded into an appropriate data stream.
Figure 2.1: Conventional transform coding. Transform coding achieves compression by exploiting the spatial correlation inherent in typical imagery.

2.1.1 Transform/Subband Representation

The choice of transform has a large influence on the visual quality of the coded image. It is generally believed that an "ideal" transform would have high compression efficiency and low computational complexity while producing high-quality coded images. The compression efficiency should be high so that the image energy is concentrated into a small number of transform coefficients. The computational complexity should be low during encoding and decoding.\(^1\) Finally, the coded images should look good, and the distortion should be only barely visible to the viewer.

Throughout this work, transform and subband methods are used to achieve a more efficient representation of the image. In the classic transform approach, the image is divided into blocks, and an orthogonal transform is applied to each block. This can be visualized as a matrix multiplication where each column of the matrix corresponds to a basis function of the transform. In the subband approach, the image is filtered with each of the basis functions, and the result is subsampled in the horizontal and vertical dimensions. The basis functions are generally designed to produce a critically sampled representation, i.e. the resulting number of coefficients equals the number of pixels in the original image. In both representations, the reconstructed image is a superposition of basis functions whose locations and amplitudes are dictated by the coefficients.

The two representations can be viewed as simple rearrangements of one another. The transform representation orders the coefficients first by frequency, and then by spatial location. In other words, each transform block contains the frequency plane of a local spatial region in the image. Meanwhile, the subband representation orders the coefficients first by spatial location, and then by frequency. Each subband contains the spatial plane of a particular frequency band. Notice that if the same basis functions are used in the two approaches, the same coefficients result, but they are simply ordered differently. An in-depth discussion on this topic can be found in [13].

A transform coder's performance depends on the characteristics of its basis functions. The lowest-frequency basis functions of a block discrete cosine transform (DCT) and a wavelet transform are shown in Figure 2.2. The block DCT has non-overlapping basis functions. Therefore, neighboring blocks are coded independently of one another. This enables simple spatially adaptive processing. However, independent coding of neighboring blocks often produces the well-known "block-

\(^{1}\)In terrestrial television broadcasting, it is more important to have low computational and memory requirements in the decoder.
ing" artifacts that afflict DCT-based coders. The wavelet transform has overlapping basis functions. Thus, neighboring blocks are coded jointly, resulting in a smoother image. However, the longer spatial extent of the basis functions can lead to visual artifacts such as ringing. This artifact is especially evident in smooth areas near sharp edges.

![Basis functions](image)

Figure 2.2: *Basis functions.* Two-dimensional basis functions for an 8 × 8 block DCT and a three-level wavelet transform.

### 2.1.2 Quantization and Entropy Coding

Quantization is the only lossy operation in the transform coder. The quantizer stepsize varies with frequency to exploit the sensitivity of the human visual system. For example, since the human visual system is less sensitive to higher frequencies, the higher-frequency coefficients are quantized more coarsely than lower-frequency coefficients. Rate control is also accomplished by adjusting the stepsizes. For example, by increasing the quantization stepsize, more coefficients are set to zero and fewer discrete amplitude levels exist; this results in greater compression at the expense of lower coded-image quality.

Quantization achieves a good deal of compression by setting many coefficients to zero. The quantized coefficients are then coded into a digital data stream as shown in Figure 2.3. After quantization, the DC (lowest-frequency) coefficients are coded separately from the AC (non-DC) coefficients because of their different nature. Only the locations and amplitudes of the nonzero quantized AC coefficients are encoded into the data stream. Entropy coding techniques such as Huff-
man coding are used to exploit the statistical properties of the amplitude and location data.\textsuperscript{2} In this framework, the final data rate of the coded data stream is given by

\begin{equation}
R = R_{\text{DC}} + R_{\text{location}} + R_{\text{amplitude}}
\end{equation}

where \(R_{\text{DC}}, R_{\text{location}},\) and \(R_{\text{amplitude}}\) are the coded data rates of the DC, AC location, and AC amplitude information.

![Diagram showing the transform coding process](image)

Figure 2.3: Conventional transform coding – a more detailed view. In conventional transform coding, the transform coefficients are quantized. After quantization, the DC coefficients and the locations and amplitudes of the nonzero AC coefficients are encoded into digital data stream.

2.1.3 Nature of Transform Coding Artifacts

Transform coding is a lossy compression technique; thus, the coded image is not identical to the original image. High degrees of compression can be achieved with lossy (as opposed to lossless) techniques. However, since visual quality is the final metric for any video coding system, care must be taken to ensure that the resulting coding artifacts are barely visible.

Transform coding artifacts stem from the quantization operation, which makes the amplitudes of the coded coefficients different from those of the original coefficients. Because the reconstructed video is a superposition of the basis functions, errors in the coefficient amplitudes translate to additive noise in the reconstructed image. However, the additive noise is not white; rather, it has a structure that results from the characteristics of the basis functions.

If the same noise level was found in all the coefficients, the resulting noise would appear “white.” However, this is seldom the case in transform coding. Rather, the quantization noise often has a nonuniform distribution in the frequency (and spatial) domain as discussed in Appendix C. This is due to the nonuniform frequency distribution of the signal energy and the nonlinear quantization operation \textsuperscript{[7]}. This results in a structured distortion that highly depends on the characteristics of the basis functions.

\textsuperscript{2}Many methods jointly code the amplitude and location data. However, since little correlation exists between these components (in particular for interframe coding), throughout this work we will assume that they are coded separately.
It is interesting to compare the distortions that afflict DCT- and wavelet-based coders. At high data rates, both coders perform quite well and the reconstructed images are nearly indistinguishable. However, at lower data rates, the effects of the basis functions become apparent. In smooth areas of the image, the wavelet-coded images perform quite well while the DCT-coded images appear blocky. This is due to the overlapping vs. nonoverlapping characteristics of the basis functions.

In smooth regions near sharp edges, the DCT- and wavelet-based coders are afflicted by "mosquito" noise and "ringing" artifacts. Since a sharp edge contains energy throughout the frequency bands, all the corresponding coefficients suffer from quantization noise. This noise appears in a region surrounding the sharp edge, the spatial extent of the noisy region corresponds to the spatial extent of the basis functions. Meanwhile, in the neighboring smooth region, only the low-frequency component is coded and a very smooth reconstruction results. The visual effect of this is a type of busyness that surrounds the sharp edges. Since the DCT basis functions are all the same size, the busy noise has a distinct boundary and is called mosquito noise. Since the wavelet basis functions have different sizes, the boundaries do not line up. The lower-frequency basis functions extend farthest into the blank areas, and lead to the ringing artifact.

### 2.1.4 Remarks

This leads to a number of interesting issues. First, compression efficiency was claimed to be a desirable property of the "ideal" transform. When designing a practical video coding system, consideration should be given to the overall system performance. For example, if a less-efficient transform is used, the resulting representation may contain more redundancy than the efficient representation. However, if the entropy coder is capable of removing this redundancy, the final system performance is not harmed in any way. A further consideration is that the resulting transform coding artifacts will have properties of the basis functions. If the communication channel causes coefficients to be dropped, the resulting artifact will resemble the characteristics of the basis functions. In this case, it may be advantageous to have a less efficient transform with better visual properties.\(^3\)

Another issue deals with the transform and subband representations of an image. While these representations are simple rearrangements of one another, practical considerations may make one representation preferable over the other. This becomes apparent when coding the locations of the nonzero quantized coefficients. In DCT-based coders, the locations of these coefficients are typically coded by zigzag scanning each transform block. The lengths of the "runs" between nonzero coefficients are then entropy coded into a digital data stream. Further coding gain is achieved by transmitting an end-of-block marker after the last nonzero coefficient.

\(^3\)This relates to the separate vs. joint source and channel coding issues discussed in Chapter 4.
The wavelet transform is typically viewed in subband form. One method of coding the locations involves scanning the subbands in a raster fashion. Since each subband is a spatial plane, neighboring coefficients are often correlated. Specifically, in subband form, many groups of coefficients are set to zero during quantization. The subband representation provides a natural framework for exploiting this spatial correlation.

2.2 Video Compression

A video sequence consists of a series of images. Typically, these images are not independent. Rather, they often exhibit a high degree of temporal correlation, i.e. consecutive frames of video are often quite similar. In fact, high temporal correlation is required to produce the illusion of continuous motion. This temporal correlation can be exploited to achieve very high compression rates.

Video compression algorithms can be divided into two classes: intraframe and interframe coding methods. Intraframe methods process each frame of video independently, thereby achieving compression by exploiting only the spatial correlation that exists in each frame just as in still-image coding. Interframe methods process multiple frames of video together, and achieve higher compression rates by exploiting the temporal correlation that exists between frames.

2.2.1 Intraframe Methods

Intraframe coding methods process each frame of video independently. Each frame is considered to be an image, and standard transform coding methods are used to encode each frame. The lower computational complexity and memory requirements make intraframe coding an attractive option. However, for applications that require higher degrees of compression, interframe coding methods may be a more appropriate choice.

2.2.2 Interframe Methods

Interframe coding methods achieve higher compression by exploiting the temporal correlation that exists between video frames. The most commonly used interframe coding method is motion-compensated transform coding. In motion-compensated transform coding, the temporal redundancy is first reduced with predictive methods applied along the temporal dimension. The remaining spatial redundancy is then reduced with conventional transform coding methods.

A block diagram of a conventional motion-compensated transform coder is shown in Figure 2.4. When coding the first frame, there is no previous information to form a prediction. Therefore, it is coded independently with a conventional transform coder. This is called the intra-coded frame. Then, the motion is estimated between the first and second frames so that a prediction can
be made of the second frame. This motion is represented by a set of motion vectors that are coded into a digital data stream. The motion vectors are used to create a motion-compensated prediction of the current frame based on the previously coded frame. The prediction error, called the motion-compensated residual, is then coded with the transform coder. The motion-compensated residual has a very different character from that of a normal image; it more resembles a difference between two images. For this reason, different quantization and entropy coding tables are used in the transform coder. However, the general method remains the same.

Figure 2.4: Motion-compensated transform coding. A high degree of compression is achieved by exploiting the temporal correlation inherent in typical video.

The most commonly used motion-estimation technique is block matching. In block matching, the goal is to find the \( n \times n \) block of the previous frame that best matches the \( n \times n \) block of the current frame. The motion or displacement between the two blocks is represented by the motion vector. The best match is determined by a criterion such as the mean absolute difference or mean squared error between blocks. At HDTV resolutions, one typically assigns one motion vector for each \( 16 \times 16 \) block. Of course, the best block size depends on many factors including the resolution and the amount of motion in the scene.

In summary, a common form of motion-compensated prediction involves the following steps. First, divide the image into blocks. Then, calculate a motion vector for each block based on the previous frame. Next, form a prediction of the current frame based on the motion vectors and the previously coded frames. Finally, code the resulting motion-compensated residual with a conventional transform coder. Motion-compensated prediction was applied to the two successive frames of video shown in Figure 2.5. The resulting motion-compensated prediction and residual are shown
in Figure 2.6. This method produces three elements that must be coded into the data stream: the motion vectors, the locations of nonzero residual coefficients, and the amplitudes of nonzero residual coefficients. With this information, the decoder can completely reconstruct the coded video.

Figure 2.5: Successive frames of original video. Each frame contains 256 × 512 pixels.

2.2.3 Remarks

A number of interesting points should be noted. First, the reconstructed video is composed of two parts: the motion-compensated prediction and the motion-compensated residual. Typically, the prediction contains a large fraction of the energy. Therefore, it is important to reconstruct the prediction accurately at the decoder. A mismatch in the predictions formed in the encoder and decoder causes a drift or loss of synchronization in the reconstructed video. In other words, the motion compensation performed in the encoder should be based only on information to which the decoder
Figure 2.6: Motion-compensated prediction and residual. The MC prediction and residual were computed for the two video frames shown in Figure 2.5. Standard block-matching techniques were used with 16 × 16 pixel non-overlapping windows.

has access. For this reason, the motion-compensated prediction is based on the previously coded frame, rather than on the original.

A separate issue lies in deciding whether the motion estimation should be performed with the coded or original version of the previous frame. Since this motion will be transmitted with a set of motion vectors, this does not cause any mismatch in the encoder and decoder. If the final quality metric is a numerical error measure, then using the previously coded frame should yield the best results since it chooses the vector that minimizes the numerical distortion based on information the decoder has. However, if visual quality is the final measure, the previous original frame may provide a more accurate estimate of the true motion, which may result in superior visual quality.

In block-based motion-compensated transform coding, the motion-compensated prediction con-
tains artificial discontinuities at the block boundaries. This is especially evident in high-motion areas. These artificial discontinuities propagate to the motion-compensated residual, and if these discontinuities are not properly coded, then blocking artifacts will also appear in the reconstructed video. One method of smoothing out the prediction and thus smoothing out the residual involves performing motion compensation with overlapped windows as opposed to the non-overlapped windows formed by the $n \times n$ blocks.

When using a DCT-based transform coder for the residual, the basis functions line up with the block boundaries of the motion-compensated prediction. However, if using a wavelet-based transform coder, the basis functions overlap the block boundaries, thus throwing extraneous energy into the transform coefficients. It was found that when using a wavelet-based transform coder, it is advantageous to use overlapped windows for motion compensation [15].

2.3 Summary

This chapter provided a brief review of still-image and video coding algorithms that are based on transform coding and motion-compensated transform coding methods.

Video coding algorithms can be classified into intraframe and interframe coding methods. Intraframe methods code each video frame independently, and therefore have lower computational and memory requirements than interframe methods. However, the interframe coding methods are capable of achieving higher compression at a given quality level by exploiting the temporal redundancy inherent in video.

If compression efficiency was the only concern, it is obvious that interframe methods are a better choice. However, in a practical video communication system it may be necessary to consider other factors, such as computational and storage requirements, delay requirements, random access capability, VCR-type functionalities, and effects of transmission errors [16].
Chapter 3

Scalable Video Coding

3.1 Introduction

The goal of scalable video coding is to efficiently encode the original video into multiple data streams of prioritized importance. In this chapter, we consider scalable methods that encode the video into a base data stream and one or more enhancement data streams as shown in Figure 3.1. The base data stream can be decoded into baseline-quality video; additional enhancement data streams can also be used to improve the decoded video quality.\(^1\) In scalable video coding, the number of decoded data streams determines the quality of the reconstructed video.

The improvement in video quality, or the type of scalability, can take on a number of forms. For example, the scalable video coder can provide improvements in the spatial, temporal, and/or amplitude resolution of the video. As another example, objects in the video can be coded in order of importance. While our results apply to many forms of scalability, we specifically address the problem of spatial scalability. In a spatially scalable system, the base data stream is decoded to form low-resolution video, and additional enhancement data streams are decoded to improve the spatial resolution. Sample frames of low-, medium-, and high-resolution video sequences are shown in Figure 3.2.

Throughout this work, we keep in mind that the final measure of a video coding system is the visual quality of the reconstructed video. The decoded video must look “good” at every resolution. Our demonstrations revealed that viewers prefer to view the lower-resolution video at full size. The full-size versions of the low, medium, and high resolution images are shown in Figure 3.3.

This chapter begins by providing an introduction to spatially scalable coding schemes for still images and video. A brief sampling of conventional approaches is described. Then, a generalized framework is introduced for layered approaches to scalable video coding, and a novel concept of

\(^1\)It is interesting to note that since the coded data streams are prioritized according to importance, scalable video coding inherently supports a form of joint source/channel coding.
conditioning is presented to combat the artifact propagation problem inherent to layered coders [17]. Finally, experimental results are shown to demonstrate the potential benefits of conditioning in spatially scalable still-image and video coding. In particular, conditioning provides a solution to the problem of layering JPEG and MPEG algorithms at non-octave subsampling ratios.

3.2 Scalable Coding Methods

3.2.1 Scalable Coding of Still Images

In the scalable coding of still images, an image is coded into prioritized data streams. A low-resolution image can be decoded from the base data stream, and higher-resolution images can be decoded from base and enhancement data streams. Common methods of spatially scalable coding are based on critically sampled and oversampled representations of the image. This section provides a description of these two approaches.
Figure 3.2: Spatial scalability. Sample frames of the “Girl” sequence are shown at low $(1/2 \times 1/2)$, medium $(3/4 \times 3/4)$, and high $(1 \times 1)$ resolutions.
Figure 3.3: Spatial scalability: Viewing full-size video. Viewers often prefer to view the lower-resolution video at full size. Sample frames of video are shown at low $(1/2 \times 1/2)$, medium $(3/4 \times 3/4)$, and high $(1 \times 1)$ resolutions.
Critically Sampled Representations

Subband coding is a very effective method of coding images [18, 19, 20, 21, 22, 23]. A block diagram of a one-dimensional subband coder is shown in Figure 3.4. The analysis $F$ and synthesis $G$ filterbanks are typically designed to satisfy a number of requirements. For example, the filterbanks are usually designed to be orthogonal and yield perfect reconstruction in the case of lossless coding. Often, the filterbanks are also designed to maximize compression efficiency and yield a critically sampled representation. The latter implies that during analysis, the filtered signals are subsampled to yield a representation that contains the same number of samples as the original signal. All these requirements impose stringent constraints on filter design, and leave few degrees of freedom for considering other factors. Most importantly, it is difficult to design these filters to behave well with respect to the human visual system.

![Block diagram of a one-dimensional subband coder](image)

**Figure 3.4: One-dimensional subband coding.** The analysis $F$ and synthesis $G$ filterbanks are typically designed to be orthogonal, maximize compression efficiency, and yield perfect reconstruction in the case of lossless coding.

Wavelet coding is a specific case of subband coding in which the subbands are non-uniform in size [24, 25, 26, 5]. The higher-frequency subbands are typically larger than the lower-frequency subbands. This results in finer frequency resolution (smaller subbands) at lower frequencies, and finer spatial resolution (narrower basis functions) at higher frequencies. Figure 3.5 shows the wavelet representation of an image for an octave-band decomposition, meaning the subbands increase in size by factors of two. The horizontal and vertical axes represent increasing horizontal and vertical spatial frequencies. The upper-left subband contains the lowest horizontal and vertical frequency components.

The subband decomposition of an image suggests an obvious method of achieving spatial scalability – group the subbands by frequency and code each group into a data stream. In this method, the subbands in group A are coded into the base data stream, the subbands in group B are coded into
Figure 3.5: Octave-band wavelet decomposition. The higher-frequency subbands are larger than the lower-frequency subbands. This results in finer spatial frequency resolution at lower frequencies, and finer spatial resolution at higher frequencies.

the first enhancement data stream; and so on. Lower-resolution images can then be reconstructed by decoding the base and an appropriate number of enhancement data streams. This method is efficient in the sense that, if the appropriate filterbank is used, the resulting representation is critically sampled so that the number of subband coefficients equals the number of image pixels. However, partial reconstruction with the prescribed synthesis filters often yields images with visually annoying artifacts due to uncanceled aliasing.

The “Girl” image was decomposed with the wavelet transform. The images were partially reconstructed with the subbands in group A; groups A and B; and groups A, B, and C as shown in Figure 3.6. The reconstructed images are shown in Figure 3.7. The resulting images have resolutions of $1/8 \times 1/8$, $1/4 \times 1/4$, and $1/2 \times 1/2$ the horizontal and vertical resolution of the original image. Notice that the set of possible spatial resolutions is dictated by the structure of the decomposition, and in conventional wavelet transforms the resolutions change by constant factors.

The partially reconstructed images are afflicted with artifacts that result from uncanceled aliasing. For example, consider the subband structure shown in Figure 3.4. If appropriate bandwidth, ideal brickwall filters are used, the subsampling would not produce aliasing. However, in image and video coding it is neither practical nor desirable to use ideal brickwall filters. Thus, the resulting subsampled signals do contain some amount of aliasing. In these cases, the filterbanks are designed such that in the case of full reconstruction and lossless coding, the aliasing in the subbands cancel out. However, during partial reconstruction the aliasing is not cancelled and leads to visual artifacts in the lower-resolution images. Notice that the uncanceled aliasing produces ripple artifacts in the lower-resolution images shown in Figure 3.7 [27, 28].
Figure 3.6: Partial reconstruction of wavelet subbands. Spatial scalability can be achieved by grouping the subbands by frequency and coding each group into the base and enhancement data streams. The images that result from partially reconstructing the subbands are shown in Figure 3.7.

Oversampled Representations

The recursive pyramid structure developed by Burt and Adelson provides another approach to achieving spatial scalability [29]. The architecture of the octave-band pyramid coder is shown in Figure 3.8. The image is recursively lowpass filtered and subsampled (resampled) to the lowest level of the pyramid. In the lowest level of the pyramid, the final lowest-resolution image is coded into the base data stream. In each higher level, a difference image is coded into an enhancement data stream. The low-resolution images can be reconstructed by decoding the base and the appropriate enhancement data streams.

A slight generalization of the recursive pyramid leads to the layered architecture shown in Figure 3.9. This structure also has two other major differences. First, the layered structure is not recursive. Rather, in each level the image is upsampled to full size and subtracted from the input. While this may be computationally inefficient, it provides added flexibility and further insight in the oversampled representation. Next, the coding error is incorporated into the feedback loop. This allows the low-frequency components that were not coded in the lower levels to be coded in the higher levels.

The pyramid representation that results from both the recursive and layered architectures is shown in Figure 3.10. The lowest band contains the lowest-frequency components. Each higher band contains higher-frequency components, in addition to the lower-frequency components that were not coded in the lower levels. The primary shortcoming of this representation is that it is oversampled or overcomplete, i.e. the number of samples exceeds the number of pixels in the original image. On the other hand, it is this property that provides extra degrees of freedom in design-
Figure 3.7: Partially reconstructed images. Octave-band wavelet transform at $1/8 \times 1/8, 1/4 \times 1/4$, and $1/2 \cdot 1/2$ resolutions.
Figure 3.8: *Octave-band recursive pyramid*. Burt and Adelson’s recursive pyramid produces natural-looking images at every resolution. The Resample 2:1 is a lowpass filter and subsample operation and the Resample 1:2 is an upsample and lowpass filter operation.

...ing the filters. These extra degrees of freedom can be used to design lowpass filters that produce natural-looking images at every resolution.

For example, with this structure, virtually any lowpass filter can be used in each level of the pyramid while still permitting perfect reconstruction. The lower-resolution images can then be obtained by decoding the the appropriate images as shown in Figure 3.11. In this scheme, it is the lowpass filters that determine the quality of each lower-resolution image. For example, sharpened Gaussian filters were used to generate the images shown in Figure 3.3. These filters can be incorporated into the pyramid structure quite easily. They can also be designed to produce any desired resolution.

Remarks

In still-image coding, spatial scalability can be achieved with either a critically sampled or an over-sampled representation. Each representation has its advantages and disadvantages, and it is not clear which is the better choice. A seemingly important consideration is the number of coefficients that must be coded in the two representations\(^2\). However, the true measure is the visual quality of the reconstructed images as it relates to the total transmission rate.

The critically sampled representation contains fewer coefficients, but stringent constraints on filter design leave few degrees of freedom to account for the visual performance of the filters. The overcomplete structure may be less efficient due to oversampling, but the extra degrees of freedom...

\(^2\)A better comparison considers the data rates that result from coding the two representations with practical methods. While the pyramid structure has more samples, these samples also contain more redundancy that can be exploited during coding.
can be used to design filters that address the visual quality of the reconstructed images. Notice that the octave-band wavelet representation is amenable to spatial resolutions at 1/8, 1/4, and 1/2 the horizontal and vertical resolutions of the original image. The recursive nature of the pyramid structure allows greater flexibility in filter design and facilitates a wide range of spatial resolutions.

It is important to recognize that the visual quality of an image largely depends on the preferences of the individual viewers. For example, some viewers would argue that the visual artifacts in the $1/2 \times 1/2$ resolution image of Figure 3.7 are not very noticeable, and that the increase in sharpness might make it preferable to the $1/2 \times 1/2$ resolution image of Figure 3.3. Meanwhile, other viewers would object to the staircase effects and therefore prefer the smoother image.

In this section, critically sampled and oversampled representations have been considered for spatially scalable coding of still images. Each has its own advantages and disadvantages, but it is unclear as to which scheme performs “best”. For this reason, when designing a spatially scalable image coder, the tradeoffs of the two representations should be understood and the final choice should be based on the requirements of the particular application.
Figure 3.10: *Octave-band pyramid representation*. The smallest band contains the lowest-frequency components. Each larger band contains additional higher-frequency components as well as the lower-frequency components that were not coded in the lower levels.

### 3.2.2 Scalable Coding of Video

In the previous section, we investigated the tradeoffs between using critically sampled and over-sampled representations for spatially scalable coding of still images. In this section, we discuss the possibility of using these representations for spatially scalable coding of video. In spatially scalable video coding, the original video is encoded into multiple data streams. The base data stream can be decoded into low-resolution video, and enhancement data streams can be decoded to improve the spatial resolution.

In a typical video scene, successive frames of video are often very similar. These similarities can be exploited to achieve high degrees of compression. Motion-compensated prediction (MCP) is a practical, highly effective method of reducing the temporal correlation that exists in video. It is commonly used in video compression algorithms, so it is natural to consider this technique for scalable video coding. A brief description of MCP can be found in section 2.2.2.

The fundamental difficulty in using MCP in spatially scalable video coding lies in the fact that it is a highly nonlinear operation. The resulting motion-compensated residual often contains energy spread throughout the frequency domain. For this reason, an efficient method of incorporating MCP in a critically sampled scalable system is not obvious.

This can be better understood by considering the wavelet representation labeled in Figure 3.5. A video sequence can be thought of as a series of individual video frames. If MCP is not used, the wavelet coefficients can be calculated for each frame, and spatial scalability can be achieved simply by coding each group of subbands into the appropriate data stream. When using MCP, the intra-frames can be treated in the same manner. However, for the inter-frames, the residual wavelet coefficients needed to reconstruct each lower resolution of video are no longer localized to the appropriate subbands. This makes it difficult to incorporate MCP in a critically sampled
Figure 3.11: Partial reconstruction of pyramid subbands. Spatial scalability can be achieved by coding each subband into a data stream. The lower-resolution images that result (using different relative bandwidths) are shown in Figure 3.3.

The oversampled nature of the pyramid structure allows MCP to be incorporated quite easily. In the lowest level of the pyramid, the lowest band of all the video frames can be coded with an MCP coder. The reconstructed video is subtracted from the original, and the residual is coded in the next level, where a higher-bandwidth lowpass filter is used to extract the larger band from all the video frames. These again can be coded with an MCP coder. This process can be repeated for any number of levels. With the oversampled structure, the frequency components of the motion-compensated residual are localized within the oversampled bands. In this way, the oversampled representation is well suited for scalable video coding with MCP.
3.3 Conventional Approaches

In recent years, there has been considerable research on spatially scalable still-image and video coding. An overview of these methods can be found in [14]. Many spatially scalable coding algorithms have been proposed; a brief sampling of these approaches intended to reflect the state of the art is described below.

3.3.1 Laplacian Pyramid

Burt and Adelson [29] developed the Laplacian pyramid coder, which decomposes an image into a series of bandpass images. One or more of these images can be used to reconstruct images of increasing resolution. The spatial decomposition is achieved by iteratively lowpass filtering the image with a Gaussian filter and subsampling the result as shown in Figure 3.8. In the lowest level of the pyramid, the lowest-resolution image is coded into a data stream. In each higher level of the pyramid, the interpolated, lower-resolution image is subtracted from the higher-resolution image, and the resulting Laplacian image (difference of two Gaussian images) is coded into an enhancement data stream. The lowest-level data stream can be used to reconstruct a low-resolution image; additional higher-level data streams can be used to reconstruct higher-resolution images.

3.3.2 Embedded Wavelet Coder

Shapiro developed the embedded zerotree wavelet (EZW) coder which encodes an image into a completely embedded data stream [30]. This data stream can be decoded up to any point, and the resulting image quality corresponds to the length of the decoded data stream. The EZW image coder is based on the wavelet transform. The image is first transformed into its wavelet representation. The wavelet data is then classified according to its precision (e.g. bit-plane encoding), amplitude, frequency band, and spatial location. This data is ordered and coded into the data stream in multiple sweeps. Adaptive arithmetic coding is used to achieve high compression that adapts to the local characteristics of the data.

3.3.3 MPEG-2

Extensive work has been done in the area of single-quality video compression. This work resulted in the development of a widely used video compression standard by the Moving Pictures Expert Group (MPEG). The MPEG standard uses block-based motion-compensated prediction and block-DCT transform coding to reduce the temporal and spatial redundancies inherent in typical video [12].
In MPEG, frames of video are either intra-frame or inter-frame coded. For intra-coded frames, the actual video frame is transform coded. For inter-coded frames, a prediction is made by motion compensating the previously coded frames, and the resulting prediction error, or residual, is transform coded. For both types of frames, transform coding involves quantizing and entropy coding the DCT transform coefficients.

MPEG-2 is the video compression standard that resulted from the second phase of activity of the Moving Pictures Expert Group [31]. The MPEG-2 standard incorporated spatial, temporal, and SNR scalable modes.

The SNR-scalable mode operates at a single spatial and temporal resolution. A base data stream results from coding the video with a typical MPEG coder. An enhancement data stream contains the refinement information that results from more finely quantizing the DCT transform coefficients. It should be noted that the motion compensation is based on the more finely quantized coefficient amplitudes. Therefore, the signal that is reconstructed with only the base data stream suffers from drift in the motion-compensation loop. In other words, the motion-compensated prediction formed in the decoder does not precisely match the prediction in the encoder.

The spatially scalable mode resembles the recursive pyramid architecture discussed in section 3.2.1. The original video is first lowpass filtered and subsampled to produce lower-resolution video that is coded with a typical MPEG coder. A prediction of the full-resolution video is formed with a method called spatio-temporal weighting. Basically, the decoded low-resolution video frame is interpolated to form a spatial prediction of the full-resolution frame, and the previously coded full-resolution video frames are motion compensated to form a temporal prediction. The final prediction is formed by weighting the spatial and temporal predictions on a locally adaptive basis.

Temporal scalability is achieved by demultiplexing successive frames of video. One group of frames is coded in the base data stream with a typical MPEG coder. The remaining frames are coded in the higher level with an MPEG-like coder. The enhancement coder can use the information in the previously coded frames to better code the remaining frames. It should be noted that MPEG inherently supports temporal scalability by its use of intra (I), predicted (P), and bidirectional (B) frames during motion-compensated coding.

3.3.4 Spatio-Temporal Pyramid

Uz, Vetterli, and LeGall developed a scalable video coder based on a spatio-temporal pyramid [32]. In the lowest level, alternate low-resolution frames of low-resolution video are coded into a base data stream. In higher levels, an enhancement signal is coded to increase the temporal and spatial resolutions. The spatial decomposition is achieved with a pyramid method similar to Burt and Adelson's Laplacian pyramid, but the coding errors are fed back in the prediction loop. The temporal decomposition is achieved by skipping frames in the lower levels of the pyramid. The coded low-
resolution video frames are upsampled and filtered (spatially interpolated) to form a prediction of the medium-resolution frames; these frames are used to form prediction of the skipped frames by temporal interpolation. The DCT transform coefficients of the spatio-temporal residuals are coded into digital data streams for transmission.

Ramchandran, Ortega, Uz, and Vetterli incorporated this scalable video coder into a digital, multiresolution HDTV system for terrestrial broadcasting [33]. The system uses joint coding methods to match the multiresolution source and channel coders. This system achieves more efficient spectrum usage by providing a stepwise-graceful degradation [34].

3.4 A Generalized Framework for SVC

3.4.1 Overview

The single-quality video coding problem has been examined for decades and is a fairly well-understood topic. However, additional issues arise when using these techniques in a scalable system. For example, the design of a single-quality system is typically motivated by the desire to minimize the perceived distortion of the coded video. However, in a layered scalable system, the coded video is not only viewed, but is also used to form a prediction for the higher levels of video. Therefore, any coding artifacts introduced in the lower levels, even if not visible, propagate to the higher levels, reducing the coding efficiency of the entire system.

In this section, we introduce the concept of conditioning, which provides a solution to the artifact propagation problem that plagues layered coders [17]. In each level of the scalable video coder, conditioning controls the propagation of artifacts into subsequent levels of the coder, thereby improving the coding efficiency of the entire system. Basically, the video entering and leaving each level of the scalable system is conditioned to remove the artifacts that are generated in lower-level coders. One could say that conditioning the lower-level signals makes the resulting higher-level signals “easier to code”.

We begin by presenting a generalized framework for layered approaches to scalable video coding. The concept of conditioning is then described, and some examples are given to illustrate its effectiveness in the context of spatially scalable still-image and video coding systems. In particular, conditioning is used to solve the problem of layering JPEG and MPEG schemes at non-octave subsampling ratios.

3.4.2 Generalized Framework

In this section, we introduce a generalized framework for scalable video coding. The framework encompasses conventional approaches to layered coding and accommodates an improved class of
coders. The improvement is due to the incorporation of pre- and post-conditioning within each level of the scalable video coder. The concept of conditioning is further discussed in the next section.

The generalized framework is shown in Figure 3.12. The overall architecture resembles a pyramid in that the video is coded in multiple levels. In the lowest level of the pyramid, baseline-quality video is coded into an appropriate base data stream. In each higher level of the pyramid, an enhancement signal is coded into an appropriate enhancement data stream. Notice that the video reconstructed in each level of the coder is subtracted from the original to form the input signal of the next higher level. This can be thought of as using the lower-level reconstructed video as a prediction for the next level; only the error in the prediction, or the residual, must be coded in the next level.

Figure 3.12: A *generalized framework for scalable video coding*. This framework encompasses current approaches and accommodates an improved class of layered scalable video coders. A novel aspect lies in the use of pre- and post-conditioning within each level of the coder. The decoder, post-conditioner, and adder in the highest level are shown for illustrative purpose; they are used only in the receiver and would not be needed in the actual encoder.

In each level of the scalable video coder, the signal is processed with a codec comprised of a pre-conditioner, encoder, decoder, and post-conditioner. The pre-conditioner extracts the signal that will be coded in the particular level. In addition, it further conditions this signal so that it is easier to code. The encoder codes the signal into an appropriate data stream that meets the requirements of the system. The decoder processes the data stream to reconstruct the same signal decoded at the
receiver. Finally, the post-conditioner processes the signal to form an appropriate prediction of the original video. In addition, it should further condition the video so that the input signals in the higher levels of the pyramid are easier to code.

This framework encompasses current approaches to layered coding. For example, in a spatially scalable video coder each pre-conditioner performs the lowpass filtering and subsampling operations, and each post-conditioner performs the upsampling and lowpass filtering operations. Narrow-bandwidth lowpass filters are used in the lowest level, and higher-bandwidth lowpass filters are used in each higher level.

At each level the fully decoded video may also be conditioned as shown in Figure 3.13. Note that this automatically occurs in the base level, but not in the higher levels where only the enhancement signal is conditioned. The opportunity to condition the completely decoded video is highly advantageous because the completely decoded video should resemble natural video. The characteristics of natural video may be better understood than the characteristics of the enhancement signal. For example, the enhancement signal reconstructed at each level may have characteristics which seem unusual when viewed alone, but seem normal when viewed in conjunction with the lower-level video. In some instances, it may be advantageous to condition the enhancement signal, and in other instances it may be advantageous to condition the fully decoded video.

3.4.3 Conditioning

Single-quality video coding is a fairly well-understood topic. However, careful consideration should be given to issues unique to scalable video coding, thereby potentially leading to significant improvements in performance. In this section, we introduce the concept of conditioning, which can lead to a number of benefits in a scalable system. Conditioning can be used to control of the artifact propagation problem that afflicts layered coders. We describe the motivation for conditioning and in the next section we illustrate its usefulness by incorporating it in two conventional scalable systems.

Maximum performance in a scalable system requires achieving high performance in each level of the coder. The interdependencies between layers makes this a difficult task. In a single-quality system, the coder must produce video that looks good. In a scalable system, the lower-level video signals must both be viewed and used to form a prediction for the higher levels. The video that is best suited for each of these purposes is not necessarily the same.

Intuitively, to achieve maximum efficiency in a layered coder, a good prediction must be formed for each higher level. More specifically, the reconstructed lower-level video will typically be afflicted with coding artifacts that are not true features of the original video. If these artifacts are not explicitly addressed, the prediction formed for the next level will also be afflicted by these artifacts. The coder in the next level would therefore have to expend capacity to cancel out these artifacts.
Figure 3.13: An improved generalized framework for scalable video coding. The fully decoded video can also be conditioned in each level of the pyramid.

This effect can occur at each level of the pyramid, leading to a significant reduction in performance. Improved performance can be achieved by removing these artifacts before they propagate to the higher levels.

This motivates the concept of conditioning in each level of the scalable video coder. The main idea is to pre-condition the video before coding to make it easier to code, and to post-condition the video after coding to prevent the propagation of artifacts. The system goal is to allocate the available capacity efficiently to the various elements that must be coded.

The function of the pre- and post-conditioners in a scalable video coder differs from the function of the pre- and post-processors in a single-quality system. In single-quality systems, the immediate goal of the pre- and post-processors is to enhance the visual quality of the output video. In scalable systems, the immediate goal of the pre- and post-conditioners is to condition the video so that the resulting signals in subsequent levels are easier to code. In this way, higher coding efficiency can be achieved for the overall system and improved video quality can be reconstructed at each receiver.
3.4.4 Example: Layered JPEG/MPEG

In this section, we use two examples to illustrate the potential benefits of conditioning. We consider two conventional spatially scalable coders, one for still-image coding and one for video coding. In both cases we employ a simple conditioning technique to improve the performance of the coders.

The two systems share the same basic two-level architecture shown in Figure 3.14. In the low level, the pre-conditioner extracts the low-resolution image/video by lowpass filtering and subsampling the image by a factor of $3/4 \times 3/4$ in the horizontal and vertical spatial dimensions. A JPEG/MPEG coder compresses the image/video into a base data stream. The base data stream is decoded to reconstruct the image/video that would be decoded at the receiver. The post-conditioner upsamples and filters this image/video to form a prediction of the original input. This prediction is subtracted from the original image/video, and the resulting residual signal is coded in the higher level.

Each of these systems is improved by incorporating a post-conditioner in the lower level of the coder. In both cases, a simple algorithm is used to reduce the blocking artifacts that afflict JPEG and MPEG coded image and video. Note that the conditioning algorithm must be incorporated in both the encoder and the decoder. Therefore, it is advantageous to use computationally simple conditioning techniques.

Figure 3.14: Conditioning example: Two-layer MPEG, with and without low-level conditioning.
Still-Image Coder

In the first example, an intra-frame, $8 \times 8$ block-DCT coder was employed in each level of the spatially scalable system. The difficulty stems from the artifacts in the low-resolution coded image. Since this coded image is used to form a prediction for the higher level, its artifacts propagate to the higher level. These artifacts must be cancelled to achieve an acceptable high-resolution output image; this cancellation requires the use of extra channel capacity.

In typical spatially scalable systems, the resolution increments are restricted to octaves, or factors of two. In this case, the DCT block boundaries of the upsampled, low-resolution image line up with the DCT block boundaries in the high-resolution codec. However, factor-of-two resolution increments are highly restrictive for scalable video coders. Therefore, we investigate the use of non-octave subsampling ratios.

The $3/4 \times 3/4$ (non-octave) subsampling ratio causes additional difficulties because the block boundaries of the two levels do not coincide. Block discontinuities that appear in the upsampled decoded low-resolution image cause similar discontinuities in the high-level residual. Since these discontinuities do not coincide with the high-level block boundaries, they throw excess energy into the high-level DCT coefficients, making the system inefficient. This problem can be alleviated with conditioning. Specifically, by post-conditioning the decoded low-resolution image to reduce the blocking artifacts, a smoother prediction can be formed for the high-level coder. This reduces the extraneous discontinuities in the higher-level residual, and thereby improves the efficiency of the overall system.

To illustrate the performance of conditioning, the $512 \times 512$ Lena image was first processed with the conventional scalable system. The reconstructed high-resolution image had a PSNR of 34.15 dB. Conditioning was applied to the decoded low-resolution image by adaptively lowpass filtering the boundary pixels across the block boundary [35]. A threshold was used to distinguish between block discontinuities and true edges. If the amplitude difference across the block boundary was below a set threshold, then a simple 3-tap lowpass filter was applied to those boundary pixels. The resulting reconstructed high-resolution image had a PSNR of 34.70 dB, achieving an increase of 0.55 dB. Furthermore, the improvement in the visual quality was quite noticeable.

Video Coder

In the second example, we considered a two-level scalable video coder which contained an MPEG coder in each level of the pyramid. Once again, coding artifacts in the coded low-resolution video propagate to the higher level, causing extraneous energy in the residual signal that must be coded in the higher level. This decreases the efficiency of the overall system. Also, because of the $3/4 \times 3/4$ subsampling ratio, the low-level MPEG block boundaries do not coincide with the high-level
MPEG block boundaries, causing further difficulty in coding.

A conventional approach to improving this system would be to post-process the fully decoded video prior to display. Our proposed approach involves post-conditioning the video in each level of the scalable video coder, prior to reconstructing the fully decoded video. In this experiment, we compare the performance of the two approaches to the performance of the conventional system without post-processing and without post-conditioning.

For illustrative purposes, a simple method of reducing blocking artifacts was used for the post-processing and post-conditioning operations. The method involved adaptively filtering the pixels located on the 8×8 block boundary. If the amplitude difference across the block boundary exceeded a predetermined threshold, a simple 3-tap filter (2,5,2) was applied across the block boundary. Neither the filter coefficients nor the threshold was optimized for the particular video sequence.

The HDTV Kodak balloon sequence was processed with three approaches to scalable video coding: the conventional approach, the conventional approach with post-processing, and the proposed approach with post-conditioning in each level of the coder. In all cases, the video was compressed to a total bit rate of .35 bits/pixel, of which 9/16 of the data was allotted to the low-level data stream and 7/16 of the data was allotted to the high-level data stream. The group of pictures contained 12 frames, with 1 I-frame, 3 P-frames, and 8 B-frames.

The improvement in PSNR for the reconstructed high-resolution video is shown for one group of pictures in Figure 3.15. Notice that with this simple filtering technique, neither approach improves the performance in coding the I-frame. This is partially due to the very high quality achieved on the I-frame by the MPEG coder and the very simple conditioning technique that was applied. In the remaining frames, the conventional post-processing technique offers about a .1 dB improvement in PSNR, while the proposed post-conditioning technique offers a .5 to .7 dB improvement in PSNR. While the quantitative improvement is small, the significance lies in the consistency of the improvement over all the frames. In addition, a substantial improvement was achieved in the visual quality of the reconstructed high-resolution video. A 200×150 pixel portion of the 4th frame of each reconstructed high-resolution video sequence is enlarged and shown in Figure 3.16. Notice that the blocking artifacts are virtually eliminated even with this simple form of conditioning.

An experiment was performed to better understand the source of the improvement achieved by conditioning. In Figure 3.17, the gain over conventional coding is compared to using post-conditioning in the following levels: low only, high only, and both low and high. This plot illustrates that most of the gain results from post-conditioning the low-level video. This can be attributed in part to the fact that most of the signal energy is contained in the low-resolution video. In addition, the identification and reduction of artifacts is best understood for the low-resolution video. As we improve our understanding of the nature of artifacts in higher-level video and/or apply more sophisticated conditioning techniques, greater gain should be achievable in the higher levels of the
Figure 3.15: *Conventional approaches vs. conditioning*. A video sequence was coded with the conventional two-level spatially scalable coder shown in Figure 3.14. The PSNRs of the reconstructed high-resolution video sequences are shown for the conventionally coded video, the conditioned video, and the post-processed video.

system. Furthermore, it is interesting to note that the improvement in the post-conditioned low-resolution video is negligible, only about .01 dB. Despite this negligible quantitative gain in the low-resolution video, it leads to a more significant gain of about .5 dB in the high-resolution video.

### 3.4.5 Summary

This chapter investigated the problem of scalable coding for still images and video. It was concluded that for spatially scalable still-image coding, both critically sampled and oversampled representations have similar performance. Therefore, the best choice depends on the system requirements of the particular application. For spatially scalable video coding, it was shown that an efficient method incorporating powerful video compression techniques like motion-compensated prediction is not obvious in a critically sampled framework, while the oversampled representation is quite amenable to such techniques. For this reason, the oversampled representation was further investigated for spatially scalable video coding.

A layered structure was shown to be a generalization of the popular recursive pyramid structure introduced by Burt and Adelson [29]. Both of these structures achieve the desired oversampled
Figure 3.16: Performance comparison. A video sequence was coded in a two-level spatially scalable video coder. An MPEG coder was used in each level. An enlarged view of the fourth frame of the reconstructed high-resolution video is shown for the conventional approach (37.22 dB, left), conventional approach with post-processing (37.33 dB, center), and proposed approach featuring conditioning (37.93 dB, right). Conditioning outperforms conventional coding and post-processing.
Figure 3.17: Effectiveness of conditioning. An experiment was performed to better understand the source of the improvement achieved by conditioning. Post-conditioning was applied in only the low level, in only the high level, and in both the low and the high levels. Results show that the majority of the gain is due to post-conditioning the low-level video.

representation. However, the layered structure provides greater flexibility and further insight into the scalable coding problem.

A generalized framework was presented for layered approaches to scalable video coding. This framework incorporates a novel concept of conditioning which is used to combat the artifact propagation problem that affects layered coders. Examples illustrated that improved performance can be achieved with little added complexity by incorporating conditioning in a layered scalable video coding system.
Chapter 4

Hybrid Video Communication for Terrestrial Broadcasting

4.1 Introduction

In terrestrial television broadcasting, a centrally located transmitter delivers a television signal to receivers distributed throughout the broadcast area. Typically, the transmitter is located downtown in a densely populated area. This results in a high concentration of receivers near the transmitter. It will be shown that the channel capacity is highest in regions near the transmitter, and is lowest at the fringes of the broadcast area.

Current high-definition television (HDTV) proposals use purely digital methods of transmission. For the most part, these systems provide the same level of video quality to all the receivers in the broadcast area, regardless of the available capacity. These systems must use a data rate dictated by the receivers operating under the worst channel conditions, i.e. those located at the fringe of the broadcast area. Therefore, they do not exploit the extra capacity that is available to close-in receivers.

The concept of hybrid transmission was proposed by Schreiber in [3]. The novelty lies in its use of analog and digital transmission methods to improve the utilization of the channel capacity available in terrestrial broadcasting. The hybrid data stream is a discrete-time signal composed of a digital discrete-amplitude component and an analog continuous-amplitude component. Within this framework, powerful digital video coding methods can be used while, at the same time, achieving the graceful degradation properties of analog transmission. Hybrid transmission achieves improved performance by exploiting the extra capacity that is available in the central regions of the broadcast area.

The goal of this chapter is to explain how video communication can benefit from hybrid trans-
mission in a terrestrial broadcasting environment. Such an explanation is essential as it provides the motivation for the source-coding development of Chapter 5.

This chapter begins by briefly describing the fundamental concepts of communication theory that are relevant to this work. This includes a discussion on the theoretical behavior of conventional, single-quality digital transmission techniques in terrestrial broadcasting. The concept of hybrid transmission is described, and the conditions under which it performs optimally is discussed. Finally, a practical video coding technique is developed for hybrid-transmission terrestrial broadcasting, and its performance is compared to an all-digital system with comparable complexity.

4.2 Fundamental Concepts

This section contains a high-level description of the fundamental concepts behind digital and hybrid transmission. This discussion is concerned with the theoretical performance achievable with these systems. An investigation into the practical performance issues can be found in [1]. In-depth treatments on the theoretical and practical issues in communication can be found in [36, 37, 38], and complete treatments of information theory can be found in [39, 40, 41].

Channel Capacity

A communication system is composed of a transmitter, a communication channel, and a receiver, as shown in Figure 4.1. The transmitter is composed of a source and channel coder, and the receiver is composed of a channel and source decoder.

\[ C = \frac{1}{2} \log_2 \left( 1 + \frac{P}{N} \right) \]

(4.1)

where SNR is the received signal-to-noise ratio with signal and noise powers \( P \) and \( N \). Figure 4.2 shows channel capacity as a function of SNR.
4.2 Fundamental Concepts

Figure 4.2: Capacity of a Gaussian channel. The channel capacity is the maximum data rate that can be delivered reliably across a channel.

Shannon’s noisy channel coding theorem states that for point-to-point communication, data can be transmitted reliably at rates up to the channel capacity. For example, if the received SNR is 40 dB, the maximum rate for reliable transmission is 6.64 bits/symbol, while if the received SNR is 12 dB, the maximum rate is only 2.04 bits/symbol.

Separate vs. Joint Source/Channel Coding

The separation theorem states that under ideal conditions, the source and channel coders can be designed separately without loss of optimality. In this case, the source coder removes all the statistical redundancy in the input signal, and the channel coder reinserts the proper amount of statistical redundancy to reliably deliver the signal to the decoder. A number of assumptions are made in using the separation theorem. These include:

1. point-to-point communication with a given SNR
2. infinite block lengths (infinite delay)
3. infinite complexity
4. equally important bits
5. lossless coding

In the ideal system, the source coder’s output signal contains no statistical redundancy. Thus, every bit of the data stream is equally important, and the data stream appears completely random.

The assumptions of the separation theorem do not hold for a practical system. A real video communication system must use practical video coding techniques. One such typical technique is motion-compensated transform coding. The resulting video data is composed of motion vectors,
Figure 4.3: *Separate source/channel coding.* The separation theorem states that the source and channel coders can be designed separately without loss of optimality. A number of conditions must be satisfied for this to hold true.

... low-frequency transform coefficients, and high-frequency transform coefficients. It is evident that these different types of video data are not equally important in reconstructing the original video. In addition, infinite delay and infinite complexity are not realizable in a practical system. Finally, the terrestrial broadcast environment is not point-to-point, and lossy compression techniques are required to meet the bandwidth requirements of the channel. For these reasons, the separation theorem is not always applicable; thus, the design of a practical video communication system can benefit from jointly designing the source and channel coders.¹

Figure 4.4: *Joint source/channel coding.* The conditions for the separation theorem are not satisfied for all types of practical video communication systems. Therefore, joint design of the source and channel coders can be beneficial.

**Terrestrial Broadcasting**

In terrestrial television broadcasting, a centrally located transmitter delivers one signal across the entire broadcast area. Close-in receivers have the highest signal power and therefore have the highest signal-to-noise ratios (SNR), while farther out receivers have lower signal powers and thus lower SNRs. Since capacity is a function of SNR, this implies that receivers closest to the transmitter have

¹Separate coding is a special case of joint coding, and thus cannot theoretically be superior.
the capacity to receive much higher data rates than those located farther from the transmitter.

Figure 4.5: Terrestrial broadcast environment. A centrally located transmitter delivers a signal to receivers distributed throughout the broadcast area. Typically, the transmitter is located downtown in a densely populated area. This results in a high concentration of receivers near the transmitter.

Bendov examined the relationship between the SNR at a receiver to its distance from the transmitter in terrestrial broadcasting [42]. Many complicated factors affect the received SNR. A theoretical analysis was performed, and a field test was conducted to measure the received SNR in a terrestrial broadcasting system over relatively flat terrain. The results of this study are shown in Figure 4.6. Notice that the relationship between channel SNR and distance is approximately linear over a large portion of the broadcast area. Thus, throughout this work, we use SNR as a rough measure of distance from the transmitter. Higher SNRs correspond to receivers closer to the transmitter, while lower SNRs correspond to receivers farther from the transmitter.

Figure 4.6: SNR vs. distance. In terrestrial broadcasting, the received SNR depends on the receiver's distance from the transmitter. (Data provided by Dr. Oded Bendov.)
Digital Transmission

In a single-quality, all-digital system, a single-rate signal is transmitted across the channel. The recovered data rate is fixed for receivers operating above threshold, and is zero otherwise. For example, consider transmitting a signal with a rate $R = 2$ bits/symbol. With ideal channel coding methods, receivers with an SNR of 11.76 dB are at the threshold of reception. Receivers with higher SNRs can perfectly decode the signal with rate $R$, while those with lower SNRs can not decode the signal at all. The receivers operating at the threshold of reception (the fringe area) are operating at their capacity, while receivers at higher SNRs are operating below their capacities. Figure 4.7 shows a plot of the decoded data rate of an ideal digital system as a function of SNR. Notice the extra capacity that is available at the higher SNRs. For example, receivers with a channel SNR of 40 dB have the capacity to receive data rates up to 6.64 bit/symbol, but with digital transmission system described above they only receive 2 bit/symbol.

![Decoded Digital Rate](image)

Figure 4.7: Decoded data rate in digital transmission. In a single-quality all-digital system, the data rate is dictated by the channel capacity of the receivers operating under the worst conditions. Notice the capacity that is wasted at the higher SNRs.

Embedded Transmission

In the scenario above, a single data stream is transmitted at rate $R$. Alternatively, two lower-rate signals with rates $R_1$ and $R_2$ can be transmitted in the same channel if $R_1 + R_2 \leq R$. This is shown in Figure 4.8. This result is often used in multiuser information theory. We use this result to indicate that the data stream can be composed of multiple components, and as long as their rates do not exceed the capacity, they can be transmitted reliably across the channel.
Consider transmitting two independent signals with powers $P_1$ and $P_2$ across a single channel. The signals can be embedded [41] such that the resulting capacities are given by

\[(4.2) \quad C_1 = \frac{1}{2} \log_2 (1 + \text{SNR}_1) \quad \text{SNR}_1 = \frac{P_1}{P_2 + N}\]

\[(4.3) \quad C_2 = \frac{1}{2} \log_2 (1 + \text{SNR}_2) \quad \text{SNR}_2 = \frac{P_2}{N}.\]

These relationships can be intuitively understood as follows. For a properly embedded signal, the "noise" seen by the first signal is both the channel noise and the second signal. Once the first signal is decoded, the noise seen by the second signal is just the channel noise.

The significance lies in the fact that proper embedding results in no loss of efficiency – the total capacity of the independent, embedded signals equals the capacity of a single signal with the same total power. In other words, if

\[(4.4) \quad P = P_1 + P_2,\]

then

\[(4.5) \quad \frac{1}{2} \log_2 \left( 1 + \frac{P}{N} \right) = \frac{1}{2} \log_2 \left( 1 + \frac{P_1}{P_2 + N} \right) + \frac{1}{2} \log_2 \left( 1 + \frac{P_2}{N} \right)\]

or

\[(4.6) \quad C = C_1 + C_2.\]

This shows that properly embedding independent signals is not wasteful of channel capacity. Figure 4.9 contains a series of plots showing the capacities of the embedded signals and of a single
signal with the same total power as a function of SNR. Each plot represents a different distribution of power between the two embedded signals. Notice that the two extremes, \( P_1 = P, \ P_2 = 0 \) and \( P_1 = 0, \ P_2 = P \), converge to the one-signal case.

![Graphs showing theoretical channel capacity of embedded data streams](image)

Figure 4.9: *Theoretical channel capacity of embedded data streams.* The various plots represent different power distributions between the two embedded components.

### 4.3 Hybrid Transmission

This section introduces hybrid transmission and describes its behavior in a terrestrial broadcasting environment. A discussion is then provided on the conditions under which hybrid transmission performs optimally. Finally, a hybrid video coding algorithm is presented for hybrid transmission, and its performance is compared to that of an all-digital system of comparable complexity.
4.3 Hybrid Transmission

4.3.1 Description

The hybrid data stream is a discrete-time signal composed of a digital discrete-amplitude component and an analog continuous-amplitude component. The digital and analog components are embedded and the resulting hybrid signal is transmitted across the broadcast area. At the receiver, the digital component is assumed to be recovered perfectly and the analog component is recovered with an SNR that depends on the fidelity of the received signal.

The hybrid signal is formed by first creating a digital signal, and then superimposing the analog signal. A simple example of a one-dimensional signal space for a hybrid data stream is shown in Figure 4.10. The signal can be decoded in the following steps. First, the digital information is determined based on the polarity of the received signal. In this step, the analog information is seen as noise. Once the digital component is known, it is extracted from the hybrid signal and the remaining amplitude represents the received analog value.

![Hybrid Signal Diagram](image)

Figure 4.10: A simple one-dimensional hybrid constellation. The hybrid signal is composed of a digital discrete-amplitude component and an analog continuous-amplitude component.

The signal power of the transmitted hybrid signal, \( P_H \), depends on the powers of its digital and analog components, \( P_d \) and \( P_a \). If the components are independent,

\[
(4.7) \quad P_H = P_d + P_a.
\]

As discussed earlier, proper embedding of the digital and analog\(^2\) components yields the following capacities and corresponding SNRs without loss of efficiency.

\[
(4.8) \quad C_d = \frac{1}{2} \log_2 (1 + \text{SNR}_d) \quad \text{SNR}_d = \frac{P_d}{P_a + N}
\]

\[
(4.9) \quad C_a = \frac{1}{2} \log_2 (1 + \text{SNR}_a) \quad \text{SNR}_a = \frac{P_a}{N}.
\]

Strictly speaking, the concept of “capacity” for the analog component is not immediately obvious.\(^3\) However, we use this term loosely because it indicates the maximum SNR with which the analog component can be received. In other words, the capacity of the analog component is defined as

---

\(^2\)The analog signal is assumed to have a Gaussian distribution.

\(^3\)One typically describes the quality of an analog signal in terms of its fidelity or SNR.
the highest rate at which a digital signal could be transmitted across a channel with SNR$_a$, which is determined by $P_a$, $P_d$, and the channel SNR [43]. For example, if the capacity of a channel is 2 bits/symbol, and the power distribution allows the analog component to have a “capacity” of 1 bit/symbol, then the analog component can be received with a maximum SNR of 6 dB. Note that the power distribution between the two embedded components can be adjusted as desired. This results in the embedded capacities discussed in section 4.2 and shown in Figure 4.9 where $P_1 = P_d$ and $P_2 = P_a$.

The received hybrid signal can be evaluated by the rate of the decoded digital component and the SNR of the decoded analog component. For example, consider the following scenario. Assume that the broadcast area encompasses the receivers whose received SNR exceeds 10 dB. The total capacity at the fringe of the broadcast area is 1.7 bits/symbol. In addition, assume the relative powers of the digital and analog components are given by $P_d = .83P_H$ and $P_a = .17P_H$, where $P_H$ is the power of the transmitted hybrid signal. At the threshold of reception, the recovered digital data rate is 1 bit/symbol and the recovered analog SNR is 2.3 dB (which corresponds to 0.7 bits/symbol). At interior regions of the broadcast area, the recovered digital data rate is the same, but the recovered analog SNR increases with the channel SNR. These results are tabulated below. The received digital data rate and analog SNR are plotted as a function of channel SNR in Figure 4.11.

<table>
<thead>
<tr>
<th></th>
<th>Power</th>
<th>SNR</th>
<th>SNR (dB)</th>
<th>R (bits/symbol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid signal</td>
<td>$P$</td>
<td>10</td>
<td>10</td>
<td>1.7</td>
</tr>
<tr>
<td>digital component</td>
<td>$.83P$</td>
<td>3</td>
<td>4.8</td>
<td>1</td>
</tr>
<tr>
<td>analog component (fringe)</td>
<td>$.17P$</td>
<td>1.7</td>
<td>2.3</td>
<td>0.7</td>
</tr>
<tr>
<td>analog component (interior)</td>
<td>$.17P$</td>
<td>$\geq$ 1.7</td>
<td>$\geq$ 2.3</td>
<td>$\geq$ 0.7</td>
</tr>
</tbody>
</table>

Table 4.1 Relevant power levels, SNRs, and data rates for a hybrid system with a threshold of operation at SNR = 10 dB.

In summary, the hybrid data stream is composed of a digital component and an analog component. In terrestrial broadcasting, the digital component is received perfectly throughout the broadcast area, and the analog component is received with some baseline SNR at the fringes of the broadcast area, and with improved SNR in the interior regions. The improvement in SNR is due to the extra capacity that is available in the central region of the broadcast area in a terrestrial broadcast environment.

### 4.3.2 Hybrid Channel Efficiency

In terrestrial broadcasting applications, it can be shown that under certain conditions hybrid transmission uses the available channel capacity more efficiently than typical single-rate digital trans-
mission methods. Specifically, by properly embedding (or superimposing) the analog data stream on the digital data stream, comparable efficiency can be achieved at the fringe area, and improved efficiency can be achieved at the higher SNRs. The improvement is achieved by exploiting the extra channel capacity that exists in the more central regions of the terrestrial broadcast area.

We will now describe the conditions under which hybrid transmission is optimal. The first assumption is that the data stream can be split into two embedded streams that can be coded independently of one another without loss of efficiency. This requires that the source data is composed of two statistically independent components that can not benefit from joint coding. Next, in hybrid transmission, the second data stream is purely analog. In an ideal all-digital scheme, this would be coded into a digital data stream. For a given SNR, transmitting the uncoded analog data is equivalent to transmitting the coded digital data stream only when the data contains no statistical redundancy and the source data has a Gaussian distribution. If these conditions are satisfied, then a hybrid system achieves identical performance to a digital system at the boundaries of the broadcast region, and improved performance in the interior regions.

Theoretically, it should be possible to achieve nearly the same performance with an all-digital system [44]. In fact, if the analog component contains statistical redundancy, a purely digital system should be able to achieve better performance with ideal coding methods. However, when comparing the performance of real systems that are based on practical coding methods, the benefits are less clear and they depend on the particular application. In the next section, we will compare a practical hybrid television broadcasting system with a digital system of comparable complexity.
4.3.3 Discussion

From the viewpoint of capacity, it seems that analog transmission yields the most efficient channel usage for broadcast channels. So, one may ask if purely analog transmission techniques can be used for HDTV broadcasting. First of all, it must be noted that the analog component of a hybrid signal is not the same as the analog signal used in a conventional analog television system. Pure analog transmission of an HDTV signal would require much greater channel bandwidths than those used in today’s coded television systems. There is no question that higher compression is possible with an all-digital television system. The difference stems from the video compression techniques that can be used in the two transmission systems – digital video compression techniques are far superior to analog video compression techniques. In the next section, it will be shown that a hybrid transmission also enables the use of the kind of powerful video compression techniques normally associated with all-digital systems.

4.4 Hybrid Video Coding

The performance of a hybrid-transmission television system is compared to that of a digital-transmission television system. In the purely digital system, a single-rate digital signal is decoded by all the receivers in the broadcast area. This results in wasted channel capacity, especially in the regions closest to the transmitter. In the hybrid system, a lower-rate digital signal is decoded by all the receivers in the broadcast area, and the analog component is recovered with an SNR that depends on the SNR in the channel. This system exploits the extra available capacity by allowing the quality of the analog component to improve in the higher-capacity areas. Of course, this improved channel usage is only beneficial if it translates to improved video quality, which is the true motivation behind this work. The hybrid video coder described in this section was designed to exploit the benefits of hybrid transmission. This system achieves comparable video quality to an all-digital system at the fringe area, and improved video quality at higher SNRs.

In this section, we consider using transform coding techniques since they are commonly used in still-image and video compression. This section begins by discussing the relative importance of the different types of video data that result from these methods. Next, a modified transform coding technique is proposed for hybrid transmission. Finally, the performance of this system is compared to its all-digital counterpart.

4.4.1 Background

Common video coding techniques include transform coding and motion-compensated transform coding. These methods were described in Chapter 2. In intraframe video coding, each frame of
video is coded independently with transform coding techniques. Interframe coding methods estimate the motion between frames and form a prediction for each frame based on previously coded frames. The motion is represented by a set of motion vectors that must be coded into a data stream. The error in the prediction, the motion-compensated residual, is then coded with transform coding techniques.

In transform coding, an orthogonal transform is used to concentrate the signal energy into a small fraction of transform coefficients. Compression is then achieved by quantizing these coefficients. After quantization, many of the low-amplitude coefficients are set to zero, and only the locations and amplitudes of the nonzero coefficients need to be coded into the data stream.

The resulting data can be classified into a number of groups. These are shown in Figure 4.12. The DC transform coefficients are quite different from the remaining AC coefficients, and therefore are coded separately. Many of the AC coefficients are set to zero during quantization, so only the amplitudes and locations of the nonzero AC coefficients must be coded into the data stream. In some conventional systems, the amplitude and location data are coded jointly. However, since there is typically little correlation between these components, there is little loss in coding them separately.

![Figure 4.12: Video coders. Transform coding and motion-compensated transform coding result in different types of data.](image)

4.4.2 Relative Importance of Different Video Data

Hybrid video coding was motivated by the fact that different types of video data have different degrees of importance. For example, the lower-frequency coefficients are more important to the reconstructed video quality than the higher-frequency coefficients because the human visual system is more sensitive to errors in the low frequencies. Furthermore, when coding the nonzero coeffi-
cents, it is more important to accurately code the coefficient locations than the coefficient amplitudes. The locations must be received perfectly; otherwise an unintelligible image will result. Meanwhile, inaccuracies in the coefficient amplitudes produce essentially the same effect as additive noise in uncoded analog video.\textsuperscript{4} This is true of both digital and hybrid systems. Finally, in motion-compensated transform coding, the motion vectors are used to form the prediction for each frame. Thus, it is very important that the motion vectors are received accurately.

The differences in importance and fidelity of the various video components should be exploited in video communication systems. The video data that results from conventional coding methods are nicely matched to the properties of hybrid transmission in terrestrial broadcasting. The digital component of the hybrid data stream contains the more important video data that must be conveyed perfectly, such as low-frequency coefficients and the location information of higher-frequency coefficients; the analog component contains the data types that can withstand noise, such as the amplitude information of higher-frequency coefficients. Of course, if motion-compensated transform coding is used, the resulting motion vectors must also be coded into the digital portion of the data stream.

Transform coding is amenable to hybrid transmission because the coefficient amplitude values are inherently analog. By transmitting this information with analog methods, the reconstructed video quality can improve gracefully with channel conditions. This can be achieved by transmitting the coefficient amplitudes in full precision (without quantization) in the analog component of the hybrid signal. The channel noise then naturally affects these values, and they are received with a fidelity (SNR) that depends on the received SNR at each receiver. Thus, these values are received with some baseline SNR at the fringes of the broadcast area, and with improved SNR in the more central regions. This naturally translates to a graceful improvement in reconstructed video quality.

### 4.4.3 Hybrid Video Coder

A conventional transform coder codes an input video signal into a digital data stream as shown in Figure 4.13. The input signal is transformed with an orthogonal transformation. The DC coefficients are quantized and coded into a digital data stream. The AC coefficients are also quantized, and the locations and amplitudes of the nonzero coefficients are coded into a digital data stream. The data rate required to code the digital data is given by

\begin{equation}
R = R_{DC} + R_{\text{location}} + R_{\text{amplitude}}
\end{equation}

\textsuperscript{4}The noise is additive, but not “white”, because only a fraction of the coefficients are coded. It also tends to be concentrated in the complex areas. In relatively blank areas, most coefficients are eliminated, also eliminating the associated noise.
where $R_{DC}, R_{location}$, and $R_{amplitude}$ are the digital data rates needed to code the DC, AC location, and AC amplitude information.

Figure 4.13: Digital transform coding. In conventional digital transform coding, the transform coefficients are quantized, and the location and amplitude information of the nonzero quantized coefficients are encoded into a digital data stream.

An intraframe hybrid video coder was designed for hybrid transmission. This coder codes the input video signal into a hybrid data stream, which is composed of a digital and an analog component as described in section 4.3. The hybrid video coder is shown in Figure 4.14. As in the conventional digital system, the input signal is transformed with an orthogonal transform and the DC transform coefficients are quantized and coded into a digital data stream. However, in the hybrid system the AC coefficients are not quantized. Rather, a number of these coefficients are selected for transmission and are coded into a hybrid data stream. The amplitudes of the selected coefficients are transmitted in full precision in the analog component of the hybrid signal, and the locations of the selected coefficients are coded into the digital component of the hybrid signal.

The hybrid signal can be characterized by the digital data rate and the received analog SNR at the fringe of reception given by

\[(4.11) \quad R = R_{DC} + R_{location} \quad \text{SNR} = \text{SNR}_{amplitude}\]

where $R_{DC}$ and $R_{location}$ are the digital data rates needed to code the DC and AC location data and SNR$_{amplitude}$ is the received SNR of AC coefficients at the fringe of reception. This SNR increases with improved channel conditions.

Figure 4.14: Hybrid transform coding. In hybrid transform coding, transform coefficients are selected for coding. The DC coefficients are coded separately into a digital data stream. The locations of the selected AC coefficients are encoded into a digital data stream, and the amplitudes are transmitted in full precision.

Notice that in the digital system, the selection process implicitly occurs during quantization.
After quantization, the coefficients with nonzero amplitudes are the selected coefficients. Selection by quantization ensures that the largest-amplitude coefficients are coded into the data stream.\(^5\) This optimizes the reconstructed video for the given number of coefficients in the least-squares sense, but it does not necessarily optimize the performance for the total data rate. Further discussion on this topic is found in Appendix B.

In hybrid video coding, the selection process is generalized — any set of coefficients can be selected for transmission as long as it meets the requirements of the system [45]. When designing a selection scheme, consideration must be given to a number of criteria, including the visual quality of the reconstructed video, the data rate needed to represent the location data, and the noise performance of the amplitude data. One such adaptive selection scheme is described in Appendix B.

### 4.4.4 Experimental Results

An experiment was designed to illustrate the behavior of a hybrid video coding system. The performance of this system was compared to that of a conventional single-quality digital system. The experiment was conservative in that it favored the digital system.

In the experiment, the original 256 \(\times\) 256 pixel image was first transformed with the block DCT. The resulting DC and AC transform coefficients were uniformly quantized with a constant weighting matrix. In both the digital and the hybrid systems, it was assumed that the quantized DC coefficients were coded and transmitted digitally, and therefore would be received reliably by all the receivers in the broadcast area.

After quantization, many of the AC coefficients are zero. As discussed earlier, only the location and amplitude information of the nonzero quantized coefficients need to be coded into a data stream. In the digital system, it was assumed that these locations and quantized amplitudes are received reliably at all receivers in the broadcast area.

In the hybrid system, it was assumed that the locations were received reliably by all receivers, and that the amplitudes were received with a noise level determined by the channel conditions at the individual receiver. In the hybrid system, the amplitudes are not quantized at the encoder. They are transmitted in full precision in the analog component of the hybrid signal. Therefore, they are received with a degree of additive noise that depends on the channel SNR at the particular receiver.

For this experiment, the same coefficients that were selected by quantization in the digital system were assumed to be selected in the hybrid system. This is one aspect in which the test is conser-

\(^5\)When weighting tables are used for quantization, the largest-amplitude weighted coefficients are selected for transmission.
ative. A more general selection scheme would improve the performance of the hybrid system. However, this assumption was made to isolate the effects of hybrid transmission.

For both systems, the same DC and AC location information were used, so it was assumed that the digital data rates \( R_{DC} \) and \( R_{location} \) were identical in the two schemes. The difference between the two schemes lies in the transmission of the amplitudes of the selected AC coefficients. In the digital system, AC amplitudes are coded into a digital data stream of rate \( R_{amplitude} \). In the hybrid system, they are coded into the analog component of the hybrid signal, and are received at the fringe of reception with \( SNR_{amplitude} \).

\( SNR_{amplitude} \) was calculated with the following method. The entropy of the nonzero quantized AC coefficients was computed. It was assumed that for the digital system, these quantized amplitudes could be coded into a data stream with a rate \( R_{amplitude} \) equal to the entropy. Of course, this assumption also favors the performance of the digital system. For the hybrid system, Gaussian noise was added to each full-precision coefficient so that the resulting SNR corresponded to the \( R_{amplitude} \) of the amplitudes of the digital case. i.e.

\[
SNR_{amplitude} = 2^{R_{amplitude}} - 1
\]

This relationship is based on the definition of capacity given in equation 4.1.

The images reconstructed by the digital and the hybrid systems described above are shown in Figure 4.15. They represent the image quality that can be decoded at the fringe of the broadcast area. Notice that the two images have similar image quality. The received analog \( SNR_{amplitude} \) increases with the SNR in the channel. The image reconstructed by a hybrid receiver operating under better channel conditions is shown in Figure 4.16. This image has higher visual quality than the images reconstructed at threshold in both the digital and hybrid systems. The improvement is modest, but it must be noted that many aspects of this experiment favored the digital system. Notice that some visual artifacts still exist at the higher SNRs. These are not caused by channel noise; rather they are due to the coefficients that were not selected for transmission.

This experiment demonstrates that in terrestrial broadcasting, a hybrid system can achieve comparable performance to a digital system at the fringes of reception and improved performance in the interior regions of the broadcast area. In this example, the improvement is modest. However, this experiment was conservative in that it favored the digital system in many ways.

### 4.4.5 Remarks

Compared to a digital video communication system, the hybrid system achieves comparable performance at the fringe area and improved performance in the higher SNR regions. While the performance in interior regions is improved, it is not necessarily "optimal." This is demonstrated with

---

6Of course, a more general selection scheme could also improve the performance of the digital system.
Figure 4.15: Comparison of digital and hybrid systems. Reconstructed images in digital receivers (top) and in hybrid receivers (bottom) located at the fringe of the broadcast area.
Figure 4.16: Hybrid system. The quality of the reconstructed video in a hybrid system varies gracefully with channel conditions. This is the image that would be reconstructed at a higher channel SNR. In a digital system, the reconstructed image would be identical to that image shown in Figure 4.15.

The following argument. The best point-to-point transform coder optimally distributes the data between the location and amplitude information, i.e. it selects the best set of coefficients for transmission at the available rate. The hybrid system improves the performance at higher SNRs by allowing the amplitudes of these selected transform coefficients to be recovered with a higher SNR. However, the selected set of coefficients does not change. On the other hand, the best transform coder for the higher SNR would select the optimal set of transform coefficients for the higher rate, and it is unlikely to match the set chosen at the lower rate. Therefore, the coder designed for the higher capacity would produce better pictures for that capacity, with the drawback that it cannot be decoded at the lower rate.
4.5 Summary

This chapter investigated the concept of using hybrid transmission for terrestrial television broadcasting. It was shown that improved efficiency can be achieved by exploiting the extra channel capacity that is available in the central regions of the terrestrial broadcast area. This can be achieved by considering the relative importance and the properties of the different video data that results from common video compression techniques. A hybrid video coding system was designed for hybrid transmission. Finally, a simple experiment demonstrated that a properly designed hybrid video coder can achieve comparable performance to a digital video coder at the fringe area, and improved performance in the more central, higher-capacity regions, at equal signal power.
Chapter 5

An ATV Terrestrial Broadcasting System

5.1 Introduction

An advanced television (ATV) system was developed for use in the terrestrial broadcast environment. The system is composed of a source coder and a channel coder that were jointly designed with the goal of providing the best possible video quality to each receiver in the broadcast area. The source coder was designed as an integral part of this thesis and is based on the ideas presented in earlier chapters. The channel coder was designed by Polley and is discussed in detail in a companion thesis [1]. Further details of the system can be found in [2, 46, 47].

The novelty of this system lies in its use of scalable and hybrid coding methods, where hybrid refers to the combination of analog and digital techniques as discussed in Chapter 4. Joint source/channel coding methods are used to transmit each video component with a robustness corresponding to its perceptual importance, thereby achieving high spectrum efficiency.

The source coder achieves its inherent scalability in video quality, video resolution, and coder/decoder complexity through its two main features: pyramid filtering and hybrid video coding. Pyramid filtering is a form of scalable video coding; it decomposes the video into streams that can be combined at the receiver to produce natural-looking video signals at multiple resolutions. It allows greater coding flexibility than its critically sampled counterparts [29]. Hybrid video coding uses adaptive coefficient selection to compress enhancement video into hybrid data streams. While a number of all-digital scalable source coders have been developed in the past [32, 48], to the authors’ knowledge this is the first scalable source coder that uses the hybrid methods described in Chapter 4.

Polley’s channel coder efficiently utilizes the available spectrum with multirate, hybrid modulation. In comparison with a conventional single-quality digital-transmission system, this system achieves higher-quality video in the central region of the broadcast area, slightly lower video resolution at the fringe area, and extended low-resolution coverage beyond the conventional broadcast
area.

We first describe the concept of hybrid transmission, since it is essential for our source-coding approach. Next, the significance of joint source/channel coding is discussed. A system description of the scalable source coder is given along with a discussion of various issues concerning scalable and hybrid video coding. Finally, we show frames of video produced by a computer simulation of an ATV terrestrial broadcasting system designed according to these ideas.

5.2 Hybrid Transmission

Conventional HDTV proposals are based on all-digital, single-resolution source coding and transmission; therefore they provide the same video quality to every receiver in the service area. The performance of these systems is limited by the receivers operating under the worst channel conditions. Hybrid transmission helps overcome this limitation.

We have designed an ATV system with the goal of providing the highest possible video quality to each receiver, thereby better exploiting the channel capacity available in the terrestrial broadcast environment [49]. In terrestrial broadcasting, the channel capacity available between the central transmitter and each receiver varies, depending on the distance and the channel conditions between the two. Intuitively, a receiver that is close to the transmitter has higher channel capacity than one farther away. Therefore, an efficient use of the channel enables closer receivers to reconstruct higher-quality video, and allows farther receivers to reconstruct natural-looking, though lower-quality video.

In hybrid transmission, both analog and digital video data are transmitted over the terrestrial broadcast channel. Hybrid transmission combines the advantages of analog and digital source and channel coding methods. Analog methods are particularly useful in the terrestrial broadcast environment because they allow the quality of the recovered data to vary gracefully as channel conditions vary; this results in efficient channel use. Digital methods are particularly useful because reliable reception of digital data allows for the use of very powerful video compression techniques. The hybrid combination of analog and digital methods allows for both high video compression ratios and channel efficiency. Further details are discussed in Chapter 4.

In the ATV system that was designed, the quality of the reconstructed video varies in two ways. First, the video can be reconstructed at several predefined levels of resolution, where each decodable level depends on the distance and channel conditions between the transmitter and the particular receiver. Second, the quality of the received analog data, and therefore the quality of the reconstructed video, improves and degrades gracefully with channel conditions. The source and channel coders described in this paper were designed jointly with the goal of achieving this multiresolution, hybrid capability.
5.3 Joint Source/Channel Coding

The ATV system shown in Figure 5.2 consists of a source coder and a channel coder. The source coder compresses the video signal and separates it into multiple data streams. The channel coder prepares the data streams for transmission over the broadcast channel. Practical video coding techniques encode several types of video components, each with a different degree of importance. For maximum spectrum efficiency, each component should be transmitted with a degree of integrity reflecting its importance. For example, motion vectors and low-frequency transform coefficients, which have high perceptual importance, should be transmitted with greater robustness than high-frequency residual transform coefficients, which have less perceptual importance. Joint source/channel coding techniques have also been developed for the analog components to distribute the available signal power based on perceptual importance.

Figure 5.2: Source and channel coders. The source coder encodes the video into three data streams. The channel coder delivers these data streams across the broadcast area.

5.4 Channel Coder

The channel coder achieves efficient spectrum utilization with multiresolution, hybrid modulation. A detailed description of the channel coder can be found in [47].

The coverage area of the system is shown in Figure 5.4. The channel coder reliably delivers a digital data stream to all the receivers in the broadcast area, which is defined as the region in
which the received signal-to-noise-ratio (SNR) exceeds 6 dB. In addition, one hybrid data stream is decodable wherever the SNR exceeds 17 dB, and a second hybrid data stream is decodable where the SNR is greater than 29 dB.

Each hybrid data stream is composed of an analog and a digital component. Each digital component has a data rate of 4 Mbps. The received SNR of the first analog component is 13 dB where it is first decodable. The received SNR of the second analog component is 21 dB where it is first decodable. The received SNRs of both these components increase with improved channel conditions.

![Diagram](image)

**Figure 5.3: Coverage area.** A digital data stream is delivered to all the receivers in the broadcast area. Receivers with higher SNRs can also decode a first hybrid data stream; and receivers with even higher SNRs can also decode a second hybrid data stream.

The channel coder is shown in Figure 5.4. The digital components are protected with standard error-correction coding techniques, and the analog components are combined with a form of spread-spectrum processing. These components are combined to form the hybrid constellation. In this constellation, the digital components are represented in the angular dimension with non-uniform phase-shift keying techniques, and the analog information is represented in the radial dimension with amplitude-modulation techniques. The resulting hybrid signal is modulated with orthogonal frequency division multiplexing.

### 5.5 Scalable Source Coder

The function of the source coder is to compress video into the three data streams (digital, hybrid-1, and hybrid-2) transmitted by the channel coder. The scalable source coders that have been developed in the past [32, 48] are based on all-digital coding methods; one novel aspect of our approach lies in the use of hybrid coding methods.

As described earlier, the channel coder enables the receivers closest to the transmitter (channel SNR greater than 29 dB) to decode the one digital and two hybrid data streams; the receivers
somewhat farther from the transmitter (channel SNR from 17 to 29 dB) to decode the digital and hybrid-1 data stream; and the receivers farthest from the transmitter (channel SNR from 6 to 17 dB) to decode only the digital data stream.

A block diagram of the source coder is shown in Figure 5.5. Pyramid filtering, conceptually similar to the Laplacian pyramid filtering scheme developed by Burt and Adelson [29], decomposes the video into multiple levels of resolution. An MPEG coder compresses the lowest-resolution video into the digital data stream; a hybrid video coder compresses the medium-resolution enhancement video into the hybrid-1 data stream; and a second hybrid video coder compresses the high-resolution enhancement video into the hybrid-2 data stream. The MPEG-coded video is used to form a prediction for higher levels. The medium-resolution enhancement video is added to the low-resolution video to form medium-resolution video; likewise, the high-resolution enhancement video is added to the coded medium-resolution video to form high-resolution video.
Figure 5.5: Scalable source coder. The scalable source coder compresses video into three data streams (Digital, Hybrid-1, and Hybrid-2) using pyramid filtering, MPEG video coding, and hybrid video coding. Receivers closest to the transmitter can use all three data streams to decode high-resolution video; receivers farther from the transmitter can use the Digital and Hybrid-1 data streams to decode medium-resolution video; and receivers farthest from the transmitter can use the Digital data stream alone to reconstruct low-resolution video.

5.5.1 Pyramid Filtering

The pyramid filtering scheme produces natural-looking video at every level of resolution. Its oversampled nature has many advantages and better flexibility as compared to the more commonly used critically sampled schemes. For example, in pyramid filtering, the video coder is given multiple chances to code the low-frequency components of the video, which are known to be very important to perceptual video quality. In addition, fewer constraints are placed on the filterbank design, so filters can be chosen for high visual quality. The layered nature also provides flexibility in the coding that can be performed within level.

The architecture of the pyramid scheme used in this system is shown in Figure 5.5. The original video is first lowpass filtered and subsampled to form the video that will be encoded in the lowest level. A standard MPEG coder encodes the low-resolution video into a digital data stream that will be transmitted by the channel coder. The MPEG-coded, low-resolution video is then upsampling and lowpass filtered to form a prediction of the original video. The error in the prediction is then
coded in the next level of the pyramid.

In the next level, the residual video is lowpass filtered and subsampled. This time, a higher-bandwidth lowpass filter is used. The resulting video contains the low-frequency components that were not coded in the lower level as well as the higher-frequency components allowed to pass through the higher-bandwidth filter. This video is encoded into a hybrid data stream with a hybrid video coder described below. The resulting hybrid data stream is sent to the channel coder for transmission. The decoded video is then used to form a prediction for the next level.

In the final level, the remaining video is encoded into a hybrid data stream with a second hybrid video coder. This hybrid data stream is also sent to the channel coder for transmission. In the last level, the input video is not filtered at all. This provides the feature of perfect reconstruction – if this video was coded perfectly, then the original video can be decoded accurately.

This architecture differs from the original recursive structure proposed by Burt and Adelson [29]. Rather than recursively subsampling the video in each level of the pyramid, the coded video is upsampled to full size at each level. This generalization has higher computational and memory requirements, but allows greater flexibility when choosing the filter bandwidths and type of processing used in each level.

The criteria used in choosing the lowpass filters for each level of the pyramid are visual quality and frequency separation. The sharpened Gaussian filters were chosen for their single overshoot in the spatial domain and their relatively smooth shape and sharp cutoff in the frequency domain. It has been shown that coding with single-overshoot filters produces fewer noticeable visual artifacts than multiple-overshoot filters [27, 28].

Another key consideration is the fact that the lowpass filtered video will be subsampled prior to coding. For this reason, the filters must be designed to work well in conjunction with particular subsampling factors. Simple experiments were performed to find the best bandwidth to use for the particular subsampling ratio. Considerations include tradeoffs in sharpness vs. ringing artifacts due to uncanceled aliasing. In addition, if one blindly uses the nominal filter coefficients of the sharpened Gaussian filter and then subsamples the result, a periodic pattern will occur in the constant-amplitude regions [50]. This occurs because the polyphase components of the filter are not normalized. In each level of the pyramid, the polyphase components of the sharpened Gaussian filters were normalized for the particular subsampling factor.

Pyramid filtering has been little used because it forms an overcomplete representation of the original image, i.e. the pyramid representation of an image contains more samples than the original image. More popular filtering schemes involve critical sampling, in which the resulting representation contains the same number of samples as the original image. However, we have found that for the purpose of multiresolution coding, the visual quality at each resolution level of the pyramid scheme is superior to that of its critically sampled counterparts.
5.5.2 Hybrid Video Coding

A hybrid video coder is used in each higher level of the pyramid to code the enhancement video into a hybrid data stream that is composed of digital and analog components. Either intraframe or interframe coding methods can be used in each level of the system. These methods were discussed in detail in Chapter 4, and therefore will only be described briefly in this section.

An intraframe hybrid video coder is shown in Figure 5.6. This coder simply contains the hybrid transform coder shown in Figure 4.14. The architecture of the coder is similar to that of a standard JPEG coder. However, rather than quantizing and entropy coding the coefficients into a digital data stream, a novel adaptive selection algorithm is used to encode the coefficients into a hybrid data stream.

![Diagram](image)

Figure 5.6: Intraframe hybrid video coder. The intraframe hybrid video coder compresses enhancement video into a hybrid data stream made up of analog and digital components. The DC coefficients are coded separately into a digital data stream; the locations and amplitudes of selected AC coefficients are encoded into a digital data stream and an analog data stream.

An interframe hybrid video coder is shown in Figure 5.7. The architecture of this coder is similar to that of a standard MPEG coder in that motion-compensated prediction is applied. However, rather than using a digital transform coder to encode residual video into a data stream, a hybrid transform coder is used to encode the residual video into a hybrid data stream.

The system requirements will determine whether intraframe or interframe hybrid video coding is the better choice. A system that uses interframe methods naturally requires greater computational complexity in both the encoder and decoder. Because the coding is performed in enhancement video, the gain from using interframe methods depends highly on system parameters such as the relative bandwidths between levels.

Interframe Hybrid Video Coding

The incorporation of intraframe methods is relatively straightforward in the hybrid environment. Therefore, in the remainder of this chapter we describe a system that uses interframe hybrid video coding in each higher level of the system.

Motion-compensated prediction (MCP) is a highly effective video compression method used
in most modern video coding systems. In MCP, each frame of video is predicted from previously coded frames.\(^1\) The prediction is completely specified by a set of motion vectors; accurate knowledge of the motion vectors is thus essential for decoding the video. For this reason, the motion vectors are coded into a digital data stream. The coding gain of MCP results because the resulting residual video typically contains much less energy than the original, and is therefore easier to compress. In the few cases where MCP performs badly and the residual is hard to code, for example during scene changes, the original video can be coded instead.

The MC-prediction error, or MC-residual, is coded with a variation of conventional transform coding methods. The residual is transformed, and a subset of the transform coefficients are selected for transmission. The amplitudes of the selected coefficients are transmitted (unquantized) in the analog component, and the coefficient locations are coded into a digital bit stream. It is assumed that the digital data stream is received perfectly, and therefore the motion vectors and locations of the residual coefficients are known. The analog signal may suffer from channel noise, and therefore the received residual coefficients may be noisy.

\(^1\)This does not imply that successive video frames must be coded consecutively. In fact, if additional coder/decoder complexity and memory requirements are acceptable, then forward, backward, and bidirectional prediction techniques should be employed for improved performance.
Rather than using the block DCT as used in MPEG, the hybrid video coder uses a wavelet transform. The wavelet transform features fine frequency resolution at lower frequencies and fine spatial resolution at higher frequencies; therefore, it is well suited for coding residual video, which is typically highpass and edgy in nature [51, 13].

The hybrid video coder uses motion-compensated prediction with overlapped windows in conjunction with the wavelet transform. The reason for using overlapped windows in the motion-compensated prediction is apparent when one considers the coding scenario. Conventional approaches to MCP are based on block matching methods, and the resulting motion-compensated (MC) residual may have amplitude discontinuities at the block boundaries. In MPEG, the MC residual is transformed with the block DCT and the boundaries of the DCT basis functions typically line up with the discontinuities in the MC residual, making the match efficient. However, the wavelet transform has basis function that overlap the discontinuities in the MC residual. This throws excess energy into the wavelet coefficients, making the system less efficient. By using overlapped-window MCP one avoids the block discontinuities and the resulting residual is smoother, making it better suited for wavelet transform coding [15].

Adaptive Selection

The adaptive selection algorithm used to select coefficients for transmission is detailed in Appendix B. The locations of selected coefficients are encoded into the digital component of the hybrid data stream, and the amplitudes of selected coefficients are transmitted in the analog component. Initially, coefficients are selected by amplitude. Then, some higher amplitude coefficients are traded for lower amplitude ones to reduce the data needed to code the selection. The coding method exploits the similarities across various subbands. Experimental results verify that adaptive selection significantly reduces the data rate needed to represent the locations of the selected coefficients. In addition, in the hybrid video coders in the upper levels of the ATV system, it produces video of similar quality to the video that results from simply transmitting the highest-amplitude coefficients.

Issues

One of the main distinctions between digital and hybrid video coding systems lies in the transmission of the residual coefficients. Digital video coding systems typically transmit quantized residual coefficients, while the hybrid video coder transmits full-precision residual coefficients in analog form. The digital systems suffer from quantization noise which is known at the encoder, while the hybrid system suffers from channel noise, which is unknown at the encoder. For this reason, it has been speculated that MCP can not be used in a hybrid system.

A series of tests have been performed to examine the feasibility of using MCP in a hybrid environment. One of these tests is detailed in Appendix A. Our tests indicate that the analog noise
performance is good at most analog SNRs in our region of interest. At the very lowest SNRs, the noise does build up after several frames, but leakage or refresh techniques can be used to control this effect.

5.6 Results

A complete computer simulation was made of the described ATV system. A number of highly complex sequences were processed to clearly demonstrate the performance of the system.

As discussed earlier, one digital and two hybrid data streams are delivered to receivers in the broadcast area. Receivers operating under the best channel conditions can decode all three data streams, and those under worse channel conditions can only decode two or one data stream. The lowest-level digital data stream is transmitted at a rate of 4 Mb/sec, as are the digital components of each of hybrid data stream. The analog components of each hybrid data stream transmit 2.5 M samples/sec.

The system uses these data streams to deliver three resolutions of video at 60 frames/sec in a 6 MHz channel. The channel SNR thresholds are similar to those quoted earlier. The highest pyramid level contains 768 × 1280 resolution video, and the medium and lower levels contain 3/4 × 3/4 and 1/2 × 1/2 the horizontal and vertical resolutions, respectively. The parameters of the ATV system are shown in Table 5.1. The parameters were chosen so that the lowest-level video would be comparable in quality to NTSC and the highest-level video should be comparable to the 720 × 1280 × 60 frame/sec standard in the Grand Alliance proposal. These goals were substantially met.

Figures 5.9–5.11 show frames of the low-, medium-, and high-resolution reconstructed video at various points in the broadcast area. The low-resolution video is decodable throughout the broadcast area, which is defined as the region where the received SNR exceeds 6 dB. The medium-resolution video is decodable by receivers operating in SNRs exceeding 17 dB. The high-resolution video is decodable by receivers operating in SNRs exceeding 29 dB. The medium- and high-resolution frames correspond to the picture that is received at channel SNRs of 19 dB and 32 dB.
<table>
<thead>
<tr>
<th>Class</th>
<th>Composition</th>
<th>Incremental Rate</th>
<th>Total Rate</th>
<th>Threshold</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Base</strong></td>
<td>MPEG stream</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>audio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ancillary data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>digital</td>
<td>4 Mb/s</td>
<td>4 Mb/s</td>
<td>6 dB SNR</td>
<td>384 × 640</td>
</tr>
<tr>
<td><strong>Enhancement 1</strong></td>
<td>enhanced motion vectors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>selection information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>additional audio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>digital</td>
<td>4 Mb/s</td>
<td>8 Mb/s</td>
<td>17 dB SNR</td>
<td>576 × 960</td>
</tr>
<tr>
<td></td>
<td>selected residual coeffs.</td>
<td>analog</td>
<td>2.5 Ms/s</td>
<td>2.5 Ms/s</td>
<td></td>
</tr>
<tr>
<td><strong>Enhancement 2</strong></td>
<td>enhanced motion vectors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>selection information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>additional audio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>digital</td>
<td>4 Mb/s</td>
<td>12 Mb/s</td>
<td>29 dB SNR</td>
<td>768 × 1280</td>
</tr>
<tr>
<td></td>
<td>selected residual coeffs.</td>
<td>analog</td>
<td>2.5 Ms/s</td>
<td>5 Ms/s</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Composition, rates, and thresholds of the data streams.
Figure 5.8: High-resolution original video. The fifth frame of the progressively scanned "Traffic-girl" sequence.
Figure 5.9: Low-resolution reconstructed video. The low-resolution video can be decoded throughout the broadcast area, which includes receivers with SNRs exceeding 6 dB.
Figure 5.10: Medium-resolution reconstructed video. The medium-resolution video can be decoded by receivers with SNRs exceeding 17 dB. This particular frame was decoded at a channel SNR of 19 dB.
Figure 5.11: High-resolution reconstructed video. The high-resolution video can be decoded by receivers with SNRs exceeding 29 dB. This particular frame was decoded at a channel SNR of 32 dB.
Chapter 6

Summary and Future Work

This thesis investigates the problem of designing an advanced television (ATV) system for terrestrial broadcasting. In particular, it focuses on relevant source coding and joint source/channel coding issues that arise in the system design.

An ATV system was designed with the goal of efficiently using the channel capacity available in the terrestrial broadcast environment. In terrestrial broadcasting, the receivers closest to the transmitter have the capacity to receive the highest data rates, and those farthest from the transmitter can only receive lower data rates. This system allows each receiver to decode video with a quality that depends on the available capacity at each receiver. The quality varies in two ways: higher spatial resolutions can be decoded in higher-capacity areas, and within each resolution the quality of the reconstructed video improves with channel conditions.

Scalable video coding methods are used to encode the video into multiple data streams. In spatially scalable video coding, the base data stream can be used to reconstruct low-resolution video; additional enhancement data streams can also be used to reconstruct higher-resolution video.

Hybrid transmission allows the quality of each enhancement data stream to improve with channel conditions. Hybrid video coding methods encode the video into a hybrid data stream that is composed of a digital and an analog component. This allows the system to use powerful video compression techniques typically associated with digital methods while achieving the graceful performance of analog transmission.

Chapter 3 investigated various approaches to spatially scalable still-image coding and spatially scalable video coding. Two possibilities include the use of critically sampled and oversampled representations, such as the wavelet and pyramid representations. These representations have similar performance for coding still images; thus, the better choice depends on the system requirements of the particular application. However, for spatially scalable video coding, the flexibility of the oversampled representation facilitates the use of motion-compensated transform coding. The nonlinearity of motion compensation makes a comparable critically sampled system less apparent.
A generalized framework was presented for layered approaches to scalable video coding. This framework accommodates an improved class of coders by means of its novel feature of conditioning. Conditioning addresses the primary problem of layered coding, error propagation. In layered coding, the video is coded in multiple levels. Lower-level coding errors, which may not be visible at normal viewing distances, propagate to higher levels where they must be uncoded. These errors decrease the efficiency of the overall system. Conditioning attempts to remove these errors before they propagate to higher levels.

Chapter 4 describes the concept of hybrid transmission. A hybrid video coder was designed to simultaneously achieve the high compression rates associated with digital methods and the graceful degradation property of analog systems. The hybrid video coder was based on transform coding methods, and its performance was compared to that of a purely digital coder of similar complexity. It was shown that hybrid coding achieves comparable performance at the fringe area of reception, and improved performance in the interior regions of the broadcast area.

The ATV system was presented in Chapter 5. The chapter focused on the source coding aspects of the system. The source coder was based on the scalable and hybrid concepts described in Chapters 3 and 4. The low-resolution video is encoded into a digital data stream with a standard MPEG coder, and the enhancement video necessary to reconstruct higher-resolution video is encoded into a hybrid data stream with the hybrid video coder.

The hybrid video coder uses a novel adaptive selection algorithm that was developed to select and encode the residual transform coefficients for hybrid transmission. The coefficients are selected so that the resulting location mask is easier to code. In the proposed system, adaptive selection was shown to significantly reduce the data rate needed to encode the coefficient location information while not noticeably degrading the visual quality of the reconstructed video.

Finally, the feasibility of the designed ATV system was demonstrated with a complete computer simulation of the system.

6.1 Future Work

This work reveals many interesting topics for future research. A few of these topics are discussed below.

Scalable Video Coding

To the author's knowledge, no set of filterbanks has been designed to explicitly optimize the visual quality of the lower resolution, partially reconstructed images. This could be a promising area for future work. It is interesting to note that while wavelets are amenable to some types of multiresolution processing, they do not allow for proper consideration of the visual artifacts associ-
ated with partial reconstruction. Specifically, human sensitivity to uncanceled aliasing artifacts changes with scale, while the wavelet basis functions are fixed across scales.

Conditioning was used to achieve improved performance in layered JPEG and MPEG coders. In the examples, simple conditioning techniques were purposely chosen to clearly demonstrate the effectiveness of conditioning. Greater improvements may be achieved by employing more sophisticated conditioning techniques, such as [52, 53] and [54] included in Appendix C.

DCT-based coders were used in the conditioning experiments because of their prevalence in current video compression schemes and because their blocking artifacts are easy to characterize. However, it should be noted that the effectiveness of conditioning is not unique to DCT coders. All lossy coding techniques suffer from coding artifacts, and all layered coders suffer from artifact propagation. By better understanding the nature of these artifacts, conditioning can be used to increase the efficiency of such scalable systems.

The effectiveness of conditioning in multiple-level systems should also be examined. Test results with a three-level spatially scalable system indicated that conditioning in the upper levels is effective, but that most of the gain results from proper conditioning of the lowest-level video. It should be noted that the enhancement video encoded in the higher levels of the pyramid is fundamentally different in nature from the video encoded in the lowest level. A better understanding of the nature of the residual video and its coding artifacts may lead to improved performance in higher-level conditioning.

This work focused on conditioning for spatially scalable systems. It should be noted that the same concepts apply to other types of scalability. For example, temporal, frequency, and SNR scalable systems could also benefit from conditioning. In each of these cases, the goal is to prevent expending extra capacity in the higher-level coders to decode the artifacts generated in lower levels.

Finally, further work should be performed in investigating the benefits of conditioning via image enhancement and restoration techniques to form better predictions in higher levels of the coder. For example, major improvement can be achieved by enhancing the resolution of the decoded low-resolution video, thereby making a better prediction and reducing the energy in the residual that must be coded in the higher levels. One such scheme employing nonlinear interpolation was proposed by Anastassiou [55].

Hybrid Video Communication for Terrestrial Broadcast

The area of video communication with hybrid methods is relatively untouched. Hybrid transmission is only now being recognized by the research community as an efficient communication method for terrestrial broadcasting. A number of interesting issues arise when jointly designing the source and channel coders of a hybrid video communication system.

An investigation should be performed to find the optimum power distribution between the
analog and digital components of the hybrid data stream. In addition, the performance of other types of video compression techniques in a hybrid terrestrial broadcasting environment should be examined.

Finally, the concept of selecting coefficients for transmission was described in a hybrid video coding framework. An adaptive selection algorithm was designed for this purpose. This algorithm has many novel concepts such as adding and removing coefficients to reduce the data rate for the coded location information and exploiting the correlation between subbands of different orientations for improved coding efficiency. Similarly, these concepts may lead to improved performance in all-digital video coders.

**An ATV Terrestrial Broadcasting System**

The system achieved very good performance on difficult video test sequences. This indicated that perhaps even higher-quality video can be delivered in each level of the coder. This can be done in a number of ways. First, the resolution of the highest-level video can be increased. Alternatively, a fourth level of resolution can be incorporated for the central area.

The system does not achieve good color rendition at the higher resolutions. This results from the poor chrominance performance of the MPEG coder in the lowest level. An appropriate tradeoff of luminance and chrominance information in the upper level data streams will further improve the overall performance.

Further work should be directed toward gaining a deeper understanding of the effectiveness of motion-compensation techniques in the higher levels of the pyramid. It is likely that these results depend on the relative bandwidths between the levels.

The novel method of coding of the location information in the adaptive selection algorithm should be extended to all-digital coders. Research has been performed on selecting individual coefficients based on rate-distortion criteria [56]. However, the method of using a coarser selection mask in the subband domain may lead to further gains in performance.
Appendix A

Motion-Compensated Prediction in a Hybrid-Transmission HDTV System

Revised internal progress report, October, 1993 [57].

A.1 Background

Many digital video coding schemes process the video temporally to achieve the high degree of compression that is necessary for most applications. By far, the most commonly used technique is motion-compensated prediction.

It is not immediately obvious that motion-compensated prediction can be used in hybrid (analog/digital) video coding systems because of the unknown channel noise that is added to the analog coefficients. This section describes an experiment that was performed to test this concept. Test results show that, indeed, the combination is possible.

In a typical digital video coding scheme, a prediction of each video frame is made from previously coded video frames. The prediction is completely specified by motion vectors that are derived in the encoder. The error in the prediction, called the motion-compensated residual, is transformed using the block discrete cosine transform (DCT). The coefficients of the residual are quantized, and the locations and amplitudes of the nonzero quantized coefficients are coded in a digital bit stream using standard entropy coding techniques. It is assumed that the digital bit stream is received perfectly, and therefore the locations and amplitudes of the nonzero quantized coefficients are decoded perfectly.

In the designed hybrid video coding scheme, once again a prediction is made of each video frame, and the error in the prediction, or the residual, is transformed using the DCT. However, the coefficients of the residual are not quantized. Rather, a percentage of the residual coefficients are
selected for analog transmission (in full-amplitude precision), and only their locations are coded in a digital bit stream. It is assumed that the digital bit stream is received perfectly, and therefore the locations of the residual coefficients are known, but that the analog signal may suffer from channel noise, and therefore the amplitudes of the received residual coefficients may be noisy.

It is now evident that one of the main distinctions between digital and hybrid video coding systems lies in the transmission of the residual coefficients. The digital video coding system typically transmits quantized residual coefficients, and therefore suffers from quantization noise that is known and accounted for in the encoder. The hybrid video coding system typically transmits residual coefficients in full-amplitude precision, but suffers from channel noise that is unknown and therefore can not be accounted for in the encoder.

The question in regard to applying motion-compensated prediction in a hybrid video coding system is outlined as follows. It is known that the received residual coefficients will be noisy. Therefore, at the decoder a video frame will be reconstructed from noisy coefficients, and this frame will be used to form a noisy prediction for the next frame. This process will continue until the frame is refreshed with an intra-coded frame. The question is: does this noise build up in the reconstructed video? What is the resulting quality of the reconstructed video?

### A.2 Experiment

The following experiment was performed to investigate the noise performance of a hybrid-transmission HDTV system based on motion-compensated prediction. The experiment used a digital video coding system developed at MIT in the Advanced Television Research Group.

#### A.2.1 Digital Video Coding System

In the digital video coding system, the first video frame is coded with standard digital intraframe coding techniques. Motion-compensated prediction is used to encode the subsequent video frames. Each subsequent frame is divided into blocks of $16 \times 16$ pixels, and a motion vector is calculated for each block. In order to estimate the motion vector, each block is compared to the blocks of the previous frame that are within a given neighborhood of the original block. The block of the previous frame which has the minimum mean square error difference from the current block is chosen to be the prediction of the current block and its corresponding motion vector is transmitted. The motion-compensated residual of the block is the difference between the original and the predicted blocks. Occasionally, the residual block may be more difficult to encode than the original block. In these cases, no prediction is made and the original intraframe block is coded. Otherwise, the residual block is encoded.
In addition to the above mentioned intraframe coded blocks, a few columns of every frame are refreshed (i.e. intraframe coded) for video acquisition. The refresh columns are swept horizontally across the video such that every region of the image is refreshed every 20 frames, which for progressively scanned video at 60 frames/sec corresponds to one refresh every three seconds.

Each residual block is encoded with an $8 \times 8$ block DCT. The DCT coefficients are quantized according to their spatial frequency using the relative step sizes shown below:

\[
\begin{array}{cccccccc}
16 & 17 & 17 & 18 & 18 & 18 & 19 & 19 \\
17 & 17 & 18 & 18 & 19 & 19 & 19 & 20 \\
17 & 18 & 18 & 19 & 19 & 20 & 20 & 20 \\
18 & 18 & 19 & 19 & 20 & 20 & 20 & 21 \\
18 & 19 & 19 & 20 & 20 & 20 & 21 & 21 \\
18 & 19 & 20 & 20 & 20 & 21 & 21 & 21 \\
19 & 19 & 20 & 20 & 21 & 21 & 21 & 21 \\
19 & 20 & 20 & 21 & 21 & 21 & 21 & 22 \\
\end{array}
\]

In other words, the lowest-frequency horizontal and vertical components are uniformly quantized with a step size of 16 and the highest frequency horizontal and vertical components are uniformly quantized with a step size of 22.

In this system, it is assumed that the digital bit stream is received perfectly, and therefore the locations and amplitudes of the quantized residual coefficients are decoded perfectly. Therefore, the error in the reconstructed video of the digital video coding system is solely due to quantization noise.

**A.2.2 Modified Digital Video Coding System**

For the experiment, the system was modified in only one way. It was no longer assumed that the amplitudes of the quantized residual coefficients were received perfectly. Rather, analog noise, with variance equal to a multiple of the quantization step size, was added to the quantized amplitudes of the coefficients. In this case, the error in the reconstructed video is due to both quantization noise and channel noise.

It is important to note that in this experiment, the performance is worse than that of an actual hybrid video coding system. In an actual hybrid system, the error in the coefficient amplitudes would only be due to channel noise – the amplitudes of the residual coefficients would be transmitted in full precision as analog values. For implementation reasons, additive channel noise could only be added to the quantized amplitudes.
A.3 Experimental Results

The digital video coding system (no channel noise) was used to code a video sequence. Then, the modified digital video coding system was run with various amounts of channel noise added to the quantized amplitudes of the residual coefficients. The standard deviation of the added noise was 0.5, 1.0, and 1.5 times the quantization step size. Noise was added to all the residual coefficients except the three lowest-frequency components. These frequency components are so important to image quality that they would be transmitted digitally in a hybrid video coding system.

```
0 0 x x x x x
0 x x x x x x
x x x x x x x
x x x x x x x
x x x x x x x
x x x x x x x
x x x x x x x
x x x x x x x
x x x x x x x
0: no added noise (digital component of the hybrid signal)
x: added noise (analog component of the hybrid signal)
```

The experiments were performed on 15 frames of the MIT girl sequence. About 10000 residual coefficients were selected for each frame. At the 0.5 and 1.0 noise levels, there was no buildup of noise. At the highest 1.5 noise level, there was a slight accumulation. However, the visual degradation only began to look somewhat objectionable in the 14th frame.

The important observation that was made in this experiment is that the noise buildup is not catastrophic as once thought. Therefore, it is concluded that motion-compensated prediction can be used in a hybrid video coding system. Once again, it is important to note that the result of a true hybrid video coding system would be better than those obtained in this report because the received amplitudes would only suffer from channel noise and not quantization noise.
Appendix B

Adaptive Selection Coding

B.1 Introduction

Transform coding requires selecting which coefficients are important, and encoding the location and amplitude information of these coefficients into a data stream. In conventional digital methods, the selection process implicitly occurs during quantization—quantized coefficients with nonzero amplitudes are automatically selected for coding.

In this section, we consider the problem of selecting coefficients in a hybrid video coder, as discussed in Chapter 4. Specifically, a novel adaptive selection algorithm was developed for the hybrid-transmission television system described in Chapter 5. A novel aspect of this method is that it was based on the following hypotheses: 1) the selection of coefficients should be performed explicitly (as opposed to being a by-product of quantization), 2) encoding the highest-amplitude coefficients is not always the best choice; i.e., for a given bit rate, improved performance can be achieved by removing isolated coefficients and using the extra capacity for other coefficients, and 3) the locations of selected coefficients are highly correlated.

B.1.1 Coefficient Selection

Conventional transform coding consists of transforming the image, quantizing the coefficients, and encoding the result, as shown in Figure B.1. After quantization, many of the low-amplitude coefficients are forced to zero, so only the locations and amplitudes of the nonzero coefficients must be encoded into a digital data stream. The total data rate required for transform coding is the sum of the rates required to code the location and amplitude information of the nonzero coefficients,

\[ R_{\text{total}} = R_{\text{location}} + R_{\text{amplitude}} \]  

Quantization guarantees that the highest-amplitude coefficients are coded into the data stream, i.e., the highest-amplitude coefficients are selected for coding. For an orthonormal transform, this
minimizes the reconstructed mean-squared error subject to a minimum number of selected coefficients. However, it does not optimize the distortion or visual quality for a given overall data rate. For example, consider the extreme case of coding an isolated, low-amplitude transform coefficient. It may cost many bits to represent the location of this coefficient, while its contribution to the final reconstructed image quality will be quite low because of its low amplitude. A more logical choice is to choose not to code this coefficient, and use the bits to represent other information.

The shortcoming of conventional transform coding techniques stems from the fact that quantization dictates the selection of coefficients, and is independent of the encoding process. This does not achieve optimal performance. While finding the optimal set of coefficients for selection would require a computationally intense process, simpler suboptimal solutions may exist.

Figure B.2 shows a more general approach to the problem: after the transforming the image, coefficients are selected for encoding. This allows the coefficients to be selected deliberately, rather than result as a by-product of quantization. With this framework, the distribution of data between the location and amplitude information can be chosen to optimize the rate/distortion performance.

In this section, we address the problem of selecting coefficients in a hybrid transform coder designed for hybrid-transmission terrestrial broadcasting. This problem was formulated in Chapter 4, where it was shown that a hybrid system can be characterized by \( R_{\text{location}} \) and \( \text{SNR}_{\text{amplitude}} \), the digital data rate required to represent the coefficient locations and the received SNR of the analog coefficient amplitudes.\(^1\) Important considerations in designing a selection algorithm include the digital data rate required to code the location information, the noise performance of the system, and the visual quality of the reconstructed images.

\(^1\)To be more precise, \( R_{\text{location}} \) and \( \text{SNR}_{\text{amplitude}} \) represent the rate and SNR of the selected AC coefficients. The DC coefficients are coded separately into the digital data stream.
B.2 Adaptive Selection Algorithm

B.1.2 Correlation Across Orientations

Different types of correlation exist in the transformed image. When viewing a transformed image in its subband form, many of the subbands appear similar. Once these similarities are understood, a good coding method can be designed to achieve high compression by exploiting the correlation between bands.

Empirical evidence indicates that it is difficult to predict the amplitude information of the selected transform coefficients. On the other hand, the locations of these coefficients do exhibit some statistical similarities that can be exploited.

There exists much correlation among the locations of unselected coefficients. This is apparent when viewing a mask of the selected transform coefficients in its subband form – many of the subbands look similar. When using a wavelet transform coding technique, zerotree methods are often used to exploit the correlation inherent across the frequencies of a particular orientation. When viewing the transformed image, it is apparent that some correlation also exists across orientations of a particular frequency.

For example, consider encoding the traffic-girl sequence with the ATV system described in Chapter 5. The residual wavelet coefficients can be thresholded so that 5% of the coefficients are retained. A part of the resulting selection mask is shown in Figure B.3. Locations represented with dark dots represent coefficients with magnitudes greater than the threshold, and white dots represent coefficients with magnitudes below the threshold. It is obvious that the white areas appear highly correlated across subbands. This correlation should be exploited when encoding the data.

B.2 Adaptive Selection Algorithm

An adaptive selection algorithm was designed for the hybrid wavelet-based transform coder used in the hybrid-transmission advanced television system described in Chapter 5. This method reduces the data rate required to code the coefficient location information by adding and removing coefficients based on neighboring coefficients. The algorithm replaces some higher-amplitude coefficients by lower-amplitude ones in order to reduce the required data rate. Further reduction is achieved by exploiting the correlation that exists in the locations of the unselected coefficients.

The adaptive selection algorithm determines which wavelet transform coefficients to select for hybrid transmission, and codes the resulting location and amplitude information into the digital and analog components of the hybrid signal. The locations of the selected coefficients can be represented by a selection mask, which is a binary image that indicates whether or not coefficients are selected.

\[ ^2 \text{Of course, visual quality should not be compromised in the process.} \]
Figure B.3: Selection mask. The traffic-girl sequence was coded with the ATV system described in Chapter 5. This is a part of the selection mask that results when selecting 5% of the highest-amplitude wavelet coefficients in the middle level.
A block diagram of the adaptive selection algorithm is shown in Figure B.4. An initial set of coefficients is chosen by comparing the magnitude of each coefficient to a set threshold. The resulting selection mask is divided into blocks that are classified according to the number of selected coefficients. The classification process adds and removes (selects and unselects) individual coefficients based on the amplitudes of their neighbors. This reduces the required data rate by making a coarser selection mask that is easier to code. Further data-rate reduction is achieved with an AND operation that exploits the correlation that exists among these classified blocks. Finally, this data is coded into a digital data stream of rate $R_{\text{location}}$.

![Block diagram of adaptive selection algorithm](image)

Figure B.4: Adaptive selection algorithm. This algorithm is used to select transform coefficients in a hybrid video coder.

In adaptive selection, an initial selection mask is obtained by comparing the magnitude of each coefficient to a set value. This initial mask contains a 1 for each coefficient that is above the threshold, and a 0 otherwise. This mask is then divided into $p \times p$ blocks. Typical values of $p$ are 2, 3, and 4. Each block is classified as empty, intermediate, or full, depending on the number of ones it contains. If the block has a low number of ones, then it is classified as empty and all the elements are set to zero. If the block has a high number of ones, then the block is classified as full and all the elements are set to one. If the block has an intermediate number of ones, then the block is classified as intermediate and the closest match in a predetermined library of masks is chosen.

Once the blocks are classified, this information must be coded into a data stream. This data stream must convey the classification [empty ($e$), intermediate ($i$), or full ($f$)] of each block, and for
each intermediate block it must convey the closest library match. In the simplest coding scheme, the $p \times p$ blocks can be coded independently. This can be greatly improved by exploiting the wavelet structure of the selection mask and the correlation that exists among subbands. An AND operation was developed to exploit this correlation.

The classification mask is labelled with the naming convention shown in Figure B.5. The lowest-frequency band is the DC band. Groups A, B, and C contain the horizontal, vertical, and diagonal components of the low-, middle-, and high-frequency bands. If the original image, and thus the wavelet-transformed image, has horizontal and vertical dimensions of $X \times Y$, then each group-A band contains $\frac{X}{8p} \times \frac{Y}{8p}$ elements, each group-B band contains $\frac{X}{4p} \times \frac{Y}{4p}$ elements, and each group-C band contains $\frac{X}{2p} \times \frac{Y}{2p}$ elements. Each element has a value of $e$, $i$, or $f$.

![Figure B.5: Naming convention for wavelet bands.](image)

The inter-band correlation in each group is exploited when coding the classification information. Different classification masks within each group often contain identical values in corresponding spatial locations. Therefore, it is often advantageous to define one AND mask for each group as shown in Figure B.6. This mask has the same dimensions as the individual bands in the group. Each element of the AND mask contains an $E$, $I$, $F$, or $X$. The mask has an $E$, $I$, or $F$ if the corresponding values in all the bands are respectively empty $e$, intermediate $i$, or full $f$; and an $X$ otherwise. For each AND mask, the value of each element must be coded. In addition, for each $X$, the values $e$, $i$, or $f$ must be specified for the corresponding values in the horizontal, vertical, and diagonal bands in the group.
Figure B.6: *AND operation*. The AND operation exploits the inter-band correlation that exists in the classification mask.

### B.3 Experimental Results

The adaptive selection algorithm was used in the hybrid video coders in the middle and upper levels of the advanced television (ATV) system described in Chapter 5. The HDTV “traffic-girl” sequence was coded with the ATV system. This is a progressively scanned sequence containing $768 \times 1280$ pixels/frame at 60 frames/sec.

In the ATV system, the transformed signal contains $576 \times 960 = 552,960$ coefficients in the middle level and $768 \times 1280 = 983,040$ coefficients in the highest level. For each upper level, selected transform coefficients may be transmitted in full precision at a rate of 2.5 Msamples/sec or 41,667 samples/frame. In addition, each upper level is allotted a digital data rate of 4 Mbits/sec or 66,667 bits/frame in which the coefficient location information and motion vectors must be conveyed. Half of the digital data rate was allocated to the location information. Thus, the adaptive selection algorithm was designed to reduce the coefficient location information to a target bit rate of 2 Mbits/sec or 33,333 bits/frame while retaining 41,667 samples/frame. Of course, this should be achieved while maintaining high visual quality in the reconstructed video.

Wavelet transform coefficients were chosen according to the adaptive selection algorithm presen-
ted earlier. First, a selection mask is created with the dimensions of the wavelet transform. The mask initially contains 1s or 0s corresponding to the coefficient amplitudes being above or below a threshold. Next, the selection mask is divided into $4 \times 4$ blocks. Each $4 \times 4$ block is classified as empty, intermediate, or full based on the number of 1s it contains. The ranges $0 – 2$, $3 – 5$, and $6 – 16$ were used for the classification of empty, intermediate, and full. The selection mask is then modified as follows. All values in the empty blocks are set to 0; all values in the full blocks are set to 1; and the values in the intermediate blocks are set to the closest match of the following 18 possibilities:

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td>0000</td>
<td>0000</td>
<td>0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0000</td>
<td>0000</td>
<td>0000</td>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1110</td>
<td>0111</td>
<td>0000</td>
<td>0000</td>
<td>0000</td>
<td>0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1110</td>
<td>0111</td>
<td>1110</td>
<td>0111</td>
<td>0000</td>
<td>0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0000</td>
<td>0000</td>
<td>1110</td>
<td>0111</td>
<td>1110</td>
<td>0111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0000</td>
<td>0000</td>
<td>0000</td>
<td>0000</td>
<td>1110</td>
<td>0111</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td>1111</td>
<td>0000</td>
<td>0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td>1111</td>
<td>1111</td>
<td>0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td>0000</td>
<td>1111</td>
<td>1111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td>0110</td>
<td>0011</td>
<td>0000</td>
<td>0000</td>
<td>1111</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following information must be coded to completely describe the selection mask. The classification of each $4 \times 4$ block must be specified as empty, intermediate, or full. In addition, for each intermediate block, the closest match must be specified as one of the 18 possibilities. The AND operation discussed above was used to reduce data rate required for the classification information.

The resulting histograms of the middle- and high-level classification masks are included at the end of this section. In both levels, the first frame is an intra-coded frame, and each subsequent frame is an inter-coded frame (a motion-compensated residual). Runlength coding and variable-length coding techniques were used to reduce the coded data rate of the classification information. The resulting data rates are shown in Table B.1. Notice that the middle-level data rates are well below the target bit rates, while the high-level data rates are very close to the target bit rates.

The mask shown in Figure B.3 results from thresholding the residual wavelet coefficients in the middle level of the ATV coder. This mask was further processed with the adaptive selection algorithm. The resulting selection mask is shown in Figure B.7. It is evident that this mask is significantly easier to code. In addition, the images reconstructed from the two masks are indistin-
B.3 Experimental Results

<table>
<thead>
<tr>
<th>Frame number</th>
<th>Middle-level data rate</th>
<th>High-level data rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>21330</td>
<td>36151</td>
</tr>
<tr>
<td>1</td>
<td>20727</td>
<td>34012</td>
</tr>
<tr>
<td>2</td>
<td>20483</td>
<td>33624</td>
</tr>
<tr>
<td>3</td>
<td>20352</td>
<td>33572</td>
</tr>
<tr>
<td>4</td>
<td>20185</td>
<td>33358</td>
</tr>
<tr>
<td>5</td>
<td>20342</td>
<td>33481</td>
</tr>
<tr>
<td>6</td>
<td>20161</td>
<td>33225</td>
</tr>
<tr>
<td>7</td>
<td>19682</td>
<td>32818</td>
</tr>
<tr>
<td>8</td>
<td>20274</td>
<td>33328</td>
</tr>
<tr>
<td>9</td>
<td>20010</td>
<td>33222</td>
</tr>
</tbody>
</table>

Table B.1 Required data rates for location information in the middle and high levels of the ATV system. These rates are very close to the target data rate of 33,333 bits/frame.

guishable. Thus, a high reduction in data rate has been achieved with virtually no loss in video quality.$^3$

An adaptive selection algorithm was developed for use in the hybrid video coders used in the upper levels of the hybrid-transmission ATV system. This algorithm was designed with the goal of reducing the coded data rate of the location information without compromising the final reconstructed video quality. The resulting algorithm was successful applied in the computer simulation of the ATV system described in Chapter 5.

---

$^3$This process was applied on an enhancement video signal in a pyramid coder and not on natural video.
Figure B.7: *Adaptive selection mask*. The mask shown in Figure B.3 resulted from simply thresholding the coefficients. This is the final selection mask that results from the adaptive selection algorithm. The data rate required to code this mask is significantly lower than that required to code the initial mask.
B.3 Experimental Results

Histogram of Classification Data

The traffic-girl sequence was coded in the computer simulation of the designed ATV system. The resulting histograms of the middle- and high-level classification masks are shown below for 10 frames of coded video. The naming convention for the bands was described in Figure B.5. Each group has a horizontal, vertical, and diagonal band as well as an AND mask. The numbers following the AND mask represent the number of positions with all-empty, all-intermediate, and all-full (E,I,F) coefficients throughout the bands in the group. The remaining positions are unspecified (X) and must be coded individually. The three numbers following the horizontal, vertical, and diagonal bands represent the remaining number of empty, intermediate, and full blocks (e,i,f) in each band.

<table>
<thead>
<tr>
<th>Dataset 0</th>
<th>Dataset 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP A</td>
<td>GROUP A</td>
</tr>
<tr>
<td>AND (14,11,189) unspec=326/540</td>
<td>AND (97,7,2) unspec=854/960</td>
</tr>
<tr>
<td>dc (2,105,219)</td>
<td>dc (713,124,17)</td>
</tr>
<tr>
<td>hor (104,136,86)</td>
<td>hor (264,363,227)</td>
</tr>
<tr>
<td>vert (64,107,155)</td>
<td>vert (168,361,325)</td>
</tr>
<tr>
<td>diag (103,107,116)</td>
<td>diag (196,316,342)</td>
</tr>
<tr>
<td>GROUP B</td>
<td>GROUP B</td>
</tr>
<tr>
<td>AND (962,44,30) unspec=1124/2160</td>
<td>AND (2089,22,8) unspec=1721/3840</td>
</tr>
<tr>
<td>hor (582,421,121)</td>
<td>hor (1173,444,104)</td>
</tr>
<tr>
<td>vert (179,634,311)</td>
<td>vert (281,1000,440)</td>
</tr>
<tr>
<td>diag (900,180,44)</td>
<td>diag (1550,116,55)</td>
</tr>
<tr>
<td>GROUP C</td>
<td>GROUP C</td>
</tr>
<tr>
<td>AND (8379,1,3) unspec=257/8640</td>
<td>AND (15105,4,1) unspec=250/15360</td>
</tr>
<tr>
<td>hor (92,124,41)</td>
<td>hor (120,100,30)</td>
</tr>
<tr>
<td>vert (147,93,17)</td>
<td>vert (122,114,14)</td>
</tr>
<tr>
<td>diag (218,25,14)</td>
<td>diag (202,20,28)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP A</td>
</tr>
<tr>
<td>AND (39,9,197) unspec=295/540</td>
</tr>
<tr>
<td>dc (32,98,165)</td>
</tr>
<tr>
<td>hor (66,139,90)</td>
</tr>
<tr>
<td>vert (29,94,172)</td>
</tr>
<tr>
<td>diag (63,124,108)</td>
</tr>
<tr>
<td>GROUP B</td>
</tr>
<tr>
<td>AND (925,45,29) unspec=1161/2160</td>
</tr>
</tbody>
</table>
hor (698,349,114)  
vert (181,653,327)  
diag (927,195,39)  
GROUP C  
AND (8448,3,2) unspec=187/8640  
hor (91,79,17)  
vert (66,107,14)  
diag (152,26,9)  

Dataset 2  
GROUP A  
AND (36,9,217) unspec=278/540  
dc (20,99,159)  
hor (54,123,101)  
vert (36,97,145)  
diag (63,137,78)  
GROUP B  
AND (1020,32,33) unspec=1075/2160  
hor (645,333,97)  
vert (177,602,296)  
diag (876,156,43)  
GROUP C  
AND (8420,6,3) unspec=211/8640  
hor (104,83,24)  
vert (89,108,14)  
diag (168,31,12)  

Dataset 3  
GROUP A  
AND (27,10,203) unspec=300/540  
dc (27,90,183)  
hor (49,157,94)  
vert (40,106,154)  
diag (64,135,101)  
GROUP B  
AND (1005,36,27) unspec=1092/2160  
hor (724,295,73)  
vert (139,625,328)  
diag (873,190,29)  
GROUP C  
AND (8450,4,2) unspec=184/8640  
hor (74,82,28)  
vert (85,84,15)  
diag (144,29,11)  

hor (927,360,93)  
vert (266,890,224)  
diag (1210,130,40)  
GROUP C  
AND (15145,3,0) unspec=212/15360  
hor (113,79,20)  
vert (103,98,11)  
diag (183,22,7)  

AND (83,28,11) unspec=838/960  
dc (492,301,45)  
hor (173,339,326)  
vert (106,384,348)  
diag (148,320,370)  
GROUP B  
AND (2485,26,15) unspec=1314/3840  
hor (863,364,87)  
vert (268,878,168)  
diag (1140,132,42)  
GROUP C  
AND (15171,2,0) unspec=187/15360  
hor (76,96,15)  
vert (100,74,13)  
diag (159,24,4)  

AND (82,24,15) unspec=339/960  
dc (470,329,40)  
hor (168,375,296)  
vert (111,369,359)  
diag (154,297,388)  
GROUP B  
AND (2460,25,13) unspec=1342/3840  
hor (945,324,73)  
vert (241,904,197)  
diag (1153,144,45)  
GROUP C  
AND (15201,4,0) unspec=155/15360  
hor (68,67,20)  
vert (75,67,13)  
diag (129,22,4)
B.3 Experimental Results

Dataset 4
GROUP A
AND (34,7,218) unspec=281/540
dc (13,83,185)
hor (60,126,95)
vert (34,94,153)
diag (66,128,87)

GROUP B
AND (1060,34,28) unspec=1038/2160
hor (601,332,105)
vert (189,575,274)
diag (835,178,25)

GROUP C
AND (8460,5,2) unspec=173/8640
hor (74,69,30)
vert (72,87,14)
diag (135,30,8)

Dataset 5
GROUP A
AND (37,11,205) unspec=287/540
dc (23,92,172)
hor (66,142,79)
vert (36,80,171)
diag (73,144,70)

GROUP B
AND (1001,29,24) unspec=1106/2160
hor (709,300,97)
vert (131,631,344)
diag (914,167,25)

GROUP C
AND (8439,7,5) unspec=189/8640
hor (100,63,26)
vert (72,94,23)
diag (150,34,5)

Dataset 6
GROUP A
AND (30,9,199) unspec=302/540
dc (19,101,182)
hor (59,142,101)
vert (39,89,174)
diag (82,129,91)

GROUP B
AND (1069,29,37) unspec=1025/2160

GROUP C
AND (2508,28,15) unspec=1289/3840
hor (837,383,69)
vert (269,824,196)
diag (1098,145,46)

GROUP B
AND (2508,28,15) unspec=1289/3840
hor (837,383,69)
vert (269,824,196)
diag (1098,145,46)

GROUP C
AND (15194,2,1) unspec=163/15360
hor (77,69,17)
vert (80,65,18)
diag (136,21,6)

GROUP C
AND (15194,2,1) unspec=163/15360
hor (77,69,17)
vert (80,65,18)
diag (136,21,6)

GROUP A
AND (90,35,10) unspec=825/960
dc (452,327,46)
hor (123,380,322)
vert (137,346,342)
diag (123,318,384)

GROUP B
AND (2461,25,10) unspec=1344/3840
hor (913,347,84)
vert (246,911,187)
diag (1156,140,48)

GROUP C
AND (15204,2,2) unspec=152/15360
hor (65,76,11)
vert (72,64,16)
diag (123,22,7)

GROUP A
AND (88,31,9) unspec=832/960
dc (459,330,43)
hor (126,378,328)
vert (128,377,327)
diag (122,296,414)

GROUP B
AND (2499,35,13) unspec=1293/3840
hor (603,327,95)  
vert (160,565,300)  
diag (845,147,33)  
GROUP C  
AND (8470,5,2)  
    unspec=163/8640  
hor (63,75,25)  
vert (74,73,16)  
diag (125,26,12)  

Dataset 7  
GROUP A  
AND (28,13,244)  
    unspec=255/540  
dc (20,92,143)  
hor (52,108,95)  
vert (46,91,118)  
diag (56,115,84)  

GROUP B  
AND (1128,28,40)  
    unspec=964/2160  
hor (537,320,107)  
vert (171,541,252)  
diag (799,143,22)  

GROUP C  
AND (8463,8,5)  
    unspec=164/8640  
hor (57,82,25)  
vert (83,58,23)  
diag (129,30,5)  

Dataset 8  
GROUP A  
AND (35,8,215)  
    unspec=282/540  
dc (30,79,173)  
hor (58,123,101)  
vert (42,85,155)  
diag (77,119,86)  

GROUP B  
AND (996,27,27)  
    unspec=1110/2160  
hor (729,293,88)  
vert (153,614,343)  
diag (892,186,32)  

GROUP C  
AND (8476,3,2)  
    unspec=159/8640  
hor (59,70,30)  
vert (81,65,13)  
diag (122,29,8)  

hor (871,353,69)  
vert (244,871,178)  
diag (1107,137,49)  
GROUP C  
AND (15202,6,1)  
    unspec=151/15360  
hor (62,75,14)  
vert (73,64,14)  
diag (122,25,4)  

AND (89,34,15)  
    unspec=822/960  
dc (425,350,47)  
hor (133,347,342)  
vert (122,355,345)  
diag (125,336,361)  
GROUP B  
AND (2604,33,20)  
    unspec=1183/3840  
hor (748,358,77)  
vert (250,793,140)  
diag (1005,143,35)  

AND (15208,5,0)  
    unspec=147/15360  
hor (67,66,14)  
vert (74,60,13)  
diag (105,34,8)  

AND (87,33,13)  
    unspec=827/960  
dc (444,333,50)  
hor (135,349,343)  
vert (134,357,336)  
diag (129,322,376)  
GROUP B  
AND (2516,31,10)  
    unspec=1283/3840  
hor (842,360,81)  
vert (249,863,171)  
diag (1096,137,50)  

GROUP C  
AND (15200,5,0)  
    unspec=155/15360  
hor (72,69,14)  
vert (71,68,16)  
diag (123,27,5)
### Dataset 9

**GROUP A**

<table>
<thead>
<tr>
<th>Operation</th>
<th>X1</th>
<th>Y1</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td>40</td>
<td>10</td>
<td>227</td>
</tr>
<tr>
<td>dc</td>
<td>24</td>
<td>101</td>
<td>138</td>
</tr>
<tr>
<td>hor</td>
<td>65</td>
<td>117</td>
<td>81</td>
</tr>
<tr>
<td>vert</td>
<td>24</td>
<td>88</td>
<td>151</td>
</tr>
<tr>
<td>diag</td>
<td>66</td>
<td>108</td>
<td>89</td>
</tr>
</tbody>
</table>

**GROUP B**

<table>
<thead>
<tr>
<th>Operation</th>
<th>X1</th>
<th>Y1</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td>1029</td>
<td>57</td>
<td>31</td>
</tr>
<tr>
<td>hor</td>
<td>523</td>
<td>395</td>
<td>125</td>
</tr>
<tr>
<td>vert</td>
<td>143</td>
<td>590</td>
<td>310</td>
</tr>
<tr>
<td>diag</td>
<td>815</td>
<td>193</td>
<td>35</td>
</tr>
</tbody>
</table>

**GROUP C**

<table>
<thead>
<tr>
<th>Operation</th>
<th>X1</th>
<th>Y1</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AND</td>
<td>8512</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>hor</td>
<td>52</td>
<td>51</td>
<td>18</td>
</tr>
<tr>
<td>vert</td>
<td>60</td>
<td>51</td>
<td>10</td>
</tr>
<tr>
<td>diag</td>
<td>92</td>
<td>21</td>
<td>8</td>
</tr>
</tbody>
</table>
Appendix C

A Wiener Filter Approach to Reducing Transform Coding Artifacts

Revised internal progress report, August, 1995 [54].

C.1 Introduction

Transform coding is used in many image and video compression systems. At high compression ratios, the reconstructed images are afflicted with visually annoying transform coding artifacts. In particular, DCT-based coders suffer from blocking effects and "mosquito" noise.\(^1\) This section examines the problem of reducing transform coding artifacts. A new perspective is presented for describing these artifacts and a solution is developed based on Wiener filtering ideas. This solution exploits the statistical properties of the signal and the quantization noise. Finally, this method is applied to a JPEG-type DCT coder. Improved qualitative as well as quantitative improvement is demonstrated.

C.2 Transform Coding Artifacts

In DCT coders, the blocking effect is usually attributed to and measured by the discontinuity that exists across block boundaries. We propose that the blocking effect and related transform coding artifacts can be further described by a more detailed analysis of the quantization noise in the spatial domain. In particular, the spatial distribution of the quantization noise is non-uniform and noise

\(^{1}\) Blocking effects result from the non-overlapping basis functions of the DCT. Mosquito noise results from the unequal noise distribution among the various frequency components.
samples are highly correlated with adjacent samples in the same block, but uncorrelated with adjacent samples in neighboring blocks.

Transform coders achieve high compression by quantizing and entropy coding the "decorrelated" transform coefficients of the image. The variance of the quantization noise on each frequency component differs depending on the signal distribution and quantization stepsize. It is interesting to note that the error variance does not always monotonically increase or decrease with frequency. This can be understood by considering the interplay of the signal distribution and the quantization stepsize in each frequency band.

In addition to viewing the quantization noise distribution in the transform domain, it is revealing to observe the average noise distribution over blocks in the spatial domain. Unlike what one might expect, this spatial-block noise distribution is not uniform. In fact, for typical DCT coders, the spatial-block noise is often distributed as an inverted pincushion.

To illustrate this point, the \(512 \times 512\) Lena image was compressed with a JPEG-type DCT coder. Portions of the original and compressed images are shown in Figure C.2. A plot of the variance of the transform-domain quantization noise as a function of frequency is shown in Figure C.2. A plot of the spatial block distribution of the quantization noise is also shown.

![Figure C.1: Original and DCT-coded images. The coded Lena image is afflicted with blocking effects.](image)
Figure C.2: Transform coding artifacts due to quantization. The average noise distribution over blocks of the DCT-coded image is shown in the transform and spatial domains. Note the inverted pincushion distribution of the quantization noise in the spatial domain.

C.3 Reduction of Artifacts

C.3.1 Previous Approaches

A number of postprocessing approaches have been proposed to reduce transform coding artifacts. Reeve and Lim reduced blocking artifacts by simply lowpass filtering boundary pixels across the block boundaries [35]. This concept of filtering across block boundaries has been extended to many spatially adaptive postprocessing methods. Higher-performance solutions have been developed based on iterative methods of image restoration. Stevenson developed an iterative method based on a stochastic model of the image and MAP estimation techniques [53]. Yang, Galatsanos, and Katsaggelos developed iterative POCS and constrained least squares methods based on “between-block” smoothness properties of the original image [52].
C.3.2 Wiener-Filter Approach

For an orthogonal transform, if the statistical information is unknown, the prescribed synthesis filters yield the optimum least-squares estimate of the original signal. However, if any statistical information is known about the signal or the noise, an improved estimate can be made. The Wiener filter provides the minimum mean-squared error estimate of the original signal while exploiting second-order statistical descriptions of the signal and the noise.

A new method of reducing transform coding artifacts has been developed based on Wiener filtering; this approach is motivated by the hypothesis that performance can be improved by flattening the spatial-block noise distribution described earlier. The optimal Wiener filter can be applied in one of two ways: 1) as an alternative set of synthesis filters in the decoder, or 2) as a post-processing operation following conventional decoding. These methods are illustrated in Figure C.3. In both methods, the optimal Wiener filter is a function of the transform as well as the covariances of the signal and quantization noise.

![Diagram](image)

Figure C.3: Wiener filtering methods for removing transform coding artifacts. The Wiener filter can be applied as an alternative set of synthesis functions in the decoder or as a post-processing operation following conventional decoding.

A Wiener filter was designed for the JPEG-compressed Lena image. The image was modelled as a first-order Markov process with a correlation coefficient of $\rho = 0.95$. An estimate of the covariance matrix of the noise was made based on the JPEG quantization table. The resulting Wiener filter was applied to the quantized transform coefficients as an alternative set of decoding synthesis filters. The resulting image is shown in Figure C.4. The inverse-DCT image clearly exhibits blocking artifacts; the Wiener-filter image produces a higher-quality image with reduced blocking artifacts. The PSNR of the reconstructed image increases from 33.67 dB to 34.20 dB. Thus, the Wiener-filter method demonstrates improved performance both qualitatively and quantitatively. The resulting spatial-block noise distribution is also shown in Figure C.4. Notice that this method flattened the spatial-block distribution of the quantization noise.
Figure C.4: Wiener filtered image and spatial-block noise distribution. The Wiener filtered image has reduced blocking effects. This results in a flattened spatial-block noise distribution.

C.4 Remarks

A number of comparisons can be made between the previous approaches and the Wiener approach. First, the developed Wiener approach outperforms the method of filtering across block boundaries; the Wiener approach flattens the quantization error over the entire spatial block, while the previous approach only reduces the error on boundary pixels. Next, the iterative image restoration approaches locally impose constraint sets to ensure convergence to a valid solution; after a number of iterations they may achieve improved performance over the Wiener method which does not employ these local checks. However, the non-iterative nature and low computational requirements of the proposed Wiener method make it an attractive solution for achieving improved quality with low complexity. Finally, the proposed Wiener method is based on the global statistics of the image.
This results in a very simple algorithm that does not require adaptive processing. While the simplicity is appealing, it should be noted that further improvement can be achieved by using spatially adaptive methods.

C.5 Summary

A new method was presented for reducing transform coding artifacts. It is based on the concept of Wiener filtering, which provides the minimum mean-squared error estimate of the original image while exploiting second-order statistical descriptions of the signal and the quantization noise. This method was applied to a DCT coder and demonstrated both a qualitative and quantitative improvement in performance. A perspective of observing the average distribution of the quantization noise over the spatial blocks was introduced. It was found that transform coding artifacts, such as the blocking effect in DCT coders, can be better understood by considering a more detailed analysis of the quantization noise in the spatial domain.
Bibliography


