Digital Signal Compression

With clear and easy-to-understand explanations, this book covers the fundamental concepts and coding methods of signal compression, while still retaining technical depth and rigor. It contains a wealth of illustrations, step-by-step descriptions of algorithms, examples, and practice problems, which make it an ideal textbook for senior undergraduate and graduate students, as well as a useful self-study tool for researchers and professionals.

Principles of lossless compression are covered, as are various entropy coding techniques, including Huffman coding, arithmetic coding, run-length coding, and Lempel–Ziv coding. Scalar and vector quantization, and their use in various lossy compression systems, are thoroughly explained, and a full chapter is devoted to mathematical transformations, including the Karhunen–Loeve transform, discrete cosine transform (DCT), and wavelet transforms. Rate control is covered in detail, with several supporting algorithms to show how to achieve it. State-of-the-art transform and subband/wavelet image and video coding systems are explained, including JPEG2000, SPIHT, SBHP, EZBC, and H.264/AVC. Also, unique to this book is a chapter on set partition coding that sheds new light on SPIHT, SPECK, EZW, and related methods.

William A. Pearlman is a Professor Emeritus in the Electrical, Computer, and Systems Engineering Department at the Rensselaer Polytechnic Institute (RPI), where he has been a faculty member since 1979. He has more than 35 years of experience in teaching and researching in the fields of information theory, data compression, digital signal processing, and digital communications theory, and he has written about 250 published works in these fields. He is a Fellow of the IEEE and the SPIE, and he is the co-inventor of two celebrated image compression algorithms: SPIHT and SPECK.

Amir Said is currently a Master Researcher at Hewlett-Packard Laboratories, where he has worked since 1998. His research interests include multimedia communications, coding and information theory, image and video compression, signal processing, and optimization, and he has more than 100 publications and 20 patents in these fields. He is co-inventor with Dr. Pearlman of the SPIHT image compression algorithm and co-recipient, also with Dr. Pearlman, of two Best Paper Awards, one from the IEEE Circuits and Systems Society and the other from the IEEE Signal Processing Society.
Digital Signal Compression
Principles and Practice

WILLIAM A. PEARLMAN
Rensselear Polytechnic Institute, New York

AMIR SAID
Hewlett-Packard Laboratories, Palo Alto, California
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>xv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>xix</td>
</tr>
<tr>
<td>1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.1 The importance of compression</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Data types</td>
<td>2</td>
</tr>
<tr>
<td>1.2.1 Symbolic information</td>
<td>2</td>
</tr>
<tr>
<td>1.2.2 Numerical information</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Basic compression process</td>
<td>4</td>
</tr>
<tr>
<td>1.4 Compression applications</td>
<td>5</td>
</tr>
<tr>
<td>1.5 Design of compression methods</td>
<td>6</td>
</tr>
<tr>
<td>1.6 Multi-disciplinary aspect</td>
<td>8</td>
</tr>
<tr>
<td>Note</td>
<td>8</td>
</tr>
<tr>
<td>References</td>
<td>9</td>
</tr>
<tr>
<td>2 Book overview</td>
<td>10</td>
</tr>
<tr>
<td>2.1 Entropy and lossless coding</td>
<td>10</td>
</tr>
<tr>
<td>2.2 Quantization</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Source transformations</td>
<td>12</td>
</tr>
<tr>
<td>2.3.1 Prediction</td>
<td>12</td>
</tr>
<tr>
<td>2.3.2 Transforms</td>
<td>13</td>
</tr>
<tr>
<td>2.4 Set partition coding</td>
<td>16</td>
</tr>
<tr>
<td>2.5 Coding systems</td>
<td>17</td>
</tr>
<tr>
<td>2.5.1 Performance criteria</td>
<td>17</td>
</tr>
<tr>
<td>2.5.2 Transform coding systems</td>
<td>18</td>
</tr>
<tr>
<td>2.5.3 Subband coding systems</td>
<td>19</td>
</tr>
<tr>
<td>2.6 Distributed source coding</td>
<td>20</td>
</tr>
<tr>
<td>Notes</td>
<td>21</td>
</tr>
<tr>
<td>References</td>
<td>22</td>
</tr>
<tr>
<td>3 Principles of lossless compression</td>
<td>23</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>23</td>
</tr>
<tr>
<td>3.2 Lossless source coding and entropy</td>
<td>23</td>
</tr>
</tbody>
</table>
Contents

3.3 Variable length codes
  3.3.1 Unique decodability and prefix-free codes 28
  3.3.2 Construction of prefix-free codes 28
  3.3.3 Kraft inequality 30
3.4 Optimality of prefix-free codes 32
  3.4.1 Sources with memory 36
3.5 Concluding remarks 37
Problems 37
References 40

4 Entropy coding techniques
  4.1 Introduction 41
  4.2 Huffman codes 41
  4.3 Shannon–Fano–Elias codes 47
    4.3.1 SFE code examples 48
    4.3.2 Decoding the SFE code 49
  4.4 Arithmetic code 50
    4.4.1 Preliminaries 50
    4.4.2 Arithmetic encoding 51
    4.4.3 Arithmetic decoding 53
  4.5 Run-length codes 55
  4.6 Alphabet partitioning: modified Huffman codes 57
    4.6.1 Modified Huffman codes 57
    4.6.2 Alphabet partitioning 58
  4.7 Golomb code 60
  4.8 Dictionary coding 63
    4.8.1 The LZ78 code 64
    4.8.2 The LZW algorithm 65
    4.8.3 The LZ77 coding method 67
  4.9 Summary remarks 72
Problems 72
Notes 75
References 76

5 Lossy compression of scalar sources
  5.1 Introduction 77
  5.2 Quantization 77
    5.2.1 Scalar quantization 77
    5.2.2 Uniform quantization 81
  5.3 Non-uniform quantization 87
    5.3.1 High rate approximations 89
  5.4 Companding 91
    5.4.1 Distortion at high rates 93
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5 Entropy coding of quantizer outputs</td>
<td>95</td>
</tr>
<tr>
<td>5.5.1 Entropy coded quantizer characteristics</td>
<td>98</td>
</tr>
<tr>
<td>5.5.2 Null-zone quantization</td>
<td>99</td>
</tr>
<tr>
<td>5.6 Bounds on optimal performance</td>
<td>101</td>
</tr>
<tr>
<td>5.6.1 Rate-distortion theory</td>
<td>102</td>
</tr>
<tr>
<td>5.6.2 The Gish–Pierce bound</td>
<td>104</td>
</tr>
<tr>
<td>5.7 Concluding remarks</td>
<td>107</td>
</tr>
<tr>
<td>5.8 Appendix: quantization tables</td>
<td>107</td>
</tr>
<tr>
<td>Problems</td>
<td>109</td>
</tr>
<tr>
<td>Note</td>
<td>113</td>
</tr>
<tr>
<td>References</td>
<td>114</td>
</tr>
<tr>
<td>6 Coding of sources with memory</td>
<td>116</td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>116</td>
</tr>
<tr>
<td>6.2 Predictive coding</td>
<td>116</td>
</tr>
<tr>
<td>6.2.1 Optimal linear prediction</td>
<td>117</td>
</tr>
<tr>
<td>6.2.2 DPCM system description</td>
<td>120</td>
</tr>
<tr>
<td>6.2.3 DPCM coding error and gain</td>
<td>121</td>
</tr>
<tr>
<td>6.3 Vector coding</td>
<td>122</td>
</tr>
<tr>
<td>6.3.1 Optimal performance bounds</td>
<td>122</td>
</tr>
<tr>
<td>6.3.2 Vector (block) quantization (VQ)</td>
<td>129</td>
</tr>
<tr>
<td>6.3.3 Entropy constrained vector quantization</td>
<td>135</td>
</tr>
<tr>
<td>6.4 Tree-structured vector quantization</td>
<td>141</td>
</tr>
<tr>
<td>6.4.1 Variable length TSVQ coding</td>
<td>144</td>
</tr>
<tr>
<td>6.4.2 Pruned TSVQ</td>
<td>145</td>
</tr>
<tr>
<td>6.5 Tree and trellis codes</td>
<td>146</td>
</tr>
<tr>
<td>6.5.1 Trellis codes</td>
<td>148</td>
</tr>
<tr>
<td>6.5.2 Encoding and decoding of trellis codes</td>
<td>150</td>
</tr>
<tr>
<td>6.5.3 Codevector alphabets</td>
<td>152</td>
</tr>
<tr>
<td>6.6 Trellis coded quantization (TCQ)</td>
<td>152</td>
</tr>
<tr>
<td>6.6.1 Entropy-coded TCQ</td>
<td>154</td>
</tr>
<tr>
<td>6.6.2 Improving low-rate performance in TCQ</td>
<td>155</td>
</tr>
<tr>
<td>6.7 Search algorithms</td>
<td>155</td>
</tr>
<tr>
<td>6.7.1 M-algorithm</td>
<td>155</td>
</tr>
<tr>
<td>6.7.2 The Viterbi algorithm</td>
<td>158</td>
</tr>
<tr>
<td>6.8 Concluding remarks</td>
<td>160</td>
</tr>
<tr>
<td>Problems</td>
<td>160</td>
</tr>
<tr>
<td>Notes</td>
<td>163</td>
</tr>
<tr>
<td>References</td>
<td>164</td>
</tr>
<tr>
<td>7 Mathematical transformations</td>
<td>166</td>
</tr>
<tr>
<td>7.1 Introduction</td>
<td>166</td>
</tr>
<tr>
<td>7.1.1 Transform coding gain</td>
<td>169</td>
</tr>
</tbody>
</table>
# Contents

7.2 The optimal Karhunen–Loeve transform  
   7.2.1 Optimal transform coding gain  
7.3 Suboptimal transforms  
   7.3.1 The discrete Fourier transform  
   7.3.2 The discrete cosine transform  
   7.3.3 The Hadamard–Walsh transform  
7.4 Lapped orthogonal transform  
   7.4.1 Example of calculation of transform coding gain  
7.5 Transforms via filter banks  
7.6 Two-dimensional transforms for images  
7.7 Subband transforms  
   7.7.1 Introduction  
   7.7.2 Coding gain of subband transformation  
   7.7.3 Realizable perfect reconstruction filters  
   7.7.4 Orthogonal wavelet transform  
   7.7.5 Biorthogonal wavelet transform  
   7.7.6 Useful biorthogonal filters  
   7.7.7 The lifting scheme  
   7.7.8 Transforms with integer output  
7.8 Concluding remarks  

Problems  
Notes  
References

8 Rate control in transform coding systems  
8.1 Rate allocation  
   8.1.1 Optimal rate allocation for known quantizer characteristics  
   8.1.2 Realizing the optimal rate allocation  
   8.1.3 Fixed level quantization  
   8.1.4 Optimal bit allocation for arbitrary set of quantizers  
   8.1.5 Building up to optimal rates for arbitrary quantizers  
   8.1.6 Transform coding gain  
8.2 Subband rate allocation  
   8.2.1 Practical issues  
   8.2.2 Subband coding gain  
8.3 Algorithms for rate allocation to subbands  
8.4 Conclusions  

Problems  
Notes  
References

9 Transform coding systems  
9.1 Introduction
## Contents

9.2 Application of source transformations

9.2.1 Model-based image transform coding

9.2.2 Encoding transform coefficients

9.3 The JPEG standard

9.3.1 The JPEG baseline system

9.3.2 Detailed example of JPEG standard method

9.4 Advanced image transform coding: H.264/AVC intra coding

9.5 Concluding remarks

Problems

Notes

References

10 Set partition coding

10.1 Principles

10.1.1 Partitioning data according to value

10.1.2 Forming partitions recursively: square blocks

10.1.3 Binary splitting

10.1.4 One-dimensional signals

10.2 Tree-structured sets

10.2.1 A different wavelet transform partition

10.2.2 Data-dependent thresholds

10.2.3 Adaptive partitions

10.3 Progressive transmission and bitplane coding

10.4 Applications to image transform coding

10.4.1 Block partition coding and amplitude and group partitioning (AGP)

10.4.2 Enhancements via entropy coding

10.4.3 Traversing the blocks

10.4.4 Embedded block coding of image wavelet transforms

10.4.5 A SPECK coding example

10.4.6 Embedded tree-based image wavelet transform coding

10.4.7 A SPIHT coding example

10.4.8 Embedded zerotree wavelet (EZW) coding

10.4.9 Group testing for image wavelet coding

10.5 Conclusion

Problems

Notes

References

11 Subband/wavelet coding systems

11.1 Wavelet transform coding systems

11.2 Generic wavelet-based coding systems

11.3 Compression methods in wavelet-based systems

© in this web service Cambridge University Press & Assessment

www.cambridge.org
## Contents

11.4 Block-based wavelet transform set partition coding 320
  11.4.1 Progressive resolution coding 321
  11.4.2 Quality-progressive coding 323
  11.4.3 Octave band partitioning 326
  11.4.4 Direct bit-embedded coding methods 328
  11.4.5 Lossless coding of quantizer levels with adaptive thresholds 329
  11.4.6 Tree-block coding 331
  11.4.7 Coding of subband subblocks 332
  11.4.8 Coding the initial thresholds 333
  11.4.9 The SBHP method 335
  11.4.10 JPEG2000 coding 336
  11.4.11 The embedded zero-block coder (EZBC) 343

11.5 Tree-based wavelet transform coding systems 347
  11.5.1 Fully scalable SPIHT 347
  11.5.2 Resolution scalable SPIHT 349
  11.5.3 Block-oriented SPIHT coding 352

11.6 Rate control for embedded block coders 354
11.7 Conclusion 356
Notes 357
References 359

12 Methods for lossless compression of images 361
  12.1 Introduction 361
  12.2 Lossless predictive coding 362
    12.2.1 Old JPEG standard for lossless image compression 362
  12.3 State-of-the-art lossless image coding and JPEG-LS 364
    12.3.1 The predictor 364
    12.3.2 The context 365
    12.3.3 Golomb–Rice coding 366
    12.3.4 Bias cancellation 366
    12.3.5 Run mode 367
    12.3.6 Near-lossless mode 368
    12.3.7 Remarks 368
  12.4 Multi-resolution methods 368
  12.5 Concluding remarks 369
Problems 370
Notes 371
References 372

13 Color and multi-component image and video coding 373
  13.1 Introduction 373
Contents

13.2 Color image representation 374
   13.2.1 Chrominance subsampling 376
   13.2.2 Principal component space 377
13.3 Color image coding 378
   13.3.1 Transform coding and JPEG 378
   13.3.2 Wavelet transform systems 380
13.4 Multi-component image coding 383
   13.4.1 JPEG2000 383
   13.4.2 Three-dimensional wavelet transform coding 384
   13.4.3 Video coding 389
13.5 Concluding remarks 395
Notes 395
References 396

14 Distributed source coding 398
   14.1 Slepian–Wolf coding for lossless compression 398
      14.1.1 Practical Slepian–Wolf coding 400
   14.2 Wyner–Ziv coding for lossy compression 404
      14.2.1 Scalar Wyner–Ziv coding 406
      14.2.2 Probability of successful reconstruction 407
   14.3 Concluding remarks 411
Problems 411
Notes 412
References 413

Index 414
This book is an outgrowth of a graduate level course taught for several years at Rensselaer Polytechnic Institute (RPI). When the course started in the early 1990s, there were only two textbooks available that taught signal compression, Jayant and Noll\(^1\) and Gersho and Gray.\(^2\) Certainly these are excellent textbooks and valuable references, but they did not teach some material considered to be necessary at that time, so the textbooks were supplemented with handwritten notes where needed. Eventually, these notes grew to many pages, as the reliance on published textbooks diminished. The lecture notes remained the primary source even after the publication of the excellent book by Sayood,\(^3\) which served as a supplement and a source of some problems. While the Sayood book was up to date, well written, and authoritative, it was written to be accessible to undergraduate students, so lacked the depth suitable for graduate students wanting to do research or practice in the field. The book at hand teaches the fundamental ideas of signal compression at a level that both graduate students and advanced undergraduate students can approach with confidence and understanding. The book is also intended to be a useful resource to the practicing engineer or computer scientist in the field. For that purpose and also to aid understanding, the 40 algorithms listed under **Algorithms** in the Index are not only fully explained in the text, but also are set out step-by-step in special algorithm format environments.

This book contains far more material than can be taught in a course of one semester. As it was being written, certain subjects came to light that begged for embellishment and others arose that were needed to keep pace with developments in the field. One example of this additional material, which does not appear in any other textbook, is Chapter 14 on “Distributed source coding.” a subject which has received considerable attention lately. The intent was to present the fundamental ideas there, so that the student can understand and put into practice the various methods being proposed that use this technique.

The two longest chapters in the book are Chapters 10 and 11, entitled “Set partition coding” and “Subband/wavelet coding systems,” respectively. They were actually the first chapters written and were published previously as a monograph in two parts.\(^4\) The versions appearing here are updated with some minor errors corrected. Being the inventors of SPIHT and proponents and pioneers of set partition coding, we felt that its fundamental principles were not expounded in the technical literature. Considering the great interest in SPIHT, as evidenced by the thousands of inquiries received by us over the years since its origin in 1995 (at this writing 94,200 hits on Google), we
Preface

were eager to publish a true tutorial on the fundamental concepts of this algorithm. We believe that Chapter 10 fulfills this intent. Other books usually present only the SPIHT algorithm, almost always by working an example without revealing the underlying principles. Chapter 11 describes more image wavelet coding systems than any other book, including the JPEG2000 standard, fully scalable SPIHT, SBHP, and EZBC. The last three are set partition coders, while JPEG2000 contains auxiliary algorithms that use set partitioning. Furthermore, this chapter explains how to embed features of scalability and random access in coded bitstreams.

Besides distributed source coding, some preliminary material are also firsts in this book. They are: analysis of null-zone quantization, rate allocation algorithms, and the link between filters and wavelets. The aforementioned link is explained in Chapter 7 on “Mathematical transformations,” in a way that requires only some knowledge of discrete-time Fourier transforms and linear system analysis. The treatment avoids the concepts of functional analysis and the use of polyphase transforms with little compromise of rigor. The intent was to make the book accessible to advanced undergraduates, who would likely not have exposure to these subjects. Also to serve this purpose, prior exposure to information theory is not a prerequisite, as the book teaches the relevant aspects needed to grasp the essential ideas.

One criticism that might be levied at this book is its emphasis on compression of images. Certainly, that reflects the main field of research and practice of the authors. However, image compression is possible only by understanding and applying the principles of compression that pertain to all source data. In fact, the material of the first eight chapters is generic and dimension independent. The notation is one-dimensional for the most part, and the generalization to higher dimensions is mostly obvious and hence unnecessary. Although the applications are mostly to images in the remainder of the book, except for the generic Chapter 14, the corresponding one-dimensional signal methods are mentioned when applicable and even included in some problems. The standards and state-of-the-art systems of image compression are treated in some detail, as they represent the most straightforward application of basic principles. The standard speech and audio coding systems require additional complications of perceptual and excitation models, and echo cancellation. Their inclusion would make the book too long and cumbersome and not add much to its primary objective. Nevertheless, the material on image compression systems in Chapters 9, 11, and 12 is comprehensive enough to meet the secondary objective of serving as a good tutorial and reference for image compression researchers and practitioners.

Chapter 12 treats the subject of lossless image compression. Lossless image compression is used only for archiving of image data. There seems to be no call for lossless compression of any sources for the purpose of multimedia communication, as the data transfer would be too slow or require too much bandwidth or both. For example, there is no compression of audio or speech in the WAV storage format for compact discs. MP3 is a compressed format for audio and speech transmission and recording; the compressed format of JPEG is standard in every digital camera for consumer use; all digital video is compressed by MPEG or ITU standard formats. Images seem to be the only sources
that are subject to lossless compression. The standard methods described in Chapter 12 serve as good examples of how basic principles are put into practice.

The book did not seem complete without a chapter on how compression is applied to three-dimensional sources, such as color images, volume medical images, and video. At the time of writing, there are no other textbooks that teach three-dimensional wavelet coding methods. Therefore, we wrote Chapter 13 with the intent to show the reader how the methods of the earlier chapters are extended to these higher dimensional sources. We purposely omitted detailed descriptions of video standards. We just explained the general framework in these systems upon which compression is applied and a little about the compression methods, which are mostly covered in detail in earlier chapters.

We urge potential buyers or browsers to read Chapters 1 and 2, which discuss the motivation to learn signal compression and take a brief tour through the book. This book turned out to be different in many respects from what was taught in the RPI course. Roughly, the coverage of that course was all of Chapters 3, 4, 5, and 6, which was deemed essential material. One can be selective in Chapter 7, for example, skipping the lapped orthogonal transform and some parts of Section 7.7, “Subband transforms,” especially the detailed development of the connection between wavelet theory and FIR filters. Likewise, in Chapter 8, some of the rate allocation algorithms may be skipped, as well as the detailed derivations of optimal rate allocation and coding gain. One can skim through Section 9.3 in the next chapter, and skip Section 9.4 on H.264/AVC intra coding, which did not exist when the course was last taught. Time may not allow anything in Chapter 10, other than set partitioning for SPIHT and the accompanying coding example, and in Chapter 11 only a sketch of the JPEG2000 coding system. Lossless image compression in Chapter 12 could be covered earlier in the course, perhaps after Chapter 4, “Entropy coding techniques.” Certainly, there is enough material to accommodate different choices of emphasis and objective.

For students with a background in information theory and signal processing or those more interested in computer science or actual implementation, an instructor may skip some of the preliminary material in the early chapters and teach all of the rate allocation algorithms in Chapter 8 and cover Chapters 10 and 11 in their entirety. Chapter 11 contains much practical material on implementation of coding systems. In fact, we think that approach would be appropriate for a short course.

The book contains many figures and problems. The problems in many cases have to be worked using a calculation and plotting program, such as MATLAB, and sometimes by making a computer program. Some datasets and basic software in C or C++ and MATLAB m-files, will be made freely available on the course website: http://www.cambridge.org/pearlman. Also freely available on the website are PowerPoint animations of the SPIHT and SPECK algorithms. Figures and problem solutions will be made available to instructors.
Notes

Acknowledgments

I wish to acknowledge my debt and gratitude over the years to all my students from whom I have learned so much. Without the experience of interacting with them, I could not have written this book. There is not enough space to list all the contributors and their contributions, as I would like. Instead, I shall mention only those who were directly involved in the content of this book. I wish to thank Sungdae Cho, who created most of the video and three-dimensional SPIHT software and who encouraged me to write this book. He also created the SPIHT animation that is available on the book’s website. I am grateful to Ying Liu, who developed the higher dimensional SBHP algorithms and graciously helped me with my programming difficulties. My gratitude goes out to Asad Islam, who developed the SPECK algorithm, and Xiaoli Tang, who extended it to higher dimensions. I also wish to thank Matthias Narroschke, who created the SPECK animation while I was visiting the University of Hannover, and Emmanuel Christophe, who developed resolution-scalable SPIHT, while he was a visiting student to RPI from the University of Toulouse and TeSA (Telecom Spatiales Aeronautiques). I also wish to thank Professor James Fowler of Mississippi State University for checking my explanation of bisection in two-dimensional block partitioning. I am grateful to Alessandro Dutra, who helped me write the solutions manual for the book’s problems.

A special acknowledgment goes to Amir Said, my co-author and long-time collaborator, who, while a doctoral student at RPI working under Professor John Anderson, came to me to do a special project on image compression. From this project, by virtue of Amir’s intellectual brilliance, creativity, and prodigious programming skills, the SPIHT algorithm was born. There are no words that adequately express my thanks to him.

Lastly, I wish to thank my wife, Eleanor, who suffered from having an often distracted and inattentive husband during more than three years while I was writing this book. This book is dedicated to her, as I could not manage without her love and support.

William A. Pearlman