COMPRESSION AND CLASSIFICATION OF IMAGERY

by

Ali Tabesh

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College of

THE UNIVERSITY OF ARIZONA

2006

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Ali Tabesh

entitled Compression and Classification of Imagery

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: August 3, 2006
Dr. Michael W. Marcellin

_____ Date: August 3, 2006
Dr. Mark A. Neifeld

_____ Date: August 3, 2006
Dr. Bane Vasic

_____ Date: _____

_____ Date: _____

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: August 3, 2006
Dissertation Director: Dr. Michael W. Marcellin

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: _____
Ali Tabesh

ACKNOWLEDGEMENTS

I am indebted to many people for their contributions to my personal, intellectual, and professional growth during the course of my doctoral work.

First, I would like to express my most sincere gratitude to my dissertation advisor, Professor Michael Marcellin, for all his support, guidance, and patience.

I am deeply indebted to my dear friend, Dr. Ali Bilgin, without whose help I would not have come this far in my studies. He and my friends, Drs. Amy Wang and Farid Masrour, never ceased to support me on personal and professional levels, particularly in times of need.

I am also grateful to my fellow graduate students in the Signal Processing and Coding Laboratory for their friendship and help.

I would like to thank Professors Mark Neifeld and Bane Vasic for accepting to be members of my defense committee, and Ms. Tami Whelan for handling the paperwork necessary for the completion of my degree.

Last, but not least, I would like to thank my parents. Pursuing my studies would have been impossible without their unconditional love and encouragement. This dissertation is dedicated to them.

To my parents,

Parvin and Hassan Tabesh

TABLE OF CONTENTS

TABLE OF CONTENTS − *Continued*

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Problems at the intersection of compression and statistical inference recur frequently due to the concurrent use of signal and image compression and classification algorithms in many applications. This dissertation addresses two such problems: statistical inference on compressed data, and rate-allocation for joint compression and classification.

Features of the JPEG2000 standard make possible the development of computationally efficient algorithms to achieve such a goal for imagery compressed using this standard. We propose the use of the information content (IC) of wavelet subbands, defined as the number of bytes that the JPEG2000 encoder spends to compress the subbands, for content analysis. Applying statistical learning frameworks for detection and classification, we present experimental results for compressed-domain texture image classification and cut detection in video. Our results indicate that reasonable performance can be achieved, while saving computational and bandwidth resources. IC features can also be used for preliminary analysis in the compressed domain to identify candidates for further analysis in the decompressed domain.

In many applications of image compression, the compressed image is to be presented to human observers *and* statistical decision-making systems. In such applications, the fidelity criterion with respect to which the image is compressed must be

selected to strike an appropriate compromise between the (possibly conflicting) image quality criteria for the human and machine observers. We present tractable distortion measures based on the Bhattacharyya distance (BD) and a new upper bound on the quantized probability of error that make possible closed form expressions for rate allocation to image subbands and show their efficacy in maintaining the aforementioned balance between compression and classification. The new bound offers two advantages over the BD in that it yields closed-form solutions for rate-allocation in problems involving correlated sources and more than two classes.

CHAPTER 1: INTRODUCTION

1.1 Motivation and Problem Statement

Various signal processing applications involve problems at the intersection of compression and statistical inference. For instance, many tasks require making inferences about the content of imagery compressed prior to transmission or storage. Information retrieval, medical diagnosis, remote sensing, and security and military applications are examples of such applications wherein inferential tasks such as detection, estimation, segmentation, and classification are to be performed. For computational and bandwidth savings reasons, it is highly desirable to be able to make these inferences using the compressed-domain information. For instance, in video surveillance applications the video is usually acquired, compressed, and transmitted continuously. If the received codestream can be processed to detect suspicious activity, decompression of the unnecessary portions can be avoided. Furthermore, significant bandwidth savings can be achieved if the portions of a codestream that are of interest to a client can be identified and delivered instead of the entire codestream.

In view of such necessities, the JPEG committee has recently issued a call for contributions for the standardization of technologies associated with searching image libraries [1]. This new effort is referred to as *JPSearch*. Compressed-domain analysis

techniques are one of the technologies that are sought by JPSearch.

Another problem at the interface of compression and statistical inference arises when one wishes to design a compression system that preserves not only the perceptually important features in the imagery, but also features vital to statistical decision-making systems. For example, in a target recognition task potential targets are first detected and classified using statistical methods to reduce the burden on the human operator, and then the human operator authenticates the potential targets by visual inspection. Similar scenarios occur in other areas such as medical imaging and remote sensing. In such applications, the fidelity criterion with respect to which the image is compressed must be selected such that an appropriate balance between the (possibly conflicting) image quality criteria for the different observers is maintained.

The mean-square error (MSE), which is the most widely used distortion measure in designing compression systems, is a suitable distortion measure for human observation at high rates. When statistical decision-making is concerned, however, the MSE is not the most appropriate criterion. This has motivated the introduction of a variety of distortion criteria for tasks involving statistical decisions [2, 3, 4, 5, 6]. To accommodate both human and statistical observers, many authors have proposed a weighted combination of the MSE and a classification criterion [2, 3, 4, 5] as the distortion measure for compression system design.

1.2 Organization and Contributions

The two problems stated above are the focus of this dissertation and are addressed

in Chapters 3, 4, and 5. Before concentrating on these problems, some background topics are reviewed in Chapter 2. First, a brief overview of the JPEG2000 standard is given with a focus on the ideas related to the work presented in Chapter 3. Next, basics of optimal rate-allocation to independent sources under the fine quantization regime are reviewed. These basics form the foundation for the work presented in Chapters 4 and 5.

In Chapter 3, we show how some features of the JPEG2000 standard make possible the development of computationally efficient algorithms to analyze the content of imagery compressed using this standard. We define the IC features as the number of bytes that the JPEG2000 encoder spends to compress the subbands and demonstrate their efficacy in two example applications, namely, texture image classification and video cut detection. IC features have not been previously defined and exploited for content analysis in a statistical framework.

Chapter 4 presents a reverse water filling-type rate-allocation algorithm for joint compression and classification using a linear combination of compression and classification distortion criteria, where the compression criterion is the MSE and the classification criterion is the BD between the class-conditional distributions. The combined criterion is closely related to that proposed in [4, 7] for transform coder design. We extend the application of the criterion to subband coders. Also, we present the correct expression for rate-allocation to independent Gaussian sources [7, 8], and generalize the expression to independent sources having generalized Gaussian distributions (GGD's).

Noting the analytical intractability of the BD for rate-allocation to correlated sources and its inherent limitation in handling problems with more than two classes, we

also propose a new bound on the quantized probability of error in Chapter 5. We use the new bound to derive closed-form expressions for rate-allocation to two sources for which the Bayes-optimal decision boundary is a piecewise linear function.

Chapter 6 concludes the dissertation with a summary, discussion of the results, and future directions.

CHAPTER 2: BACKGROUND

In this chapter, we briefly review some of the technical background needed for the material presented in Chapters 3, 4, and 5. In the next section, we present an overview of JPEG2000 focusing on concepts related to the ideas presented in Chapter 3. A comprehensive treatment of JPEG2000 is provided in [9]. In Section 2.2, we provide a review of optimal rate-allocation to independent sources under the fine quantization regime. A more detailed review of this topic is given in [9].

2.1 Overview of JPEG2000

The block diagram of a representative JPEG2000 encoder is given in Figure 2.1. The first stage of encoding consists of (optionally) dividing the input image into non-overlapping rectangular *tiles*. For multi-component images, e.g., color images, an

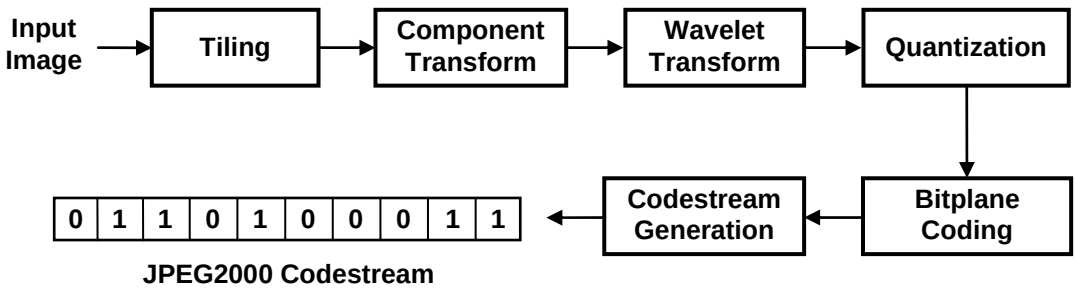


Figure 2.1. Block diagram of a representative JPEG2000 encoder.

optional component transform can be applied to decorrelate the components. The transformed components of each tile are referred to as *tile-components*. A wavelet transform is then applied to each tile-component and the resulting wavelet subband coefficients are partitioned into rectangular blocks called *codeblocks*. After being quantized, the wavelet coefficients in each codeblock are entropy coded independently of other codeblocks. Entropy coding is carried out via context-dependent, binary arithmetic coding of bitplanes. The bitplane coder makes three passes over each bitplane of a codeblock. These passes are referred to as *coding passes*. Finally, the encoder forms a codestream by including coding passes selected based on a desired rate-distortion criterion.

The structure of a simple JPEG2000 codestream is given in Figure 2.2. This structure is explained via the notions of *precinct* and *packet*. A precinct is formed by grouping together the codeblocks that correspond to a particular spatial location at a given resolution. Compressed data from each precinct are arranged to form a packet. Each packet contains a *header* and a *body*. The packet header contains information about

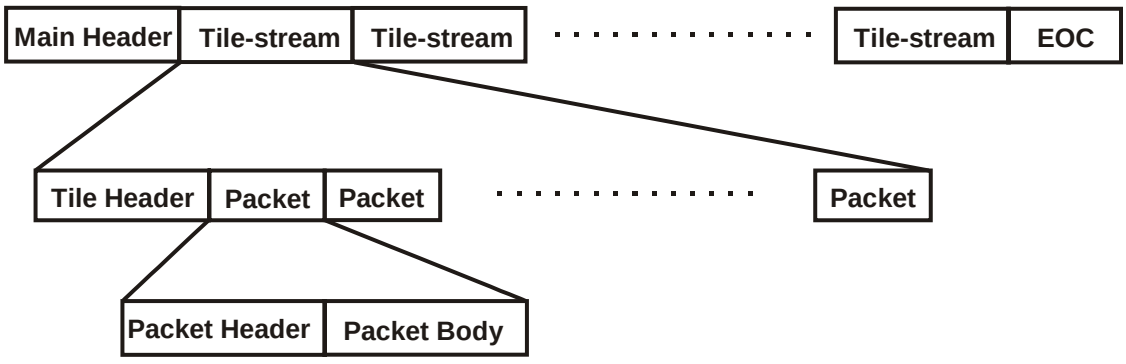


Figure 2.2. A simple JPEG2000 codestream.

the contribution of each codeblock in the precinct to the packet, while the body contains compressed coding passes from the codeblocks. Packets that belong to a particular tile are grouped together to form a *tile-stream*, and tile-streams are grouped together to form the JPEG2000 codestream. Similar to packets, tile-streams are composed of a header and a body. The EOC marker indicates the end of the codestream.

Motion JPEG2000 is the extension of JPEG2000 for video compression [10]. Once each video frame is compressed independently using JPEG2000, the JPEG2000 codestreams may be wrapped to form a single motion JPEG2000 file.

2.2 Review of Optimal Rate Allocation [9]

Let D denote the MSE between a vector source (image), $\mathbf{x}$, and its quantized version, $\tilde{\mathbf{x}}$. For transform (and subband) coders, the relationship between D and the MSE for the individual transform coefficients, $D_i(R_i)$, $i = 1,...,N$, can be written as

$$D(R) = \sum_{i=1}^{N} g_i D_i(R_i), \tag{2.1}$$

where R_i denotes the rate allocated to coefficient i and N is the number of coefficients. Synthesis (inverse transform) gain g_i associated with coefficient i is one if the transform is orthonormal. For wavelet transforms used in JPEG2000, g_i are given in [9]. The constraint on the total rate is given by

$$\sum_{i=1}^{N} R_i \le NR. \tag{2.2}$$

where R is the average rate available to the transform coefficients. The rate-allocation

problem can be stated as

$$\text{minimize (2.1) subject to (2.2).} \tag{2.3}$$

For large R_i, $D_i(R_i)$ is given by

$$D_i(R_i) = \varepsilon_i^2 \sigma_i^2 2^{-2R_i}, \tag{2.4}$$

where ε_i^2 is a constant determined from the distribution of coefficient i, and σ_i^2 is the variance of coefficient i. Note that $D_i(R_i)$ is a linear function of σ_i^2, while it is an exponential function of $-R_i$. In other words, doubling σ_i^2 increases $D_i(R_i)$ two-fold, whereas increasing R_i by half a bit cuts $D_i(R_i)$ in half.

For efficient entropy-coded uniform scalar quantization, i.e., for R_i approaching the entropy of $\tilde{x}_i$, $H(\tilde{x}_i)$, the quantization step size Δ_i is given by

$$\Delta_i = 2^{h(x_i)-R_i}. \tag{2.5}$$

Here, $h(x_i)$ denotes the differential entropy of x_i in bits defined as

$$h(x_i) = -\int p(x_i) \log_2 p(x_i) dx_i, \tag{2.6}$$

where $p(x_i)$ denotes the distribution of x_i. Given that $D_i = \Delta_i^2/12$ for high rates, (2.4) and (2.5) yield

$$\varepsilon_i^2 \sigma_i^2 = \frac{1}{12} 2^{2h(x_i)}. \tag{2.7}$$

The minimization problem (2.3) can be solved for R_i using the method of Lagrange multipliers. Set

$$\frac{\partial}{\partial R_i}\left(\sum_{i=1}^{N} g_i \varepsilon_i^2 \sigma_i^2 2^{-2R_i} + \lambda \sum_{i=1}^{N} R_i\right) = 0\,, \quad i = 1,...,N\,, \tag{2.8}$$

where λ is the Lagrange multiplier. Taking the derivative, (2.8) yields

$$g_i D_i(R_i) = \frac{\lambda}{2\ln 2}\,, \quad i = 1,...,N\,. \tag{2.9}$$

That is, at the optimal solution, the weighted distortions $g_i D_i(R_i)$ for the quantized coefficients should be equal. This is referred to as *reverse water-filling*.

After a few manipulations, (2.9) yields

$$R_i = R + \frac{1}{2}\log_2 \frac{g_i \varepsilon_i^2 \sigma_i^2}{\left(\prod_{j=1}^{N} g_i \varepsilon_i^2 \sigma_i^2\right)^{1/N}}\,. \tag{2.10}$$

Note that (2.10) holds only when R is large enough to guarantee $R_i \geq 0$. When this assumption does not hold, with the assumption that the distortion model (2.4) is still valid, the solution is given by

$$R_i(\lambda) = \max\left\{0, \frac{1}{2}\log_2 \frac{2\ln 2\, g_i \varepsilon_i^2 \sigma_i^2}{\lambda}\right\}, \tag{2.11}$$

where the value of λ must be adjusted to accommodate (2.2). This can be achieved in an iterative fashion.

CHAPTER 3: CONTENT ANALYSIS OF JPEG2000-COMPRESSED IMAGERY

3.1 Previous Work

Myriad algorithms have been developed in the past for content analysis of images and video compressed using the JPEG and MPEG standards [11]. Although codestreams produced by these standards are structurally different than JPEG2000 codestreams, the essence of some ideas can still be used. Our work was motivated by [12] wherein the *coding cost*, i.e., the number of bits spent to encode image blocks, is used for segmentation of JPEG-compressed compound documents into text, graphics, half-tones, continuous tone images, and background. Note, however, that the coding cost is not readily available in the JPEG codestream and must be stored with the coded data at compression time.

Little work has been done on the analysis of JPEG2000 codestreams. In [13, 14] wavelet-domain features have been proposed for image retrieval. The proposed features can be computed at compression time. However, they do not take advantage of the particular information available in JPEG2000 codestream headers.

In [15], two representations have been proposed for indexing JPEG2000-compressed images. One representation is based on the information about the

significance status of wavelet coefficients. For the lowest resolution lowpass subband, the significance map of the wavelet coefficients at all bitplanes is used as an index. The histogram $h(b,r)$ of the number of significant bits at a bitplane b for subbands constituting a resolution level r is used as another index. This representation requires decompressing the significance propagation and cleanup passes of the codestream, if the RESTART marker, which allows for the identification of individual coding passes in the compressed domain, is used at compression time. Otherwise, the entire codestream must be decompressed. The second representation proposed in [15] uses as index the means and variances of the number of nonzero bitplanes in the codeblocks of each subband. This technique takes advantage of some, but not all, of the information in the JPEG2000 codestream headers for image description.

In [16], similar ideas to those presented in this chapter have been proposed for image scaling and cropping for image display applications. The scaling and cropping parameters are determined such that the most important portions of the image to be displayed are retained. The importance of the codeblocks is measured by the number of bits allocated to them, which can be determined from the packet headers of the compressed codestream.

In [17], the difference between the number of bits allocated to horizontal and vertical detail subbands of the highest resolution level, which indicates the level of texture directionality, is used for detecting barcode candidates in JPEG2000-compressed document images. The codeblocks in the horizontal and vertical detail subbands of the highest resolution level corresponding to barcode candidates are consequently

decompressed and the barcode candidates are subjected to further verification and refinement in the wavelet domain.

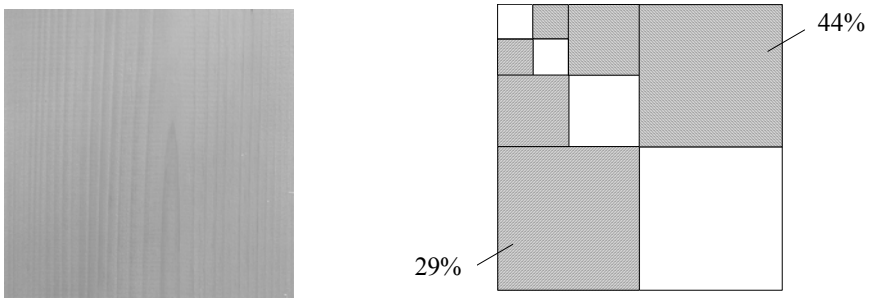
The chapter is organized as follows. In the next section, we define the IC of a subband and intuitively relate it to image texture characteristics. Next, we describe statistical frameworks for detection and classification of imagery using IC features. Finally, we present experimental results using these frameworks for texture image classification and cut detection in video.

3.2 Information Content of Subbands

We define the IC of a wavelet subband as the number of bytes that the JPEG2000 entropy coder spends on encoding that subband. As described in the previous section, the JPEG2000 codestream consists of a series of packets together with additional header information. Each packet header contains information about the coding passes included in the packet, such as the number of bytes that each codeblock within the precinct contributes to a given packet. Packet headers can be sequentially identified starting from the packet header immediately following the start-of-data (SOD) marker, which comes after the tile header. Packet headers are compressed using tag tree coding. Once the beginning of a packet header is identified, it needs to be tag tree-decoded [9]. The contributions of codeblocks to the packet are then obtained from the decoded packet headers. The contribution information is then used to skip the packet body, which contains the arithmetically coded data, and reach the beginning of the next packet header. This procedure is continued until the end of the codestream is reached. Finally, the

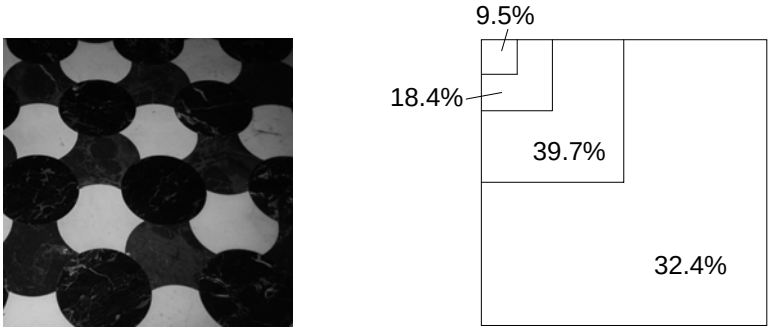
contribution information is aggregated for all codeblocks within each subband to obtain the IC of that subband. Thus, the IC of a subband can be obtained in a fast and computationally efficient manner by simply reading the headers of the packets for the corresponding resolution, and accumulating the size information for all the codeblocks within the subband. It is worth reiterating that obtaining such information from the packet headers does not require arithmetic decoding of any data.

Intuitively, the IC of subbands convey information about the texture characteristics of the image. Two characteristics captured by the subband IC's are texture orientation and texture coarseness. This is demonstrated using the IC data for texture images from the Vision Texture (VisTex) database [18] with different orientation and coarseness characteristics. The images were compressed using the Verification Model (VM) version 9.0 [19] implementation of JPEG2000. All images were compressed at 1 bit/pixel, using 64×64 codeblocks, and three wavelet decomposition levels. Figure 3.1 shows the vertically oriented texture image Wood.0002 and the proportion of IC in each of its subbands relative to the total IC of the image. It can be seen that a higher proportion of image IC is in the horizontal detail subbands (44%) than the vertical detail subbands (29%). Figure 3.2 shows the coarse texture image Tile.0003 and the fine texture image Fabric.0018 and the proportion of IC for each of the four resolution levels. The fineness of Fabric.0018 is reflected by the larger proportion of its IC in the highest resolution level, whereas the coarseness of Tile.0003 is indicated by a larger proportion of its IC in the lowest resolution level.

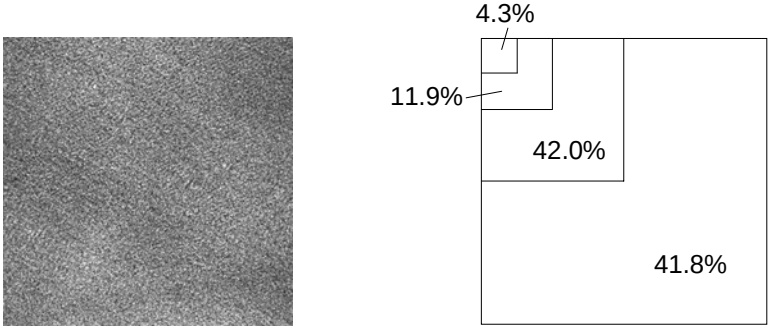


(a) (b)

Figure 3.1. (a) Texture image Wood.0002; and (b) the proportion of IC in the horizontal and vertical detail subbands of its wavelet transform.



(a) (b)



(c) (d)

Figure 3.2. (a) Texture image Tile.0003; and (b) the proportion of IC in each of the resolution levels of its wavelet transform. (c) Texture image Fabric.0018; and (d) the proportion of IC in each of the resolution levels of its wavelet transform.

3.3 Statistical Inference Using Information Content Features

The examples given in the previous section provide intuitive evidence that the IC of subbands can be used to characterize the texture properties of JPEG2000-compressed imagery. In this section, we present example algorithms that make use of this information to carry out detection and classification tasks. The specific applications of these algorithms are texture image classification and video cut detection. However, any other application that can be cast as detection or classification is addressable using these and other algorithms. Note that the objective in this research is to demonstrate the utility of IC features in inferential tasks, and that the following algorithms may not necessarily be optimal for a specific task and/or dataset.

3.3.1 Application to Detection

Let $\mathbf{X}$ denote the $N \times 1$ vector of IC features, where $N = 3L + 1$, and L is the number of wavelet decomposition levels used for compressing the imagery. The objective in detection is to decide which one of the null H_0 or alternative H_1 hypotheses generated $\mathbf{X}$. For instance, in cut detection, H_0 and H_1 constitute the absence or presence of a cut between two consecutive video frames. Various assumptions can be made regarding the nature and structure of H_i, each giving rise to a different detector. We consider two possibilities here.

One possibility is to assume that H_0 has a parametric structure, i.e.,

$$H_0 : \mathbf{X} \sim p(\mathbf{X}; \Theta_0), \tag{3.1}$$

where $p(.)$ is a known distribution parameterized by Θ_0, which is estimated from the

image data, and

$$H_1 : \mathbf{X} \not\sim p(\mathbf{X};\Theta_0) . \tag{3.2}$$

In this case, the detector can be expressed as

"do not reject H_0, if $p(\mathbf{X};\Theta_0) > t$; reject H_0 otherwise,"

where the threshold t controls the trade-off between false positives and false negatives. The smaller t is, the larger the probability of false positives becomes. Note that throughout this work, and for a given false positive/false negative trade-off, we use a global t value for all $\mathbf{X}$.

In this work, we consider the case where $p(.)$ is Gaussian, i.e.,

$$p(\mathbf{X};\boldsymbol{\mu}_0,\mathbf{R}_0) = \frac{1}{(2\pi)^{N/2}\,|\mathbf{R}_0|^{1/2}} \exp(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_0)^T \mathbf{R}_0^{-1}(\mathbf{X}-\boldsymbol{\mu}_0)) , \tag{3.3}$$

where $\boldsymbol{\mu}_0$ and $\mathbf{R}_0$ are the mean vector and covariance matrix of $\mathbf{X}$ under H_0. Note that while $p(.)$ may be non-Gaussian, the Gaussian model still works well in many non-Gaussian cases [25]. Intuitively, it is reasonable to compare the normalized distance $(\mathbf{X}-\boldsymbol{\mu}_0)^T \mathbf{R}_0^{-1}(\mathbf{X}-\boldsymbol{\mu}_0)$ to a properly selected threshold to perform detection.

The other possibility is to view $\mathbf{X}$ as a probability distribution (after normalization), signifying the likelihood of allotting a bit to the subbands. In this case,

$$\begin{aligned} H_0 &: \mathbf{X} = p_0 , \\ H_1 &: \mathbf{X} \neq p_0 , \end{aligned} \tag{3.4}$$

where p_0 is a distribution estimated from the data corresponding to H_0. The detector in this case becomes

"do not reject H_0, if $d(\mathbf{X}, p_0) < t$; reject H_0 otherwise,"

where the test statistic $d(.,.)$ measures the deviation of $\mathbf{X}$ from p_0 and can be any of many possible goodness-of-fit measures. We use the ubiquitous two-sample χ^2 statistic given by [20]

$$\chi^2 = \sum_i \frac{(\sqrt{N_p / N_X}\, X_i - \sqrt{N_X / N_p}\, N_{pi})^2}{X_i + N_{pi}}, \tag{3.5}$$

where N_{pi}, $i = 1,...,N$, denotes the unnormalized version of p_{0i}, i.e.,

$$N_{pi} = p_{0i} \sum_j N_{pj},$$

$N_p = \sum_i N_{pi}$, and $N_X = \sum_i X_i$. Note that p_i inherently has a finite number of bins, which makes the χ^2 statistic a suitable choice. Also, the two-sample statistic is used to accommodate the fact that p_0 is estimated from the image data.

The chi-square detector is different from the parametric detector in two aspects. First, it assumes that $\mathbf{X}$ is a distribution rather than a random vector. Second, there is no prior assumption about the structure of p_0.

3.3.2 Application to Classification

The IC distribution of an image can be used for classification as well. Let $\Omega = \{c_i;\ i = 1,...,C\}$ denote the set of possible class labels for $\mathbf{X}$, where C is the number of classes. We classify $\mathbf{X}$ according to the MAP rule given by

$$i = \arg\max_i p(c_i)\, p(\mathbf{X} \mid c_i), \tag{3.6}$$

where $p(c_i)$ is the prior probability of c_i and $p(\mathbf{X} \mid c_i)$ is the distribution of $\mathbf{X}$ given c_i. Equivalently, we can use $p(c_i \mid \mathbf{X})$ in the above rule, depending on which estimate is

more convenient to obtain. The values of $p(c_i)$ and $p(\mathbf{X}\,|\,c_i)$ are estimated from the training set. The value of $p(c_i)$ is estimated simply as the proportion of samples in the training set that belong to class c_i. For estimating $p(\mathbf{X}\,|\,c_i)$, we can use either a parametric or non-parametric form. In our experiments, we considered the Gaussian distribution and the *k*-nearest neighbor (*k*NN) estimate as a parametric and a non-parametric estimate of $p(\mathbf{X}\,|\,c_i)$, respectively. The Gaussian distribution is given by

$$p(\mathbf{X}\,|\,c_i) = \frac{1}{(2\pi)^{N/2}\,|\,\mathbf{R}_i\,|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^T \mathbf{R}_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i)\right), \tag{3.7}$$

where $\boldsymbol{\mu}_i$ and $\mathbf{R}_i$ are the mean vector and covariance matrix of class c_i.

The *k*NN estimate is given by [25]

$$p(c_i\,|\,\mathbf{X}) = \frac{n_i}{k}, \tag{3.8}$$

where n_i is the number of samples belonging to class c_i that fall within the set of k *nearest* samples to $\mathbf{X}$. Note that we are using the equivalent form of the MAP rule. In order to define "nearness" we need to use a distance measure $d(.,.)$ between $\mathbf{X}$ and the samples $\mathbf{Y}_i$ in the training set. In our experiments, we use the weighted Euclidean distance given by

$$d(\mathbf{X},\mathbf{Y}_i) = \sum_{j=1}^{N} \frac{(X_j - Y_{ij})^2}{s_j^2}, \tag{3.9}$$

where s_j^2 denotes the sample variance of Y_{ij}. Scaling the features to the same standard deviation prevents features with larger variances from dominating the Euclidean distance.

3.4 Results

We assessed the efficacy of IC features in video cut detection and texture image classification. The VM version 9.0 [19] implementation of JPEG2000 was used in the experiments. The cut detection algorithms were implemented using MATLAB [21] and the texture classification algorithms were realized using Tooldiag [22] and MATLAB.

3.4.1 Video Cut Detection

The data for cut detection consisted of a segment from the movie *Batman Returns*, as well as five sequences from the TRECVID-2001 database [24]. The ground truth for *Batman Returns* was established by visual examination of the frames. For the TRECVID-2001 database, the ground truth provided in [24][1] was used. Information on the video sequences used in the experiments is listed in Table 3.1.

The video sequences were losslessly compressed using five wavelet transform levels, 32×32 codeblocks, and the reversible color transform. In evaluating all cut detection algorithms, only the Y (intensity) component of the frames was used.

For the χ^2 statistic, p_0 was set to the IC distribution of the frame preceding the cut candidate. The χ^2 statistic was computed by comparing p_0 to the IC distribution of the frame following the cut candidate.

---

[1] We believe there are errors (typos, false positives/negatives) in the supplied ground truth files. However, for purposes of reproducibility and fairness in comparison, we used the files without any corrections. We regarded as "cuts" transition entries in the files whose beginning and end indices were different by one frame. This was more accurate than using the "CUT" identifier in the entries.

Table 3.1. Video sequences used for cut detection experiments.

Title	Source	Frame size	# of frames	# of cuts
Batman	*Batman Returns*	640×480	20000	334
anni005	[24]	320×240	11363	38
anni009	[24]	320×240	12306	39
bor03	[24]	352×240	48450	230
bor08	[24]	352×240	50568	378
senses111	[24]	352×240	86788	292

* Data for video sequences from [24] are according to the ground truth files supplied with the video files.

For the Gaussian statistic, $\boldsymbol{\mu}_0$ and $\mathbf{R}_0$ were obtained from W frames preceding the cut candidate. The likelihood of the frame following the cut candidate was obtained given the distribution parameterized by $\boldsymbol{\mu}_0$ and $\mathbf{R}_0$. We considered four choices for $\mathbf{R}_0$, namely, $\boldsymbol{\Sigma}$, $\text{diag}(\boldsymbol{\Sigma})$, $(\text{tr}(\boldsymbol{\Sigma})/N)\mathbf{I}$, and $\mathbf{I}$, where $\boldsymbol{\Sigma}$ is the full covariance matrix of the IC features in the window of size W, $\text{diag}(.)$ denotes a matrix with the same main diagonal elements as those of the argument and zeros elsewhere, $\text{tr}(.)$ is the matrix trace, N is the number of IC features, and $\mathbf{I}$ is the identity matrix.

For $\mathbf{R}_0 = \mathbf{I}$, W was set to 1, and for other choices of $\mathbf{R}_0$ the values $\{2,\ 5,\ 10, 15\}$ were considered for W. The experiments showed that $\mathbf{R}_0 = \text{diag}(\boldsymbol{\Sigma})$ together with $W = 10$ yielded the best results.

Note that when $\mathbf{R}_0$ was rank-deficient, a regularized version was used. For $\mathbf{R}_0 = \text{diag}(\boldsymbol{\Sigma})$, the regularized matrix was

$$\mathbf{R}_0 = (1-\lambda)\,\text{diag}(\boldsymbol{\Sigma}) + \lambda(\text{tr}(\boldsymbol{\Sigma})/N)\mathbf{I}, \tag{3.10}$$

where the regularization parameter λ was set to 0.001.

In the experiments, we observed that excluding the IC features for higher resolution subbands from analysis improved results, particularly for the TRECVID-2001 video sequences. For TRECVID-2001 sequences, this is probably due to the MPEG compression noise visible in the sequences, which render the IC of higher resolution subbands ineffective. For each detector, we found the number of lower resolution levels L out of a total of six levels at which the accuracy of the detector was maximized. Table 3.2(a) shows the minimum number of errors (total number of false negatives and false positives) in cut detection for the χ^2 and the Gaussian detectors using the optimal L, given in parentheses. The minimum error count was found by varying the decision threshold between the lowest and highest statistic values for the sequence. The table indicates that the Gaussian detector consistently outperforms the χ^2 detector for all the video sequences, implying that the Gaussian model is a more suitable choice for modeling the IC features than the χ^2 model.

For comparison, cut detection was also performed using three baseline, pixel-domain statistics. The statistics considered are the following [11]:

- *Pixel-wise MSE between frames:* The MSE e_n between the intensities of frames $n-1$ and n, is given by

$$e_n = \frac{1}{N_r N_c} \sum_{i,j} (I_{ij}^n - I_{ij}^{n-1})^2 , \tag{3.11}$$

where I_{ij}^n denotes the intensity of frame n at pixel (i, j) and N_r and N_c are the dimensions of the frame.

- *Block-wise MSE between frames:* This statistic, which is particularly suited for the MPEG compressed domain, is calculated as the MSE between averaged intensity values of video frames, where averaging is performed on non-overlapping 8×8-pixel blocks.
- χ^2 statistic between the gray-level histograms of frames: In this method, the χ^2 statistic given by

$$c_n = \sum_i \frac{(m_i^n - m_i^{n-1})^2}{m_i^n + m_i^{n-1}} \tag{3.12}$$

is computed between the intensity histograms of consecutive frames, where m_i^n denotes the number of pixels in frame n having intensity value i. Note that (3.12) has a simpler form than (3.5) due to the fact that the histograms to be tested in (3.12) have the same total number of samples.

The above statistics were used in the following procedure to detect cuts. If $T_n > t$, where T_n is the value of the statistic for frame n and t is a global threshold, a cut is declared between frames $n-1$ and n. Otherwise, no cut is declared. The cut detection results using the baseline statistics are given in Table 3.2(a). Overall, the MSE detector has the highest performance among the pixel-domain detectors, followed closely by the block MSE detector. The pixel-domain χ^2 detector performs poorly on the video sequences. Its poor performance may be attributed to the non-local nature of frame histograms from which the χ^2 statistic is obtained.

Previous research (e.g., [23]) has indicated that imposing temporal constraints on cut candidates significantly improves detection accuracy. We used the temporal

constraints suggested in [23] with a slight modification:

Declare a cut between frames $n-1$ and n if $T_n > t$ and

1. T_n is equal to the maximum of T_i values within a symmetric window of size $2m-1$ about frame n, i.e., $i = n-m+1,...,n+m-1$, and
2. T_n is larger than l times the second largest maximum in the window.

Parameter m is set to the minimum duration allowed between two consecutive cuts. We chose to restrict the minimum duration to half a second, thus setting $m = 15$ for a video frame rate of 30 frames/second. The value of l was set to 3 following [23]. Table 3.2(b) shows the results after imposing the temporal constraints. Note that the performance of all detectors benefited from the temporal constraints. However, the Gaussian detector made the least gain from the constraints. This is due to the fact that the Gaussian detector already uses a window for computing the covariance matrix $\mathbf{R}_0$, incorporating the local temporal information into the detector decisions.

The better performance of pixel-domain detectors compared to that of detectors using IC features may be attributed to the spatially localized nature of MSE and block MSE features. This explanation is based on our previous observation regarding the poor performance of pixel-domain χ^2 features.

Table 3.2. Experimental results for cut detection. (a) Number of errors and corresponding threshold values for IC features with optimal levels of decomposition and baseline statistics, without time constraints; (b) number of errors after incorporating temporal constraints.

(a)

Title	# of errors / t_{opt}									
	IC (χ^2) ($L=3$)		IC (Gaussian) ($L=5$; $W=10$)		MSE		Block MSE		χ^2	
Batman	224	1.241	121	154.1	144	1109	147	865.9	238	33760
anni005	20	1.331	7	359.4	4	2963	4	3497	28	61250
anni009	24	1.359	6	177.3	3	2915	6	2714	30	57480
bor03	170	1.234	50	154.0	17	1525	15	1027	134	16930
bor08	209	1.030	86	174.5	122	2224	129	1629	125	21560
senses111	124	0.7596	27	193.6	7	2194	8	1575	33	9692
Total	771	n/a	297	n/a	297	n/a	309	n/a	588	n/a

(b)

Title	# of errors / t_{opt}									
	IC (χ^2) ($L=4$)		IC (Gaussian) ($L=5$; $W=10$)		MSE		Block MSE		χ^2	
Batman	93	1.733	115	154.1	94	873.7	97	804.0	84	0
anni005	12	0	7	359.4	3	2943	3	1354	9	0
anni009	8	0	6	177.3	2	2434	2	1432	10	0
bor03	57	0	44	154.0	11	1303	7	472.4	45	10260
bor08	113	0	86	134.6	62	974.6	51	267.0	55	0
senses111	61	0	16	103.2	3	970.6	3	852.3	5	7000
Total	344	n/a	274	n/a	175	n/a	163	n/a	208	n/a

We also compared the performance of the detectors on the TRECVID-2001 sequences to that of the system designed at IBM, which achieved the highest overall performance in the TRECVED-2001 benchmarking effort. The comparison was performed using the software SBSoft provided in [24]. SBSoft tolerates misalignments of up to five frames between detected cuts and the ground truth, leading to slightly smaller

Table 3.3. Numbers of errors using IC Gaussian, block MSE, and IBM detectors. IC Gaussian and block MSE detectors are the same as the ones in Table 2(b). Error counts were obtained using SBSoft [24].

Title	# of errors		
	IC (Gaussian) ($L = 5$; $W = 10$)	Block MSE	IBM
anni005	7	3	3
anni009	5	2	5
bor03	42	7	18
bor08	83	52	17
senses111	14	1	1
Total	151	65	77

error counts than for regular error counting. The error counts for the IBM detector were computed from the precision and recall values reported in [24]. The results are compared in Table 3.3. Note that the block MSE detector outperformed the IBM detector. This is perhaps due to the fact that the two detectors operate in different parts of the ROC curve, i.e., they maintain different false positive/false negative balances.

3.4.2 Texture Classification

We performed texture image classification experiments on a set of 480 128×128 8-bit gray-scale texture images belonging to 30 classes, with 16 images per class. The images were obtained by splitting 30 512×512 images from the VisTex database [18]. The list of texture images is given in Table 3.4. The images were compressed losslessly, as well as at 0.25, 0.5, 1, and 2 bits/pixel using $L = 3$ and 16×16 codeblocks, yielding $N = 3L + 1 = 10$ IC features per image. We used smaller codeblocks to lower the deviations of the actual compression rates from target rates, caused by the small size of

the images.

To improve classification accuracy, feature selection was performed on the IC features using *exhaustive search* [25]. Exhaustive search entails evaluating all possible feature subsets and picking the subset achieving the best optimality criterion. The leave-one-out (LOO) estimate of classification accuracy [25] on the training set was used as the optimality criterion. In LOO, one sample is excluded from the training set, the classifier is trained on the remaining samples, and the trained classifier is tested on the excluded sample. This process is repeated for all samples in the training set and the total number of misclassified samples is used to estimate the LOO error rate on the training set.

The LOO classification accuracy estimate was also used for selecting the best k value for the kNN classifier. For each choice of k, a feature subset maximizing the LOO accuracy was selected using exhaustive search. Then, the LOO accuracies for different choices of k were compared. The combination of the k value and the corresponding feature subset that yielded the highest LOO accuracy was selected. The values considered for k were $\{1, 2, 5, 10, 20, 50, 100\}$.

The probability of error (P_e) was estimated via K-fold cross-validation (CV) [26]. This involves splitting the dataset into K equal subsets. Each subset is used once as the

Table 3.4. Texture images from the VisTex database used in the classification experiments.

Bark.0000	Bark.0004	Bark.0006	Bark.0008	Bark.0009	Brick.0001
Brick.0004	Brick.0005	Fabric.0000	Fabric.0004	Fabric.0007	Fabric.0009
Fabric.0011	Fabric.0013	Fabric.0016	Fabric.0017	Fabric.0018	Food.0000
Food.0002	Food.0005	Food.0008	Grass.0001	Sand.0000	Stone.0004
Tile.0001	Tile.0003	Tile.0007	Water.0006	Wood.0001	Wood.0002

test set and the rest of the subsets are pooled together and used as the corresponding training set. For each of the *K* training sets, feature selection and classifier design are performed. The classifier is then evaluated on the corresponding test set and an estimate of the error rate is obtained for that particular test set. The value of P_e is then estimated as the average of the error rate estimates over the *K* test sets. In this study, *K* was set to 5.

For comparison against IC features, subband variances σ_l^2, given by

$$\sigma_n^{\ 2} = \frac{1}{N_r N_c} \sum_i \sum_j x_{ij}^{(n)2} - \left(\frac{1}{N_r N_c} \sum_i \sum_j x_{ij}^{(n)}\right)^2 \tag{3.13}$$

were also considered as baseline features for classification, where $x_{ij}^{(l)}$, $i = 1,...,N_r$, $j = 1,...,N_c$, $n = 1,...,N$, denotes a wavelet coefficient in subband n of size $N_r N_c$, and $N = 10$ as in the case of IC features. In our experiments, it was found that the logarithm of variance features performed better than the variance features for some compression rates. Therefore, we will report the results for both cases[1].

Table 3.5 summarizes the classification results using IC and variance features. For lossless compression, the IC features perform as well or even better than variance features. However, for lossy compression, the performance of IC features deteriorates more significantly than that of variance features. This decline is caused in part by the fact that IC features lose one degree of freedom compared to variance features, due to the fact that the total bit budget is identical for all images. We can alleviate this loss by storing

---

[1] Other transformations of the features are possible, which may yield better accuracies. The use of the log transformation was motivated by the logarithmic relationship between variance and rate in the high-rate quantization regime.

side-information in the comment segment of the main header of the compressed codestream at the time of compression. One useful piece of side-information is inspired by the high-rate, reverse water-filling rate allocation equation given by[1]

$$r_i = R + \frac{1}{2}\log_2 \frac{n_i \sigma_i^2}{G}, \qquad G = \left(\prod_{j=1}^{N} n_j \sigma_j^2\right)^{1/N}, \tag{3.14}$$

where R denotes the average bit budget, and n_i and σ_i^2 are the number and variance of the wavelet coefficients in subband i, respectively. Note that if we were to find σ_i^2 using r_i, we would need to know G. This motivates the use of the knowledge of G as side information. Using this side-information, we estimated σ_i^2 and used the estimates as features for classification. We noted in the classification experiments that the logarithm of σ_i^2 performed better than σ_i^2. Table 3.5 shows P_e for the log-of-variance features. The table indicates that the use of G as side-information improves the classification accuracies to the level of accuracies using variance (or log-of-variance) features.

We compared the classification results using IC features to those of histogram and co-occurrence features reported in [27]. To the best of our knowledge, [27] reports the highest classification accuracies for the same set of images. To make the results comparable, we split the texture images into 64×64 subimages, and used an LOO

---

[1] We assume for simplicity that coefficients in all subbands follow the same distribution shape and that the distributions are only different in their variances.

Table 3.5. Five-fold CV estimates of the PE using IC and variance features for texture images compressed at different rates.

Rate (bpp)	Classifier	Five-fold CV P_e (%)			
		IC	Log(est var)	Variance	Log(variance)
Lossless	Gaussian	2.50	N/A	10.21	5.83
	*k*NN	6.25	N/A	8.96	5.00
2	Gaussian	23.33	13.96	10.62	10.62
	*k*NN	19.37	9.38	7.92	7.08
1	Gaussian	26.67	15.00	14.38	8.33
	*k*NN	20.00	11.46	6.87	5.83
0.5	Gaussian	31.04	14.79	15.00	10.00
	*k*NN	23.75	11.46	9.38	17.08
0.25	Gaussian	38.75	19.58	19.37	25.21
	*k*NN	42.08	13.96	21.25	37.71

estimate for both classifier training and estimating P_e. The subimages were losslessly compressed using the same compression parameters as before. The classification results are shown in Table 3.6. Note also that our estimate of P_e using variance features is very close to that reported in [27] for energy features, confirming the validity of the proper duplication of [27].

We verified the statistical meaningfulness of our results obtained from 480 samples by finding the size of the 95% confidence interval for the highest classification error obtained from the best performing compressed- and decompressed-domain features per each rate. Using the Gaussian approximation to the binomial distribution, the 95% confidence interval (CI) for P_e is given by

$$\pm 1.96\sqrt{\frac{P_e(1-P_e)}{n}}, \tag{3.15}$$

Table 3.6. Error rates for IC, variance, energy, and histogram and co-occurrence features.

Feature set	Classifier	LOO P_e (%)
IC	Gaussian	4.32
	*k*NN	9.90
Log(variance)	Gaussian	4.58
	*k*NN	9.84
Energy [27]	Gaussian	N/A
	*k*NN	10.9
Histogram and co-occurrence [27]	Gaussian	N/A
	*k*NN	1.2

where n is the total number of samples. For $P_e = 0.2125$ and $n = 480$, the CI is ±3.7%. The CI is becomes smaller for $P_e < 0.2125$. The validity of our results for 480 images and lossless compression (Table 3.5) is further confirmed by the reasonably comparable accuracies obtained with 1920 images (Table 3.6).

3.4.3 Computational Complexity Comparison

In this section, we demonstrate the computational advantage of using IC features for compressed-domain content analysis over decompressed-domain analysis. Due to the variety of features possible in the decompressed domain, we did not compare the complexity of feature computations between compressed and decompressed domains. Rather, we compared the complexity of decoding the entire compressed codestream to that of decoding only its headers. Computing almost any set of features from the decompressed imagery would incur a higher computational cost than that of computing IC features from the decompressed headers. Thus, the above-mentioned comparison strategy yields a lower bound on the difference between the computational complexity of

compressed- and decompressed-domain analysis.

We compared the decoding time for the Batman video sequence described in Table 3.1 and compressed according to the parameters mentioned in Section 3.4.1 to that of decoding only its headers. For this study, we used Kakadu version 5.1 [28], which provides a highly optimized implementation of JPEG2000. The study was run on a 2.8GHz Xeon platform running the Fedora release of Linux. The running time for decoding only the headers was 179 seconds vs. 4200 seconds for decoding the entire codestream. Hence, extracting IC features is more than 23 times faster than obtaining pixel-domain features[1]. Also, extracting IC features is more than 3.7 times faster than real-time video playback, suggesting that IC features can be extracted (and analyzed) for real-time applications, such as video surveillance.

---

[1] Note that for wavelet-domain analysis, the inverse wavelet transform need not be performed. However, the wavelet transform constitutes a relatively small fraction of JPEG2000 decoding complexity.

CHAPTER 4: RATE ALLOCATION FOR JOINT COMPRESSION AND CLASSIFICATION

4.1 Previous Work

Previous work on joint compression and classification has involved the design of two types of compression algorithms. In [2, 3, 5], vector quantization (VQ) algorithms have been proposed, and in [4], a transform coder has been presented. For the design of VQ's, Lloyd-type algorithms [2, 5] and a learning vector quantization-based algorithm [3] have been developed.

In [4], the Karhunen-Loeve transform has been shown to be locally optimal for joint compression and classification of class-conditional distributions in the form of Gaussian mixtures, and a rate allocation scheme has been derived to minimize the loss in the Chernoff distance between the class-conditional distributions. The rate allocation scheme is based on the high-rate analysis of Poor [6], which approximates the loss in the Chernoff distance as a WMSE criterion whose weights depend on the class-conditional distributions of the transform coefficients.

The JPEG2000 standard permits the minimization of a WMSE distortion measure on the wavelet subbands. We exploit this feature for joint image compression and classification. Our aim in classification is to distinguish between images belonging to one

of two possible classes. Classification is performed based on the wavelet subband coefficients. The joint distortion measure is defined as a linear combination of the MSE in the image domain and the loss in the BD between the class-conditional distributions in the wavelet domain. Similar to [4], we use Poor's result [6] to approximate the loss in the BD using the WMSE criterion. Thus, the joint criterion to be minimized also takes the form of the WMSE. The weights in the WMSE criterion are derived under the assumption that the subband coefficients follow GGD's. The choice of the GGD for modeling subband coefficients is supported by experimental studies of the statistics of natural images [29, 30].

The chapter is organized as follows. In the next section, we derive the WMSE criterion as an approximation to the BD and show that its minimum is achieved using reverse water-filling. Then, we present experimental results using synthetic images.

4.2 Distortion Function for Joint Compression and Classification

4.2.1 Classification, Detection, and the Bhattacharyya Distance

Bayes probability of error is the most widely used measure of performance in classification problems. However, with the exception of a few cases, it is often impossible to express and analyze mathematically. We thus turn to the more tractable Bhattacharyya upper bound on the Bayes error. Let $\mathbf{x}$ denote the wavelet transform coefficients of an input image $\mathbf{I}$, re-arranged to form an $N \times 1$ vector. Let $p_i(\mathbf{x}) = p(\mathbf{x} \mid c_i)$, $i = 0, 1$, denote the conditional probability of $\mathbf{x}$ given class c_i. The Bayes probability of error, P_ε, is given by

$$P_\varepsilon = \int \min(P_0 p_0(\mathbf{x}), P_1 p_1(\mathbf{x}))\, dx^N \tag{4.1}$$

where P_i denotes the prior probability of class i. Using the trivial inequality

$$\min(a,b) \le \sqrt{ab}\,,$$

for $a,b \ge 0$, we get the Bhattacharyya bound, B, and the BD, $d(p_0, p_1)$, given by [25]

$$B = \sqrt{P_0 P_1} \int \sqrt{p_0(\mathbf{x}) p_1(\mathbf{x})}\, dx^N = \sqrt{P_0 P_1}\, e^{-d(p_0, p_1)}\,, \tag{4.2}$$

where $B \ge P_\varepsilon$. We can express B in an alternative form as

$$B = E_0\{l(\mathbf{x})^{1/2}\} \tag{4.3}$$

where $l(\mathbf{x}) = p_1(\mathbf{x}) / p_0(\mathbf{x})$ is the likelihood ratio, and $E_0\{.\}$ denotes expectation with respect to $p_0(\mathbf{x})$.

The area under the receiver-operating characteristic (ROC) curve (AUC) is an important performance measure in detection problems. The AUC has been related to the BD through several bounds and approximations [31, 32]. Barrett *et al.* [31] have shown that for equally likely classes,

$$\text{AUC} \ge 1 - \frac{1}{2}\exp(-2d(p_0, p_1))\,, \tag{4.4}$$

and

$$\text{AUC} \approx 1 - Q(2\sqrt{d(p_0, p_1)})\,, \tag{4.5}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{t^2}{2})\, dt\,. \tag{4.6}$$

An attractive property of the BD is that, under the assumption of independence, it is

additive with respect to the elements of $\mathbf{x}$, i.e.,

$$d(p_0, p_1) = \sum_{i=1}^{N} d(p_{0i}(x_i), p_{1i}(x_i)), \tag{4.7}$$

where $p_j(\mathbf{x}) = \prod_{i=1}^{N} p_{ji}(x_i)$, $j = 0, 1$. We will take advantage of this property in the following subsection.

4.2.2 Bhattacharyya Distance and Fine Quantization

Let $\tilde{\mathbf{x}} = Q(\mathbf{x})$, where $Q(.)$ denotes uniform quantization, and define $\tilde{p}_i = p(\tilde{\mathbf{x}} \mid c_i)$, $i = 0, 1$. Let $d_0 = d(p_0, p_1) = -\ln E_0\{l(\mathbf{x})^{1/2}\}$ and $d_\Delta = -\ln E_0\{(E_0\{l(\mathbf{x}) \mid \tilde{\mathbf{x}}\})^{1/2}\}$ be the BD before and after uniform quantization, respectively, where the conditional expectation $E_0\{l(\mathbf{x}) \mid \tilde{\mathbf{x}}\}$ estimates the likelihood ratio given $\tilde{\mathbf{x}}$. It can be shown that [6]

$$\begin{aligned} -\ln E_0\{(E_0\{l(\mathbf{x}) \mid \tilde{\mathbf{x}}\})^{1/2}\} &\le -\ln E_0\{E_0\{l(\mathbf{x})^{1/2} \mid \tilde{\mathbf{x}}\}\} \\ &= -\ln E_0\{l(\mathbf{x})^{1/2} \mid \tilde{\mathbf{x}}\}, \end{aligned} \tag{4.8}$$

where the inequality is Jensen's inequality [33] and the equality is the iterative property of conditional expectations. That is to say that the BD for quantized data cannot exceed the BD for unquantized data. The loss $d_0 - d_\Delta$ due to fine uniform quantization can be characterized using the high-rate analysis of Poor [6] as

$$d_0 - d_\Delta \approx \ln(1 + \alpha\Delta^2) \tag{4.9}$$

where Δ denotes the quantization step size and

$$\alpha = \frac{1}{96} \frac{E_0\{\|\nabla \ln l(\mathbf{x})\|^2 \, l(\mathbf{x})^{1/2}\}}{d_0}. \tag{4.10}$$

For $\alpha\Delta^2 << 1$, we can use the Taylor series to further approximate the loss as

$d_0 - d_\Delta \approx \alpha\Delta^2$.

At high rates, the MSE for each element x_i of $\mathbf{x}$ can be expressed as $\mathrm{MSE}_i = \Delta^2/12$. Thus, the loss in the BD can be related to MSE_i as $d_0 - d_\Delta \approx 12\alpha\,\mathrm{MSE}_i$. When the x_i are independent, we can write

$$d_0 - d_\Delta = \sum_{i=1}^{N}(d_{0i} - d_{\Delta i}) \approx \sum_{i=1}^{N}\alpha_i\Delta_i^2 = \sum_{i=1}^{N}12\alpha_i\mathrm{MSE}_i, \tag{4.11}$$

where

$$\alpha_i = \frac{1}{96}\frac{E_0\{(d\ln l_i(x_i)/dx_i)^2 l_i(x_i)^{1/2}\}}{d_{0i}} \tag{4.12}$$

and $l_i(x_i) = p_{1i}(x_i)/p_{0i}(x_i)$. Equation (4.11) has the form of a WMSE function with weights $12\alpha_i$.

To gain insight into the meaning of α_i, consider the case of Gaussian class-conditional densities $p_j(\mathbf{x}) \sim N(\boldsymbol{\mu}_j, \Sigma_j)$, $j = 0, 1$, where $\Sigma_j = \mathrm{diag}(\sigma_{ji}^2)$. We have

$$\alpha_i = \frac{1}{96}\left[B_i C_i^2 + (A_i C_i + D_i)^2\right] \tag{4.13}$$

where

$$A_i = \frac{\mu_{0i}\sigma_{1i}^2 + \mu_{1i}\sigma_{0i}^2}{\sigma_{0i}^2 + \sigma_{1i}^2}, \quad B_i = \frac{2\sigma_{0i}^2\sigma_{1i}^2}{\sigma_{0i}^2 + \sigma_{1i}^2}, \quad C_i = \frac{1}{\sigma_{0i}^2} - \frac{1}{\sigma_{1i}^2}, \quad \text{and } D_i = \frac{\mu_{1i}}{\sigma_{1i}^2} - \frac{\mu_{0i}}{\sigma_{0i}^2}. \tag{4.14}$$

In the case $\Sigma_0 = \Sigma_1 = \mathrm{diag}(\sigma_i^2)$, (4.13) simplifies to

$$\alpha_i = \frac{1}{96}\frac{(\mu_{0i} - \mu_{1i})^2}{\sigma_i^4}, \tag{4.15}$$

and in the case $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1 = \boldsymbol{\mu}$, it simplifies to

$$\alpha_i = \frac{1}{48}\frac{(1/\sigma_{0i}^2 - 1/\sigma_{1i}^2)^2}{1/\sigma_{0i}^2 + 1/\sigma_{1i}^2}. \tag{4.16}$$

The above expression corrects the result given by Jana and Moulin [7, 8].

Equations (4.15) and (4.16) behave in a similar way as the BD for normal distributions, given by [25]

$$d(p_{0i}, p_{1i}) = \frac{1}{8}\frac{(\mu_{0i} - \mu_{1i})^2}{\sigma_i^2}, \tag{4.17}$$

for equal covariance matrices, and

$$d(p_{0i}, p_{1i}) = \frac{1}{2}\ln\frac{\sigma_{0i}^2 + \sigma_{1i}^2}{2\sqrt{\sigma_{0i}^2\sigma_{1i}^2}}, \tag{4.18}$$

for equal mean vectors. This indicates that the elements of $\mathbf{x}$ that contribute more significantly to the BD have a larger α_i. This implies that under a rate constraint, such elements must be quantized at a higher rate than others to maximize the classification information retained in $\tilde{\mathbf{x}}$.

In many image processing tasks such as compression and classification, the GGD has been shown to be a suitable model for representing the distribution of wavelet and discrete cosine transform coefficients (See, e.g., [29, 30]). The GGD $G(\mu, a, b)$ is given by

$$p(x) = \frac{b}{2a\Gamma(1/b)}\exp(-(\frac{|x-\mu|}{a})^b) \tag{4.19}$$

where μ, a, and b denote the mean, width parameter, and shape parameter of the distribution, and $\Gamma(.)$ denotes the Gamma function. The parameters of the GGD

distribution can be estimated via moment-matching or maximum-likelihood (ML) estimation [29, 30]. For the general case where

$$p_j(\mathbf{x}) \sim \prod_{i=1}^{N} G(\mu_{ji}, a_{ji}, b_{ji}), \; j = 0, 1, \tag{4.20}$$

parameter α_i should be determined by numerical evaluation of (4.12). However, in the special case where $\mu_{0i} = \mu_{1i} = \mu_i$ and $b_{0i} = b_{1i} = b_i$, α_i has a closed form given by

$$\alpha_i = \frac{1}{96} b_i^2 \frac{\Gamma(2 - 1/b_i)}{\Gamma(1/b_i)} \left(\frac{1}{a_{1i}^{b_i}} - \frac{1}{a_{0i}^{b_i}}\right) \left(\frac{2}{\frac{1}{a_{0i}^{b_i}} + \frac{1}{a_{1i}^{b_i}}}\right)^{\frac{2(b_i - 1)}{b_i}}. \tag{4.21}$$

When $a_{ji} = \sqrt{2}\sigma_{ji}$, $j = 0, 1$, and $b_i = 2$, $i = 1, ..., N$, (4.21) reduces to (4.16).

4.2.3 Joint Distortion Function

To minimize the MSE in the image domain, we need to minimize the WMSE criterion given by

$$\text{WMSE} = \sum_{i=1}^{N} g_i \text{MSE}_i \tag{4.22}$$

in the wavelet domain, where g_i denotes the synthesis gain for subband i [9]. The joint compression and classification distortion function is thus formed as a linear combination of (4.11) and (4.22) given by

$$J = \sum_{i=1}^{N} ((1 - \lambda) g_i + 12\lambda\alpha_i) \text{MSE}_i = \sum_{i=1}^{N} \beta_i \text{MSE}_i \tag{4.23}$$

where $\beta_i = (1 - \lambda) g_i + 12\lambda\alpha_i$, and the parameter λ controls the trade-off between the MSE and the loss in the BD. When $\lambda = 0$, J reduces to the MSE criterion, and when

$\lambda = 1$, J reduces to the loss in the BD.

4.2.4 Optimal Rate Allocation

The rate allocation problem consists of finding rates R_i for quantizing x_i subject to two constraints on the rate and can be stated as

$$\text{minimize } J \text{ such that } \frac{1}{N}\sum_{i=1}^{N} R_i \le R, \text{ and } R_i \ge 0 \text{ for all } i. \tag{4.24}$$

Under a high-rate regime, MSE_i and R_i are related via $\text{MSE}_i = \varepsilon_i^2 2^{-2R_i}$, where $\varepsilon_i^2 = (1/12)e^{2h(x_i)}$ and $h(x_i)$ is the differential entropy of x_i [9]. The minimizer of (4.24) is given by

$$R_i = R + \frac{1}{2}\log_2 \frac{\varepsilon_i^2 \beta_i}{\left(\prod_{j=1}^{N} \varepsilon_j^2 \beta_j\right)^{1/N}}, \tag{4.25}$$

assuming that $R_i \ge 0$, $i = 1,...,N$. This is the reverse water-filling equation for rate allocation to independent Gaussian random variables with variances $\varepsilon_i^2 \beta_i$ (cf. Section 2.2).

4.3 Simulation Results

We assessed the merit of the rate allocation scheme in (4.25) through simulations on synthetic images. The Kakadu [28] implementation of JPEG2000 was used in the simulations. The wavelet transform coefficients of each image are assumed to be independent within and between subbands, and identically distributed within subbands.

The subband coefficients for image i in subband j, $j = 1,...,J$, follow the GGD $G(0, a_{ij}, b_{ij})$. These assumptions are reasonably accurate for most natural images, provided that the image has zero mean. The above assumptions also allow us to estimate the parameters of the joint distribution of the wavelet coefficients for each class i, $i = 0, 1$, using a single, sufficiently large simulated image i for that class. The simulated images have 512×512 pixels, are quantized to 16-bit signed integers in the pixel domain, and are transformed using one level of wavelet transformation ($J = 4$).

We considered two cases in the simulations, namely, $b_{0j} = b_{1j}$ (Dataset 1) and $b_{0j} \neq b_{1j}$ (Dataset 2). To achieve more realistic results, we selected the b_{ij} according to those of two natural texture images in the MIT VisTex database [18], namely, Grass.0001 (class 0) and Sand.0000 (class 1). We estimated the shape parameters of these images using the ML method as described in [30] and implemented in [34] and generated the synthetic images according to the resulting estimates. The shape parameters were once estimated under the $b_{0j} = b_{1j}$ assumption, and once without this assumption. The ML method in [30] was slightly modified to accommodate the case $b_{0j} = b_{1j}$. The resulting shape parameters as well as the arbitrarily selected width parameters used in generating the synthetic images are listed in Table 4.1.

Once the images were synthesized, the shape and width parameters of the GGD's underlying each of their subbands were estimated using the ML criterion as described above. Parameters α_i were then obtained from these estimates. For Dataset 1, α_i were determined via (4.21). For Dataset 2, α_i were obtained once via (4.21) under the $b_{0j} = b_{1j}$

assumption and once via (4.12) under the $b_{0j} \neq b_{1j}$ assumption. The weights β_i were subsequently determined using (4.22). The BD after compression d_Δ was estimated directly from the compressed images using (4.2) as an approximation to the definition of d_Δ given at the beginning of Section 4.2.2. These two estimates of d_Δ are close in the high-rate regime.

Figure 4.1, Figure 4.2, and Figure 4.3, show the MSE and the loss in the BD for Dataset 1 and Dataset 2, using three choices of λ and different rates. The results agree with the theory for almost all choices of rate and λ. Increasing the rate decreases both the MSE and the BD loss as expected. Varying λ controls the trade-off between the MSE and BD loss. Increasing λ increases the MSE while reducing the BD loss. The effect of λ on the BD loss is most visible at moderately high rates. Note that usually the MSE is several orders of magnitude larger than the BD loss. Therefore, λ must be a

Table 4.1. Parameters of the GGD distributions for (a) Dataset 1, and (b) Dataset 2.

(a)

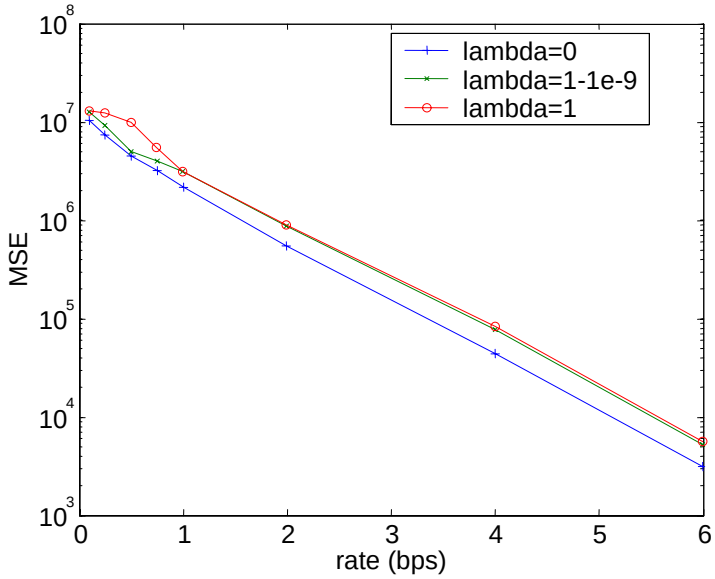
	Class 0		Class 1	
Subband	Shape	Width	Shape	Width
LL	2.7	12000	2.7	4000
HL	1.3	5000	1.3	1500
LH	1.3	5000	1.3	1000
HH	1.3	2500	1.3	500

(b)

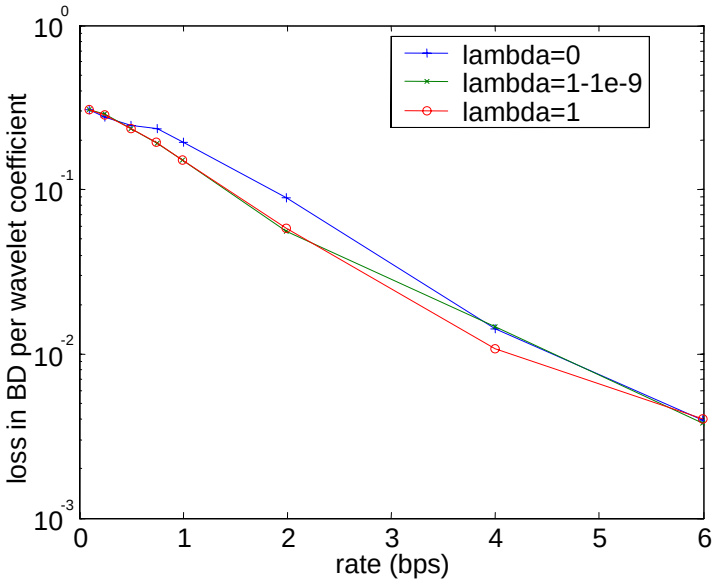
	Class 0		Class 1	
Subband	Shape	Width	Shape	Width
LL	6.9	12000	2.0	4000
HL	1.0	5000	1.7	1500
LH	1.0	5000	1.6	1000
HH	1.1	2500	1.6	500

number very close to 1 in order for the BD loss term to figure in J .

Figure 4.2, and Figure 4.3 show the results for Dataset 2 obtained from equal and unequal shape parameter assumptions, respectively. As the Figures suggest, the equal shape parameter assumption in this example did not affect the results significantly. However, in general this assumption should be used with caution.

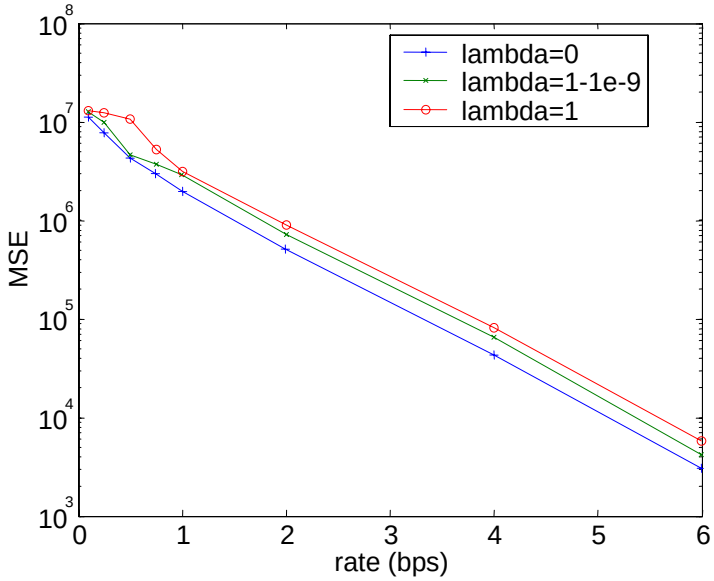


(a)

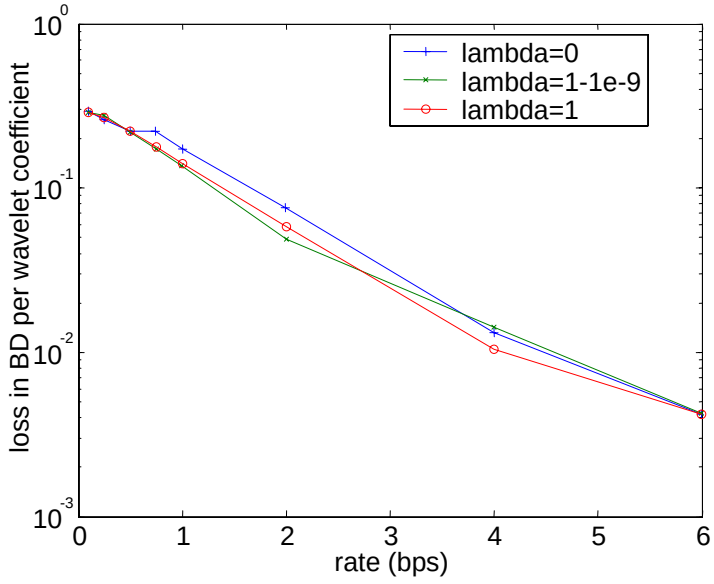


(b)

Figure 4.1. (a) MSE and (b) loss in BD for Dataset 1.

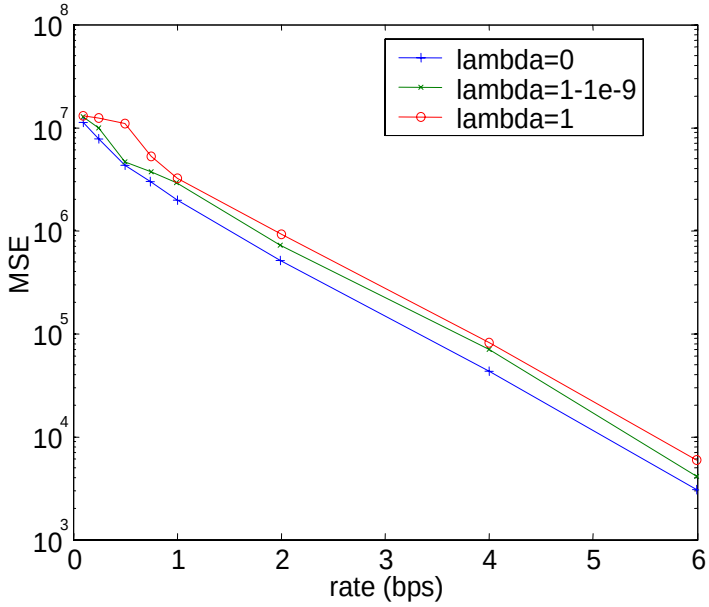


(a)

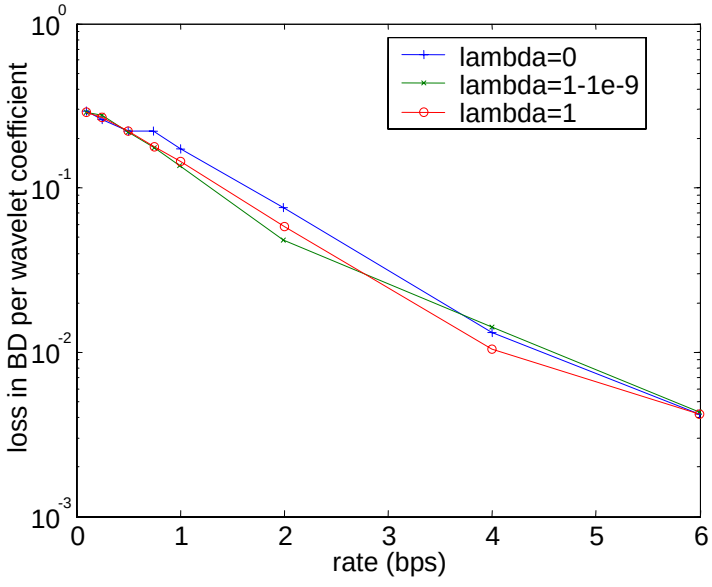


(b)

Figure 4.2**.** (a) MSE and (b) loss in BD for Dataset 2 assuming equal shape parameters.



(a)



(b)

Figure 4.3. (a) MSE and (b) loss in BD for Dataset 2 assuming unequal shape parameters.

CHAPTER 5: RATE ALLOCATION FOR NONBINARY CLASSIFICATION AND DEPENDENT SOURCES

The developments in Chapter 4 and [4, 35] are based on the assumption that $p_j(\mathbf{x}) = \prod_{i=1}^{N} p_{ji}(x_i)$, $j = 0, 1$, i.e., the elements of $\mathbf{x}$ are independent conditioned on class j. Unfortunately, closed-form rate-allocation schemes based on the BD are possible only under this conditional independence assumption. For dependent sources, iterative search strategies in the rate space are needed and may lead to local minima of the loss in the BD. Moreover, the BD is only applicable to binary hypothesis testing problems.

In general, the rate-allocation problem with the goal of minimizing the quantized Bayes error is intractable and finding the optimal rates requires exhaustive search of the rate space.

Hu and Blum [36] considered the rate allocation problem for dependent sources under nonbinary hypothesis testing. They showed that there is a maximum rate necessary to communicate the output of a given source, and increasing the rate beyond this maximum does not improve the overall quantized Bayes error. The value of this maximum depends on the rates allocated to the other sources. They used this result to eliminate some of the candidates in the exhaustive search for optimal integer rates.

This chapter presents a rate-allocation scheme based on a new bound on the quantized Bayes error. The primary advantages of this scheme are that it allows for an intuitive, closed-form rate-allocation scheme for dependent sources and nonbinary hypothesis testing when the Bayes decision boundary is piecewise linear. We consider continuous rates similar to the previous chapter and [4, 35].

In the next section, we present our new upper bound on the quantized Bayes error. We use the bound to address the rate-allocation problem for two sources for which the unquantized Bayes decision boundary is linear. An important instance of this scenario is the detection of known signals in correlated Gaussian noise. We extend this result to the piecewise linear decision boundary. Moreover, we show an extension of the rate-allocation scheme to the case where a joint compression and classification criterion is to be optimized, where the compression criterion is the WMSE. Finally, simulation examples are presented.

5.1 New Upper Bound on Quantized Bayes Error

Consider the classification problem shown in Figure 5.1. The observations are uniformly quantized with the quantization cell boundaries at -3, -2, …, 6. The unquantized Bayes decision boundary δ^0 is shown with a dotted line. The quantized Bayes boundary δ^Δ (solid line) consists of one of the quantization cell boundaries for the cell that encompasses the unquantized Bayes boundary. Note that the observations that fall into quantization cells other than this cell will be classified to the same class both before and after quantization. Thus, quantization of such observations does not result in

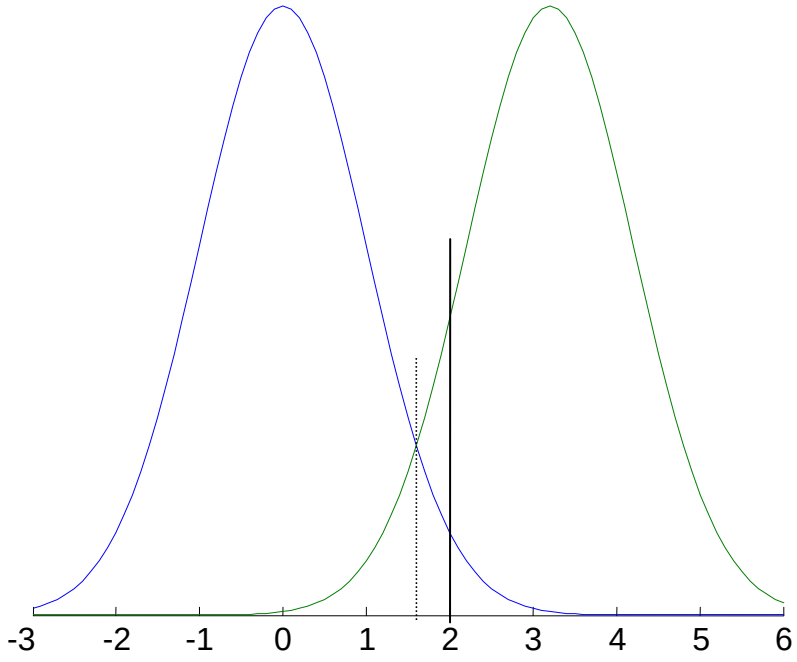


Figure 5.1. Illustrative classification example. The dotted vertical line represents the unquantized Bayes decision boundary and the solid line shows the quantized Bayes boundary. The observations are quantized to integer values.

additional misclassifications. The only observations quantization of which may incur additional misclassifications are those that fall into the quantizer cell *encompassing* δ^0.

Let P_ε^0 and P_ε^Δ denote the unquantized and quantized Bayes errors, respectively. We can write the above observation as

$$P_\varepsilon^\Delta = P_\varepsilon^0 + \sum_{i \in B} \varepsilon_i \tag{5.1}$$

where B is the set of quantizer cells that encompass δ^0, and ε_i is the contribution of quantizer cell i to P_ε^Δ, where $i \in B$. We can further write

$$P_\varepsilon^\Delta \le P_\varepsilon^0 + \frac{\varepsilon_{i^*}}{V_{i^*}} \sum_{i \in B} V_i \tag{5.2}$$

where V_i is the volume of the quantizer cell i, and $i^* = \arg\max_{i \in B} \varepsilon_i / V_i$ denotes the index of the cell with the largest ε_i / V_i value. For uniform quantization, we have $V_i = V$, $\forall i \in B$. Thus, (5.2) simplifies to

$$P_\varepsilon^\Delta \le P_\varepsilon^0 + N \varepsilon_{i^*} \tag{5.3}$$

where N denotes the number of quantizer cells[1] that encompass δ^0. Equation (5.3) holds for nonbinary problems as well. In that case, B denotes the set of quantizer cells that encompass all segments of the Bayes decision boundary between all pairs of classes.

5.2 Rate-Allocation for the Bayes Linear Decision Boundary

Of the terms on the right-hand side of (5.3), we can control N via rate-allocation. To illustrate this, consider the unquantized linear Bayes boundary in Figure 5.2. We can write

$$\begin{aligned} N &= \left\lceil \frac{U_1}{\Delta_1} \right\rceil + \left\lceil \frac{U_2}{\Delta_2} \right\rceil - 1 \\ &\approx \sum_{s=1}^{2} \frac{U_s}{\Delta_s} \end{aligned} \tag{5.4}$$

where U_s and Δ_s, $s = 1, 2$, denote the extent of the boundary and quantization step size for source s, respectively, and $\lceil . \rceil$ is the ceiling function. The rate-allocation problem for

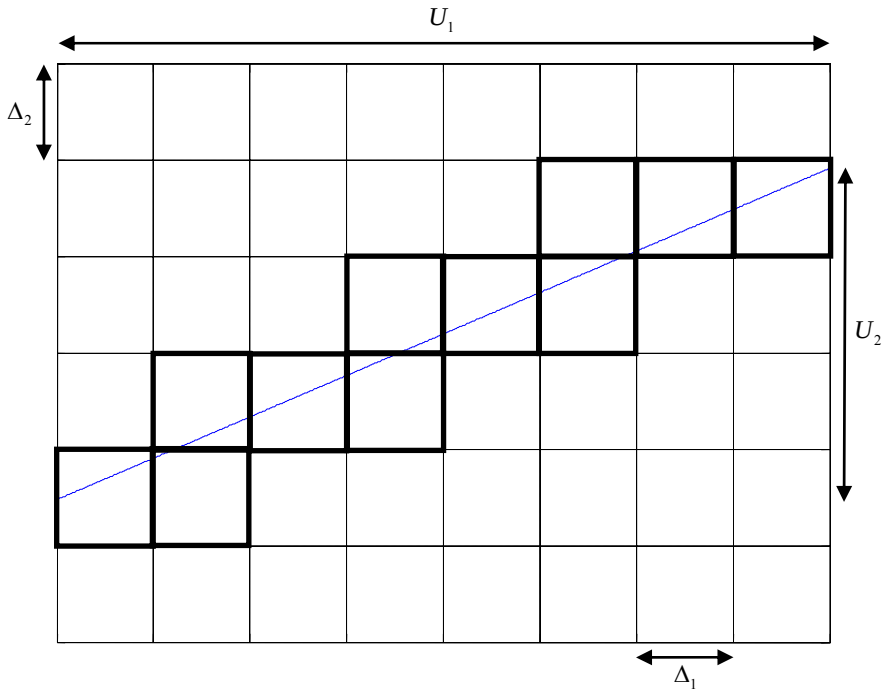


Figure 5.2. A linear Bayes boundary for two sources and its encompassing quantization cells marked in bold.

fine, uniform quantization can be then expressed as

$$\begin{aligned} &\text{minimize } N \approx \sum_{s=1}^{2} \frac{U_s}{\Delta_s} \\ &\text{such that } \sum_{s} \log \Delta_s \geq K, \end{aligned} \tag{5.5}$$

where K is a function of total bit budget, expressed as the log-area of quantizer cells. Note that K is also an explicit function of the class-conditional densities due to its dependence on the differential entropy of the source observations. However, N does not

---

[1] Without loss of generality and to simplify presentation, we assume that the decision boundary has finite support and thus N is finite. The results are easily generalized to the case of infinite support.

explicitly depend on the class-conditional densities; the only assumption that weakly links N to the class-conditional densities is that δ^0 is linear.

Applying the method of Lagrange multipliers to (5.5), the relationship between Δ_1 and Δ_2 is found as

$$\frac{\Delta_2}{\Delta_1} = \frac{U_2}{U_1}. \tag{5.6}$$

Note that relationship in (5.6) does not uniquely determine Δ_1 and Δ_2. Among the uncountable solutions to (5.6), the one that satisfies the constraint in (5.5) should be chosen.

The above result suggests that the source observations that contribute more to class separation should be encoded more finely. To further illustrate this, consider the case in Figure 5.2 when δ^0 approaches a vertical line. In this case, source 1 provides all the classification information and source 2 does not provide any separation. Intuition suggests that all of the bit budget should be allotted to source 1 and source 2 should be completely ignored. Indeed, for δ^0 approaching a vertical line, we have U_1 approaching 0 and thus (5.6) prescribes Δ_1 to be as small as the rate constraint permits, which confirms our intuition.

It is interesting to compare the step sizes above to those obtained by the minimization of the BD loss as prescribed in Chapter 4. Let $p_i(\mathbf{x}) = p(\mathbf{x} \mid c_i)$, $i = 0, 1$, denote the class-conditional densities for the 2×1 observation vector $\mathbf{x}$ given class c_i. Assuming Gaussian densities $p_i(\mathbf{x}) \sim N(\boldsymbol{\mu}_i, \mathbf{I})$, $i = 0,1$, where $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$, and $\mathbf{I}$ is

the identity matrix, the rate-allocation problem to minimize the BD loss can be written as

$$\begin{aligned} &\text{minimize } \sum_{s=1}^{2} \alpha_s \Delta_s^2 \\ &\text{such that } \sum_{s} \log \Delta_s \geq K \end{aligned} \tag{5.7}$$

where $\alpha_s \propto \mu_s^2$. Using the method of Lagrange multipliers, we get

$$\frac{\Delta_2}{\Delta_1} = \frac{\mu_1}{\mu_2}, \tag{5.8}$$

which is identical to (5.5). Thus, for the binary hypothesis testing problem involving two sources with independent Gaussian class-conditional densities, minimizing the BD loss and the new bound result in the same rate allocation to the sources.

5.3 Extensions

5.3.1 Piecewise Linear Bayes Decision Boundary

Equation (5.4) can be generalized to the case where the Bayes decision boundary is piecewise linear. In that case,

$$\begin{aligned} N &= \sum_{l=1}^{L} \left\lceil \frac{U_{l1}}{\Delta_1} \right\rceil + \left\lceil \frac{U_{l2}}{\Delta_2} \right\rceil - 1 \\ &\approx \sum_{s=1}^{2} \frac{1}{\Delta_s} \sum_{l=1}^{L} U_{ls} \end{aligned} \tag{5.9}$$

where L denotes the number of segments in the decision boundary, and U_{ls} denotes the extent of segment l for source s. The rate-allocation problem can again be expressed as in (5.5) and solved using the method of Lagrange multipliers to get

$$\frac{\Delta_2}{\Delta_1} = \frac{\sum_l U_{l2}}{\sum_l U_{l1}}. \tag{5.10}$$

The right-hand side of (5.6) can be thought of as the "average" slope of the line segments that constitute the Bayes boundary.

5.3.2 Joint Compression and Classification

We use a linear combination of the WMSE as the fidelity criterion for human observation and the bound in (5.3) to characterize classification performance to form a joint distortion criterion J. Criterion J is given by

$$J = (1-\lambda)\sum_{s=1}^{2} w_s \mathrm{MSE}_s + \lambda N, \tag{5.11}$$

where MSE_s and w_s are the MSE and associated weight for source s. Parameter $\lambda \in [0, 1]$ controls the trade-off between the WMSE and the classification criterion. When $\lambda = 0$, J reduces to the WMSE criterion, and when $\lambda = 1$, J reduces to the classification criterion.

Minimizing (5.11) subject to a rate constraint yields

$$\frac{1-\lambda}{6}\left(w_1\Delta_1^2 - w_2\Delta_2^2\right) - \lambda\left(\frac{U_1}{\Delta_1} - \frac{U_2}{\Delta_2}\right) = 0. \tag{5.12}$$

The above equation is a forth-order polynomial and can be solved via standard root-finding techniques. Note that only one of the four roots of (5.12) corresponds to the minimum of (5.11).

5.4 Examples

In this section, we show by numerical examples the efficacy of the proposed rate-allocation scheme.

5.4.1 Binary Hypothesis Testing in Dependent Gaussian Noise

This example deals with binary hypothesis testing where observations follow Gaussian class-conditional distributions with identical covariance matrices, but different means. It can be thought of as known signal detection in uncorrelated or correlated Gaussian noise. Assume $p_i(\mathbf{x}) \sim N(\boldsymbol{\mu}_i, \mathbf{C})$, $i = 0,1$, where $\boldsymbol{\mu}_0 = \mathbf{0}$, $\boldsymbol{\mu}_1 = [2 \quad 0.2]^T$, and $\mathbf{C} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$. The unquantized Bayes decision boundary is given by

$$\delta^0 : \ \boldsymbol{\mu}_1^T \mathbf{C}^{-1} \mathbf{x} + a = 0, \tag{5.13}$$

where a is a constant. From (5.6), we have

$$\frac{\Delta_2}{\Delta_1} = \left| \frac{2 + 0.2r}{2r + 0.2} \right|. \tag{5.14}$$

Table 5.1 shows P_ε^Δ for the rate constraint $\sum_s \log_2 \Delta_s \geq 2$ and four choices of quantization step sizes:

- $\Delta_1 = \Delta_2$, which minimizes the MSE in expressing $\mathbf{x}$;
- $\Delta_1 = (\mu_2 / \mu_1)\Delta_2 = \Delta_2 / 10$, which minimizes the loss in BD assuming independent observations (cf. (5.8)); and
- step sizes from (5.14).

Two values $r = 0, -0.7$ were considered in the simulations, for which (5.14) yields

Table 5.1. Bayes probability of error P_ε^Δ for the example in Section 5.4.1; (a) $r = 0$, $P_\varepsilon^0 =$ 0.1573; (b) $r = -0.7$, $P_\varepsilon^0 =$0.0666.

(a)

	$\Delta_1 = \Delta_2$	$\Delta_1 = \Delta_2 / 10$
average P_ε^Δ	0.1943	0.1624

(b)

	$\Delta_1 = \Delta_2$	$\Delta_1 = \Delta_2 / 10$	$\Delta_1 = (20/31)\Delta_2$
average P_ε^Δ	0.1447	0.1396	0.1350

$\Delta_1 = \Delta_2 / 10$ (identical to BD loss minimizer) and $\Delta_1 = (20/31)\Delta_2$, respectively. Noting that the choice of quantization grid placement affects P_ε^Δ, we considered shifting the grid locations in increments of $\Delta_s / 5$ for source s. The table shows the average of P_ε^Δ taken over all shifted grids.

For correlated data the proposed scheme offers a small improvement over the BD loss-minimizing scheme. However, the MSE due to the latter scheme (3.3667) is much larger than that of the new scheme (0.7317). Thus, when joint compression and classification is concerned, the advantage of the proposed scheme becomes very significant.

5.4.2 Three-Class Hypothesis Testing in Independent Gaussian Noise

In this example, we consider a 3-ary detection problem with uncorrelated Gaussian noise. Let $p_i(\mathbf{x}) \sim N(\boldsymbol{\mu}_i, \mathbf{I})$, $i = 0, 1, 2$, where $\boldsymbol{\mu}_0 = -\boldsymbol{\mu}_1 = \begin{bmatrix} 3 & 0 \end{bmatrix}^T$, $\boldsymbol{\mu}_2 = [0 \quad 1]^T$. The Bayes decision boundary consists of three line segments with slopes 3, -3, and $-\infty$ with

Table 5.2. Quantized Bayes error P_ε^Δ for the example in Section 5.4.2; P_ε^0=0.0759.

	$\Delta_1 = \Delta_2$	$\Delta_1 = (2/9)\Delta_2$
average P_ε^Δ	0.1146	0.0920

respect to the x_1-axis. Equation (5.10) yields $\Delta_1 = (2/9)\Delta_2$. Table 5.2 shows P_ε^Δ for $\Delta_1 = \Delta_2$ and $\Delta_1 = (2/9)\Delta_2$, where the rate constraint is $\sum_s \log_2 \Delta_s \geq 2$. As expected, the proposed scheme offers a substantial improvement in average P_ε^Δ.

5.4.3 Joint Compression and Classification in Independent Gaussian Noise

Consider the problem described in Section 5.4.1 with $r = 0$. The goal in this example is to minimize (5.11) with $w_1 = w_2 = 1$, $U_1 = 0.2$, $\lambda = 0, 0.5, 1$, and $\sum_s \log_2 \Delta_s \geq 2$. From the previous example, we have $\Delta_1 = \Delta_2$ and $\Delta_1 = \Delta_2/10$, $\lambda = 0, 1$. For $\lambda = 0.5$, the root of (5.12) is found at $\Delta_1 = 0.504\Delta_2$. Table 5.3 shows the MSE and P_ε^Δ for the above choices of λ. Both the MSE and P_ε^Δ for $\lambda = 0.5$ fall between the corresponding values for $\lambda = 0, 1$, as expected. Note that achieving the lowest P_ε^Δ ($\lambda = 1$) comes at a very high MSE cost. Thus, $\lambda = 0.5$ may be a more sensible choice in terms of achieving a reasonable trade-off between the two criteria.

Table 5.3. MSE and quantized Bayes error P_ε^Δ for the example in Section 5.4.3.

	$\lambda = 0$ $(\Delta_1 = \Delta_2)$	$\lambda = 0.5$ $(\Delta_1 = 0.504\Delta_2)$	$\lambda = 1$ $(\Delta_1 = \Delta_2 / 10)$
MSE*	0.6667	0.8294	3.3667
average P_ε^Δ	0.1943	0.1773	0.1624

* From high-rate analysis.

CHAPTER 6: CONCLUSIONS

This dissertation addressed two problems at the intersection of compression and statistical inference, namely, statistical inference on compressed data, and rate-allocation for joint compression and classification.

6.1 Content Analysis of JPEG2000-Compressed Imagery

In Chapter 3, we proposed the use of the IC of subbands for texture characterization of JPEG2000-compressed images and video. The primary advantage of IC features is that they can be directly extracted from codestream headers, avoiding the computationally demanding decompression of the arithmetically encoded image data. Our results indicate that this approach achieves accuracy comparable to that of baseline statistics computed from decompressed imagery on a texture classification task. On video cut detection, IC features perform reasonably well, but not as well as baseline features.

The methods reported in Chapter 3 can be improved upon in many ways. One approach is to exploit the IC features more efficiently. For example, the IC of codeblocks rather than that of subbands can be used. This may be particularly useful for cut detection, as effective cut detection hinges on a spatially more localized texture change characterization across frames than that provided by the IC of entire subbands. In the case of color imagery, including the IC of chrominance components in the analysis may be

beneficial.

The IC features can be augmented by other information obtainable from packet headers, such as the number of zero bitplanes in codeblocks, as suggested in [15]. If the codestream contains multiple quality layers, or when the RESTART marker is used, other information becomes available that can supplement the already-mentioned features.

The comment segments of the main and tile headers of the JPEG2000 codestream allow for storing side information about the imagery. This side information, which can be easily retrieved from the compressed codestream without arithmetic decoding, can be used to complement other features. An example of such side information is G given in (3.14).

Content analysis based on IC features can serve as a preliminary stage to more detailed analysis in the decompressed domain. In video cut detection, for example, cut candidates can be identified in the compressed-domain and then frames in the vicinity of the candidates can be subjected to more elaborate analysis in the decompressed domain.

We believe that a hybrid compressed-decompressed-domain content analysis method will achieve accuracy approaching that of the state-of-the-art decompressed-domain analysis techniques at a fraction of the computational cost.

6.2 Rate Allocation for Joint Compression and Classification

In Chapter 4, we presented a rate allocation scheme, given in (4.25), for joint compression and classification in wavelet image coders. This scheme can be applied to wavelet image coders such as JPEG2000 that allow for minimizing a WMSE criterion on

the subbands. The scheme is based on Poor's high-rate analysis [6], which relates the BD between the classes to the quantization step size, and is closely related to that proposed in [4, 7] for transform coder design. Our primary contribution was to extend the application of the criterion to subband coders. Moreover, we presented the correct expression for rate-allocation to independent Gaussian sources [7, 8], and generalized the expression to independent sources whose statistics follow GGD's. Through simulations, we demonstrated the utility of this equation in controlling the trade-off between compression and classification in JPEG2000. Even though it was derived under the high-rate assumption, our simulation results indicate that (4.25) can be used under moderate rates as well.

The above-mentioned high-rate approximation of the BD has two limitations. For correlated sources, the BD loses its additive property and becomes analytically intractable. Moreover, the BD is inherently restricted to two-class classification problems. We proposed a new bound (Equation (5.3)) on the quantized probability of error in Chapter 5 that addresses both of these limitations. We used the new bound to derive closed-form expressions for rate-allocation to two sources for which the unquantized Bayes decision boundary is a piecewise linear function (Equation (5.10)). Moreover, the scheme was extended to accommodate joint compression and classification (Equation (5.12)). Our simulations confirm the benefits of new rate-allocation scheme.

Future work on the new rate-allocation scheme may consist of three primary directions. First, the scheme may be extended to handle more than two sources. Our intuition is that the relationship between the quantization step sizes and the slope of the

decision boundary will still hold.

Second, the scheme may be generalized to deal with more complex decision boundaries. This may be accomplished via taking the limit of the expression for rate-allocation to piecewise linear decision boundaries as the number of linear segments approaches infinity.

Finally, the new bound may be beneficial as a surrogate to the Bayes error in the design and analysis of other compression systems, particularly VQ's.

REFERENCES

[1] "JPSearch Scope and Requirements 1.0," ISO/IEC JTC1/SC29/WG1, Document Number 3373, July 2004.

[2] K. L. Oehler and R. M. Gray, "Combining image compression and classification using vector quantization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 461-473, 1995.

[3] J. S. Baras and S. Dey, "Combined compression and classification with learning vector quantization," *IEEE Trans. Info. Theory*, vol. 45, pp. 1911-1920, 1999.

[4] S. Jana and P. Moulin, "Optimal transform coding of Gaussian mixtures for joint classification/reconstruction," in *Proc. IEEE Data Comp. Conf.*, Snowbird, UT, 2003, pp. 313-322.

[5] R. Gupta and A. O. Hero, "High rate vector quantization for detection," *IEEE Trans. Info. Theory*, vol. 49, pp. 1951-1969, 2003.

[6] H. V. Poor, "Fine quantization in signal detection and estimation," *IEEE Trans. Info. Theory*, vol. 34, pp. 960-972, 1988.

[7] S. Jana and P. Moulin, "Optimal design of transform coders and quantizers for image classification," in *Proc. IEEE Int. Conf. Image Proc.*, Vancouver, BC, 2000, pp. 841-844.

[8] S. Jana, personal communication, 2003.

[9] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Practice, and Standards*. Boston, MA: Kluwer, 2002.

[10] "Information technology -- JPEG 2000 image coding system -- Part 3: Motion JPEG 2000," ISO/IEC 15444-3, 2002.

[11] M. K. Mandal, F. Idris, and S. Panchanathan, "A critical evaluation of image and video indexing techniques in the compressed domain," *Image and Vision Computing*, vol. 17, pp. 513-529, 1999.

[12] R. L. de Queiroz and R. Eschbach, "Fast segmentation of the JPEG compressed documents," *J. Elec. Imag.*, vol. 7, pp. 367-77, 1998.

[13] J. Bhalod, G. F. Fahmy, and S. Panchanathan, "Region based indexing in the JPEG2000 framework," in *Proc. SPIE Conf. Internet Multimedia Management Systems II*, vol. 4519, 2001, pp. 91-96.

[14] Z. Xiong and T. S. Huang, "Wavelet-based texture features can be extracted efficiently from compressed-domain for JPEG2000 coded images," in *Proc. IEEE Int. Conf. Image Proc.*, Rochester, NY, 2002, pp. I-481-I-484.

[15] M. K. Mandal and C. Liu, "Efficient image indexing techniques in the JPEG2000 domain," *J. Elec. Imag.*, vol. 13, pp. 182-187, 2004.

[16] R. Neelamani and K. Berkner, "Adaptive representation of JPEG 2000 images using header-based processing," in *Proc. IEEE Int. Conf. Image Proc.*, Rochester, NY, 2002, pp. I-381-I-384.

[17] X. Feng and M. J. Gormish, "Locating barcodes using JPEG 2000 compressed data," in *Proc. SPIE Conf. Visual Communications and Image Processing*, vol. 5960, Beijing, China, 2005, pp. 908-916.

[18] "Vision texture," http://vismod.www.media.mit.edu/vismod/imagery/VisionTexture/.

[19] "VM 9.0 software" ISO/IEC JTC1/SC29/WG1, Document Number 2131, April 2001.

[20] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, UK: Cambridge University Press, 1992.

[21] "The MathWorks," http://www.mathworks.com.

[22] "Tooldiag pattern recognition toolbox," http://www.inf.ufes.br/~thomas/home/tooldiag.html.

[23] B. L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Trans. Circ. Syst. Vid. Tech.*, vol. 5, pp. 533-544, 1995.

[24] "TREC video retrieval evaluation," http://www-nlpir.nist.gov/projects/trecvid/.

[25] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[26] R. O. Duda, R. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[27] G. Van de Wouwer, P. Scheunders, and D. Van Dyck, "Statistical texture characterization from discrete wavelet representations," *IEEE Trans. Image Proc.*, vol. 8, pp. 592-598, 1999.

[28] "Kakadu software," http://www.kakadusoftware.com.

[29] K. A. Birney and T. R. Fischer, "On the modeling of DCT and subband image data for compression," *IEEE Trans. Image Processing*, vol. 4, pp. 186-193, 1995.

[30] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Trans. Image Proc.*, vol. 11, pp. 146-158, 2002.

[31] H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions," *J. Opt. Soc. Am. A.*, vol. 15, pp. 1520-1535, 1998.

[32] J. H. Shapiro, "Bounds on the area under the ROC curve," *J. Opt. Soc. Am. A.*, vol. 16, pp. 53-57, 1998.

[33] T. A. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[34] "Matlab implementation of the moment matching and maximum likelihood estimators for the generalized Gaussian density," http://www.ifp.uiuc.edu/~minhdo/ software/gg_mle.tar.gz.

[35] C. T. Yu and P. K. Varshney, "Bit allocation for discrete signal detection," *IEEE Trans. Commun.*, vol. 46, pp. 173-175, 1998.

[36] J. Hu and R. S. Blum, "On the optimality of finite-level quantization for distributed signal detection," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1665-1671, 2001.