Entropy Encoding in Wavelet Image Compression

Myung-Sin Song[1]

Department of Mathematics and Statistics, Southern Illinois University
Edwardsville `msong@siue.edu`

Summary. Entropy encoding which is a way of lossless compression that is
done on an image after the quantization stage. It enables to represent an image
in a more efficient way with smallest memory for storage or transmission. In
this paper we will explore various schemes of entropy encoding and how they
work mathematically where it applies.

1 Introduction

In the process of wavelet image compression, there are three major steps
that makes the compression possible, namely, decomposition, quanti-
zation and entropy encoding steps. While quantization may be a lossy
step where some quantity of data may be lost and may not be re-
covered, entropy encoding enables a lossless compression that further
compresses the data. [13], [18], [5]

In this paper we discuss various entropy encoding schemes that are
used by engineers (in various applications).

1.1 Wavelet Image Compression

In wavelet image compression, after the quantization step (see Figure 1)
entropy encoding, which is a lossless form of compression is performed
on a particular image for more efficient storage. Either 8 bits or 16 bits
are required to store a pixel on a digital image. With efficient entropy
encoding, we can use a smaller number of bits to represent a pixel in
an image; this results in less memory usage to store or even transmit
an image. Karhunen-Loève theorem enables us to pick the best basis
thus to minimize the entropy and error, to better represent an image
for optimal storage or transmission. Also, Shannon-Fano entropy (see

section 3.3), Huffman coding (see section 3.4), Kolmogorov entropy (see section 3.2) and arithmetic coding (see section 3.5) are ones that are used by engineers. Here, *optimal* means it uses least memory space to represent the data. i.e., instead of using 16 bits, use 11 bits. Thus, the best basis found would make it possible to represent the digital image with less storage memory. In addition, the choices made for entropy encoding varies; one might take into account the effectiveness of the coding and the degree of difficulty of implementation step into programming codes. We will also discuss how those preferences are made in section 3.
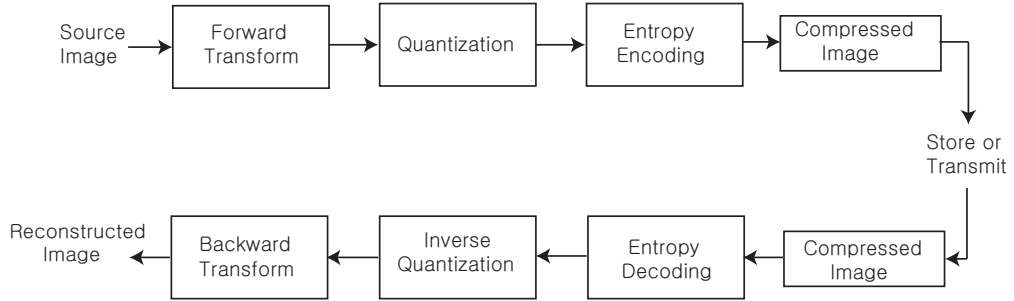

Fig. 1. Outline of the wavelet image compression process. [13]

1.2 Geometry in Hilbert Space

While finite or infinite families of nested subspaces are ubiquitous in mathematics, and have been popular in Hilbert space theory for generations (at least since the 1930s), this idea was revived in a different guise in 1986 by Stéphane Mallat, then an engineering graduate student. In its adaptation to wavelets, the idea is now referred to as the multiresolution method.

What made the idea especially popular in the wavelet community was that it offered a skeleton on which various discrete algorithms in applied mathematics could be attached and turned into wavelet constructions in harmonic analysis. In fact what we now call multiresolutions have come to signify a crucial link between the world of discrete wavelet algorithms, which are popular in computational mathematics and in engineering (signal/image processing, data mining, etc.) on the one side, and on the other side continuous wavelet bases in function spaces, especially in $L^2(\mathbb{R}^d)$. Further, the multiresolution idea closely mimics how fractals are analyzed with the use of finite function systems.

But in mathematics, or more precisely in operator theory, the underlying idea dates back to work of John von Neumann, Norbert Wiener, and Herman Wold, where nested and closed subspaces in Hilbert space were used extensively in an axiomatic approach to stationary processes, especially for time series. Wold proved that any (stationary) time series can be decomposed into two different parts: The first (deterministic) part can be exactly described by a linear combination of its own past, while the second part is the opposite extreme; it is *unitary*, in the language of von Neumann.

von Neumann's version of the same theorem is a pillar in operator theory. It states that every isometry in a Hilbert space $\mathcal{H}$ is the unique sum of a shift isometry and a unitary operator, i.e., the initial Hilbert space $\mathcal{H}$ splits canonically as an orthogonal sum of two subspaces $\mathcal{H}_s$ and $\mathcal{H}_u$ in $\mathcal{H}$, one which carries the shift operator, and the other $\mathcal{H}_u$ the unitary part. The shift isometry is defined from a nested scale of closed spaces V_n, such that the intersection of these spaces is $\mathcal{H}_u$. Specifically,

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset V_n \subset V_{n+1} \subset \cdots$$

$$\bigwedge_n V_n = \mathcal{H}_u, \text{ and } \bigvee_n V_n = \mathcal{H}.$$

An important fact about the wavelet aplication is that then $\mathcal{H}_u = \{0\}$

However, Stéphane Mallat was motivated instead by the notion of scales of resolutions in the sense of optics. This in turn is based on a certain "artificial-intelligence" approach to vision and optics, developed earlier by David Marr at MIT, an approach which imitates the mechanism of vision in the human eye.

The connection from these developments in the 1980s back to von Neumann is this: Each of the closed subspaces V_n corresponds to a level of resolution in such a way that a larger subspace represents a finer resolution. Resolutions are relative, not absolute! In this view, the relative complement of the smaller (or coarser) subspace in larger space then represents the visual detail which is added in passing from a blurred image to a finer one, i.e., to a finer visual resolution.

This view became an instant hit in the wavelet community, as it offered a repository for the fundamental father and the mother functions, also called the scaling function φ, and the wavelet function ψ. Via a system of translation and scaling operators, these functions then generate nested subspaces, and we recover the scaling identities which initialize the appropriate algorithms. What results is now called the family of pyramid algorithms in wavelet analysis. The approach itself

is called the multiresolution approach (MRA) to wavelets. And in the meantime various generalizations (GMRAs) have emerged.

Haar's work in 1909–1910 had implicitly the key idea which got wavelet mathematics started on a roll 75 years later with Yves Meyer, Ingrid Daubechies, Stéphane Mallat, and others—namely the idea of a multiresolution. In that respect Haar was ahead of his time. See Figures 2 and 3 for details.
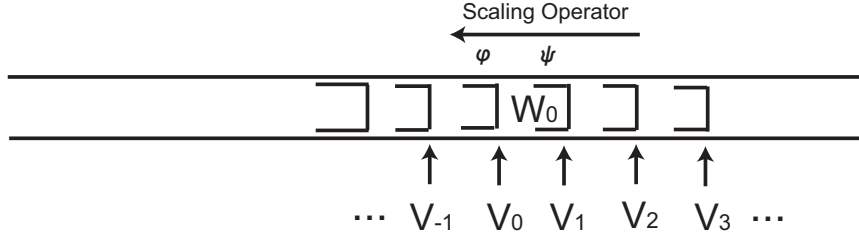
Fig. 2. Multiresolution. $L^2(\mathbb{R}^d)$-version (continuous); $\varphi \in V_0$, $\psi \in W_0$.
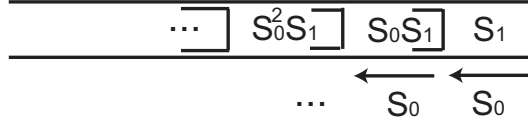
Fig. 3. Multiresolution. $l^2(\mathbb{Z})$-version (discrete); $\varphi \in V_0$, $\psi \in W_0$.

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots, \ V_0 + W_0 = V_1$$

The word "multiresolution" suggests a connection to optics from physics. So that should have been a hint to mathematicians to take a closer look at trends in signal and image processing! Moreover, even staying within mathematics, it turns out that as a general notion this same idea of a "multiresolution" has long roots in mathematics, even in such modern and pure areas as operator theory and Hilbert-space geometry. Looking even closer at these interconnections, we can now recognize scales of subspaces (so-called multiresolutions) in classical algorithmic construction of orthogonal bases in inner-product spaces, now taught in lots of mathematics courses under the name of the Gram–Schmidt algorithm. Indeed, a closer look at good old Gram–Schmidt reveals that it is a matrix algorithm, Hence new mathematical tools involving non-commutativity!

If the signal to be analyzed is an image, then why not select a fixed but suitable *resolution* (or a subspace of signals corresponding to a selected resolution), and then do the computations there? The selection of a fixed "resolution" is dictated by practical concerns. That idea was key in turning computation of wavelet coefficients into iterated matrix algorithms. As the matrix operations get large, the computation is carried out in a variety of paths arising from big matrix products. The dichotomy, continuous vs. discrete, is quite familiar to engineers. The industrial engineers typically work with huge volumes of numbers.

In the formulas, we have the following two indexed number systems $\mathbf{a} := (h_i)$ and $\mathbf{d} := (g_i)$, $\mathbf{a}$ is for averages, and $\mathbf{d}$ is for local differences. They are really the input for the DWT. But they also are the key link between the two transforms, the discrete and continuous. The link is made up of the following scaling identities:

$$\varphi(x) = 2 \sum_{i \in \mathbb{Z}} h_i \varphi(2x - i);$$

$$\psi(x) = 2 \sum_{i \in \mathbb{Z}} g_i \varphi(2x - i);$$

and (low-pass normalization) $\sum_{i \in \mathbb{Z}} h_i = 1$. The scalars (h_i) may be real or complex; they may be finite or infinite in number. If there are four of them, it is called the "four tap", etc. The finite case is best for computations since it corresponds to compactly supported functions. This means that the two functions φ and ψ will vanish outside some finite interval on a real line.

The two number systems are further subjected to orthgonality relations, of which

$$\sum_{i \in \mathbb{Z}} \bar{h}_i h_{i+2k} = \frac{1}{2} \delta_{0,k} \tag{1}$$

is the best known.

Our next section outlines on how the whole wavelet image compression process works step by step.

In our next section we give the general context and definitions from operators in Hilbert space which we shall need: We discuss the particular orthonomal bases (ONBs) and frames which we use, and we recall the operator theoretic context of the Karhunen-Loève theorem [1]. In approximation problems involving a stochastic component (for example noise removal in time-series or data resulting from image processing) one typically ends up with correlation kernels; in some cases as frame kernels; see [10]. In some cases they arise from systems of vectors in

Hilbert space which form frames (see [10]). In some cases parts of the frame vectors fuse (fusion-frames) onto closed subspaces, and we will be working with the corresponding family of (orthogonal) projections. Either way, we arrive at a family of selfadjoint positive semidefinite operators in Hilbert space. The particular Hilbert space depends on the application at hand. While the Spectral Theorem does allow us to diagonalize these operators, the direct application the Spectral Theorem may lead to continuous spectrum which is not directly useful in computations, or it may not be computable by recursive algorithms.

The questions we address are optimality of approximation in a variety of ONBs, and the choice of the "best" ONB. Here "best" is given two precise meanings: (1) In the computation of a sequence of approximations to the frame vectors, the error terms must be smallest possible; and similarly (2) we wish to minimize the corresponding sequence of entropy numbers (referring to von Neumann's entropy). In two theorems we make precise an operator theoretic Karhunen-Loève basis, which we show is optimal both in regards to criteria (1) and (2). But before we prove our theorems, we give the two problems an operator theoretic formulation; and in fact our theorems are stated in this operator theoretic context.

2 How it works

In wavelet image compression, wavelet decomposition is performed on a digital image. Here, an image is treated as a matrix of functions where the entries are pixels. The following is an example of a representation for a digitized image function:

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} f(0,0) & f(0,1) & \cdots & f(0, N-1) \\ f(1,0) & f(1,1) & \cdots & f(1, N-1) \\ \vdots & \vdots & \vdots & \vdots \\ f(M-1,0) & f(M-1,1) & \cdots & f(M-1, N-1) \end{pmatrix}. \quad (2)$$

After the decomposition quantization is performed on the image. The quantization maybe a lossy (meaning some information is being lost) or lossless. Then a lossless means of compression, entropy encoding is done on the image to minimize the memory space for storage or transmission. Here the mechanism of entropy will be discussed.

2.1 Entropy Encoding

In most images their neighboring pixels are correlated and thus contain redundant information. Our task is to to find less correlated representation of the image, then perform redundancy reduction and irrelavancy reduction. Redundancy reduction removes duplication from the signal source (for instance a digital image). Irrelavancy reduction omits parts of the signal that will not be noticed by the Human Visual System (HVS).

Entropy encoding further compresses the quantized values in lossless manner which gives better compression in overall. It uses a model to accurately determine the probabilities for each quantized value and produces an appropriate code based on these probabilities so that the resultant output code stream will be smaller than the input stream.

Some Terminology

(i) Spatial Redundancy : correlation between neighboring pixel values.
(ii) Spectral Redundancy : correlation between different color planes or spectral bands.

When a digital image is 1-level wavelet decomposed from the matrix representation in section 2

Inside the paper we use (φ_i) and (ψ_i) to denote generic ONBs. However, in wavelet theory, [4] there is a tradition for reserving φ for the father function and ψ for the mother function. A 1-level wavelet transform of an $N \times M$ image can be represented as

$$\mathbf{f} \mapsto \begin{pmatrix} \mathbf{a}^1 & | & \mathbf{h}^1 \\ -- & & -- \\ \mathbf{v}^1 & | & \mathbf{d}^1 \end{pmatrix} \tag{3}$$

where the subimages $\mathbf{h}^1, \mathbf{d}^1, \mathbf{a}^1$ and $\mathbf{v}^1$ each have the dimension of $N/2$ by $M/2$.

$$\mathbf{a}^1 = V_m^1 \otimes V_n^1 : \varphi^A(x,y) = \varphi(x)\varphi(y) = \sum_i \sum_j h_i h_j \varphi(2x-i)\varphi(2y-j)$$
$$\mathbf{h}^1 = V_m^1 \otimes W_n^1 : \psi^H(x,y) = \psi(x)\varphi(y) = \sum_i \sum_j g_i h_j \psi(2x-i)\varphi(2y-j)$$
$$\mathbf{v}^1 = W_m^1 \otimes V_n^1 : \psi^V(x,y) = \varphi(x)\psi(y) = \sum_i \sum_j h_i g_j \varphi(2x-i)\psi(2y-j)$$
$$\mathbf{d}^1 = W_m^1 \otimes W_n^1 : \psi^D(x,y) = \psi(x)\psi(y) = \sum_i \sum_j g_i g_j \psi(2x-i)\psi(2y-j)$$
$$\tag{4}$$

where φ is the father function and ψ is the mother function in sense of wavelet, V space denotes the average space and the W spaces are the

difference space from multiresolution analysis (MRA) [4]. h and g are both low-pass and high-pass filter coefficients.

$\mathbf{a}^1$: the first averaged image, which consists of average intensity values of the original image. Note that only φ function, V space and h coefficients are used here.

$\mathbf{h}^1$: the first detail image of horizontal components, which consists of intensity difference along the vertical axis of the original image. Note that φ function is used on y and ψ function on x, W space for x values and V space for y values; and both h and g coefficients are used accordingly.

$\mathbf{v}^1$: first detail image of vertical components, which consists of intensity difference along the horizontal axis of the original image. Note that φ function is used on x and ψ function on y, W space for y values and V space for x values; and both h and g coefficients are used accordingly.

$\mathbf{d}^1$: the first detail image of diagonal components, which consists of intensity difference along the diagonal axis of the original image. The original image is reconstructed from the decomposed image by taking the sum of the averaged image and the detail images and scaling by a scaling factor. It could be noted that only ψ function, W space and g coefficients are used here.

See [19], [17].

This decomposition not only limits to one step but it can be done again and again on the averaged detail depending on the size of the image. Once it stops at certain level, quantization (see [15], [13], [18]) is done on the image. This quantization step may be lossy or lossless. Then the lossless entropy encoding is done on the decomposed and quantized image as Figure 6.

The above figure illustrates on how mathematically wavelet image decomposition is done. An example would illustrate how average, horizontal, vertical and diagonal details are obtained through the wavelet decomposition of a digital image of an octagon.

There are various means of quantization and one commonly used one is called thresholding. Thresholding is a method of data reduction where it puts 0 for the pixel values below the thresholding value or something other 'appropriate' value. Soft thresholding is defined as follows:

$$T_{soft}(x) = \begin{cases} 0 & \text{if } |x| \leq \lambda \\ x - \lambda & \text{if } x > \lambda \\ x + \lambda & \text{if } x < -\lambda \end{cases} \tag{5}$$
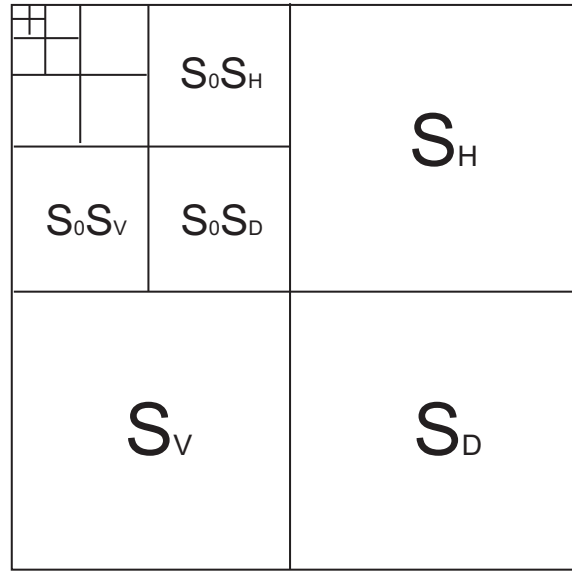
Fig. 4. How the subdivision works.
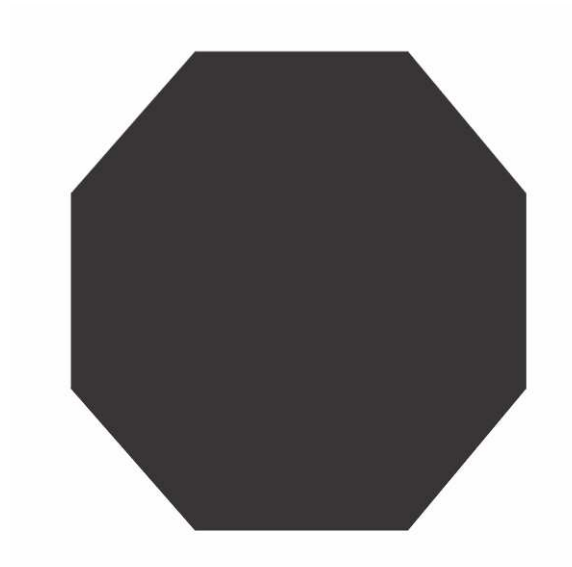
and hard thresholding as follows:

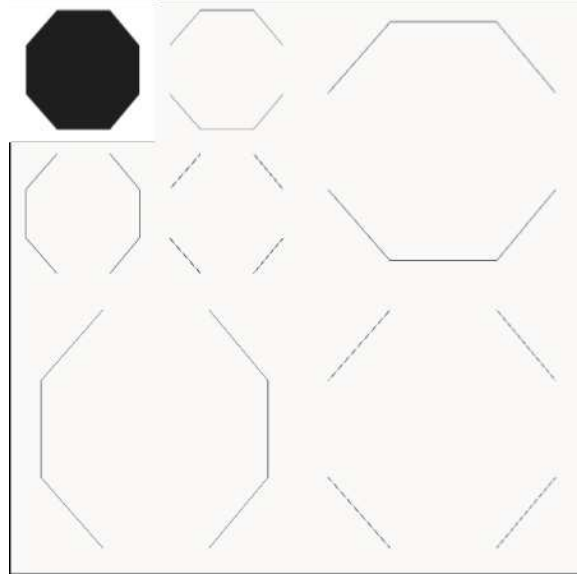

Fig. 5. Original octagon image.

Fig. 6. Octagon after 2-level decomposition

$$T_{hard}(x) = \begin{cases} 0 & \text{if } |x| \leq \lambda \\ x & \text{if } |x| > \lambda \end{cases} \tag{6}$$

where $\lambda \in \mathbb{R}_+$ and x is a pixel value. It could be observed by looking at the definitions, the difference between them is related to how the coefficients larger than a threshold value λ in absolute values are handled. In hard thresholding, these coefficient values are left alone. Where else, in soft thresholding, the coefficient values area decreased by λ if positive and increased by λ if negative [20]. Also, see [19], [8], [17].

Another way of quantization is as follows:

Definition 1. *Let X be a set, and K be a discrete set. Let Q and D be mappings $Q : X \rightarrow K$ and $D : K \rightarrow X$. Q and D are such that*

$$\|x - D(Q(x))\| \leq \|x - D(d)\|, \qquad for\ all \qquad d \in K$$

Applying Q to some $x \in X$ is called quantization, and $Q(x)$ is the quantized valued of x. Likewise, applying D to some $k \in K$ is called dequantization and $D(k)$ is the dequantized value of k. [15]

During the quatization process, the number of bits needed to store the wavelet transformed coefficients by reducing the precision of the values. This is a many-to-one mapping, meaning that it is a lossy process resulting in lossy compression.

Entropy encoding further compresses the quantized values in lossless manner which gives better compression in overall. It uses a model to accurately determine the probabilities for each quantized value and produces an appropriate code based on these probabilities so that the resultant output code stream will be smaller than the input stream.

2.2 Benefits of Entropy Encoding

One might think that the quantization step suffices for compression. It is true that the quantization does compress the data tremendously. After the quantization step many of the pixel values are either eliminated or replaced with other suitable values. However, those pixel values are still represented with either 8 or 16 bits. See 1.1. So we aim to minimize the number of bits used by means of entropy encoding. Karhunen-Loève transform or PCAs makes it possible to represent each pixel on the digital image with the least bit representation accoding to their probability thus yields the lossless optimized representation using least amount of memory.

3 Various entropy encoding schemes

In this section we discuss various entropy encoding schemes on how they work, the mathematics behind it.

3.1 The Kahunen-Loève transform

Karhunen-Loève transform also known as Principal Components Analysis (PCA) allows us to better represent each pixels on the image matrix using the smallest number of bits. It makes us possible to assign the smallest number of bits for the the pixel that has the highest probability, then the next number to the pixel value that has second highest probabilty, and so forth; thus the pixel that has smallest probability gets assigned the highest value among all the other pixel values.

An example with letters in the text would better depict how the mechanism works. Suppose we have a text with letters a, e, f, q, in order of probabilty. That is, 'a' shows up most frequently and 'q' shows up least frequently. Then we would assign 00 to 'a', then 01 to 'e', 100 to 'f', and 101 to 'q'.

In general, one refers to a *Karhunen-Loève transform* as an expansion in Hilbert space with respect to an ONB resulting from an application of the Spectral-Theorem.

The Algorithm

Our aim is to reduce the number of bits needed to represent an image by removing redundancies as much as possible.

The algorithm for entropy encoding using Karhunen-Loève expansion can be described as follows:

1. Perform the wavelet transform for the whole image. (ie. wavelet decomposition.)
2. Do quantization to all coefficients in the image matrix, except the average detail.
3. Subtract the mean: Subtract the mean from each of the data dimensions. This produces a data set whose mean is zero.
4. Compute the covariance matrix.

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

5. Compute the eigenvectors and eigenvalues of the covariance matrix.
6. Choose components and form a feature vector(matrix of vectors),

$$(eig_1, ..., eig_n)$$

   Eigenvectors are in order of their eigenvalues. Eigenvalues found in step 5 are different in values. The eigenvector with highest eigenvalue is the principle component of the data set.
7. Derive the new data set. $FinalData = RowFeatureVector \times RowDataAdjust$ where $RowDataAdjust$ is the mean-adjusted data transposed. [15]

Starting with a matrix representation for a particular image, we then compute the covariance matrix using the steps from (3) and (4) in algorithm above. We then compute the Karhunen-Loève eigenvalues. Next, the eigenvalues arranged in decreasing order. The corresponding eigenvectors are arranged to match the eigenvalues with multiplicity. The eigenvalues mention here are the same eigenvalues λ_i in this section, thus yielding smallest error and smallest entropy in the computation.

In computing probabilities and entropy, Hilbert space serves as a helpful tool. For example, take a unit vector f in some fixed Hilbert space $\mathcal{H}$, and an orthonormal basis (ONB) ψ_i with i running onto an indext set I. With this we now introduce to families of probability measures one $P_f(\cdot)$ indexed by $f \in \mathcal{H}$, and a second P_T indexed by a class of operators $T : \mathcal{H} \to \mathcal{H}$.

Definition 2. *Let $\mathcal{H}$ be a Hilbert space. Let (ψ_i) and (ϕ_i) be orthonormal bases (ONB), with index set I. Usually*

$$I = \mathbb{N} = \{1, 2, ...\}. \tag{7}$$

If $(\psi_i)_{i \in I}$ is an ONB, we set $Q_n :=$ the orthogonal projection onto $span\{\psi_1, ..., \psi_n\}$.

We now introduce a few facts about operators which will be needed in the paper. In particular we recall Dirac's terminology [**?**] for rank-one operators in Hilbert space. While there are alternative notation available, Dirac's bra-ket terminology is especially efficient for our present considerations.

Definition 3. *Let vectors u, $v \in \mathcal{H}$. Then*

$$\langle u|v \rangle = inner\ product \in \mathbb{C}, \tag{8}$$

$$|u\rangle\langle v| = rank\text{-}one\ operator, \mathcal{H} \to \mathcal{H}, \tag{9}$$

where the operator $|u\rangle\langle v|$ acts as follows

$$|u\rangle\langle v|w = |u\rangle\langle v|w\rangle = \langle u|w\rangle u, \quad for\ all\ w \in \mathcal{H}. \tag{10}$$

Consider an ensemble of a large number N of similar objects, of which Nw^α, $\alpha = 1, 2, ..., \nu$ where the relative frequency w^α satisfies the probability axioms:

$$w^\alpha \geq 0, \ \sum_{\alpha=1}^{\nu} w^\alpha = 1$$

Assume that each type specified by a value of the index α is represented by $f^\alpha(\xi)$ in a real domain $[a, b]$, which we normalize by

$$\int_a^b |f^\alpha(\xi)|^2 d\xi = 1$$

Let $\{\psi_i(\xi)\}$, $i = 1, 2, ...$, be a complete set of orthonomal base functions defined on $[a, b]$ Then any function $f^\alpha(\xi)$ can be expanded as

$$f^\alpha(\xi) = \sum_{i=1}^{\infty} x_i^{(\alpha)} \psi_i(\xi) \tag{11}$$

with

$$x_i^\alpha = \int_a^b \psi_i^*(\xi) f^\alpha(\xi) d\xi. \tag{12}$$

Here, x_i^α is the component of f^α in ψ_i coordinate system. With the normalization of f^α we have

$$\sum_{i=1}^{\infty} |x_i^\alpha|^2 = 1 \tag{13}$$

Then substituting (12) in (11) gives

$$f^\alpha(\xi) = \int_a^b f^\alpha(\eta)[\sum_{i=1}^{\infty} \psi_i^*(\eta)\psi_i(\eta)]d\eta = \sum_{i=1}^{\infty} \langle \psi_i(\eta)|f^\alpha\rangle \psi_i \tag{14}$$

in definition of ONB.

Let $\mathcal{H} = L^2(a,b)$. $\psi_i : \mathcal{H} \to l^2(\mathbb{Z})$ and $U : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$ where U is a unitary operator

Note that the distance is invariant under a unitary transformation. Thus, using another coordinate system $\{\phi_j\}$ in place of $\{\psi_i\}$, would not change the distance.

Let $\{\phi_j\}$, $j = 1,2,...$, be another set of ONB functions instead of $\{\psi_i(\xi)\}$, $i = 1,2,...,$. Let y_j^α be the component of f^α in $\{\phi_j\}$ where it can be expressed in terms of x_i^α by a linear relation $y_j^\alpha = \sum_{i=1}^{\infty}\langle \phi_j, \psi_i\rangle x_i^\alpha = \sum_{i=1}^{\infty} U_{i,j}x_i^\alpha$ where $U : l^2(\mathbb{Z}) \to l^2(\mathbb{Z})$, U is a unitary operator matrix $U_{i,j} = \langle \phi_j, \psi_i\rangle = \int_a^b \phi_j^*(\xi)\psi_i(\xi)d\xi$ Also, x_i^α can be written in terms of y_j^α under the following relation $x_i^\alpha = \sum_{j=1}^{\infty}\langle \psi_i, \phi_j\rangle y_j^\alpha = \sum_{j=1}^{\infty} U_{i,j}^{-1} y_j^\alpha$ where $U_{i,j}^{-1} = \overline{U_{i,j}}$ and $\overline{U_{i,j}} = U_{j,i}^*$

$$f^\alpha(\xi) = \sum_{i=1}^{\infty} x_i^\alpha(\xi)\psi_i(\xi) = \sum y_i^\alpha(\xi)\phi_i(\xi)$$

So $U(x_i) = (y_i)$ and $\sum_{i=1}^{\infty} x_i^\alpha \psi_i(\xi) = \sum_{j=1}^{\infty} y_j^\alpha \phi_j(\xi)$

$$x_i^\alpha = <\psi_i, f^\alpha> = \int_a^b \psi_i(\xi)f^{(\alpha)}(\xi)d\xi$$

The squared magnitude $|x_i^{(\alpha)}|^2$ of the coefficient for ψ_i in the expansion of $f^{(\alpha)}$ can be considered as a good measure of the average in the ensemble

$$Q_i = \sum_{\alpha=1}^{n} w^{(\alpha)}|x_i^{(\alpha)}|^2$$

can be considered as the measure of importance of $\{\psi_i\}$.

$$Q_i \geq 0, \quad \sum_i Q_i = 1$$

Then the entropy function in terms of the Q_i's is defined as

$$S(\{\psi_i\}) = -\sum_i Q_i \log Q_i.$$

We are interested in minimizing the entropy, that is if $\{\Theta_j\}$ is one such optimal coordinate system, we shall have

$$S(\{\Theta_j\}) = min_{\{\psi_j\}} S(\{\psi_i\})$$

Let $G(\xi,\xi') = \sum_\alpha w^\alpha f^\alpha(\xi) f^{\alpha*}(\xi')$. Then G is a Hermitian matrix and $Q_i = G(i,i) = \sum_\alpha w^\alpha x_i^\alpha x_i^{\alpha*}$ where normalization of $\sum Q_i = 1$ give us trace $G = 1$ where the trace means the diagonal sum.

Then define a special function system $\{\Theta_k(\xi)\}$ as the set of eigen-functions of G, i.e.

$$\int_a^b G(\xi,\xi')\Theta_k(\xi)d\xi' = \lambda_k \Theta_k(\xi). \tag{15}$$

so $G\Theta_k(\xi) = \lambda_k \Theta_k(\xi)$.

When the date are not functions but vectors v^αs whose components are $x_i^{(\alpha)}$ in the ψ_i coordinate system, we have

$$\sum_{i'} G(i,i') t_{i'}^k = \lambda_k t_i^k \tag{16}$$

where t_i^k is the $i - th$ component of the vector Θ_k in the coordinate system $\{\psi_i\}$. So we get $\psi : \mathcal{H} \to (x_i)$ and also $\Theta : \mathcal{H} \to (t_i)$. The two ONBs result in

$$x_i^\alpha = \sum_k c_k^\alpha t_i^k \text{ for all } i, c_k^\alpha = \sum_i t_i^{k*} x_i^\alpha$$

which is Karhunen-Loève expansion of $f^\alpha(\xi)$ or vector v^α. Then $\{\Theta_k(\xi)\}$ is the K-L coordinate system dependent on $\{w^\alpha\}$ and $\{f^\alpha(\xi)\}$. Then we arrange the corresponding functions or vectors in the order of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1} \geq \lambda_k \geq \ldots$.

Now, $Q_i = G_{i,i} = \langle \psi_i G \psi_i \rangle = \sum_k A_{ik} \lambda_k$ where $A_{ik} = t_i^k t_i^{k*}$ which is a double stochastic matrix. Then

$$G = U \begin{pmatrix} \lambda_1 & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & \cdots & \lambda_k \end{pmatrix} U^{-1}$$

3.2 Kolmogorov Entropy

This is an example of hard implementation into coding. Thus, is not very commonly used in industry compared to other methods mentioned.

Implementation

Let X be a metric space with distance function ρ. If $f \in X$ and $r > 0$, let

$$\mathbf{B}(f, r) := \mathbf{B}(f, r)x := \{g \in X : \rho(f, g) < r\}.$$

be the open ball with radius r centered at f. For $K \subset X$ compact, there is a finite collection of balls $\mathbf{B}(f_i, \epsilon)$, $i = 1, \ldots, n$, for each $\epsilon > 0$, which cover K: $K \subset \bigcup_{i=1}^{n} \mathbf{B}(f_i, \epsilon)$. Then the covering number $N_\epsilon(K) := N_\epsilon(K, X)$ is the smallest integer n for which there is such an ϵ−covering of K.

Definition 4. *The Kolmogorov ϵ−entropy of K is defined as*

$$H_\epsilon(K) := H_\epsilon(K, X) := \log N_\epsilon(K), \ \epsilon > 0.$$

where the log is the logarithm to the base two. [3]

3.3 Shannon-Fano Entropy

For each data on an image, ie. pixel, a set of probabilities p_i is computed, where $\sum_{i=1}^{n} p_i = 1$. The entropy of this set gives the measure of how much choice is involved, in the selection of the pixel value on average.

Definition 5. *Shannon's entropy $E(p_1, p_2, \ldots, p_n)$ which satisfy the following:*

- *E is a continuous function of p_i.*
- *E should be steadily increasing function of n.*
- *If the choice is made in k successive stages, then $E =$ sum of the entropies of choices at each stage, with weights corresponding to the probabilities of the stages.*

$E = -k \sum_{i=1}^{n} p_i \log p_i$. k *controls the units of the entropy, which is* "bits." *logs are taken base 2.* [2, 14]

Shannon-Fano entropy encoding is done according to the probabilities of data and the method is as follows:

- The data is listed with their probabilities in decreasing order of their probabilities.
- The list is divided into two parts that has roughly equal probability.
- Start the code for those data in the first part with a 0 bit and for those in the second part with a 1.
- Continue recursively until each subdivision contains just one data. [2, 14]

3.4 Huffman Coding

This was developed by Huffmand shortly after Shannon's work. This gives a greater compression compared to Shannon entropy encoding. Huffman coding is done as follows:

- The data is listed with their probabilities.
- The two data with the smallest probabilities are located.
- The two data are replaced by a single set containing both, whose probability is the sum of the individual probabilities.
- These steps are repeated until the list is left with only one member.

See [2].

3.5 Arithmetic Coding

This is one of the latest and popular encoding scheme. In arithmetic coding, symbols are restricted in such a way that tranlation is done into an integral number of bits, thus making the coding more efficient. In this coding, the data is represented by an interval of real numbers between 0 and 1. As the data becomes larger, the interval required for representation becomes smaller, and the number of bits required to specify that interval increases. Successive symbols of the data reduce the size of the interval according to the probabilities of the symbol generated by the model. The data that is more likely has more reduced ranged compared to the unlikely data, thus fewer bits are used. [2, 22]

The above mentioned entropy encoding schemes are chosen in application in wavelet image compression with the preference of coding simplicity, effectiveness in minimization of entropy and the lossless compression ratio.

Acknowledgement

Peter Massopust, Dorin Dutkay, Simon Alexander, for helpful discussions, and Professor Victor Wickerhauser for suggesting [1, 7], Professor Brody Johnson for suggesting [21], .

References

1. Ash RB (1990) Information theory. Corrected reprint of the 1965 original. Dover Publications, Inc., New York
2. Bell TC, Cleary JG, Witten IH (1990) Text Compression. Prentice Hall, Englewood Cliffs
3. Cohen A, Dahmen W, Daubechies I, DeVore R (2001) Tree Approximation and Optimal Encoding. Applied Computational Harmonic Analysis 11:192–226
4. Daubechies I (1992) Ten Lectures on Wavelets. SIAM
5. Donoho DL, Vetterli M, DeVore RA, Daubechies I (Oct. 1998) Data Compression and Harmonic Analysis. IEEE Trans. Inf. Theory, 44 (6):2435–2476
6. Dirac PAM (1947) The Principles of Quantum Mechanics. 3d ed Oxford, at the Clarendon Press
7. Effros M, Feng H, Zeger K (Aug. 2004) Suboptimality of the Karhunen-Loève Transform for Transform Coding. IEEE Trans. Inf. Theory, 50 (8):1605–1619
8. Field DJ (1999) Wavelets, vision and the statistics of natural scenes. Phil. Trans. R. Soc. Lond. A 357:2527–2542
9. Gonzalez RC, Woods RE, Eddins SL (2004) Digital Image Processing Using MATLAB. Prentice Hall
10. Jorgensen PET (2006) Analysis and Probability Wavelets, Signals, Fractals. Springer, Berlin Heidelberg New York
11. Jorgensen PET, Song M-S (2007) Entropy Encoding using Hilbert Space and Karhunen-Loève Transforms. preprint
12. Pierce JR (1980) An Introduction to Information Theory Symbols, Signals and Noise. 2nd Edition Dover Publications, Inc., New York
13. Schwab C, Todor RA (2006) Karhunen-Loève approximation of random fields by generalized fast multipole methods. Journal of Computational Physics 217, Elsevier
14. Skodras A, Christopoulos C, Ebrahimi T (Sept. 2001) JPEG 2000 Still Image Compression Standard. IEEE Signal processing Magazine 18:36–58
15. Shannon CE, Weaver W (1998) The Mathematical Theory of Communication. University of Illinois Press, Urbana and Chicago
16. Smith LI (2002) A Tutorial on Principal Components Analysis. http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
17. Song M-S (2005) Wavelet image compression. Ph.D. thesis The University of Iowa, Iowa

18. Song M-S (2006) Wavelet image compression. Operator theory, operator algebras, and applications Contemp. Math. 414:41–73, Amer. Math. Soc., Providence, RI
19. Usevitch BE (Sept. 2001) A Tutorial on Modern Lossy Wavelet Image Compression: Foundations of JPEG 2000. IEEE Signal processing Magazine 18:22–35
20. Walker JS (1999) A Primer on Wavelets and their Scientific Applications. Chapman & Hall, CRC
21. Walnut DF (2002) An Introduction to Wavelet Analysis. Birkhäuser
22. Watanabe S (1965) Karhunen-Loève Expansion and Factor Analysis Theoretical Remarks and Applications Transactions of the Fourth Prague Conference on Information Theory Statistical Decision Functions Random Process. Adademia Press
23. Witten IH, Neal RM, Cleary JG, (June 1987) Arithmetic Coding for Data Compression. Communications of the ACM 30 (6):520–540