1998

Lossless Set Compression of Correlated Information.

Oleg Stanislavovich Pianykh
Louisiana State University and Agricultural & Mechanical College

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

LOSSLESS SET COMPRESSION OF CORRELATED INFORMATION

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Computer Science

by
Oleg S. Pianykh
M.S., Moscow State University, 1994
August 1998

To my parents

Acknowledgments

I would like to thank all faculty, present and former students of the Louisiana State University who directly or implicitly contributed to this work, and primarily

Dr. John M. Tyler, my major professor, for making my research both creative and enjoyable,

My defense committee members:

Dr. Warren N. Waggenspack, Jr.

Dr. S. Kundu,

Dr. J. Hurrelbrink,

Dr. S.Q. Zheng

for making my work more accurate,

Computer Science Department of Louisiana State University for giving me the knowledge and the opportunity to apply it,

SPIE Society for helping me to present my research,

Ochsner Clinic of Baton Rouge for filling the gaps in my medical background,

M.D. Anderson Cancer Center for providing all data used in this compression study.

Oleg S. Pianykh

TABLE OF CONTENTS

Abstract

Set compression allows the compression a set of similar (correlated) images more efficiently than compressing the same images independently. Currently, set compression is performed with different inter-image predictive models, that forecast the common image properties from a few reference images. With sufficient inter-image correlation, one can predict any database image from a few templates, hence avoiding inter-image redundancy and achieving much improved compression ratios. This research focused on two major aspects of this technique: the practical limits of the predictive set compression, and the theoretical estimates of the compression efficiency. This includes a review of the previous work in set compression area, a discussion of the more important statistical and informational aspects involved in predictive set compression, practical observations and measurements for medical (CT and MR) data, and theoretical analysis of lossless similar image compression. This research proposes new and more reliable approaches for lossless set compression, as well as their extensions to more general lossy set compression.

Introduction: The Similar Image Compression Problem

In modern science and technology, the amount of information produced, analyzed and stored is increasing and has created a constant quest for improving data compression techniques. *Data compression* means storing and transmitting information in its most compact form, achievable through removal of data redundancies. All compression techniques can be subdivided into two groups: *lossy* (irreversible) and *lossless* (reversible) compression,depending on the reversibility of this removal. *Lossy* compression can achieve high compression ratios (the ratio of the original to the compressed information size), usually sacrificing the least important details. Lossy compression is commonly used to compress images and sounds, usually with compression ratios varying from 3 to more than 100 depending on the level of detail preserved. In contrast, *lossless* compression does not sacrifice any information. This permits a complete recovery of the original data from its compressed form, but yields more moderate compression ratios of 1.5-2. Lossless compression is required when information cannot be lost nor altered; e.g., compression of text or medical images.

Historically, both lossy and lossless techniques were developed to compress single data items like single images, signals, data files, etc. However, many modern data-producing applications such as medical imaging create large sets of very similar, rather than independent, data. For instance, in a set of computer tomography (CT) brain image scans $V = \{v_1, ..., v_n\}$ [1] , the average inter-image correlation $\overline{\rho}(v_i, v_j)$ typically exceeds 0.75, which is accompanied by similar image patterns in shape and

[1] Where v_i represents an image stored as a sequence of pixel intensities.

intensity. This appearance of common image structures across all database images demonstrates the presence of inter-image redundancy, leading to the idea of the compression. *Set compression*, either lossy or lossless, assumes that a set of similar data items can be compressed more efficiently than compressing each item separately, if the inter-item redundancy is exploited. Careful removal of repeating patterns from a set of similar entities can lead to substantial information reduction which, followed by traditional single entity compression, produces higher compression ratios than compressing all the similar entities independently.

A primary goal of this research was to investigate the relation between inter-image similarities and resulting set compression efficiency. Major previously used techniques such as inter-frame prediction were studied first, and their limitations in similar database compression were demonstrated. Then the similarities between images and image-compressing transforms are investigated, which lead to the alternative "common transform" approach to similar data compression. We illustrate this approach with common autoregressive (CAR) compression for CT (computer tomography) and MR (magnetic resonance) image databases, which in this research improved the lossless compression ratio from 2:1 for a single image to more than 3:1 for a database. Next, an information theory for integer correlated sources is proposed. Information redundancy between integer correlated sources is analyzed as a function of their correlation. Finally, some more general extensions to lossless set compression are discussed, and conclusions are given.

Image Entropy

The lossy image compression ratio has virtually no upper bound: the choice of the amount of information one wants to sacrifice is always subjective and depends on the particular application. Conversely, the entropy H (informational content) of the information source v puts a lower boundary on the losslessly compressed source size. If v is a sequence of numbers from the set $\{n_1, n_2, ..., n_k\}$, and the probability of each number n_j to appear in v is $p_j = P(n_j)$, then Shannon entropy [1] of v is

$$H(v) = -\sum_{j=1}^{k} p_j \log_2(p_j). \tag{1}$$

By definition, $H(v)$ measures the amount of information provided by an observation of v [2]. It is also often interpreted as an averaged uncertainty about v, the "randomness" of v, or the average number of bits necessary to code v as a memoryless Markov source (the average number of bits in Huffman code table[2] for v).

As a function of probabilities p_j, $H(v)$ is continuous, positive and a concave mapping from $[0,1]^k \in R^k$ into $[0,1]$ (see Figure 1 for $k = 2$ and 3). Since for any $p \in [0,1]$, $(-p\log_2(p)) \in [0,1]$ as well[3] , the only case when $H(v) = 0$ is when all p_j except one are equal to zero. We list below a few other useful and less obvious entropy properties that can be found with their proofs in [3] and [4]:

[2] Huffman compression encodes each number n_i replacing it with a unique codeword c_i of length $\left\lceil \log_2 \frac{1}{p_i} \right\rceil$, which results in the average (expected) code length equal to $H(v) = \sum_{j=1}^{k} p_j \text{length}(c_i)$. With Huffman encoding, compression is achieved because the most probable numbers n_i (i.e., n_i with highest p_i) will be replaced with the shortest codewords c_i.

[3] $\lim_{p \to 0} p \log_2 p = 0$.

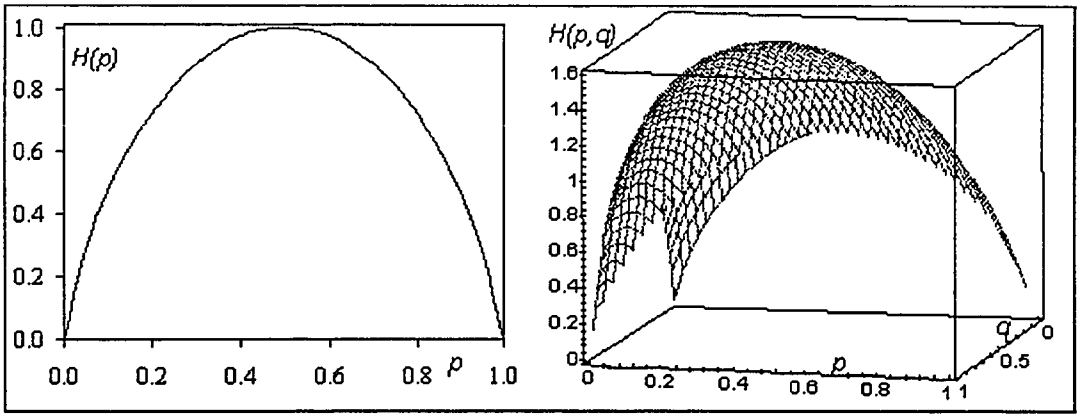

Figure 1: Entropy function in 2D and 3D.

For any random variables:

1. $H(u|v) \leqslant H(u)$ (where $u|v$ is a conditional probability distribution for u given v) - additional information never increases average uncertainty.

2. $H(u_1,u_2,\ldots\ u_n) \leqslant \sum_{i=1}^{n} H(u_i)$ - the entropy of an event consisting of several random events u_i never exceeds the sum of their entropies. The equality holds only if all events u_i are independent.

For probability distributions:

1. $H(p_1,p_2,...,p_n) \leqslant H(\frac{1}{n},\frac{1}{n},...\frac{1}{n}) = \log_2(n)$ - entropy is maximized by uniform distribution.

2. $H(p_1,p_2,p_3...,p_n) = H(p_1+p_2,p_3...,p_n) + (p_1+p_2)H(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}) \geqslant H(p_1+p_2,p_3...,p_n)$ - recursivity. Combining any two cases into one decreases the entropy.

In digital signal processing, all images are represented and analyzed as numerical sequences of their intensity values [5]. In particular, if an image v has all intensities $v[i]$ equal to the same value, then $k=1$ and $p_1=1$. This yields $H(v)=0$ in (1) and can be interpreted as no information contained in image v. Inversely, if v contains

multiple intensities, its entropy becomes positive, producing rich information. The higher $H(v)$, the more information v has, and the more difficult it is to compress. Lossless compression techniques such as Huffman compression or arithmetic coding produce compression ratios close to the source entropy; therefore we will use "information" and "entropy" interchangeably, and will also refer to the entropy as the measure of the compressed image size.

Intuitive Definition of Image Similarity

In our research we primarily used test sets of similar computer tomography (CT) and magnetic resonance (MR) images, some of which are shown on Figure 2.

One can see that in general all images of the same class look very similar. Moreover, it is possible by seeing only a few images to form a general "CT image pattern" or "MR image pattern" and to recognize if any other image belongs to this class or not. On the other side, all test images shown on Figure 2 have been taken from different people and hence contain many individual details, resulting in the presence of local inter-image dissimilarities. Therefore we intuitively define *similar images* as images:

1. Displaying the same object of specific shape (e.g., human brain scan, Boeing 707, etc.).

2. Produced on the same device or with the same technology (e.g., same computer tomography scanner).

This definition corresponds to the intuitive human perception of similarity and explains the presence of redundant patterns to be removed with set compression. For

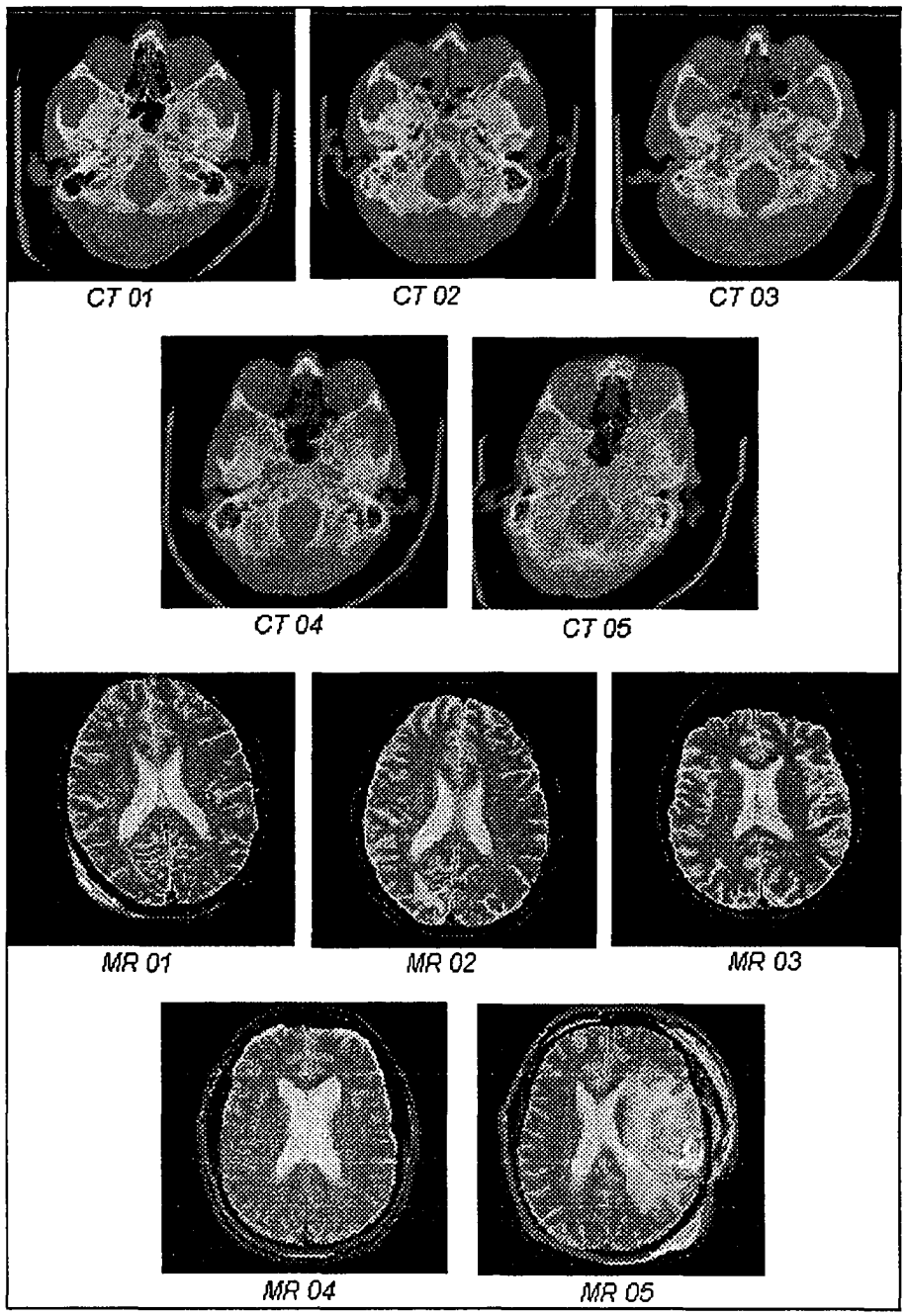

Figure 2: Similar images.

instance, the presence of the same shape creates shape redundancy: all CT brain images are similar because they are oval with two concave areas on top. It also filters out objects with no predefined shape, such as clouds or water. The second requirement guarantees that common image patterns will not look different because of different picturing techniques: CT image of a human brain is different from an MR image of the same brain, and there is almost no intensity correlation between an apple as we see it and the same apple displayed in infrared light.

Informational Definition of Similar Images

The main goal of image compression is to express the maximum information in the shortest possible form. In particular, lossless image compression attempts to replace an original image v with another image v' such that:

1. Entropy $H(v') < H(v)$.

2. There is a well-defined reversible function (known as compressing transform) $f : v = f(v')$.

This allows the formulation of image similarity into more general and accurate "image compression language":

Definition 1. Images u and v are similar if there exists a reversible transform f such that:

1. $u = f(v)$.

2. $H(f) < \min(H(u), H(v))$.

This definition will be used for the remainder of our study, but we include the following comments. First, the reversibility of $f()$ implies the symmetry of similarity

property: if u is similar to v, then v is similar to u. Second, the transform f entropy $H(f)$ is the entropy of the symbolic expression for f. For instance, if $f(v) = v+r$ (difference compression described later), then $H(f) = H(\{'+', r\})$. With this definition, the main problem of similar image compression is finding a reversible f_0 :

$$H(f_0) = \min_{f:\ u=f(v)} H(f).$$

For practical purposes, the efficiency of computing f_0 also becomes very important. Therefore many set compression techniques are based on linear transforms $f()$, which guarantee both reversibility and simplicity. Moreover, the theory of linear data transformations has been developed in statistical analysis with linear regression. The latter finds f_0 such that:

$$\sigma(u - f_0(v)) = \min_{\text{linear } f:\ u=f(v)} \sigma(u - f(v)),$$

where σ stands for the variance operator. This link between informational $H()$ and variance $\sigma()$ measures produce the question of how information redundancy in a similar image set depends upon their common statistical properties. Our research focused on various aspects of this relation and started with analysis of the previously used set compression approaches, given in the following section.

This section covers major similar compression techniques currently in use [18], [33], [31]. It also demonstrates how these techniques can be applied to our test CT and MR data. Finally, this overview is used to develop the criteria that any similar image compression technique must satisfy, and to study the range of their applicability.

Inter-Image Prediction

Inter-image prediction [9], [32], [33], [28] assumes that high image correlation alone implies image similarity. Consequently, high correlation among several images means that the images are almost linearly dependent, i.e., some part of them can be efficiently predicted with linear combinations of the others. If $V = \{v_1, v_2, .., v_n\}$ is a set of similar images v_i, each v_i is highly correlated with the other images in the set represented as $V^{(i)} = V \backslash \{v_i\}$ and v_i can be expressed as

$$v_i = \sum_{j \neq i} \beta_j v_j + r^{(i)} = \beta^{(i)} V^{(i)} + r^{(i)} = \widehat{v}_i + r^{(i)}, \tag{2}$$

where the constant vector $\beta^{(i)} = (\beta_1, \beta_2, .., \beta_{i-1}, \beta_{i+1}, .., \beta_n)$. Equation (2) is a simple linear regression model where each image represented as a vector of its pixel intensities, and $r^{(i)}$ is the error (residual) term [36]. Residual $r^{(i)}$ is viewed as the image v_i decorrelated with respect to the other similar images. With high inter-image correlation ρ, $\rho = \rho(v_i, V^{(i)}) \rightarrow 1$, the residual $r^{(i)}$ becomes small: $||r^{(i)}|| = (1 - \rho)\,||v_i|| \rightarrow 0$. If $r^{(i)}$ is small and can be neglected, one can store only a few

predicting parameters $\beta^{(i)}$ instead of an image v_i, and approximately recover the compressed v_i as a linear combination of other images: $\widehat{v}_i = \beta^{(i)} V^{(i)}$.

This view fits the lossy compression paradigm with remarkable compression results, but it always leads to losing the most dependent part of the data. In lossless compression, none of the data can be sacrificed, which makes many lossy techniques either unsuitable or impractical. The best attempt to modify this approach for lossless similar images compression is to rewrite regression equation (2) as

$$v_i = \left\lfloor \sum_{j \neq i, j=1}^{j=n} \beta_j v_j \right\rfloor + r^{(i)} = \left\lfloor \beta^{(i)} V^{(i)} \right\rfloor + r^{(i)} = \widehat{v}_i + r^{(i)} \tag{3}$$

where $\lfloor \cdot \rfloor$ stands for integer truncation. In this case $r^{(i)}$ becomes an integer vector as all v_i, and can be compressed with traditional image compression techniques. Then, to make this compression lossless, one has to store both coefficients $\beta^{(i)}$ and compressed residual $r^{(i)}$ to completely recover v_i using (3). With 5-10 predicting images in $V^{(i)}$ the overhead to store 5-10 constant numbers $\beta^{(i)}$ is negligeable with respect to the typical 0.5-2 Megabyte image v_i. However, reversibly replacing v_i by $\{\beta^{(i)}, r^{(i)}\}$ with low variance $r^{(i)}$, $\left|\left|r^{(i)}\right|\right| = (1-\rho)\left|\left|v_i\right|\right| \ll \left|\left|v_i\right|\right|$, it was expected that any compression algorithm applied to $r^{(i)}$ will give better results than the same compression applied to v_i. If true, storing residuals instead of images would lead to improved set compression when compared to compressing the same images independently.

The simplest modification of this approach with $n = 1$, $\beta_1 = 1$ would be predicting one similar image from the other one as

$$v_i = v_j + r^{(i)} = v_j + d_{ij} \tag{4}$$

known as *difference compression.* Another modification is *centroid compression* which predicts an image v_i as an average of the other $(n-1)$ images in $V^{(i)}$ (i.e., $\beta_j = \frac{1}{n-1}$):

$$v_i = \left\lfloor \frac{1}{n-1} \sum_{j \neq i} v_j \right\rfloor + r^{(i)} = \left\lfloor \overline{V^{(i)}} \right\rfloor + r^{(i)} = \widehat{v}_i + r^{(i)}. \tag{5}$$

We tested (3) on 50 CT images; predicting the first image v_1 from five sets $V^{(1)} = \{v_2\}$ (chosen as the most correlated to v_1), $V^{(2)} = \{v_5\}$ (chosen as the least correlated to v_1), $V^{(3)} = \{v_2, ..., v_5\}$, $V^{(4)} = \{v_2, ..., v_{10}\}$ and $V^{(5)} = \{v_2, ..., v_{50}\}$. The original image, predicted images $\widehat{v}_1^{(i)}$ with their correlation ρ to v_1, and the residual images $r^{(i)}$ with their variances σ are shown on the Figures 3 and 4. The entropy H is also given for each image.

As one can observe, inter-image prediction on these data seems to provide quick and simple set compression: the entropy of residual images $r_1^{(i)}$ in this example is smaller than that of the original v_1 by $10-18\%$. This means that if we had a choice between Huffman compression of v_1 before regression (3) and after, the second choice would have a $10-18\%$ higher compression ratio. However, the real applicability of this approach is severely limited by several problems:

1. The predicted image $\widehat{v}_1^{(i)}$ on the Figures 3 and 4 is blurred and looks more like several predicting images overlapped rather than a good approximation to the original. It does not capture any local details in v_1. However, all sharp details from v_1

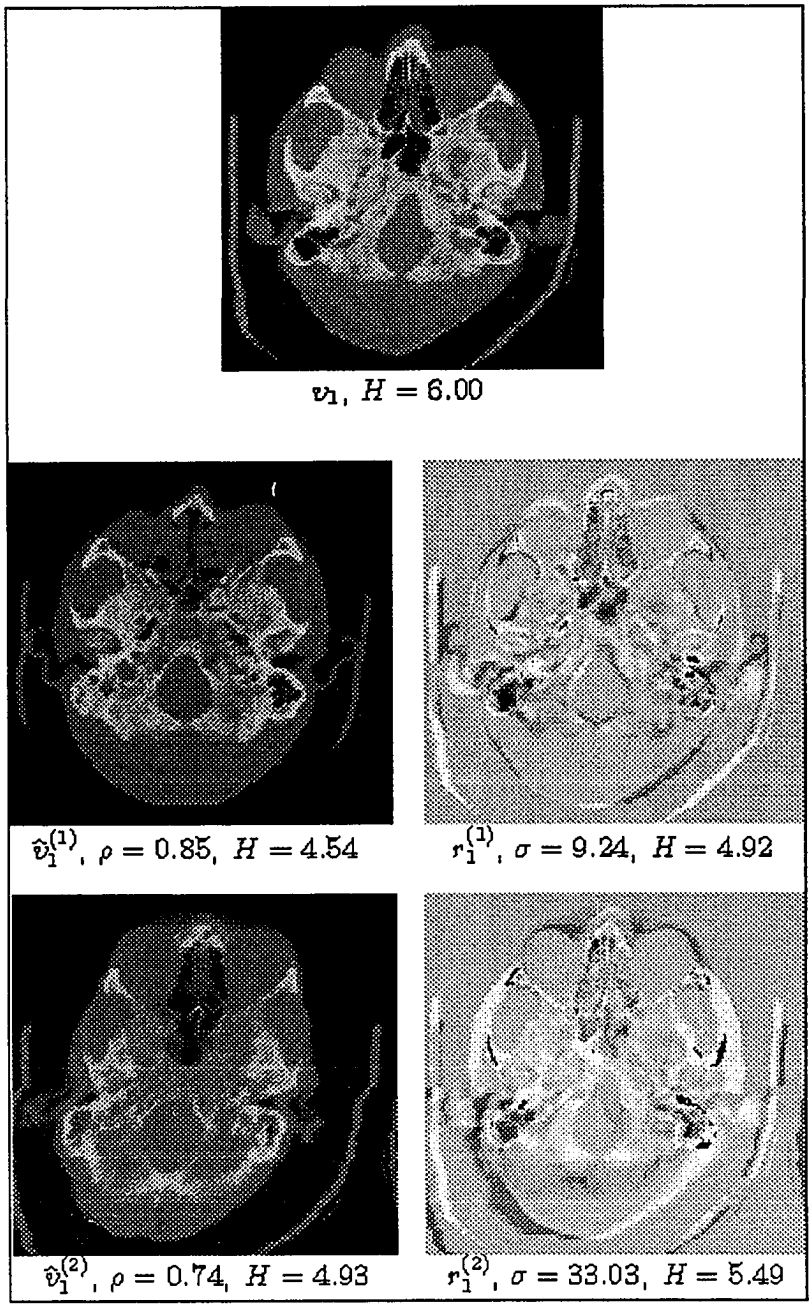

Figure 3: Inter-image prediction from a single image.

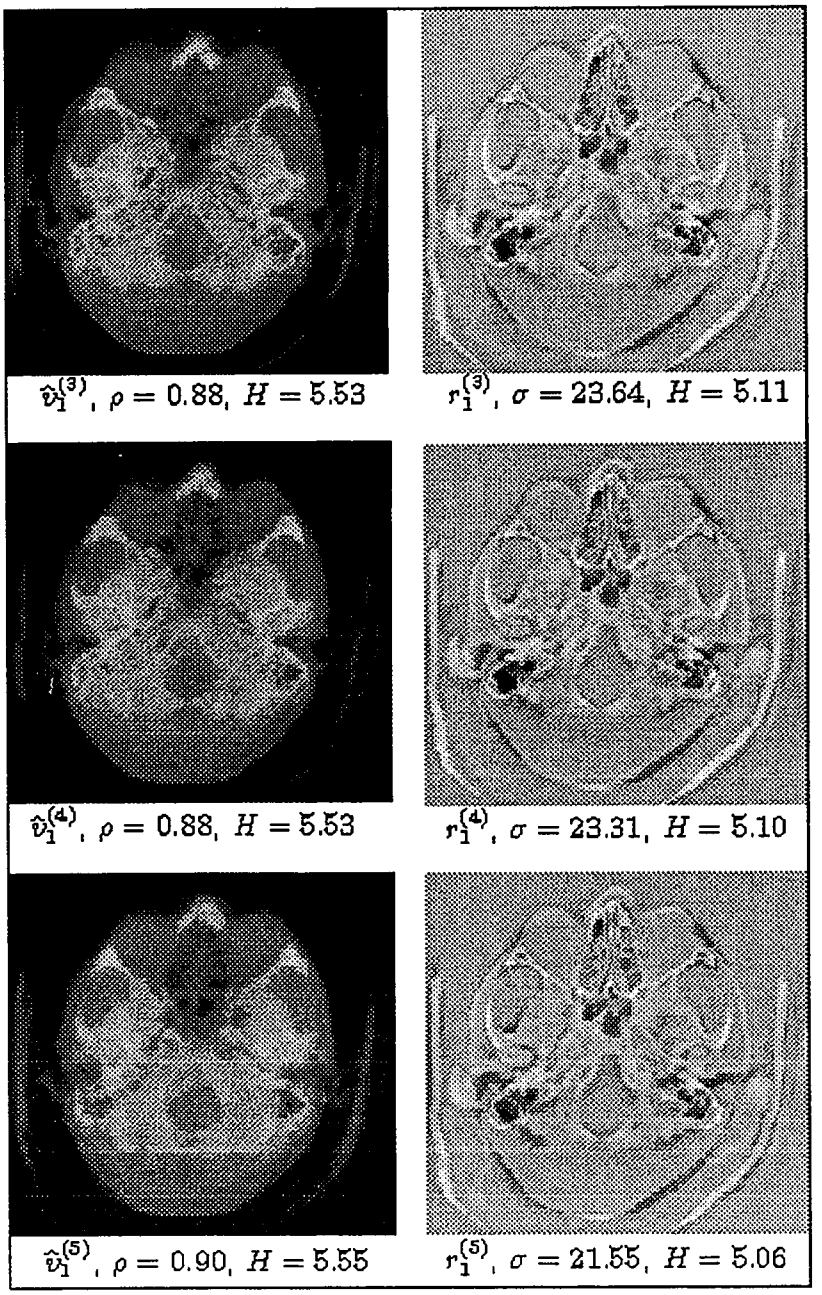

Figure 4: Inter-image prediction from several images.

and its predictors tend to accumulate in its residual $r_1^{(i)}$, making it very informative and difficult to compress.

2. The choice of predictors becomes very important: even though all predictor sets except $V^{(2)}$ include v_2, predicting v_1 from v_2 alone provides a much better result. Therefore, increasing the number of predictors can decrease the efficiency of compression, which means that predictors must somehow be clustered and carefully chosen for each image. For some images there may not even exist a predictive set which would reduce the image entropy (we will give an example later in this subsection). Including a real time database clustering into image compression and maintaining all "predictor-predicted" relations is not a trivial task.

3. Correlation between similar images can easily be destroyed if we translate or rotate one image with respect to the others (we will discuss this in more detail later). For instance, a 45-degree rotated CT image is correlated to its original (0-rotated) copy with ρ as small as 0.4. This is too small to be used for inter-image prediction, but rotation does not make the images less similar ! Therefore before applying (3), all the images must be aligned or *registered*, which is also a computationally intensive and error-prone problem, often performed manually. If one develops a similar image compressing algorithm, it must be insensitive to all transforms which do not destroy the image similarity.

4. Finally, it is easy to prove that even for highly correlated and visually similar images, predictive compression (3) may result in *increased* [28], rather than decreased, database entropy. Below, we demonstrate this both theoretically and numerically. The theoretical proof of inter-image prediction inefficiency lies in the following lemma.

Lemma 1. For any small δ, $0 < \delta < 1$, and large M there exist two images v_1, v_2 such that:

1. $\rho = \rho(v_1, v_2) > 1 - \delta$.
2. difference entropy $H(v_1 - v_2) = H(d_{12}) > M$.

Proof.

Consider an image v_1 with intensity average $\overline{v}_1 = 0$ and variance $\sigma(v_1, v_1) = v_1^T v_1 = \Psi^2$. We choose difference image $d_{12} = v_1 - v_2$ as a normal noise with 0 mean and variance σ : $d_{12} \sim N(0, \sigma)$, therefore v_2 is defined as $v_1 - d_{12}$. Then:

1. $\sigma(v_1, v_2) = v_1^T v_2 = v_1^T(v_1 - d_{12}) = \Psi^2$ (v_1 and d_{12} are not correlated).
2. $\sigma(v_2, v_2) = v_2^T v_2 = (v_1 - d_{12})^T(v_1 - d_{12}) = (v_1^T - d_{12}^T)(v_1 - d_{12}) = v_1^T v_1 + d_{12}^T d_{12} = \Psi^2 + \sigma^2$.
3. $\rho(v_1, v_2) = \sigma(v_1, v_2)/\sqrt{\sigma(v_1, v_1)\sigma(v_2, v_2)} = \frac{\Psi}{\sqrt{\Psi^2 + \sigma^2}}$.
4. The entropy of the normal source d_{12} is known to be $H(d_{12}) = \frac{1}{2} + \ln(\sqrt{2\pi}\sigma) = M$.

Given M and δ, one can always choose $\sigma > \frac{1}{\sqrt{2\pi}} e^{M - \frac{1}{2}}$ to satisfy $H(v_1 - v_2) = H(d_{12}) > M$, and then $\Psi > \sigma \frac{1-\delta}{\sqrt{1-(1-\delta)^2}}$ to satisfy $\rho(v_1, v_2) > 1 - \delta$, which proves the lemma. ■

This lemma demonstrates that the difference entropy in its absolute value can be arbitrarily large even for highly correlated images[4] . Note that with lossless compression, the entropy of an image gives the lower bound for the compressed image size. Therefore the lemma proves that with choice of $\delta \to 0$, two images v_1

[4]This does not say anything about the relative entropy $H(v_1 - v_2)/H(v_1)$, and this question will be addressed later.

and v_2 can be made as correlated as possible, and yet the difference d_{12} between them may have an arbitrarily high entropy. In particular, the difference (as well as residual) entropy can greatly exceed the entropy of image v_1, even though the difference variance $||d_{12}|| = (1-\rho(v_1,v_2))\,||v_2|| \to 0$. Moreover, replacing in the proof v_2 with $\widehat{v}_1$, the same conclusion follows for the general predictive set compression given by (3). In this case the general inter-image predictive method (3) will fail to produce any improvement and in fact may even substantially increase the total database entropy.

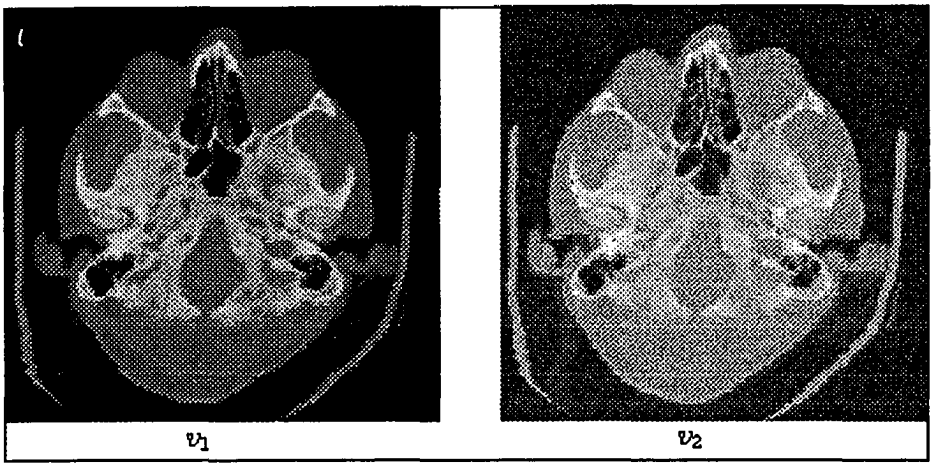

Figure 5: Original and noisy CT images.

Figure 5 illustrates this result numerically. We used a CT image as v_1, and introduced normal noise d_{12} to produce $v_2 = v_1 + d_{12}$. The amount of noise (variance σ) was chosen such that the difference entropy $H(d_{12})$ slightly exceeds $H(v_1)$. Therefore predicting v_1 as $v_1 = v_2 - d_{12}$ increases the total entropy: $H(v_2) + H(d_{12}) > H(v_2) + H(v_1)$. However, in this example $\rho(v_1,v_2) = 0.98$, and both images still look very

similar. Thus, it is enough to transmit an image through a noisy channel or reproduce it on a different device to cause predictive set compression failure.

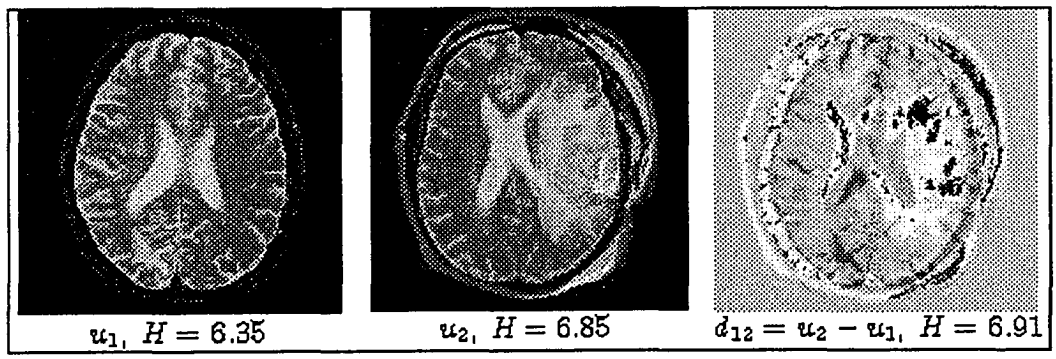

Figure 6: Bad MR set compression.

If one argues that this illustration is too simulated, one may find the same example in an actual MR database. The correlation between two MR images u_1 and u_2 on Figure 6 is $\rho(u_1, u_2) = 0.83$. However, we tested that subtracting or regressing one image with the other produces difference and residual images like the one on the right with entropies higher than those of the original images. If we had only these two MR images u_1 and u_2 in a database, any attempt to use predictive set compression would increase the database entropy.

Multiple experiments, conducted with more complicated and nonlinear predictors (logistic, polynomial up to 10-th degree, rational, neural-network based and predictors with logic operators), did not result in any crucial improvement. We found no evidence that these problems, which are related to the nature of the method and not to the choice of the images, can be avoided with any inter-image prediction modification This again proves that *visible similarity of two images does not guarantee the success of predictive inter-image set compression.*

Principal Component Prediction

Principal components [16] $P = \{p_1, p_2, ..., p_n\}$ for a set of n vectors $V = \{v_1, v_2, .., v_n\}$ are defined as

$$p_i = e_i V = \sum_{k=1}^{n} e_i^k v_k$$

where vector $e_i = (e_i^1, e_i^2, .., e_i^n)$ is the i-th eigenvector of the $n \times n$ covariance matrix $\Lambda = (\Lambda_{km})_{k,m=1}^{n}$ with $\Lambda_{km} = \sigma(v_k, v_m)$. If all eigenvalues λ_i of Λ are in decreasing order, $\lambda_1 \geqslant \lambda_2 \geqslant .. \geqslant \lambda_n$, the first k, $k \leqslant n$, principal components $P^{(k)} = \{p_1, p_2, ..., p_k\}$ form the best k-vector predictor set for the n vectors in V in terms of preserved variance. This means that predicting all vectors in $V = \{v_1, v_2, .., v_n\}$ from a reduced set $P^{(k)} = \{p_1, p_2, ..., p_k\}$ will produce the least variance loss, which has been proven [16] to be

$$c_{k+1} = \sum_{i=k+1}^{n} \lambda_i. \tag{6}$$

Because of this property, principal components are often used for *dimension reduction* - reducing number of variables through removal of the most linearly dependent of them. This is directly related to image set compression, and previous attempts to apply principal components to image analysis exist [23], [8], [7], [11], [10].

We applied principal components analysis to a set of 50 similar CT images. Figure 7 represents the ratios c_k/c_1 and λ_k/c_1, and as one can observe, most of the variance for this set can be expressed with the first few principal components. From these 50 images, the first CT image v_1 was chosen to be predicted from 4 sets $P^{(1)} = \{p_1\}$,

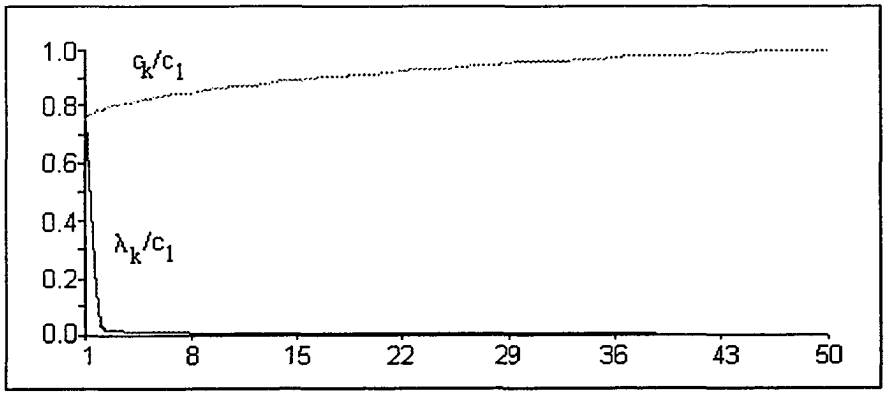

Figure 7: Principal component analysis for CT database.

$P^{(5)} = \{p_1, p_2, ..., p_5\}$, $P^{(10)} = \{p_1, p_2, ..., p_{10}\}$ and $P^{(25)} = \{p_1, p_2, ..., p_{25}\}$. The images on Figures 8 and 9 show the original image v_1, four respective predicted images $\widehat{v}_1^{(1)}$, $\widehat{v}_1^{(5)}, \widehat{v}_1^{(10)}, \widehat{v}_1^{(25)}$ with their correlation ρ to the corresponding predicting set, and the residual (error) images $r_1^{(1)}$, $r_1^{(5)}, r_1^{(10)}$ and $r_1^{(25)}$ with their variances σ. Entropies H are also indicated for each image.

Several important observations made from this numerical principal component analysis are:

1. Principal components can be used for lossy prediction: 25 components are adequate to make the lost residual vector $r_1^{(25)}$ virtually invisible (Figure 9). Successful results have been reported in this area, as well as successful use of principal components to underline the differences between similar images[5] .

2. Principal components are not convenient for lossless prediction. First, there is no conceptual difference between principal component prediction and inter-image regression studied in the previous subsection, plus all problems that we outlined be-

[5]The first principal component is typically the average of all images, while the remaining components serve as contrast vectors among several similar subclusters in the original image set.

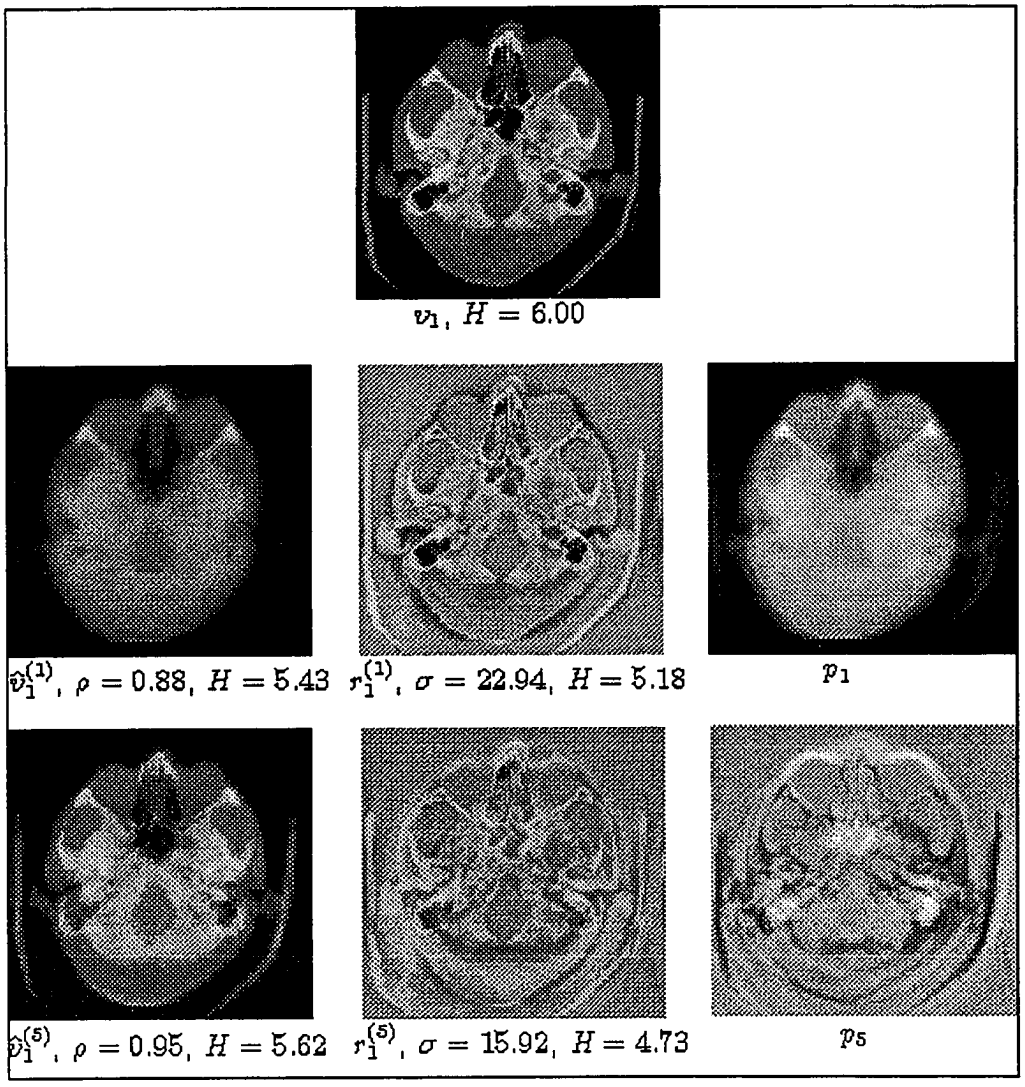

Figure 8: Principal Component images for $P^{(1)}$ and $P^{(5)}$.

fore apply to the principal component case. In particular, it is still possible for an increase to occur in the total database entropy with this predictive model. Second, with principal components one has to store both residual images and principal components instead of original set of images, and recompute all principal components if at least one image was changed. This is inefficient, doubles the number of images and inevitably increases the size of the database. The decreasing variance of principal component $\sigma(p_i) = \lambda_i$ does not result in less entropy: in fact, images like p_{25} contain much detail and are as difficult to compress as v_i (tend to have the same entropy). Thus, for lossless compression, these problems make principal component compression even less practical than simple inter-image prediction.

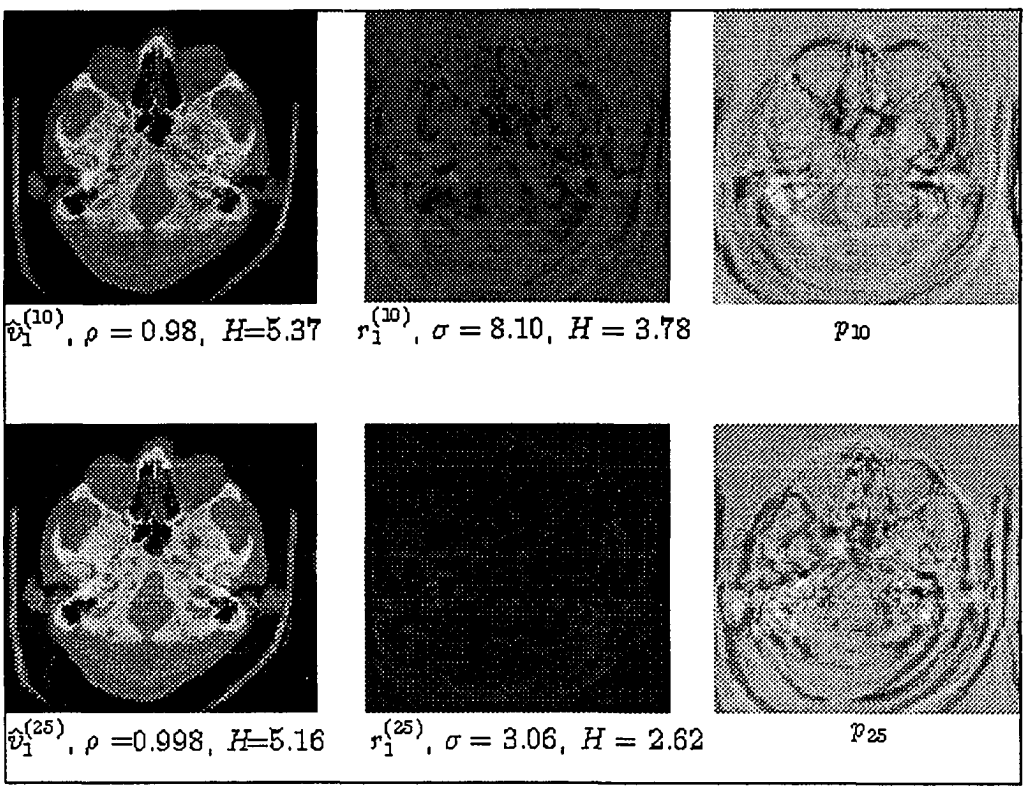

Figure 9: Principal Component images for $P^{(10)}$ and $P^{(25)}$.

2D and 3D Autoregressive (AR) Models

This technique, also known as differential pulse code modulation (DPCM) [24], [6], [27], [29], [12], [19], includes image shifts in the predictive model (3) [34], [31], [29], [21], [22]. This is justified by the fact that in any image, neighboring pixels tend to have close, i.e., correlated, intensity values. Therefore, one may forecast a pixel intensity from the intensities of its surrounding pixels. For example, the typical second-order AR model for the 2D image $u[i,j]$ is

$$u[i,j] = \lfloor \beta_1 u[i-1,j] + \beta_2 u[i,j-1] \rfloor + r = \lfloor (\beta_1 B + \beta_2 L)u \rfloor + r = \widehat{u} + r, \qquad (7)$$

where L and B are left and bottom shift operators respectively. The residual $r[i,j] = u[i,j] - \widehat{u}[i,j]$ represents the part that cannot be predicted. The AR model of the k-th order is

$$u[i,j] = \left\lfloor \sum_{m=1}^{k} \beta_m u[i-a_m, j-b_m] \right\rfloor + r = \left\lfloor \sum_{m=1}^{k} \beta_m B^{a_m} L^{b_m} u \right\rfloor + r = \lfloor \beta u_s[i,j] \rfloor + r, \qquad (8)$$

where β_m, a_m and b_m are optimally chosen constants (a_m and b_m are integers). We use the notation u_s for all left L and bottom B shifts of the image u used in a particular model, since any AR model predicts an image from its own translations. To build an AR model one must choose a_m and b_m, and use a linear regression similar to (2) to determine an optimal β.

The compression ratio $H(u)/H(r)$ produced by this model increases with the model order. Unfortunately, computing optimal model coefficients β for large model

orders k becomes a computationally intensive task and prohibits the use in any compression. Besides, 2D AR models do not consider any inter-image similarities.

For certain data, 2D models can be extended to 3D models which include inter-image relations. 3D AR models are combinations of (3) and (8), i.e., prediction of an image from its own translations and other similar images. This approach proved to be efficient for 3D volumetric data, when L and B operators correspond to translations in the (x, y) plane, and when the prediction from other similar images is viewed as prediction from translations in z plane:

$$u = \left\lfloor \sum_j \alpha_j v_j + \sum_{m=1}^{k} \beta_m B^{a_m} L^{b_m} u \right\rfloor + r = \left\lfloor \beta u_s + \sum_j \alpha_j v_j \right\rfloor + r, \qquad (9)$$

where v_j are images similar to u. When all v_j represent some close slices of the same 3D object, this model efficiently decorrelates the data.

However, for our test data inter-image prediction does not yield any valuable compression improvement, due to the lack of positional correlation and consistency in the z direction. Therefore, for general sets of similar images, 3D AR models may perform worse than separate 2D models[6] .

Inter-Image Prediction with Region Matching

It was soon understood that image similarity does not imply similar images must almost coincide if properly overlapped. The complexity of mapping which matches one similar image to another can be overwhelming, and almost never results in simple transforms such as rigid-body translations and rotations. Therefore some techniques

[6]The same is true for emerging 3D wavelet transforms, efficiently compressing 3D medical data, but providing very poor performance for uncorrelated 2D images.

emerged trying to improve the inter-image prediction accuracy destroying the integrity of the image to be predicted [13], [14]. A typical example can be found in [17], when each pixel $u[x,y]$ in image u is not predicted from the similar overlapped region of the image v, but from the region in v which has the closest intensity match to some neighborhood of $u[x,y]$.

This approach does not require registration because it becomes a part of the compression procedure. This allows more accurate image matching and an improved compression ratio, but a high price must be paid:

1. The process of searching for the best predictive region in the predictor image is extremely computationally expensive.

2. Storing information about regional correspondences adversely affects the compression ratio.

To study how far one can go with this type of compression, we conducted a numerical experiment illustrated on the Figure 10. Two $256 \times 256 \times 8$ CT images, u and v, were chosen as the most correlated from the 50-image CT database (we used these images as $CT01$ and $CT02$ on Figure 2). Image u was scanned pixel by pixel. As shown on Figure 10, for each pixel $u[x,y]$ a pixel $v[p,q]$ in the v image was found such that 8-pixel neighborhoods of these minimize the matching error $|a_0 - a_1| + ... + |h_0 - h_1|$, and the value of $u[x,y]$ was forecast from image v as $v[p,q]$. In case of a tie, $u[x,y]$ was predicted as the average of all $v[p,q]$ with the same minimal error.

The residual image r (bottom Figure 10), $r[x,y] = u[x,y] - v[p(x,y),q(x,y)]$, resulting from this prediction had entropy $H(r) = 3.75$. This is a considerable

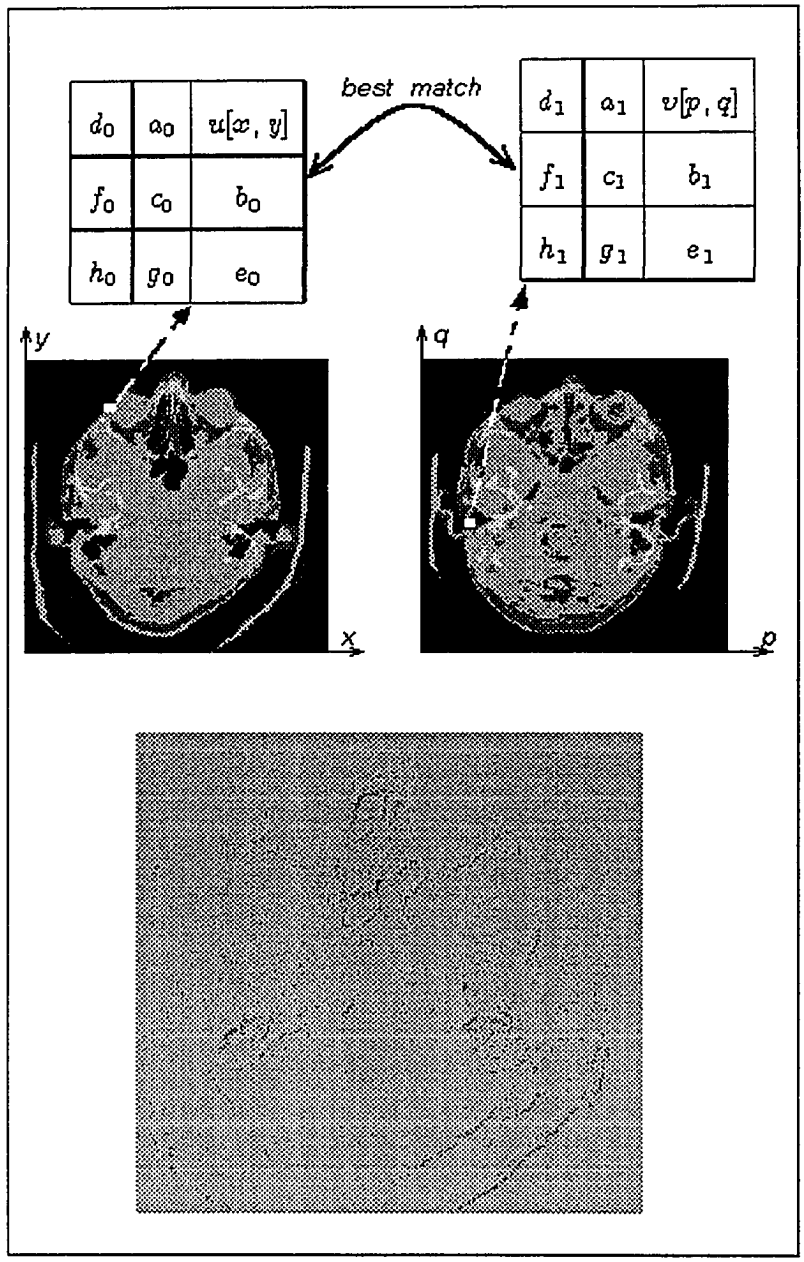

Figure 10: Optimal regional matching.

improvement over the $H = 4.92$ entropy obtained after simple regression of u on v. However, the simple regression took about 8 minutes on an Alpha 4/233 workstation, and to compute the complete best region match search took 2 days. Accelerating this approach is a compromise between the quality of prediction (registration in this case) and compression speed, and does not improve the cumbersome nature of the method. This type of image registration seems to be prohibitive when included in any database compression technique, unless some very special cases exist. In general, this approach must be abandoned because it is inefficient.

Combining Several Compression Techniques

It is possible to compress similar images with several different compressing transforms, applying them one after another. This may lead to compression ratio improvement, typically within 5%. The discussion of this approach can be found in [20] and is beyond the scope of our research. We are primarily interested in single transform methods.

Correlation vs. Information

The question of how efficiently similar images can be compressed cannot be answered without a study of the numerical similarity measures and their effect on image compressibility. In image compression, the numerical *similarity* between two images u and v was usually assessed from pixel-to-pixel correspondences in several different ways. *Absolute difference* $||r_{\text{dif}}||$ (where $r_{\text{dif}} = u - v$), often used in inter-frame compression, is applicable only if the intensities in u and v are on the same scale. *Correlation* $\rho = \rho(u, v)$ accounts for all linear transformations in the intensity domain. Maximizing the correlation means minimizing the mean squared error $\sigma^2 = ||u - \beta_0 - \beta_1 v||^2$, a common error measure for the lossy compression. The corresponding residual $r = u - \beta_0 - \beta_1 v$ generalizes r_{dif} with $||r|| \leqslant ||r_{\text{dif}}||$. Finally, *residual entropy* $H(r)$ is proportional to the lower bound of the compressed file size. Since most lossless compression techniques result in producing and storing decorrelated image residuals, we studied the behavior of residual entropy $H(r)$ with respect to inter-frame image correlation ρ.

Experimenting with CT and MR images, we found that inter-image prediction methods may not even result in decreased database entropy. We have already seen examples when $H(r)$ increases when inter-image correlation $\rho = \sqrt{1 - \sigma^2}$ is being maximized or residual deviation $\sigma = ||r||$ is being minimized. The problem is that minimal $H(r)$ requires all residual intensities $r[x, y]$ be compactly distributed near some constant value; e.g., have as many $r[x, y] = 0$ as possible. The compactness, rather than the variance, of the residual distribution reduces the entropy. The least

squares difference, $r = u - \beta_0 - \beta_1 v$, minimizes $r[x,y]$ *on average*, so that all $r[x,y]$ can be small but still almost uniformly distributed around 0. Since for a given range of residual values, a uniform distribution maximizes the entropy, a least squares residual r may have higher entropy than the images u and v, and be harder to compress than the original images. Fortunately, there is still a connection between *small* $H(r)$ and *small* $||r||$. In compression the residuals are rounded to integer values, and for discrete integers $r[x,y]$, $\sigma^2 = ||r||^2 = \frac{1}{N}\sum_{x,y} r^2[x,y] \to 0$, will inevitably make as many $r[x,y]$ equal to 0 as possible, which will cause $H(r)$ to decrease. However, this decrease becomes apparent only after the variance $\sigma = ||r||$ falls below some threshold, and we estimated the corresponding inter-image correlation threshold from theoretical and practical models.

Normal Estimate of Entropy Behavior in Predictive Model

For preliminary theoretical analysis we assumed that highly correlated least squares predictors result in the almost normally distributed residual intensity values $r[x,y] \sim N(0,\sigma)$ (it was also observed in practice [7]).

Lemma 2. I

1. $u \sim N(0,\sigma)$, and $r \sim N(0,\cdot)$ are independent random variables,

2. Correlation $\rho(u-r,r) = \rho$

Then

$$\frac{H(r)}{H(u)} = HR(\rho) = \frac{\log\sqrt{2\pi e\sigma^2(\frac{1}{\rho^2}-1)}}{\log\sqrt{2\pi e\sigma^2}}. \qquad (10)$$

[7] We used SAS® Data Analysis tools to analyze typical residual distributions for CT and MR images, which were found to be normal at a 5% confidence level.

Proof.

If $\sigma(r) = \delta^2$, then from 1. and 2.

$\rho(u-r,r) = \rho = \frac{\sigma(u-r,r)}{\sqrt{\sigma(u-r)\sigma(r)}} = \frac{\sigma^2}{\sqrt{(\sigma^2+\delta^2)\sigma^2}}$, and from here

$\delta^2 = \sigma^2(\frac{1}{\rho^2} - 1)$

Then the entropy of the normal[8] [26] r

$$H(r) = \log\sqrt{2\pi e\sigma(r)} = \log\sqrt{2\pi e\sigma^2(\tfrac{1}{\rho^2} - 1)},$$

which proves the lemma. ■

If one considers image $v = \frac{1}{\rho}(u - r)$ as ρ-correlated predictor for image u, then ratio (10) gives the normalized entropy of the residual image r with respect to the original image u. The *entropy ratio function* $HR(\rho)$ is also proportional to the inverse compression ratio produced by the model $u = \rho v + r$.

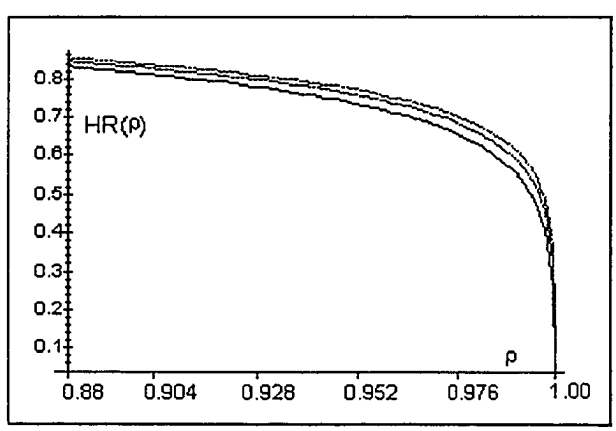

Figure 11: HR(ρ) for the normal distribution model.

[8] Normal distribution has the maximum entropy over all distributions with the same variance. Thus, our estimate for the residual entropy is in fact the *worst-case* model.

We observed that the typical variance σ^2 for our test CT and MR images always lies between 500 and 2000. Figure 11 shows the behavior of $HR(\rho)$ for $\sigma^2 \in \{500, 1200, 2000\}$ (smaller σ correspond to lower curves). From this theoretical model, a 50% entropy reduction $HR(\rho) = 0.5$ corresponds to the correlation

$$\rho_{0.5} = \frac{1}{\sqrt{1 + \frac{1}{\sqrt{2\pi e\sigma^2}}}},$$

which for $\sigma^2 \in [500, 2000]$ belongs to [0.995, 0.997]. From (10) one can find the behavior of ρ as a function of compression ratio $C = \frac{1}{HR(\rho)}$ and variance σ. Solving

$$\frac{\log\sqrt{2\pi e\sigma^2(\frac{1}{\rho^2} - 1)}}{\log\sqrt{2\pi e\sigma^2}} = \frac{1}{C}$$

for ρ yields

$$\rho = \sqrt{\frac{2\pi e\sigma^2}{2\pi e\sigma^2 + \exp\left(\frac{\ln(2\pi e\sigma^2)}{C}\right)}}.$$

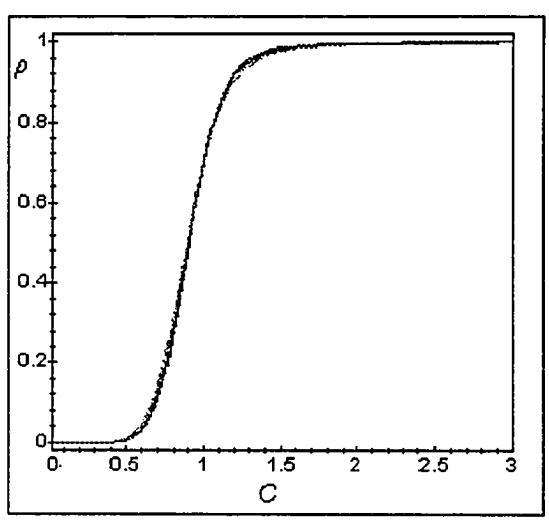

Figure 12: Correlation vs. compression ratio.

Figure 12 demonstrates $\rho(C)$ sufficient to ensure a given compression ratio C for $\sigma^2 \in \{500, 1200, 2000\}$ (the three curves almost coincide).Note how fast $\rho(C)$

increases in all three cases: for $C > 1$, when compression starts being practical, a small increase in compression ratio can be achieved only with substantial increase in inter-image correlation ρ, and $C > 2$ becomes possible only for virtually identical images (ρ becomes very close to 1).

The following theoretical result provides an upper boundary for the entropy reduction function $HR(\rho)$ (see Figure 11) :

Lemma 3. For any two images u and v function $HR(\rho) = \frac{H(u-v)}{H(u)}$ satisfies

$$HR(\rho) \leqslant 1 + \frac{H(v)}{H(u)},$$

with equality possible only for independent u and v.

Proof.

For any random variables u_i entropy $H(u_1, u_2, \ldots u_n) \leqslant \sum_{i=1}^{n} H(u_i)$, and equality holds iff all u_i are independent[9] . In the case of two random variables $H(u, v) \leqslant H(u) + H(v)$. Consider random variable $r = u - v$. It has probability distribution $\mathcal{P}^r = \{P^r(r = r_0)\}$, where

$$P^r(r = r_0) = \sum_{u_0 - v_0 = r_0} P^{(u,v)}\left((u, v) = (u_0, v_0)\right),$$

which means that $\mathcal{P}^r$ values are obtained as sums of probabilities in the distribution $\mathcal{P}^{(u,v)}$ of the image pair (u, v). Also, merging probabilities cannot increase the entropy:

$$H(p_1, p_2, p_3 \ldots, p_n) \geqslant H(p_1 + p_2, p_3 \ldots, p_n).$$

[9] See the properties of the entropy function listed in the introduction.

Since $\mathcal{P}^r$ is obtained merging (adding) some of $\mathcal{P}^{(u,v)}$, we conclude $H(r) \leqslant H(u,v)$ and consequently $H(r) \leqslant H(u) + H(v)$. The entropy ratio function is defined as $HR(\rho) = H(r)/H(u)$, therefore

$$HR(\rho) \leqslant \frac{H(u) + H(v)}{H(u)} = 1 + \frac{H(v)}{H(u)}.$$

■

Corollary 1. Compression ratio $C(\rho)$ obtained replacing image u by $r = u - v$ satisfies

$$C(\rho) \geqslant \left(1 + \frac{H(v)}{H(u)}\right)^{-1}.$$

Proof.

Follows directly from the previous lemma because $C(\rho) = 1/HR(\rho)$.

■

Corollary 2. If image u is chosen such that $H(u) \geqslant H(v)$, then $HR(\rho) \leqslant 2$, and $C(\rho) \geqslant 0.5$.

Any compression is practical only if the ratio $C > 1$; we have already demonstrated that replacing one image with its difference from another lower entropy image cannot increase the entropy more than 100%. In other words, if one applies the difference model to compress the more information rich of the two images, the entropy reduction function cannot be arbitrarily high.

A Numerical Estimate of the Entropy Behavior in Predictive Models

The theoretical estimate was made with the assumption of a normal continuous intensity distribution in the images and residuals. This is almost always true for

residuals, but only approximately true for typical CT or MR intensity distributions. Moreover, in images the intensity range is always limited, which was not considered with the normal model. To investigate realistic problems using brain images (where $r[x,y]$ is discrete, integer and limited to a certain intensity range) we considered a set of 50 256×256×8 CT images [10] . All 50 principal components p_k , $k = 1, .., 50$, of this set were computed and the consecutive subsets $P^{(k)} = \{p_1, ..., p_k\}$ of first $k = 1$, 2, ..., 50 principal components were used to predict the same randomly chosen image u_1. After applying each of these 50 models, the residual values $r[x,y]$ were almost normally distributed on the (−127, 128) interval, and the residual entropy $HR(\rho)$ decreased as shown on Figure 13. Except for some small deviations, one can observe the remarkable similarity between the theoretical and numerical estimates of $HR(\rho)$. From the numerical estimate, a good image predictor must correlate with the image in at least the $0.95 - 0.99$ range, and 50% compression improvement, $HR(\rho) = 0.5$, corresponds to $\rho_{0.5} \approx 0.995$.

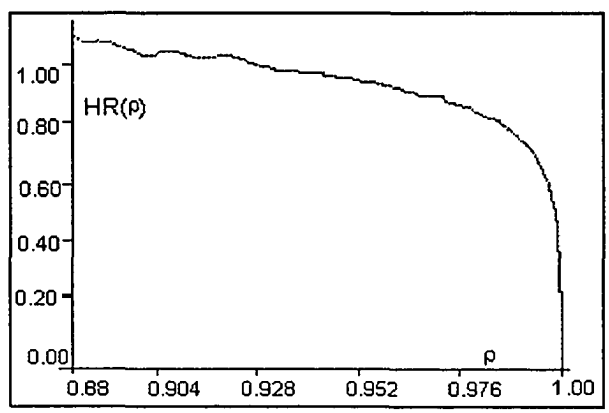

Figure 13: HR(ρ) from 50-image CT database.

[10] Images were taken from different patients, but registered by a radiologist. Images of the same scene can be efficiently registered [25].

This high correlation level is impossible between two different CT or MR images, when $0.7 \leqslant \rho \leqslant 0.8$. Therefore both theoretical and numerical estimates demonstrate that a simple prediction of one image from the others in a correlated image database, may not provide better compression and in practice may even increase in size after the data is "compressed". This phenomenon can be observed on the Figure 13: $HR(\rho \approx 0.88) > 1$, which is an *increase* in residual entropy compared to the original image entropy[11] . We started plotting with the first principal component, which corresponds to the averaged brain image, which demonstrated that computing the residual as the difference between an image and the average (centroid) c of several database images can *increase* the entropy. We illustrate this same conclusion with the first 10 images u_i in Table 1. The average entropy for this 10-image database is $e_0 = 5.975$. If we apply a simple *difference model* storing $\{u_1, u_2 - u_1, ..., u_{10} - u_1\}$ instead of the original image set, it results in *increasing* the database entropy by 4%. For the *centroid model*, we have to compute the centroid image c and store all eleven images $\{c, u_1 - c, ..., u_{10} - c\}$, which increases the database entropy by 12%.

The origin of this problem is insufficient inter-image correlation, which causes failure in simple difference-based predictors. As one can observe from the Table 1, *difference and residual images can have higher entropies than the original images.* Therefore, the need for a more reliable and efficient lossless similar image compression method is apparent, and this model can be built only through the better correlation of similar images.

[11] While the residual variance σ is much smaller compared to that of the image: $\sigma \approx (1 - 0.88)var(u)$.

Table 1: Inter-frame prediction

Image	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	average
Original entropy	6.00	5.59	5.84	6.11	6.26	5.90	6.05	5.86	5.81	6.37	5.975 (100%)
u_i-u_1 entropy, i>1	6.00	5.69	6.13	6.22	6.56	6.27	6.42	6.31	6.04	6.45	6.209 (104%)
u_i-c entropy	5.99	6.01	6.00	5.88	6.21	5.98	6.14	6.03	6.03	5.99	6.693 (112%) [12]
Correlation to u_1	1.00	.84	.77	.78	.74	.77	.75	.74	.75	.78	.77 [13]

"Intra-image" vs. "Inter-image" Correlation

The previous discussion demonstrated that the correlation ρ between images ("inter-image" correlation) can be used for efficient compression only if it exceeded ~ 0.95, which is improbable for both CT and MR images. However, the correlation among several neighboring pixels in the same brain image ("intra-image" correlation) is typically more than 0.95, which makes it a good choice for residual entropy reduction. Does this also mean that inter-image correlation is valueless? As an answer to this question, we will prove that there is essentially no difference between these two correlation types, and a relatively low inter-image correlation ρ can still guarantee the presence of a high intra-image correlation α in all similar images. This result leads to the proposal of a new approach for correlated data compression.

We use the subscript s to indicate the shift operator: $u_s[i] = u[i-1]$. We also assume that for any two similar images u and v $||u|| = ||u_s|| = ||v|| = ||v_s||$ [14] .

[12] Including the centroid image c with entropy 6.67 (note the increase in centroid entropy).

[13] u_1 excluded

[14] A natural assumption for similar images, which have very close statistical characteristics. We may also assume without any loss of generality that all images are centered: $\overline{u} = \overline{v} = \overline{u_s} = \overline{v}_s = 0$.

Lemma 4. Linear regression models are defined as (see Figure 14):

$u = \alpha u_s + r$ ($\alpha \geqslant 0$ is the intra-image correlation for image u),

$v = \beta v_s + q$ (β is the intra-image correlation for image v),

$v = \rho u + \delta$ ($\rho \geqslant 0$ is the inter-image correlation between images u and v),

$\delta = p\delta_s + \varepsilon$ ($p \geqslant 0$ is the intra-image correlation for the shifted residual).

Then the lower bound for the correlation between v and v_s is:

$\beta = \rho(v, v_s) \geqslant \rho^2\alpha - \frac{\rho}{\alpha}\sqrt{1-\alpha^2}\sqrt{1-\rho^2}\left[|\alpha - p| + \sqrt{1-p^2}\right] + p(1-\rho^2) = M_L(\rho, \alpha, p)$,

and the average estimate is

$\beta = \rho(v, v_s) \approx \rho^2 m + p(1-\rho^2) = M_A(\rho, \alpha, p)$.

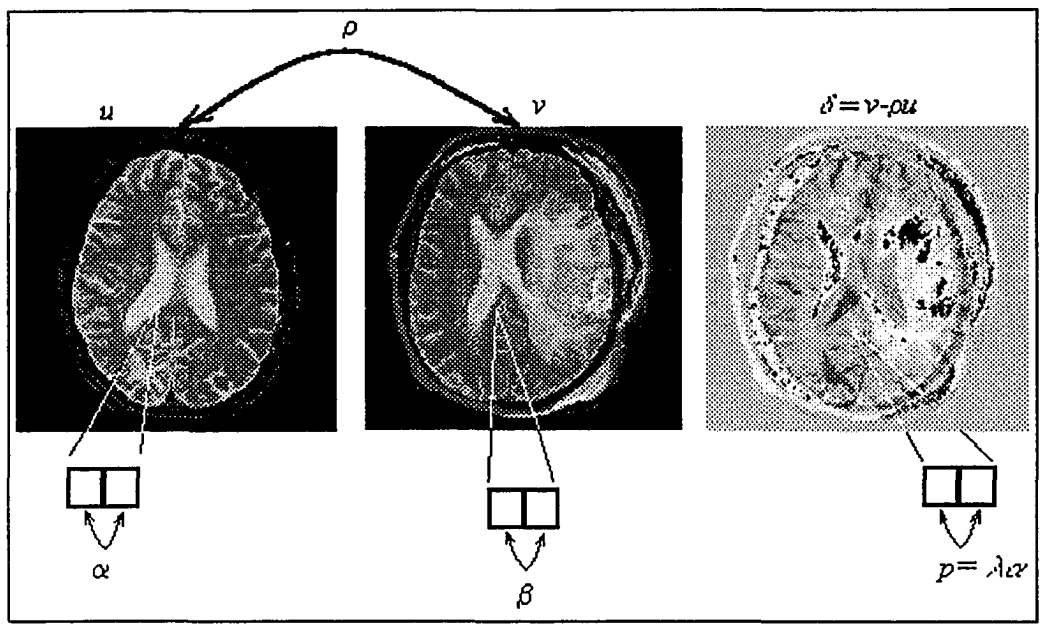

Figure 14: Correlation Lemma.

Proof.

Shifting $v = \rho u + \delta$ we obtain $v_s = \rho u_s + \delta_s$, and $v = \rho u + \delta = \rho\alpha u_s + \rho r + \delta$.

Then:

$v_s^T v = (\rho u_s^T + \delta_s^T)(\rho\alpha u_s + \rho r + \delta) = \rho^2\alpha u_s^T u_s + \rho^2 u_s^T r + \rho u_s^T \delta + \rho\alpha\delta_s^T u_s + \rho\delta_s^T r + \delta_s^T \delta.$

From the orthogonality of linear regression estimators $u_s^T r = \delta_s^T u_s = 0$, yielding:

$v_s^T v = \rho^2\alpha u_s^T u_s + \rho u_s^T \delta + \rho\delta_s^T r + \delta_s^T \delta.$

From linear regression equations:

1. $\rho u_s^T \delta + \rho\delta_s^T r = \rho(u_s^T \delta + \delta_s^T r) = \rho(\frac{1}{\alpha}(u^T - r^T)\delta + r^T\delta_s) = \rho(-\frac{1}{\alpha}r^T\delta + r^T\delta_s)$

$= \rho r^T(-\frac{1}{\alpha}\delta + \delta_s) = \rho r^T(-\frac{1}{\alpha}(p\delta_s + \varepsilon) + \delta_s) = \rho((1 - \frac{p}{\alpha})r^T\delta_s - \frac{1}{\alpha}r^T\varepsilon),$

and

$\left|\rho u_s^T \delta + \rho\delta_s^T r\right| \geqslant -\rho\left|1 - \frac{p}{\alpha}\right|\left|r^T\delta_s\right| - \frac{\rho}{\alpha}\left|r^T\varepsilon\right|$

$\geqslant (-\rho\left|1 - \frac{p}{\alpha}\right|\sqrt{1-\alpha^2}\sqrt{1-\rho^2} - \frac{\rho}{\alpha}\sqrt{1-\alpha^2}\sqrt{1-\rho^2}\sqrt{1-p^2})\,||u||^2$

This is the lower bound for $\rho u_s^T \delta + \rho\delta_s^T r$. On average,

$\left|\rho u_s^T \delta + \rho\delta_s^T r\right| = \left|\rho(-\frac{1}{\alpha}r^T\delta + r^T\delta_s)\right|$

$\leqslant (\rho\left|1 - \frac{p}{\alpha}\right|\sqrt{1-\alpha^2}\sqrt{1-\rho^2} + \frac{\rho}{\alpha}\sqrt{1-\alpha^2}\sqrt{1-\rho^2}\sqrt{1-p^2})\,||u||^2.$

Therefore we chose the middle interval value 0 to estimate[15] $\rho u_s^T \delta + \rho\delta_s^T r \approx 0$

2. $||\delta_s^T \delta|| = ||p\delta_s^T\delta_s + \delta_s^T\varepsilon|| = ||p\delta_s^T\delta_s|| = p||\delta^2|| = p(1-\rho^2)||u||^2.$

Because $||u|| = ||v||$,

$\frac{v_s^T v}{||v||^2} \geqslant \rho^2\alpha - \rho\left|1 - \frac{p}{\alpha}\right|\sqrt{1-\alpha^2}\sqrt{1-\rho^2} - \frac{\rho}{\alpha}\sqrt{1-\alpha^2}\sqrt{1-\rho^2}\sqrt{1-p^2}$

$+p(1-\rho^2),$

$\frac{v_s^T v}{||v||^2} \approx \rho^2\alpha + p(1-\rho^2)$ ■

One can consider a function $R_L(\rho,\alpha,p) = M_L(\rho,\alpha,p)/\rho$, which gives the lower bound of the ratio β/ρ. Similarly, $R_A(\rho,\alpha,p) = M_A(\rho,\alpha,p)/\rho$ estimates the ratio β/ρ on average. We found numerically for CT and MR images $\rho \approx 0.75$, $\alpha \approx 0.95$,

[15]The validity of this estimate was verified numerically with CT and MR images. Analytically, we neglected the term of the lowest order for $\alpha \to 1$.

and $p \approx \alpha$. Figure 15 left shows the worst, and Figure 15 right - the average behavior of the contours $\beta/\rho = 1$ for choices of $\lambda = p/\alpha = \{0.8, 1.0, 1.2\}$. Thus, for each choice of λ, the corresponding curve shows the worst-case, i.e., lowest (Figure 15 left) and average (Figure 15 right) intra-image correlation $\alpha = \alpha(\rho)$ in *at least one* database image u, sufficient to guarantee that for any other image v, $\rho(u,v) = \rho$, the intra-image correlation β in v will exceed the inter-image correlation ρ. The region above each curve corresponds to $\beta/\rho > 1$, when the presence of only one highly intra-correlated image u ensures that given ρ, all other similar images v, $\rho(u,v) = \rho$, will have intra-image correlation β greater than ρ. Figure 15 has CT and MR clusters to show that this condition is satisfied.

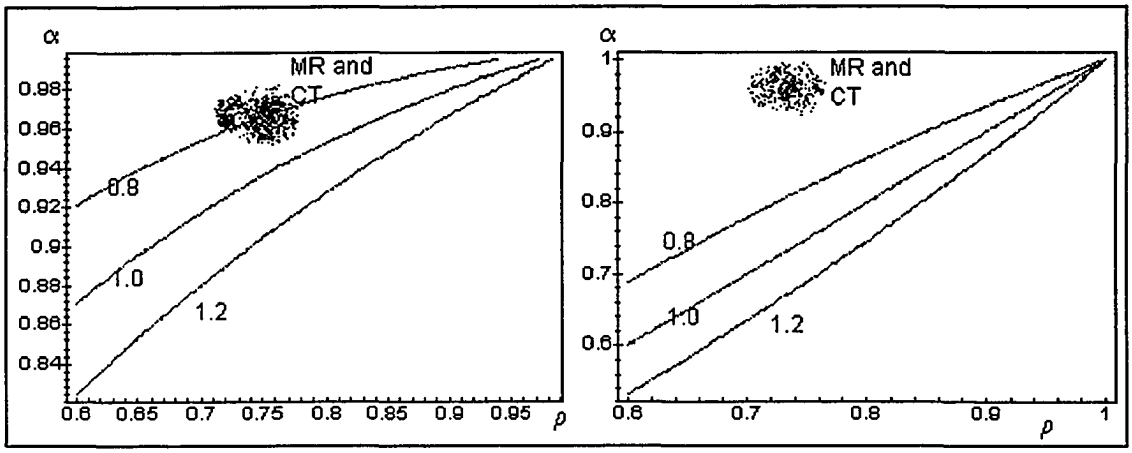

Figure 15: Worst and average-case intra-image correlation α.

Corollary 3. For $\forall p \geqslant 0.5$, $\forall$ inter-image correlation $\rho \quad \exists\, \alpha = \alpha(\rho)$, $0 < \alpha < 1$ such that if **at least one** image u has intra-image correlation α, all images ρ-correlated to u will have intra-image correlation $\beta > \rho$. For $\forall p$, if $\rho \longrightarrow 1$, then $\beta \longrightarrow \alpha$.

The corollary proves an intuitively clear idea that high intra-image correlation will propagate into all inter-correlated images.

Proof.

For any ρ and p, $0 \leqslant \rho, p < 1$, $R_L(\rho, \alpha, p)$ increases as a function of α, therefore it is sufficient to prove the result for $\alpha = 1$. $R_L(\rho, 1, p) = R_A(\rho, 1, p) = \rho + \frac{p}{\rho}(1-\rho^2) > 1$ yields $\rho^2 + p(1-\rho^2) > \rho$, or $(\rho - 1)(\rho - \frac{p}{1-p}) > 0$. Since $\rho < 1$ leads to $\rho < \frac{p}{1-p}$. If $p > 0.5$, $\frac{p}{1-p} > 1$ and for any $\rho < 1$, $\quad \rho < \frac{p}{1-p}$ is always true. ■

CT and MR images satisfy this condition with their typical $\rho \approx 0.75$, $\alpha \approx 0.95$, $p \approx \alpha$. If p is less than 0.5, the corollary will still be true for $\forall \rho < \frac{p}{1-p}$.

We can transpose this conclusion by defining the (x, y) *pixel-set* as the set of all pixels from the image database corresponding to the (x, y) image coordinate. The images are assumed to be of the same size, so we have as many pixel-sets as pixels in each image, and the number of pixels in the pixel-sets is the number n of images in the database. In a matrix V with each image v_i as a column, each pixel-set corresponds to a row. If we transpose this matrix, each image becomes a row, each pixel-set becomes a column. One can repeat this reasoning treating pixel-sets as new images and images as new pixel-sets. After transposing the V matrix the intra-correlation will become an inter-correlation, and vice versa, or

Corollary 4. For $\forall p \geqslant 0.5$, $\forall$ intra-correlation $\alpha \quad \exists\ \rho = \rho(\alpha)$, $0 < \rho < 1$ such that if **at least one** (x, y) pixel-set has inter-image correlation ρ, all images with intra-correlation α have inter-image correlation $\rho^* > \alpha$. For $\forall p$, if $\alpha \rightarrow 1$, then $\rho^* \rightarrow \rho$.

In other words, high inter-image correlation will also propagate for all pixel coordinates given high intra-image correlation.

This demonstrates that high intra-image correlation and inter-image correlation rarely occur separately, i.e., if one is large in at least one similar image, the other will also be large. Since both correlation types are related, we may expect that removing one of them may essentially result in removing both. However, this conclusion is more general than its intuitive implication. Note, in our proofs, we did not use any specific pixel-based image interpretation, i.e., the same reasoning will be true if we represent the images as vectors of their Fourier or DCT coefficients, as nonlinear transformed images, etc. It is the generality of this result that leads to the similar data compression technique described in the next subsection.

Functional Approach to Similar Image Compression

When $\rho \rightarrow 1$, $\beta \rightarrow \alpha$; or, *for any image representation*, a function removing α intra-image correlation from one of the several highly correlated images will result in intra-image correlation reduction in all of them. This allows the introduction of the concept of functional decorrelation for the redundant image data set, which is expressed in the following:

1. Given a set of similar images v_i, choose any representative v_0 and construct some transform $f_0()$ which removes the intra-image correlation α from v_0. This will decrease the entropy of v_0, i.e., provide improved compression for v_0.

2. Then apply *the same* transform $f_0()$ to any other v_i. With a high inter-image correlation ρ, an intra-image correlation β for any v_i will tend to be close to α. Therefore, if $f_0()$ was chosen as optimal for removing α, it should be nearly optimal for removing β as well. It will become optimal for all images as $\rho \rightarrow 1$.

Thus, any decorrelating transform, efficiently decreasing the entropy of *at least one image*, will tend to decrease the entropy of all similar images. If we define a *compression-similar* set as the set of similar vectors (images) efficiently compressed by the same algorithm, then we prove that *highly correlated images form a compression-similar set.* Moreover, the existence of this common compression function for several correlated images can be generalized in the following way: after applying the same transform to each image in a correlated set, the transform parameters are also expected to be correlated for the images (for instance, Fourier transforms of each of the correlated images, as the "intra-image" predicting model, are also correlated for the images). This is important for similar images like CT or MR scans, when the pixel-based correlation ρ fails to guarantee sufficient compression quality: *the inter-image correlation between transform parameters,* shown in the next section, *can be higher and more fault-tolerant than inter-image pixel correlation.*

The functional approach to compressing sets of similar data is very beneficial in large data sets, when we do not have the luxury of finding an optimal compressing transform for each element in the set. In this case, nearly optimal set compression can be achieved by finding the optimal compressing transform for a single element (image) and using this transform to compress the entire set.

Functional Approach Using Common Autoregressive Models (CARs)

Introduction

Previously, similar image compression approaches tend to equate image similarity with high correlation, and remove inter-image redundancy somehow predicting part of the images from the others [9],[32],[33]. However, we demonstrated that for certain classes of similar images this may not lead to any compression. Another problem with this approach is its sensitivity; for instance, the correlation between two images can be substantially decreased, and the mismatch increased, if one image is shifted or rotated with respect to the other, which does not reduce the visual similarity of these images. Attempts to solve this problem with image registration (proper alignment) generally lead to complicated nonlinear algorithms, which are a burden to any data set compression.

We propose in this research using image resemblance in functional, rather than correlated, context: a set of images is similar if a compression transform optimally chosen for one will be optimal or nearly optimal for any other one in the set. The degree of this "near optimality" also becomes a reliable similarity measure. Hence, a class of similar images has only one common compression transform, which should efficiently compress any image from this class and perform poorly for any image outside the class. In this research we choose autoregressive (AR) models to illustrate the existence and properties of common autoregressive (CAR) model for compressing a set of similar images.

Common AR Model Applications

Practical Tests

We illustrate our results numerically with computer tomography (CT) scans of human brains and an AR compressing model. For CT images, intra-image correlation between neighboring pixels is typically as high as 0.95, which is greater than their average 0.75 inter-image correlation and suggests the use of AR compressing transforms [27],[34],[31]. For example, the typical second-order AR model for the 2D image $u[i,j]$ is $u[i,j] = \beta_1 u[i-1,j] + \beta_2 u[i,j-1] + r = (\beta_1 B + \beta_2 L)u + r = f(u) + r$, where L and B are left and bottom shift operators respectively [27],[30],[34],[29]. The residual $r = u - f(u)$ represents the part that cannot be predicted with $f()$. Least-squares AR models of increasing order can efficiently compress smooth images. However, building them for each image is extremely time consuming, which makes large image-dependent predictors unpopular for image compression.

We are not reintroducing the AR models for singular image compression. Instead, we are proposing AR model correlation as a much better method to compress similar images than the traditional pixel-to-pixel or region-to-region correlation. With our proposal, the time and memory costs for image-dependent AR predictors can be dramatically reduced for a set of similar images. These results are illustrated in Table 2 for the 5^{th} order autoregressive model[16] $u[i,j] = (\beta_1 B + \beta_2 L + \beta_3 BL + \beta_4 B^2 + \beta_5 L^2)u + r = f(u) + r$ applied to 10 similar CT brain scans (compare these results to Table 1 on page 35). The average database entropy is $e_0 = 5.975$ is

[16] Found to be sufficient to outperform previous compression techniques based on image correlation, including image-independent AR models.

proportional to the Huffman-compressed database size. We chose the CAR model as the optimal 5^{th} order autoregressive transform $f_1^{ls}()$ for the first image u_1, with $\beta_1 = 0.968950$, $\beta_2 = 0.764156$, $\beta_3 = -0.471855$, $\beta_4 = -0.169767$, $\beta_5 = -0.097293$ (these coefficients differ from image-independent AR models [27]). This CAR was applied to each u_i, yielding an average entropy $e_1 = 2.644 = 0.4425e_0$. Finally, we compressed (decorrelated) each image with its own optimal 5^{th} order autoregression $f_i^{ls}()$, yielding an average entropy $e_2 = 2.624 = 0.4392e_0 = 0.9925e_1$, which is remarkably close to the CAR entropy e_1.

Table 2: CAR model with least squares.

Image	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	average
Original entropy	6.00	5.59	5.84	6.11	6.26	5.90	6.05	5.86	5.81	5.81	5.975 (100%)
Entropy after $f_1^{ls}()$	2.81	2.74	2.72	2.74	2.74	2.40	2.67	2.46	2.59	2.57	2.644 (44.2%)
Entropy after $f_i^{ls}()$	2.81	2.73	2.72	2.73	2.72	2.36	2.63	2.42	2.59	2.54	2.624 (43.9%)
Correlation (u_1, u_i)	1.	.84	.77	.78	.74	.77	.75	.74	.75	.78	.77
Correlation (f_1^{ls}, f_i^{ls})	1.	.999	.998	.999	.999	.998	.994	.997	.998	.999	.998 [17]

Thus,using only a u_1−based decorrelating AR model, we achieved an almost optimal 56% compression improvement[18] with no additional time or memory requirement. Technically, only one optimal 5^{th} order compressing autoregressive transform was determined instead of 10, and we stored only 5 β-coefficients for the entire database. Since all 10 test images were randomly selected from a larger CT database, the

[17] u_1 excluded

[18] In terms of compressed file size being proportional to e_0. In terms of the original 8-bit gray level image size we achieved $\frac{8}{2.644} = 3.0257$ compression ratio.

same improvement would be expected if the $f_1^{ls}()$ transform is applied to any number of the similar CT images.

CAR model database compression has several other advantages. First, since only neighboring pixels are involved in this prediction, the model will not be affected by any transforms such as shifts and rotations (transforms that often make brain images less correlated), and can be equally useful for registered or non-registered sets of images. This can save a large amount of time, because image registration is challenging, computationally expensive, and a bottleneck for previous predictive approaches. Furthermore, similar internal properties of the images, such as almost the same intra-image correlation, enable the construction of $f_i^{ls}()$-like transforms which are more similar than the images themselves. The two last rows in Table 2 compare the inter-image pixel correlation (u_1, u_i) with the inter-function correlation $(f_1^{ls}(), f_i^{ls}())$ (computed for the correlation between the optimal 5^{th} order $\beta-$coefficients derived from each image). One can observe why "common-function" compression outperforms inter-image predictive models. It is because the inter-function average correlation is as high as 0.998, and so we can essentially use the same compressing function. This is why $f_1^{ls}()$ efficiently compresses all similar images. Thus, instead of relying on relatively low and error-sensitive pixel-to-pixel image correlation, the functional nature of image similarity takes advantage of the similarities in the images of the same type.

Moreover, images can be added or removed from the database without recomputing the compressing function, centroid images, inter-frame predictive coefficients or any image clusters. After a CAR compressing transform is built from one database representative, each image is compressed independently, which makes this technique

very suitable for parallel architectures. Essentially, the CAR predicting model for brain images can predict any brain image of approximately the same size and orientation, or a CAR predicting model for a knee can be applied to any knee scan of the same nature to achieve an improved compression ratio, etc.

Finally, it was interesting to observe that after $f_1()$ is used to efficiently remove all intra-image correlation, the inter-image correlation among the residuals decreases to almost zero (in our experiment as low as 0.003). This again shows that the intra-image correlation was actually induced by the inter-image correlation, and vice versa. Removing one essentially removes both, and the database becomes essentially completely uncorrelated.

The following subsections provide the theoretical support for the CAR method.

Assumptions

We based our studies on CT, MR and aerial surveillance images. For each class of images, all representatives display objects of the same nature (e.g., human brain) and look very similar, but their similarities are more structure-based than region-based. In particular, unlike MPEG compression, these images do not have any closely coinciding regions even after mapping them onto each other. For our study, we only assumed image smoothness (positive correlation between neighboring pixels) and similarly a smoothness of image differences. The latter means that if we subtract one smooth similar image from any linear combinations of other smooth similar images, the residual will remain smooth (positively autocorrelated). This last assumption is intuitively clear and can be supported with our numerical results.

Simple Study

In the previous chapter we studied how intra image correlation β in any similar database image v depends upon average inter-image correlation $\rho = \rho(u, v)$ and intra-image correlation α in some chosen "reference" image u. One can perform very similar derivations to analyze how the proximity between α and β affects the autoregressive compression quality of these images as opposed to compressing the same set of images with inter-image predictors. We use the following notations: u and v for vectors (images), r for the predictive residual in regression $v = \rho u + r$, α and β for intra-image autocorrelation coefficients, and the subscript s for image shifts in the coordinate domain. Thus we can express the first order autoregressive (AR) model for the u image as $u = \beta u_s + r$, where β is a constant and s corresponds to 1-pixel shift in the "past" direction: $u_s[i] = u[i-1]$ [30], [34]. A similar model for the v image is $v = \alpha v_s + t$. Without any loss of generality we can assume that all images are centered and normalized: $\overline{u} = \overline{v} = 0$ (averages) and $||u|| = ||v|| = 1$ (least square norms). In this case, α, β and ρ are correlation coefficients between images and their predictors in the respective models.

Remark: $q = v - \beta v_s$, the residual after applying the u-model to v. Then $||q||^2 = (\alpha - \beta)^2 + 1 - \alpha^2$.

Proof.

$||q||^2 = q^T q = (v^T - \beta v_s^T)(v - \beta v_s) = v^T v - \beta v^T v_s - \beta v_s^T v + \beta^2 v_s^T v_s$

$= 1 - \beta\alpha - \beta\alpha + \beta^2 = (\alpha - \beta)^2 + 1 - \alpha^2$.

■

In particular, if ρ is the inter-frame correlation, $\rho = u^T v$, and r is the inter-frame prediction residual, $r = v - \rho u$, compressing all database images with the same AR model (CAR) will be more efficient when compared to the inter-frame predictors if $||q|| < ||r||$ or $(\alpha - \beta)^2 + 1 - \alpha^2 < 1 - \rho^2$, or

$$\Delta = |\alpha - \beta| < \sqrt{\alpha^2 - \rho^2}, \text{ or}$$

$$\alpha - \sqrt{\alpha^2 - \rho^2} < \beta < 1$$

Example: We observed that for the registered CT images ρ rarely exceeds 0.8, the average α is often above 0.95. Thus for any other CT image with intra-frame ("auto") correlation $\beta > 0.95 - \sqrt{0.95^2 - 0.8^2} = 0.44$. This means it will be more efficiently compressed with the same AR model, rather than inter-frame predictor. Since in practice β, as α, is much higher than 0.44, CAR compression for CT images works better than traditional inter-frame prediction.

Considering Inter-frame Correlation

The previous simplified result does not consider that α, β and ρ depend on each other: ρ is the correlation between images with α and β intra-image correlations. In particular, if ρ and α become large, β cannot be small, and $\Delta = \alpha - \beta$ depends on ρ as:

$$\Delta = \alpha - \beta = v_s^T v - u_s^T u = v_s^T v - \tfrac{1}{\rho}(v_s^T - r_s^T)\tfrac{1}{\rho}(v - r)$$

$$= v_s^T v - \tfrac{1}{\rho^2}(v_s^T - r_s^T)(v - r) = v_s^T v(1 - \tfrac{1}{\rho^2}) + \tfrac{1}{\rho^2} v_s^T r + \tfrac{1}{\rho^2} r_s^T v - \tfrac{1}{\rho^2} r_s^T r$$

$$= \alpha(1 - \tfrac{1}{\rho^2}) + \tfrac{1}{\rho^2}(v_s^T r + r_s^T v - r_s^T r),$$

where $\left|v_s^T r + r_s^T v\right| \leqslant 2\sqrt{1-\rho^2}$ and $\left|r_s^T r\right| \leqslant 1-\rho^2$. The term $v_s^T r + r_s^T v - r_s^T r$ can be estimated as follows:

$$v_s^T r + r_s^T v - r_s^T r = \sum_i (r[i]v[i-1] + r[i]v[i+1] - r[i]r[i-1])$$

$$= \sum_i (r[i](v[i-1] + v[i+1]) - r[i]r[i-1]),$$

and since we know that

$$v[i+1] = \alpha v[i] + t[i],$$

introducing a similar AR model for r as

$$r[i+1] = \gamma r[i] + \delta[i],$$

we obtain

$$v_s^T r + r_s^T v - r_s^T r = \sum_i (r[i](v[i-1] + v[i+1]) - r[i]r[i-1])$$

$$= \sum_i (r[i]\left\{\alpha v[i] + t[i] + \tfrac{1}{\alpha}(v[i] - t[i-1])\right\} - r[i](\gamma r[i] + \delta[i]))$$

$$= \sum_i (\left[\alpha + \tfrac{1}{\alpha}\right] r[i]v[i] + r[i](t[i] - \tfrac{1}{\alpha}t[i-1]) - \gamma r^2[i]).$$

We observed that the inter-frame residual r and AR residual t in practice appear to be almost uncorrelated, and so we can assume[19] $\sum_i r[i]t[i] = \sum_i r[i]t[i-1] \approx 0$. Then

$$v_s^T r + r_s^T v - r_s^T r \approx \sum_i (\left[\alpha + \tfrac{1}{\alpha}\right] r[i]v[i] - \gamma r^2[i]) = \left[\alpha + \tfrac{1}{\alpha}\right] r^T v - \gamma r^T r$$

$$= \left[\alpha + \tfrac{1}{\alpha}\right] r^T(\rho u + r) - \gamma r^T r = \left[\alpha + \tfrac{1}{\alpha}\right] r^T r - \gamma r^T r = (\alpha + \tfrac{1}{\alpha} - \gamma) r^T r$$

$$= (\alpha + \tfrac{1}{\alpha} - \gamma)(1-\rho^2).$$

Finally,

$$\Delta = \alpha - \beta = \alpha(1 - \tfrac{1}{\rho^2}) + \tfrac{1}{\rho^2}(v_s^T r + r_s^T v - r_s^T r)$$

$$\approx \alpha(1 - \tfrac{1}{\rho^2}) + \tfrac{1}{\rho^2}(\alpha + \tfrac{1}{\alpha} - \gamma)(1-\rho^2) = (\tfrac{1}{\alpha} - \gamma)(1-\rho^2)$$

[19] One can also neglect this term as having the lowest order of magnitude with respect to the others.

For relatively smooth (intra-correlated) images like CT or MR, the residual r that results after removing one similar image from the other, inherits this intra-image correlation, i.e., $\gamma > 0$ (our assumption of residual smoothness). In fact, we experimented with CT and MR images and discovered that $0 \leqslant \gamma \leqslant \alpha$, often with $\gamma \rightarrow \alpha$. Therefore

$$\Delta \approx (\tfrac{1}{\alpha} - \alpha)(1 - \rho^2),$$

so the condition for the CAR model to provide better decorrelation than the inter-frame predictor becomes

$$(\tfrac{1}{\alpha} - \alpha)(1 - \rho^2) < \sqrt{\alpha^2 - \rho^2}.$$

If the image intra-correlation α increases, i.e., $\alpha \rightarrow 1$, then $\frac{1}{\alpha} - \alpha \rightarrow 0$, and the above condition will always be satisfied *for any* $\rho > 0$. The condition $\alpha \rightarrow 1$ can be met with higher order CAR models: 10-th order models for our test images accounted for 0.995 of intra-image correlation. Therefore,

Corollary 5. For a set of similar ($\rho > 0$) images with high intra-image correlation, $\alpha \rightarrow 1$ **for at least one image** v, the AR model derived from this image will decorrelate **any** other similar image better than the inter-frame decorrelation.

Residual Smoothness

One may argue that the assumption $\gamma \rightarrow \alpha$ (inter-frame residual autocorrelation approaches image autocorrelation) is too strong to always be true. In general, it is always true that, $0 \leqslant \gamma \leqslant \alpha$. Therefore, we can introduce a parameter $\lambda = \frac{\gamma}{\alpha}$, and

$0 \leqslant \lambda \leqslant 1$. The condition for the AR model to outperform inter-frame compression becomes:

$$(\tfrac{1}{\alpha} - \alpha\lambda)(1 - \rho^2) < \sqrt{\alpha^2 - \rho^2}.$$

We choose five values of λ, $\{0.2, 0.4, 0.6, 0.8, 1.0\}$, and plotted curves for $(\frac{1}{\alpha} - \alpha\lambda)(1 - \rho^2) = \sqrt{\alpha^2 - \rho^2}$ on the Figure 16.

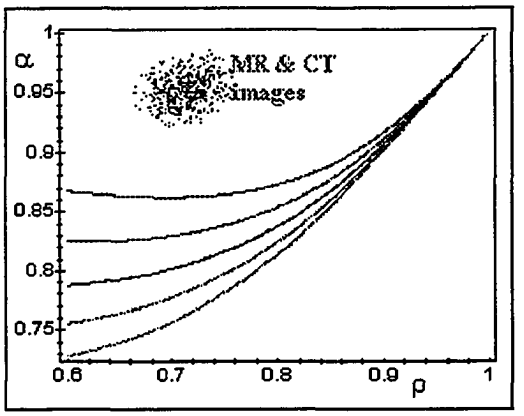

Figure 16: Intra (α) vs inter (ρ) image correlation.

The lowest curve corresponds to $\lambda = 1$ ($\gamma \rightarrow \alpha$), the highest curve to $\lambda = 0.2$. Each curve shows how the intra-image autoregressive correlation α increases with increasing inter-image correlation ρ, and the region above the curve that corresponds to $(\frac{1}{\alpha} - \alpha\lambda)(1 - \rho^2) < \sqrt{\alpha^2 - \rho^2}$, is the case when a one-image AR outperforms an inter-frame decorrelation. Two important observations can be made:

1. AR intra-image correlation α increases faster than the inter-frame correlation ρ, both in terms of values and functions[20] .

2. All of the CT, satellite and MR images studied, with $(\alpha, \rho) \approx (0.95, 0.80)$, fall in the "above the curve" region even for small values of λ. This result indicates

[20] $\alpha(\rho)$ increases approximately hyperbolically for $\rho < 1$, and is almost linear only when $\rho \rightarrow 1$.

that a CAR model will compress the similar database much better than inter-frame correlation.

Therefore, the residual r smoothness assumption can be relaxed to $\gamma = corr(r, r_s) > 0$ without changing the validity of our results.

The Existence of Common AR Models

Do AR models perform well simply because they compress any image efficiently or because they really correspond to some underlying image similarities ? To demonstrate the connection between similar image AR transforms and provide an answer to this question, we used 21 test images: 10 CT, 10 MR and Lena, 256x256x8 bits each. The images were plotted in 2D using optimal principal component mapping. One can observe on Figure 17 that these images fall into 3 distinct clusters: MR images are different from CT, and both classes are quite different from the Lena image.

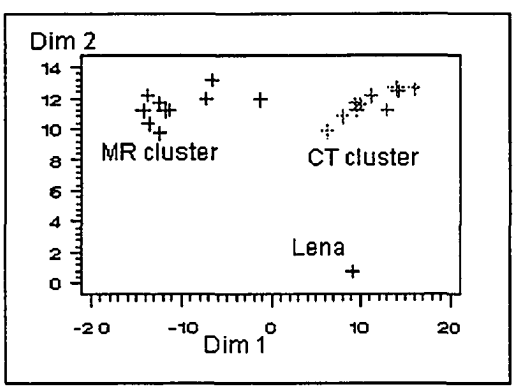

Figure 17: Similar image classes in 2D.

Then we computed AR models of orders 2 through 10 for each image and optimally mapped the model coefficients β into 2D. The models of order 2 and 3 do not differentiate between CT and MR clusters (Figure 18, left), but already outlay the

Lena image. As the order increases to 10 (Figure 18, right), the distinction between the three clusters becomes obvious.

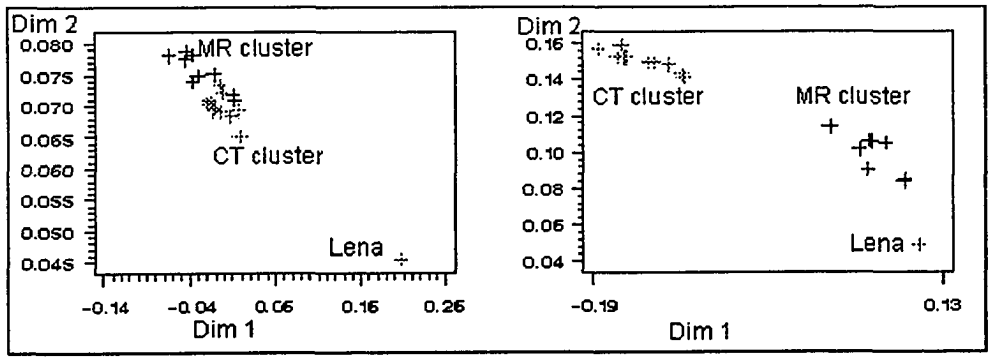

Figure 18: 3^{rd} (left) and 10^{th}(right) order AR models.

Concluding that AR models do correspond to similar image classes, we also cross-validated the efficiency of each AR model within each class. Figure 19 represents our results visually. All 21 AR models have been applied to compress all 21 images, and all 21×21 variances of compressed (decorrelated) images have been determined. Figure 19 represents this variance matrix, where the n-th row corresponds to the AR model derived and therefore is optimal for the n-th test image, and k-th column corresponds to the k-th image. The images (models) from 1 to 10 are CT images (models), the images (models) from 11 to 20 are MR, and #21 corresponds to Lena. For each (n, k) cell the intensity of the cell is proportional to the residual variance after applying the model extracted from the n-th image to the k-th image (lighter shade means a higher error).

The three dark diagonal clusters correspond to compressing CT images with CT AR models, MR images with MR AR models, and Lena image with Lena AR model. In this case, the compression ratio is high. The light areas below and above the

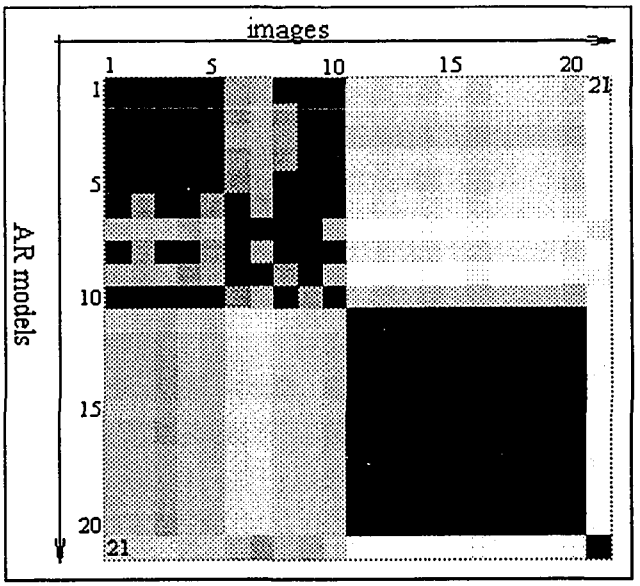

Figure 19: AR model crossvalidation.

main diagonal show that AR models from one class applied to images from the other similar image class perform poor: in this experiment, applying AR models to images from different similar classes yielded a 20-50% residual variance increase. This image nature-sensitive model performance plus the model coefficient clustering points to a conclusion that AR models in fact capture similarities in the nature of the images. Therefore, it appears to be valid to talk about a general AR model for CT images, which is different from an AR model for MR images and so on[21] .

AR models for similar images tend to be more correlated than the images themselves and are able to reproduce image similarities in a very concise form. This leads to a proposal for model-based compression for large data sets of similar images as follows:

[21] Of course, assumption of image simularity includes an assumption that images are taken using the same modality, same type of imaging device etc.

1. Given a large database of similar images, choose one (typical) representative and construct its optimal n-th order AR model.

2. Use this same model to achieve nearly optimal compression for any other image from the same database.

AR models are built from local image properties (pixel neighborhoods) and therefore do not require images to be registered, coincide over some regions, etc. In next section we will examine how the similarities between AR models can be affected by image transforms.

AR Model Tolerance

Theoretical Estimate

Many similar image database compressing techniques are very sensitive to image translations and rotations, and therefore require all images to be accurately matched (registered) before compression. The process of image registration is computationally expensive in large image databases. Therefore, there has always been a need for transform-insensitive compression approaches. Common AR models provide an efficient and transform-tolerant method to compress similar images. The 5-th order AR model was used to derive an AR model tolerance estimate with respect to image translations and rotations. With the operators B and L representing 1-pixel bottom and left image shifts respectively, the 5-th order AR model can be represented as:

$$u = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)u + r, \tag{11}$$

where $L^m B^n u[i,j] = u[i-m, j-n]$. Optimal values for $\beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]^T$ are found with traditional least squares regression as:

$$\beta = (v^T v)^{-1}(v^T u),$$

where matrix v has five columns, and all are shifts of the vector u:

$$v = (Lu, Bu, LBu, L^2 u, B^2 u).$$

It follows from the definition of an AR model that image translation $T_{\mu\nu}(u) = L^\mu B^\nu u$, applied to the entire image, does not have any effect on the model coefficients β (at least for $\mu, \nu \ll$size u). Therefore, only rotations must be analyzed. A rotation by p degrees R_p will transform the model (11) into

$$u = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2) R_p(u) + q, \tag{12}$$

where $(L^m B^n R_p)u[i,j] = u[i - m\cos(p) + n\sin(p), j - m\sin(p) - n\cos(p)]$. To find how coefficients β of the rotated image model depend on rotation angle p, it was assumed that (12) must be satisfied by all polynomial functions of order less than three. Thus, substituting into (12) $u_1[i,j] = 1$, $u_2[i,j] = i$, $u_3[i,j] = j$, $u_4[i,j] = ij$, $u_5[i,j] = i^2$, and $u_6[i,j] = j^2$, and simplifying the expression it was found that:

$$A\beta = V,$$

where:

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -\cos p & \sin p & -\cos p + \sin p & -2\cos p & 2\sin p \\ -\sin p & -\cos p & -\sin p - \cos p & -2\sin p & -2\cos p \\ \frac{1}{2}\sin 2p & -\frac{1}{2}\sin 2p & \cos 2p & 2\sin 2p & -2\sin 2p \\ \cos^2 p & \sin^2 p & -\sin 2p + 1 & 4\cos^2 p & 4\sin^2 p \\ \sin^2 p & \cos^2 p & \sin 2p + 1 & 4\sin^2 p & 4\cos^2 p \end{pmatrix},$$

$$V = [1,0,0,0,0,0]^T, \text{ and } \beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]^T.$$

The overdefined system $A\beta = V$, which has more equations than variables, was solved with least squares to produce:

$$\beta = \beta(p) = \frac{1}{2\cos^4 p - 2\cos^2 p + 7} \begin{bmatrix} 2\cos^3 p \sin p + 2\cos^4 p - \cos p \sin p - 2\cos^2 p + \frac{9}{2} \\ -2\cos^3 p \sin p + 2\cos^4 p + \cos p \sin p - 2\cos^2 p + \frac{9}{2} \\ -2\cos^4 p - 1 + 2\cos^2 p \\ -\cos^3 p \sin p + \frac{1}{2}\cos p \sin p - 1 \\ \cos^3 p \sin p - \frac{1}{2}\cos p \sin p - 1 \end{bmatrix},$$

Therefore, the correlation between original and rotated models was:

$$corr(p) = corr(\beta(0), \beta(p)) = \frac{16\cos^4 p - 16\cos^2 p + 33}{3\sqrt{121 + 113\cos^4 p - 109\cos^2 p + 4\cos^8 p - 8\cos^6 p}},$$

and relative model error

$$e(p) = \frac{||V - A\beta(p)||}{||V - A\beta(0)||} = \sqrt{\frac{7}{2\cos^4 p - 2\cos^2 p + 7}}.$$

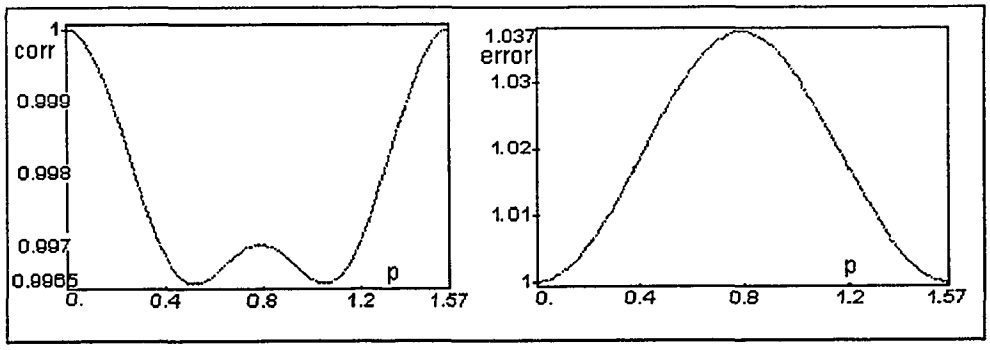

Figure 20: Functions $corr(p)$ and $e(p)$.

Solving for $corr(p)$ and $e(p)$ extrema, one can prove that for any p, $0 \leqslant p \leqslant \frac{\pi}{2}$, (see Figure 20)

$$\sqrt{\tfrac{17776}{17901}} \leqslant corr(p) \leqslant 1, \text{ and } 1 \leqslant e(p) \leqslant \sqrt{\tfrac{14}{13}}.$$

This means that the decrease in model correlation will not exceed 0.4%, and the error introduced by model rotation will not increase by more than 4%. In practice, as will be shown in the next section, these changes will be more visible; however, this still demonstrates the potential of the AR models and their insensitivity to translations and rotations of an image.

Numerical Estimate of Tolerance

CT and MR images may not be properly registered, i.e., may be shifted and rotated with respect to each other. Even a relatively small change in image orientation affects traditional set compression techniques based on pixel or region correlation.

We measured how these orientation changes can affect AR models in practice and discovered that AR models are much more fault tolerant than the images themselves. For instance, the left graph on Figure 21 shows two correlation curves: correlation

$\rho_{image}(ct_1, R_\alpha(ct_2))$ between 2 CT images ct_1 and ct_2, as the second image ct_2 was rotated from 0 to 90 degrees (the decreasing curve), and the correlation between the 8-th order AR models $\rho_{model}(m_{ct_1}, m_{R_\alpha(ct_2)})$ of the original and rotated image. One can observe that AR models exhibit, in general, more similarities than the corresponding images for small rotation angles, and is very similar to our theoretical prediction.

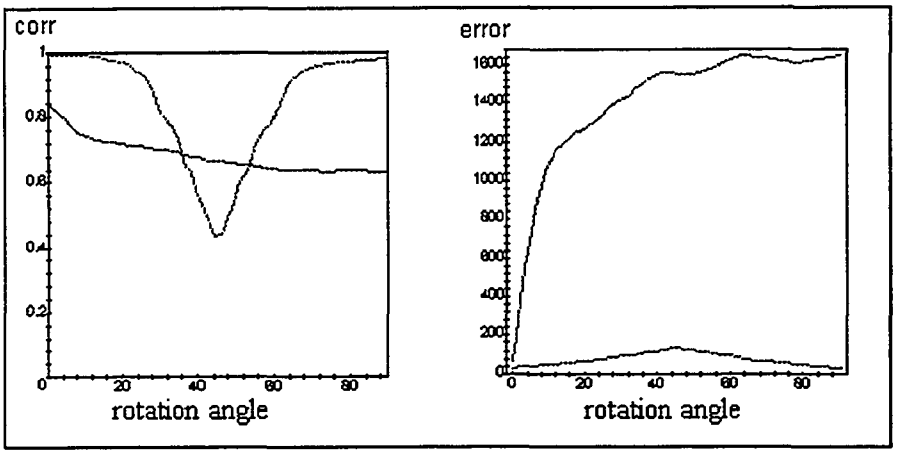

Figure 21: Rotational tolerance for correlation (left) and error.

Similarly, the higher curve in the right graph on Figure 21 shows how mean squared error $MSE_{image}(ct_1, R_\alpha(ct_2))$ between images increases with rotation. This curve grows almost exponentially even for small rotation angles showing that pixel-based interpretation of image similarity is unstable in terms of image transforms. However, if we continuously apply the first image AR model m_{ct_1} to the second rotated image, the lower curve on the right graph demonstrates very stable decorrelation and significantly outperforms pixel-based decorrelation.

We conducted a similar experiment for image translation. Instead of rotation, one CT image ct_2 was shifted in 2D (x and y directions) with respect to the other

one[22] . Figure 22, left, shows that AR model correlation (flat surface) $\rho_{model}(m_{ct_1}, m_{T_{dx,dy}(ct_2)})$, as expected, does not have any noticeable changes. Unlike AR correlation, pixel correlation $\rho_{image}(ct_1, T_{dx,dy}(ct_2))$ is not very stable and is translation dependent. Consequently, the pixel registration error $MSE_{image}(ct_1, T_{dx,dy}(ct_2))$ increases on Figure 22, right, (the top surface). However, MSE corresponding to the application of the same 8-th order AR model derived from ct_1 to rotated ct_2 (the other surface), remains almost insensitive to any translations and demonstrates the efficiency of similar image decorrelation.

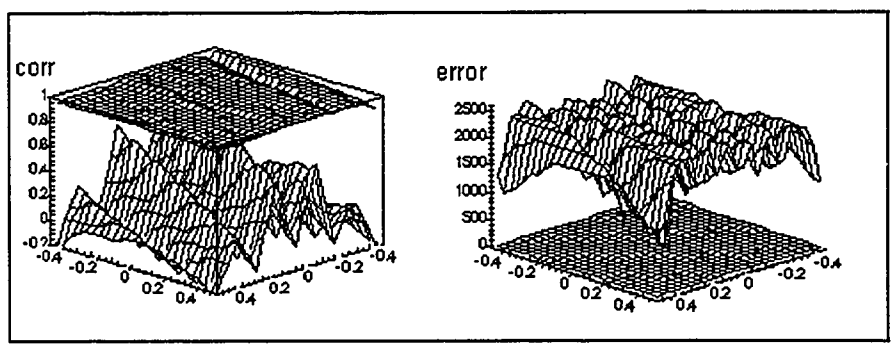

Figure 22: Translational tolerance for correlation (left) and error.

Conclusion

The study of CAR models can be summarized as follows:

a. CAR models corresponding to similar images are highly correlated for images of the same nature and modality, and are distinct for dissimilar images.

b. A CAR model derived from one image will efficiently compress all similar images, and in general will perform poorly for any dissimilar image.

c. CAR models are very insensitive to global image transforms.

[22]Translations varied from -50% to 50% of the image size.

d. As Figure 19 suggests, CAR models can be also used for image classification, since they are capable of distinguishing the images taken from different similarity classes.

Because CAR models are very efficient in time, memory and also fault tolerant, one can conclude that they provide an efficient approach to similar image database compression.

We used correlation between images or transform coefficients as a principal measure of their respective similarity. However, when it comes to predictive set compression, it is the entropy of the difference image, rather than correlation or variance, which is proportional to compression ratio. Some preliminary estimates for this relation between entropy (as an informational measure) and correlation (as the measure of fitness for linear transforms) were already given, pages 27-34. Now we will develop a more complete and general theory relating these statistical and informational similarity measures, which can be applied to any integer data (digitized images, quantized transform coefficients, etc.)

For any linear regression model, predicting v_i from a set of similar v_j

$$v_i = \left\lfloor \sum_{j\neq i, j=1}^{j=n} \beta_j v_j \right\rfloor + r^{(i)} = \left\lfloor \beta^{(i)} V^{(i)} \right\rfloor + r^{(i)} = \widehat{v}_i + r^{(i)}, \tag{13}$$

the model residual $r^{(i)}$ can be viewed as the difference

$$r^{(i)} = v_i - \widehat{v}_i$$

between the original image v_i and its predictor $\widehat{v}_i$. Using this view, we will limit our discussion to the basic difference compression models, when image v_i is replaced by $r^{(i)}$ with compression ratio

$$C = \frac{H(v_i)}{H(r^{(i)})} = \frac{H(v_i)}{H(v_i - \widehat{v}_i)}.$$

Information theory will be used to develop results which allow conclusions about C based on image-to-predictor correlation $\rho = \rho(v_i, \widehat{v}_i)$. This will be done step by step

in the several sections, and the difference model with more complicated extensions will be used as a basis to study information dependency of two correlated sources. This theoretical approach will be illustrated and supported with various numerical results.

Relativity of Correlation

In general, the presence of correlation between two information sources does not uniquely define their entropies nor the entropy of their difference. For instance, consider two pair of sources (vectors)

$$u_1 = (0,1,2),\ v_1 = (0,2,1), \text{ and } u_2 = (0,1,3),\ v_2 = (0,3,1).$$

All these vectors possess the same probability distribution of their components (uniform), and the same entropies. The difference vectors $r_i = u_i - v_i$, $i = 1,2$, also have the same uniform distributions and same entropies ($\log_2 3$). Therefore, entropy coding of these two pairs and the differences within each pair will be identical. However, replacing symbol 2 in the pair (u_1, v_1) by 3 to form (u_2, v_2) did affect the vector correlation: $\rho(u_1, v_1) = 0.5$, and $\rho(u_2, v_2) = 0.14$. Since image entropy and coding deals primarily with the distributions of different image intensities, ignoring the intensity values, and correlation is computed based on the intensity values, using image correlation to judge the inter-image information similarity would be ambiguous.

To eliminate this ambiguity[23] , we chose the same signal alphabet which assumes that all image intensities must be integer and distributed within the same range of n integer values, from 1 to some integer $n \geqslant 2$.

[23]Since shifting (increasing by the same value) all intensity values neither effects entropy nor correlation, no assumptions are needed about minimum or maximum intensities.

Due to the image-related nature of this research, we will use the concepts of "image" and "discrete integer source" interchangeably.

Continuity of Entropy as a Function of Inter-Image Correlation

When $\rho(u,\widehat{u}) = 1$, u and $\widehat{u}$ are identical, and entropy $H(r) = H(u-\widehat{u}) = 0$. However, in practice $\rho(u,\widehat{u}) < 1$, and we prefer to know how increasing $\rho(u,\widehat{u})$ affects $H(r)$. In the following two subsections, a theory relating difference entropy $H(r) = H(u-\widehat{u})$ and correlation $\rho(u,\widehat{u})$ will be developed.

Lemma 5. For any discrete integer source u with variance $\sigma(u)$ and a constant number of states n $\sigma(u) \to 0$ implies $H(u) \to 0$.

Proof.

Let's assume u has n different intensity levels (states): $u[i] \in \{1,2,...,n\}$. The probability of the $i-$th intensity is p_i, and the average intensity is $\overline{u} = \sum_{i=1}^{n} ip_i \in [1,n]$. We choose integer k as the closest integer intensity level to the average intensity: $k = \lfloor \overline{u} + \frac{1}{2} \rfloor$, $k \in [1,n]$. If all probabilities except p_k are zero, both $\sigma(u) = 0$ and $H(u) = 0$, and the lemma is proved. Otherwise the image variance $\sigma^2(u) = \sum_{i=1}^{n}(i-\overline{u})^2 p_i \geqslant \sum_{i=1,i\neq k}^{n}(i-\overline{u})^2 p_i \geqslant \sum_{i=1,i\neq k}^{n}(\frac{1}{2})^2 p_i = \frac{1}{4}\sum_{i=1,i\neq k}^{n} p_i$, and $\sigma(u) \to 0$ implies $\sum_{i=1,i\neq k}^{n} p_i \to 0$ and $p_i \to 0$ for any $i \neq k$. Since $\lim_{p\to 0}(-p\log_2(p)) = 0$, and n is constant, also requires $\sum_{i=1,i\neq k}^{n}(-p_i\log_2(p_i)) \to 0$. On the other side $\sum_{i=1}^{n} p_i = 1$, therefore $\sum_{i=1,i\neq k}^{n} p_i \to 0$ leads to $p_k \to 1$, and $-p_k\log_2(p_k) \to 0$. Therefore the entropy $H(u) = \sum_{i=1}^{n}(-p_i\log_2(p_i)) = \sum_{i=1,i\neq k}^{n}(-p_i\log_2(p_i)) - p_k\log_2(p_k) \to 0$. ■

The requirement of discreteness is essential. An obvious counterexample is the continuous normal distribution with entropy $H(N(0,\sigma)) = \log_2\sqrt{2\pi e\sigma^2} \to -\infty$ as

$\sigma \to 0$. The other "hidden" use of the discreteness is that the number of image intensities n does not increase as $\sigma(u) \to 0$ (otherwise one could consider a discrete binomial distribution, which converges to a normal distribution as $n \to \infty$, and this will also contradict $H(u) \to 0$).

Lemma 6. For two images u and v with equal variances $\sigma(u) = \sigma(v)$ and a constant number of intensities n

$\rho(u,v) \to 1$ implies $H(r) = H(u-v) \to 0$.

Proof.

The difference $r = u - v$ can have at most $n' = 2n-1$ intensity levels. Its variance $\sigma^2(r) = \sigma^2(u-v) = \sigma^2(u) + \sigma^2(v) - 2cov(u,v) = 2\sigma^2(u) - 2\rho(u,v)\sigma^2(u) = 2\sigma^2(u)(1-\rho(u,v)) \to 0$ as $\rho(u,v) \to 1$, and apply the previous lemma. ■

Assumption $\sigma(u) = \sigma(v)$ is not so restrictive as it seems because we can always multiply all intensities in v by $\frac{\sigma(u)}{\sigma(v)}$ to satisfy this. It simply means that difference compression model makes sense only if the images are on the same scale. A counterexample is $u = (-1,1)$, $v = (-2,2) : \rho(u,v) = 1$, but $r = u - v = (1,-1) = -u$, and $H(r) = H(u) = 1$. Moreover, the assumption of equal variances is very natural for similar image databases, where images are inclined to have very close statistical properties. Therefore, we accepted this assumption for our study of difference model compression performance.

For the difference model, $u = v + r$, we reintroduce the entropy ratio function as

$$HR(u,v,\rho(u,v),n) = \frac{H(u-v)}{H(u)} = \frac{1}{C(u,v,\rho(u,v),n)}, \tag{14}$$

where $\sigma(u) = \sigma(v)$; n is the maximum number of intensity levels in u and v; and $C(u, v, \rho(u, v), n)$ is the compression ratio that we achieve replacing image u by its $v-$difference $r = u - v$. Both $H(u)$ and therefore $HR(u, v, \rho(u, v), n)$ and $C(u, v, \rho(u, v), n)$ are positive continuous functions of probabilities with image intensity distributions, and we already proved that given any image u_0, $\sigma_0 = \sigma(u)$, and integer n

$$\lim_{\rho(u,v)\to 1} HR(u, v, \rho(u, v), n) = 0,$$

over all v with at most n intensity levels and $\sigma(v) = \sigma_0$. However, it is still unclear how this convergence to 0 may be affected by the choice of u; for instance, what happens if $H(u) \to 0$? In other words, is there a finite lower (worst-case) bound on compression ratio $C(u, v, \rho(u, v), n)$ (upper bound on $HR(u, v, \rho(u, v), n)$) for two $\rho-$correlated images with n intensity levels, which only depend on inter-image correlation ρ ?

These questions will be answered in the following sections. Before addressing them, notice that $HR(u, v, \rho(u, v), n)$ is a function of the inter-image bivariate probability distribution $\mathcal{P} = \{P_{ij} = P(u = i \ \& \ v = j)_{1 \leqslant i,j \leqslant n}\}$ between intensity levels in images u and v because all other parameters in (14) can be expressed in terms of $\mathcal{P}$ as:

1. Distribution of u intensities $\mathcal{P}^u = \{P_i^u = P(u = i)_{1 \leqslant i \leqslant n}\}$ is a marginal distribution of $\mathcal{P}$ with $P_i^u = \sum_{j=1}^{n} P_{ij}$; similarly for the image v, $\mathcal{P}^v = \{P_j^v = P(v = j)_{1 \leqslant j \leqslant n}\}$ with $P_j^v = \sum_{i=1}^{n} P_{ij}$.

2. Average intensity $\overline{u} = \sum_{j=1}^{n} iP_i^u = \sum_{i,j=1}^{n} iP_{ij}$ (similarly for v).

3. Variance $\sigma^2(u) = \sum_{j=1}^{n}(i-\overline{u})^2 P_i^u = \sum_{i,j=1}^{n}(i - \sum_{i,j=1}^{n} iP_{ij})^2 P_{ij}$ (similarly for v).

4. Covariance $cov(u,v) = \sum_{i,j=1}^{n}(i-\overline{u})(j-\overline{v})P_{ij}$.

5. Correlation $\rho(u,v) = cov(u,v)/\sigma^2(u)$ (since $\sigma(u) = \sigma(v)$).

6. Entropy $H(u) = -\sum_{i=1}^{n} P_i^u \log_2 P_i^u$.

7. Difference distribution for $r = u - v$ is $\mathcal{P}^r = \{P_k^r = P(r=k)_{-n+1 \leqslant k \leqslant n-1}\}$, where $P_k^r = \sum_{i-j=k} P_{ij}$.

8. Difference entropy $H(u-v) = -\sum_{k=-1+n}^{n-1} P_k^r \log_2 P_k^r$.

Thus:

$$HR(u,v,\rho(u,v),n) = HR(\mathcal{P},n). \tag{15}$$

Therefore the worst and the best case information reduction over all ρ-correlated images with n intensity levels, obtained after subtracting one image from the other, are respectively

$$HR^{\sup}(n,\rho) = \sup_{\mathcal{P}:\ \sigma(\mathcal{P}^u)=\sigma(\mathcal{P}^v),\ \rho(u,v)=\rho} HR(\mathcal{P},n), \tag{16}$$

$$HR^{\inf}(n,\rho) = \inf_{\mathcal{P}:\ \sigma(\mathcal{P}^u)=\sigma(\mathcal{P}^v),\ \rho(u,v)=\rho} HR(\mathcal{P},n), \tag{17}$$

and the corresponding worst-case/best-case compression ratios are

$$C^{\inf}(n,\rho) = \frac{1}{HR^{\sup}(n,\rho)}, \tag{18}$$

$$C^{\sup}(n,\rho) = \frac{1}{HR^{\inf}(n,\rho)}. \tag{19}$$

The problem with definitions (16) and (18) is that we do not know if $HR^{\sup}(n,\rho)$ is finite, or how large it can be[24] . The next section will clarify the concept of $HR^{\sup}(n,\rho)$ $(HR^{\inf}(n,\rho))$ and $C^{\inf}(n,\rho)$ $(C^{\sup}(n,\rho))$ and illustrate how these functions can be introduced and computed for $n = 2$. Then we will consider the general case for variable n.

The notation $HR^{ext}(n,\rho)$ will be used to refer to any of $HR^{\sup}(n,\rho)$ or $HR^{\inf}(n,\rho)$, as well as $C^{ext}(n,\rho)$ for any one of $C^{\inf}(n,\rho)$ or $C^{\sup}(n,\rho)$.

[24] We proved earlier that $HR^{\sup}(n,\rho) \leqslant 2$ if it is assumed that $H(u) \geqslant H(v)$, but here we do not use this assumption.

Binary Images: Complete Entropy-Correlation Study

A powerful entropy-correlation theory can be developed for the images u and v with only $n = 2$ intensity values: $\{1, 2\}$[25] . In this case the inter-image bivariate intensity probability has $n^2 = 4$ values:

$$\mathcal{P} = \{P(u = i \text{ and } v = j)_{1 \leqslant i,j \leqslant 2}\} = \{p_1, p_2, p_3, p_4\},$$

where

$$p_1 = P(u = 1 \text{ and } v = 1),\ p_2 = P(u = 2 \text{ and } v = 1),\ p_3 = P(u = 1 \text{ and } v = 2),$$

$$p_4 = P(u = 2 \text{ and } v = 2).$$

The marginal distribution for the image u is $\mathcal{P}^u = \{P(u = 1) = p_1 + p_3,\ P(u = 2) = p_2 + p_4\}$, for image v, $\mathcal{P}^v = \{P(v = 1) = p_1 + p_2,\ P(v = 2) = p_3 + p_4\}$, and for the difference $r = u - v$, $\mathcal{P}^r = \{P(r = -1) = p_3,\ P(r = 0) = p_1 + p_4, P(r = 1) = p_2\}$[26] . We impose the following constraints on p_i :

1. Probabilities, $p_i \geqslant 0$ and $\sum_{i=1}^{4} p_i = 1$, and
2. Variance similarity: $\sigma^2(u) = \sigma^2(v)$.

We can determine the functions $HR^{ext}(n, \rho)$ and $C^{ext}(n, \rho)$ for $n = 2$, using the above constraints for the probability values, and substitute these into (16), (17), (19) and (18). From 1, $p_4 = 1 - p_1 - p_2 - p_3$. Then $\overline{u} = 1(p_1 + p_3) + 2(p_2 + p_4) = 2 - p_1 - p_3$, and $\overline{v} = 1(p_1 + p_2) + 2(p_3 + p_4) = 2 - p_1 - p_2$. We also find $\sigma^2(u)$ and $\sigma^2(v)$ using results from the previous section, and substitute these into 2:

[25] One can use any two integers without affecting the results that follow.

[26] The difference of binary images is not a binary image !

$$\sigma^2(u) = p_1 - p_1^2 - 2p_1p_3 + p_3 - p_3^2 = \sigma^2(v) = p_1 - p_1^2 - 2p_1p_2 + p_2 - p_2^2,$$

which, when solved for p_3, yields:

a) $p_3 = p_2$, or

b) $p_3 = -p_2 - 2p_1 + 1$.

We will consider in detail a) since it will lead to numerical values for $HR^{ext}(2,\rho)$ and $C^{ext}(2,\rho)$. The second case, $p_3 = -p_2 - 2p_1 + 1$, was evaluated exactly the same way with $p_3 = p_2$, but it did not result in any improved values for $HR^{ext}(2,\rho)$ when compared to those numerically determined for $p_3 = p_2$ in the next section.

Case $p_3 = p_2$

In this case inter-image correlation can be expressed as

$$\rho(u,v) = \frac{p_1 - p_1^2 - 2p_1p_2 - p_2^2}{p_1 + p_2 - p_1^2 - 2p_1p_2 - p_2^2}.$$

Solving this for $p_2 = p_2(p_1,\rho)$ yields[27] :

$$p_2 = \frac{-\rho + 2\rho p_1 - 2p_1 + \sqrt{\rho^2 + 4\rho p_1 - 4p_1}}{2(1-\rho)}.$$

Now

$$p_3 = p_2 = \frac{-\rho + 2\rho p_1 - 2p_1 + \sqrt{\rho^2 + 4\rho p_1 - 4p_1}}{2(1-\rho)},$$

and

$$p_4 = 1 - p_1 - p_2 - p_3 = \frac{-1 + \rho p_1 - p_1 + \sqrt{(\rho^2 + 4\rho p_1 - 4p_1)}}{\rho - 1},$$

[27] p_2 has two roots, but only one is positive.

completely expressing the inter-image distribution $\mathcal{P}$ in terms of p_1 and $\rho(u,v)$. These probabilities are substituted into (14), producing

$$HR(\mathcal{P},2) = \frac{H(u-v)}{H(u)} = \frac{\log_2(p_2)+\log_2(p_1+p_4)+\log_2(p_3)}{\log_2(p_1+p_3)+\log_2(p_2+p_4)} = HR(p_1,\rho,2)$$

$$= \frac{2\ln 2 - 2\ln\left(\rho - 2\rho p_1 + 2p_1 - \sqrt{(\rho^2+4\rho p_1-4p_1)}\right) + 3\ln(\rho-1) - \ln\left(2\rho p_1 - 2p_1 - 1 + \sqrt{(\rho^2+4\rho p_1-4p_1)}\right)}{2\ln 2 - \ln\left(\rho - \sqrt{(\rho^2+4\rho p_1-4p_1)}\right) + 2\ln(\rho-1) - \ln\left(\rho + \sqrt{(\rho^2+4\rho p_1-4p_1)} - 2\right)}$$

and $HR^{\sup}(2,\rho)$ from (16) is

$$HR^{\sup}(2,\rho) = \sup_{p_1} HR(p_1,\rho,2),$$

with constraint $\min(p_1,p_2,p_3,p_4) \geqslant 0$. Similarly

$$HR^{\inf}(2,\rho) = \inf_{p_1} HR(p_1,\rho,2),$$

with the same constraint.

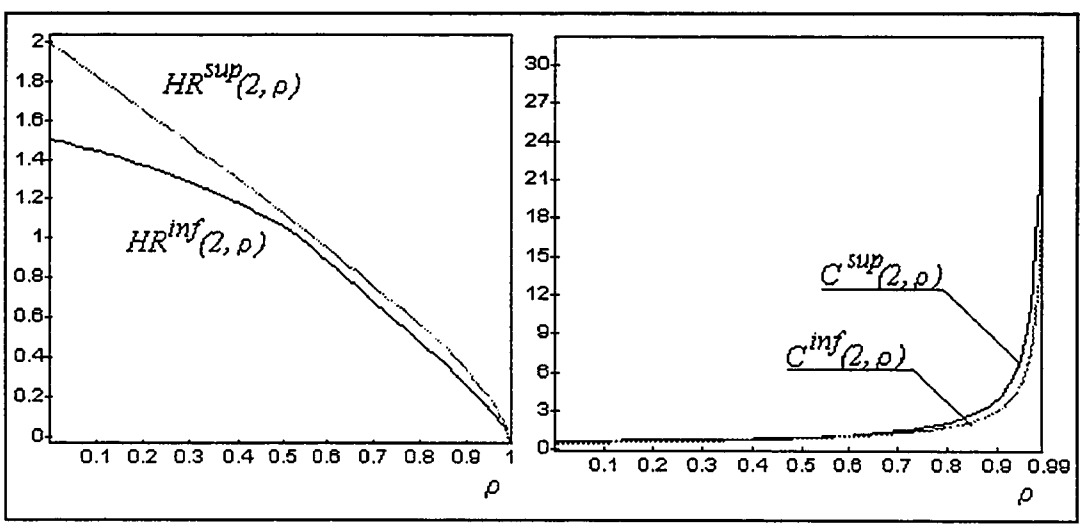

Figure 23: Functions $HR^{ext}(2,\rho)$ and $C^{ext}(2,\rho)$.

The analytical complexity of $HR(p_1,\rho,2)$ does not permit an exact solution for its extrema. Therefore the optimization was performed numerically and presented

on Figure 23 for $HR^{ext}(2,\rho) = \{HR^{\sup}(2,\rho),\ HR^{\inf}(2,\rho)\}$ as functions of inter-image correlation ρ.

Table 3: $HR^{\sup}(2,\rho)$ and $HR^{\inf}(2,\rho)$.

ρ	$HR^{\inf}$	$HR^{\sup}$	ρ	$HR^{\inf}$	$HR^{\sup}$	ρ	$HR^{\inf}$	$HR^{\sup}$
.00	1.5000	2.0000	.36	1.2243	1.3731	.70	0.6879	0.7649
.02	1.4897	1.9636	.38	1.2031	1.3384	.72	0.6467	0.7269
.04	1.4788	1.9298	.40	1.1812	1.3036	.74	0.6056	0.6884
.06	1.4674	1.8929	.42	1.1587	1.2688	.76	0.5639	0.6495
.08	1.4553	1.8578	.44	1.1354	1.2338	.78	0.5218	0.6099
.10	1.4427	1.8229	.46	1.1114	1.1988	.80	0.4792	0.5689
.12	1.4295	1.7881	.48	1.0867	1.1637	.82	0.4361	0.5264
.14	1.4158	1.7534	.50	1.0612	1.1284	.84	0.3925	0.4821
.16	1.4013	1.7187	.52	1.0350	1.0930	.86	0.3482	0.4359
.18	1.3865	1.6841	.54	1.0056	1.0574	.88	0.3032	0.3874
.20	1.3709	1.6496	.56	0.9667	1.0217	.90	0.2573	0.3363
.22	1.3547	1.6150	.58	0.9276	0.9857	.92	0.2104	0.2822
.24	1.3380	1.5805	.60	0.8883	0.9496	.94	0.1622	0.2243
.26	1.3206	1.5460	.62	0.8488	0.9132	.96	0.1123	0.1614
.28	1.3026	1.5115	.64	0.8090	0.8766	.98	0.0597	0.0907
.30	1.2840	1.4769	.66	0.7689	0.8397	1.0	0.0000	0.0000
.32	1.2648	1.4424	.68	0.7286	0.8025			
.34	1.2449	1.4078						

Values of $HR^{ext}(2,\rho) = \{HR^{\sup}(2,\rho),\ HR^{\inf}(2,\rho)\}$ are also presented in the Table 3. Since $p_2 = p_3$, images u and v have identical probability distributions (histograms) and $H(u) = H(v)$. Therefore $HR(\rho) \leqslant 2$, with equality only for independent (uncorrelated) u and v.

Applications of $HR^{ext}(n=2,\rho)$

The values of $HR^{\sup}(2,\rho)$ and $HR^{\inf}(2,\rho)$ can be used for many interesting conclusions about compressing binary, and more general n-ary, images with predictive models. First, $HR^{\sup}(n,\rho)$ is an inverse to the compression ratio; the worst (lowest) compression ratio one can achieve with $\rho-$correlated $n-$intensity level images is

$$C^{\inf}(n,\rho)=\frac{1}{HR^{\sup}(n,\rho)}.$$

From our numerical results, for example, we determined that $HR^{\sup}(n=2,\rho=0.83)=HR^{\inf}(n=2,\rho=0.79)=0.5$. Therefore, to guarantee a 50% compression ratio $C(n=2,\rho)=2$ needs to occur *for any two binary images* based on the difference method, the inter-image correlation must be at least $\rho_{0.5}^{(2)}=0.83$. To achieve this compression for *some carefully chosen* pairs of binary images, the correlation ρ between them must not be below 0.79. Since any higher number of image intensities n includes the case $n=2$, the lower bound on the minimal correlation ρ for the same compression ratio will increase with n. So, *to guarantee a 50% compression for any two images with $n\geqslant 2$ intensity levels, the correlation between them must be at least* 0.83.

Another application of $HR^{ext}(2,\rho)$ is to predicting the worst-case and best-case compression ratios for a correlated database. If one has a database of correlated images where the correlation between any two images varies from ρ_1 to ρ_2, the average worst-case compression ratio for the database can be estimated as

$$C^{\inf}(n,[\rho_1,\rho_2])=\frac{1}{\rho_2-\rho_1}\int_{\rho=\rho_1}^{\rho=\rho_2}\frac{1}{HR^{\sup}(n,\rho)}d\rho.$$

One can compute from this formula that compressing an arbitrary database of binary images with correlations in the range $[0.7, 0.9]$ will *always yield an average compression ratio at least* 1.88. Similarly, from

$$C^{\sup}(n, [\rho_1, \rho_2]) = \frac{1}{\rho_2 - \rho_1} \int_{\rho=\rho_1}^{\rho=\rho_2} \frac{1}{HR^{\inf}(n, \rho)} d\rho$$

with the best-case scenario we compute $C^{\sup}(2, [0.7, 0.9]) = 2.27$ and conclude that *it is impossible to compress a database of binary images with correlations uniformly distributed in the [0.7,0.9] range with an average ratio above 2.27.* Increasing the number of intensity levels n increases the number of degrees of freedom (independent variables) in $HR(\mathcal{P}, n)$, therefore

$$HR^{\sup}(n_1, \rho) \geqslant HR^{\sup}(n_2, \rho) \text{ if } n_1 \geqslant n_2.$$

This means that compressing a database of $n-$ary images with correlations in the range $[0.7, 0.9]$ produces an overall compression ratio less than 1.88.

Additional examples using $HR^{\sup}(2, \rho)$ will be given later.

Examples of the Worst-Case Binary Inter-Image Distributions

Of interest is which binary distributions minimize and maximize compression ratio between two binary images for the given inter-image correlation ρ. We determined these extremal probability values numerically and show some of them in the Table 4 (due to the symmetry of the problem we have in fact two extremal distributions for each value of ρ : the second one is obtained swapping p_1 and p_4) :

Table 4: Worst-case distributions for $C^{\text{inf}} = \frac{1}{HR^{\text{sup}}}$ and $C^{\text{sup}} = \frac{1}{HR^{\text{inf}}}$.

ρ	HR^{sup}	C^{inf}	p_3 / p_1	p_4 / p_2	for C^{inf}	HR^{inf}	C^{sup}	p_3 / p_1	p_4 / p_2	for C^{sup}
0.0	2	0.50	.0005	.0000		1.5	0.66	.2500	.2500	
			.999	.0005				.2500	.2500	
0.1	1.82	0.55	.0019	0.0002		1.44	0.69	.2249	.2249	
			.996	.0019				.2760	.2249	
0.2	1.65	0.61	.0053	0.0014		1.36	0.72	.2000	.2000	
			.988	.0053				.3000	.2000	
0.3	1.48	0.68	.0102	0.0046		1.28	0.77	.1749	.1749	
			.975	.0102				.3260	.1749	
0.4	1.30	0.77	.0162	0.0116		1.18	0.84	.1500	.1500	
			.956	.0162				.3500	.1500	
0.5	1.13	0.88	.0235	0.027		1.06	0.94	.1250	.1250	
			.926	.0235				.3750	.1250	
0.6	.95	1.05	.0331	0.0578		.88	1.12	.0002	.0002	
			.876	.0331				.9990	.0002	
0.7	.76	1.31	.0453	0.1404		.68	1.45	.0002	.0002	
			.769	.0453				.9991	.0002	
0.8	.57	1.75	0.050	.450		.47	2.08	.0002	.0002	
			.450	0.050				.9992	.0002	
0.9	.34	2.94	.0250	0.475		.25	3.88	.0002	.0002	
			.475	.0250				.9992	.0002	

Tabulated values for $HR^{ext}(2, \rho)$ and corresponding distributions have many practical interpretations. For example, it follows from the table that for correlated binary

images, difference compression makes sense only when $\rho \geqslant 0.6$ (i.e., compression ratio $C > 1$). Since increasing the number of intensity levels n decreases $C(n, \rho)$, the difference compression in general should never be applied to images with $\rho < 0.6$. Another less straightforward example is answering a question like "How much compression can one achieve after subtracting two binary images, given that overlapping these images results in 91% pixel value match ?". Questions like this often occur in image registration. From the above Table 4 $100\% - 91\% = 9\%$ pixel value mismatch corresponds to $\rho = 0.7$: the probability of a mismatch in this case is $2 * .0453 \approx 9\%$. This results in a $C = 1.31$ (compression ratio), always guaranteed in this case.

Asymptotic Behavior of $HR(2, \rho)$

Before doing the general case of $HR^{ext}(n, \rho)$, it was enlightening to observe the behavior of entropies involved in the definition (16) of $HR^{ext}(2, \rho)$. In the binary case they can be easily visualized with two variables: p_1 and ρ. Figures 24, 25 and 26, are graphs for $H(u)$, $H(r)$ and $HR = H(r)/H(u)$ for the binary case as functions of p_1 and ρ.

If we were interested in the question: What happens to $HR(u, v) = \frac{H(u-v)}{H(u)}$ if $H(u) \rightarrow 0$? One can observe from these plots, in the binary case, this does not lead to any infinite values; moreover, for any positively correlated binary images u and v, $HR(u, v) \leqslant 2$. This also means that subtracting one binary image from another never results in more than a 100% entropy increase. The value $HR(u, v) = 2$ corresponds to two uncorrelated ($\rho = 0$) binary images.

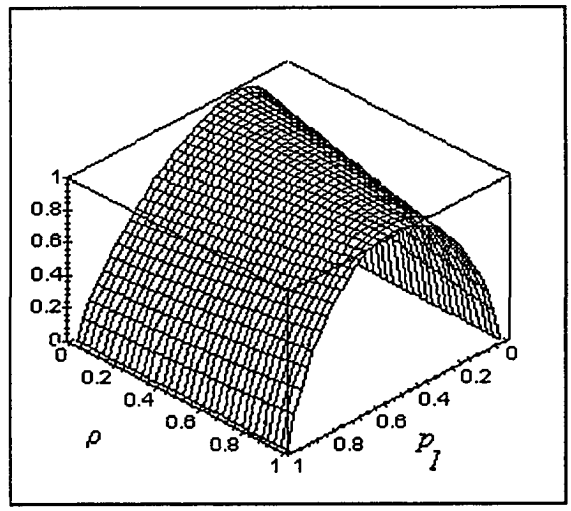

Figure 24: $H(u)$.

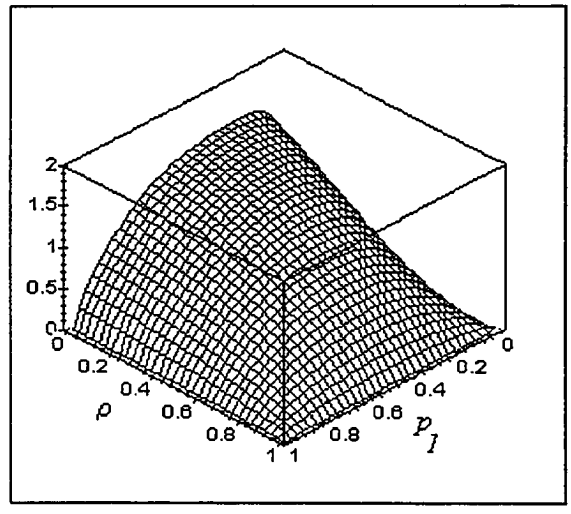

Figure 25: $H(u-v) = H(r)$.

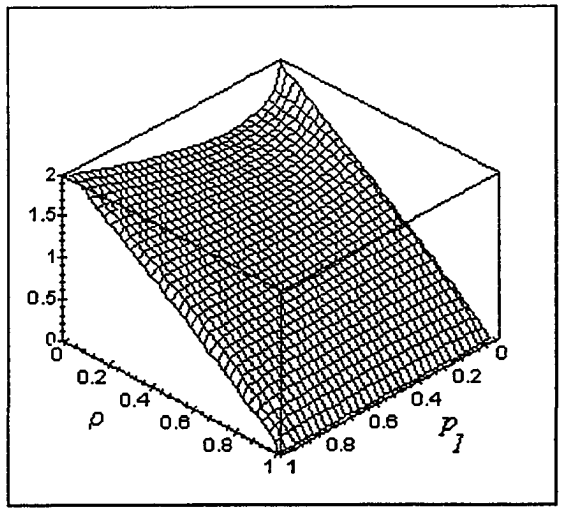

Figure 26: $HR = H(r)/H(u)$.

This conclusion is not trivial because even in this binary case increasing the number of pixels and deliberately decreasing $H(u)$ to 0 could produce an arbitrarily high $HR(u, v)$; however, this does not happen. This is the main reason that permits the introduction and study of the functions $HR^{ext}(n, \rho)$ for the binary ($n = 2$) case.

Applying Binary Model to n-ary Images

For general $n \geqslant 2$, the function (15) contains n^2 variables (bivariate probabilities P_{ij}) with only 2 constraints ($\sum P_{ij} = 1$ and $\sigma(u) = \sigma(v)$). This means the functions $HR^{ext}(n, \rho)$ in (16) and (17) must be determined as extremal values over the space of $n^2 - 2$ independent variables. For an image with $n = 256$ intensity levels, this causes an extrema search over $256^2 - 2 = 65534$ independent variables, challenging any current computer.

One way to approach this problem is to extend our binary study to certain classes of n-ary images. Many images can be considered approximately binary, if they have

at most two distinct intensity clusters. Typically these are images with a sharp distinction between background and foreground intensities, this is true for most medical images containing light foreground detail on dark (black) film background. This is illustrated with the histograms for CT and MR images shown on Figure 27.

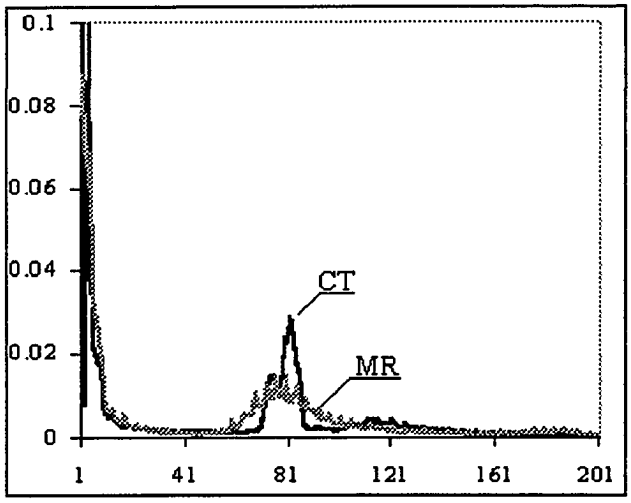

Figure 27: CT and MR histograms.

One can observe on Figure 27 that both CT and MR image intensity levels are clustered into either background (intensity values below 10-20) or foreground (intensity values from 60 to 100) regions, which are also spatially distinct on the images with foreground intensities concentrated in the central part of an image. This distinction allows the use of many binary model results and, in particular, the worst-case compression ratio estimates for the n-ary images with two distinct intensity levels.

We will not pursue the binary approach any further; instead, we will develop some general $HR^{ext}(n,\rho)$ theory and less complicated extrema models for computing $HR^{ext}(n,\rho)$ values.

Consider the general case of any integer $n \geqslant 2$. The question we would originally like to answer is whether it is possible for a given compression ratio C, to find a correlation threshold $\rho = \rho(C)$, such that for *any* two images u and v with discrete intensities from $\{1, 2, ..., n\}$ their correlation $\rho(u, v) > \rho(C)$ implies $H(u-v)/H(u) < 1/C$. That is, if subtracting these images produces a compression ratio greater than C. In the previous section, this question was answered positively for $n = 2$ when we found numerically two functions $HR^{\sup}(2,\rho)$ and $HR^{\inf}(2,\rho)$, yielding the upper and lower bounds on $1/C$, and used these to make predictions about the efficiency of difference compression. In this subsection, we will prove that this same result is possible for any number of image intensities n, and determine some functional estimates for $HR^{\sup}(n,\rho)$ and $HR^{\inf}(n,\rho)$.

Lemma 7. For an image u with integer intensities from $\{1, 2, ..., n\}$, $n \geqslant 2$, distributed with probabilities $\{p_1, p_2, .., p_n\}$ and variance $\sigma = \sigma(u) < \sqrt{\frac{2}{n}}(n-1)$, there is at least one intensity k, $1 \leqslant k \leqslant n$, such that

$$\frac{\sigma^2}{2(n-1)^2} \leqslant p_k \leqslant 1 - \frac{\sigma^2}{2(n-1)^2}.$$

Proof.

The lemma will be proven by contradiction: assume that for $\delta = \frac{\sigma^2}{(n-1)^2}$, and for any intensity i either $p_i < \delta$ or $p_i > 1 - \delta$.

First, $\delta = \frac{\sigma^2}{2(n-1)^2} < (n-1)^2 \frac{2}{n} \frac{1}{2(n-1)^2} = \frac{1}{n}$, i.e., $\delta < \frac{1}{n} \leqslant \frac{1}{2}$.

Since $\sum_{i=1}^{n} p_i = 1$, and $\delta < \frac{1}{2}$, there is at most one intensity level k such that $p_k > 1 - \delta$ (otherwise we have two probabilities p_k above $\frac{1}{2} + \frac{1}{2} = 1$). Consider first

the case when such k exists. The average intensity is

$$\overline{u} = \sum_{i=1}^{n} ip_i = \sum_{i=1, i\neq k}^{n} ip_i + kp_k,$$

and

1. $\overline{u} - k = \sum\limits_{i=1, i\neq k}^{n} ip_i + k(p_k - 1) < n \sum\limits_{i=1, i\neq k}^{n} p_i + k(p_k - 1)$

$= n(1 - p_k) + k(p_k - 1) = (n - k)(1 - p_k) < (n - 1)(1 - p_k),$

2. $k - \overline{u} = k(1 - p_k) - \sum\limits_{i=1, i\neq k}^{n} ip_i < k(1 - p_k) - \sum\limits_{i=1, i\neq k}^{n} p_i$

$= k(1 - p_k) - (1 - p_k) = (k - 1)(1 - p_k) < (n - 1)(1 - p_k).$

Thus $(k - \overline{u})^2 < (n - 1)^2(1 - p_k)^2$. Then the variance

$\sigma^2 = \sum\limits_{i=1}^{n}(i - \overline{u})^2 p_i = \sum\limits_{i=1, i\neq k}^{n} (i - \overline{u})^2 p_i + (k - \overline{u})^2 p_k$

$< (n - 1)^2 \sum\limits_{i=1, i\neq k}^{n} p_i + (n - 1)^2(1 - p_k)^2 p_k$

$= (n - 1)^2(1 - p_k) + (n - 1)^2(1 - p_k)^2 p_k$

$= (n - 1)^2(1 - p_k)(1 + (1 - p_k)p_k) < (n - 1)^2(1 - p_k)(1 + 1)$

$< 2(n - 1)^2\delta = \sigma^2$, or $\sigma^2 < \sigma^2$, which is a contradiction.

If k does not exist, then all $p_i < \delta$, and

$\sum\limits_{i=1}^{n} p_i < \delta n < 1$, which is a contradiction as well.

■

Lemma 8. For an image u with integer intensities from $\{1, 2, ..., n\}$ and variance $\sigma = \sigma(u) < \sqrt{\frac{2}{n}}(n - 1)$ the entropy is

$$H(u) \geqslant \frac{1}{2}\left(\frac{\sigma^2}{2(n-1)^2}\log_2\frac{2(n-1)^2}{\sigma^2} + (1 - \frac{\sigma^2}{2(n-1)^2})\log_2\frac{2(n-1)^2}{2(n-1)^2 - \sigma^2}\right).$$

Proof.

This follows from the previous lemma: consider k such that $\frac{\sigma^2}{2(n-1)^2} \leqslant p_k \leqslant 1 - \frac{\sigma^2}{2(n-1)^2}$. Then

$$H(u) = \sum_{i=1}^{n} p_i \log_2(\tfrac{1}{p_i}) \geqslant p_k \log_2(\tfrac{1}{p_k})$$

$$\geqslant \max\left(\tfrac{\sigma^2}{2(n-1)^2} \log_2 \tfrac{2(n-1)^2}{\sigma^2}, (1 - \tfrac{\sigma^2}{2(n-1)^2}) \log_2 \tfrac{2(n-1)^2}{2(n-1)^2-\sigma^2}\right)$$

$$\geqslant \tfrac{1}{2}\left(\tfrac{\sigma^2}{2(n-1)^2} \log_2 \tfrac{2(n-1)^2}{\sigma^2} + (1 - \tfrac{\sigma^2}{2(n-1)^2}) \log_2 \tfrac{2(n-1)^2}{2(n-1)^2-\sigma^2}\right).$$

■

The implication of this result is, besides the image entropy $H(u)$, the image variance $\sigma(u)$ is also a measure of image randomness. Therefore if $\sigma(u) > 0$, then $H(u)$ cannot be arbitrarily low. We will now prove a similar result for the converse, when $\sigma(u)$ becomes small, it corresponds to a decrease of image randomness, and $H(u)$ cannot be arbitrarily high. Since for any image with n intensities $H(u) \leqslant \log_2 n$, we will improve this upper bound and prove that it must vanish as $\sigma(u) \to 0$.

Lemma 9. For an image u with integer intensities from $\{1, 2, ..., n\}$ and variance $\sigma = \sigma(u) < \frac{1}{4}$ entropy

$$H(u) \leqslant \sum_{i=1}^{n} \frac{4\sigma^2}{i^2} \log_2 \frac{i^2}{4\sigma^2} + (1 - 8\sigma^2) \log_2 \frac{1}{1 - 8\sigma^2}.$$

Proof.

Let's choose an integer k as the closest to the average intensity $\overline{u}$: $k = \left\lfloor \overline{u} + \frac{1}{2} \right\rfloor$. Then for any other $i \neq k$ we have $|i - \overline{u}| \geqslant \frac{1}{2}$, and

$$\sigma^2 = \sum_{i=1}^{n} (i - \overline{u})^2 p_i \geqslant \sum_{i=1, i\neq k}^{n} (i - \overline{u})^2 p_i. \tag{20}$$

Consider the numbers $s_i = |i - \overline{u}|$. At most one is less than $\frac{1}{2}$ (this corresponds to $i = k$), therefore at least one is greater than $\frac{1}{2}$ (for $i = k + 1$), and at least one is greater than 1 (for $i = k - 1$), etc. Therefore, we can renumber our probabilities p_i as p_{i_m} such that $|i_m - \overline{u}| \geqslant \frac{m}{2}$, $m = 1, .., n$. Inequality (20) implies $\sigma^2 \geqslant (i - \overline{u})^2 p_i$, therefore

$$p_{i_m} \leqslant \frac{\sigma^2}{(i_m - \overline{u})^2} \leqslant \frac{4\sigma^2}{m^2},$$

where index i_m includes all values from 1 to n except k, as does m. The largest of these numbers is $\frac{4\sigma^2}{1^2} = 4\sigma^2 < 4 * \frac{1}{16} < \frac{1}{2}$, therefore all are included in the region where the function $x * \log_2 \frac{1}{x}$ increases and

$$\sum_{i=1, i \neq k}^{n} p_i \log_2 \frac{1}{p_i} \leqslant \sum_{i=1, i \neq k}^{n} \frac{4\sigma^2}{i^2} \log_2 \frac{i^2}{4\sigma^2}.$$

Then $p_k = 1 - \sum\limits_{i=1, i \neq k}^{n} p_i \geqslant 1 - \sum\limits_{i=1, i \neq k}^{n} \frac{4\sigma^2}{i^2} > 1 - \sum\limits_{i=1}^{n} \frac{4\sigma^2}{i^2} = 1 - 4\sigma^2(1 + \sum\limits_{i=2}^{n} \frac{1}{i^2})$. We can estimate

$$\sum_{i=2}^{n} \frac{1}{i^2} < \sum_{i=2}^{n} \frac{1}{i(i-1)} = \sum_{i=2}^{n} \left(\frac{1}{i-1} - \frac{1}{i} \right) = 1 - \frac{1}{n}.$$

Hence $p_k > 1 - 4\sigma^2(1 + 1 - \frac{1}{n}) > 1 - 8\sigma^2 > \frac{1}{2}$ (because $\sigma < \frac{1}{4}$), and in this region $x \log_2 \frac{1}{x}$ decreases, leading to

$$p_k \log_2 \frac{1}{p_k} < (1 - 8\sigma^2) \log_2 \frac{1}{1 - 8\sigma^2}.$$

The entropy $H(u) = \sum\limits_{i=1}^{n} p_i \log_2 \frac{1}{p_i} \leqslant \sum\limits_{i=1}^{n} \frac{4\sigma^2}{i^2} \log_2 \frac{i^2}{4\sigma^2} + (1 - 8\sigma^2) \log_2 \frac{1}{1-8\sigma^2}$.

■

We also need the proof of the following:

Lemma 10. For any constants $a_i, b_i, c_i, d_i > 0$

$$L = \lim_{x \to 0} \frac{\sum\limits_i a_i x \log_2 \frac{1}{a_i x} + \sum\limits_k (1 - c_k x) \log_2 \frac{1}{(1 - c_k x)}}{\sum\limits_j b_j x \log_2 \frac{1}{b_j x} + \sum\limits_m (1 - d_m x) \log_2 \frac{1}{(1 - d_m x)}} = \frac{\sum\limits_i a_i}{\sum\limits_j b_j}$$

Proof.

First, for $x \to 0$

$$(1 - cx) \log_2 \frac{1}{(1-cx)} = (1 - cx)(cx + O(x^2))$$

Then

$$\sum_i a_i x \log_2 \frac{1}{a_i x} + \sum_k ((1 - c_k x) \log_2 \frac{1}{(1 - c_k x)}$$
$$= -\sum_i a_i x (\log_2 a_i + \log_2 x) + \sum_k ((1 - c_k x)(c_k x + O(x^2))$$
$$= -x \sum_i a_i \log_2 a_i - x \log_2 x \sum_i a_i + x \sum_k ((1 - c_k x) c_k + O(x^2) \sum ((1 - c_k x),$$

and

$$L = \lim_{x \to 0} \frac{-x \sum\limits_i a_i \log_2 a_i - x \log_2 x \sum\limits_i a_i + x \sum\limits_k ((1 - c_k x) c_k + O(x^2) \sum\limits_k ((1 - c_k x)}{-x \sum\limits_j b_j \log_2 b_j - x \log_2 x \sum\limits_j b_j + x \sum\limits_m (1 - d_m x) d_m + O(x^2) \sum\limits_m (1 - d_m x)}$$

$$= \lim_{x \to 0} \frac{-\frac{1}{\log_2 x} \sum\limits_i a_i \log_2 a_i - \sum\limits_i a_i + \frac{1}{\log_2 x} \sum\limits_k ((1 - c_k x) c_k + \frac{1}{\log_2 x} O(x) \sum\limits_k ((1 - c_k x)}{- - \frac{1}{\log_2 x} \sum b_j \log_2 b_j - \sum\limits_j b_j + \frac{1}{\log_2 x} \sum\limits_m (1 - d_m x) d_m + \frac{1}{\log_2 x} O(x) \sum\limits_m (1 - d_m x)}$$

$$= \frac{\sum\limits_i a_i}{\sum\limits_j b_j}.$$

■

These lemmas lead to the following statement:

Theorem 1. For any integer $n \geqslant 2$ the function $HR(\mathcal{P}, n)$ defined in (15) has a finite upper bound over all $n \times n$ distributions $\mathcal{P}$.

Proof.

By definition, $HR(\mathcal{P},n)=\frac{H(\mathcal{P}^r,n)}{H(\mathcal{P}^u,n)}$. Given n, subdivide all $\mathcal{P}$ into 2 regions:

Region A: $\sigma_u=\sigma(\mathcal{P}^u)\geqslant\sigma_0(n)=\min\left(\frac{1}{8},\sqrt{\frac{2}{n}}(n-1)\right)$, and complementary

Region B: $\sigma_u=\sigma(\mathcal{P}^u)<\sigma_0(n)$.

Region A is n^2-2 dimensional compact with respect to bivariate probabilities $\mathcal{P}=\{P_{ij}\}$. On this region $H(\mathcal{P}^u,n)$ is a continuous function such that always $H(\mathcal{P}^u,n)>0$. Therefore there is a positive constant $H_0>0$ such that over A, $H(\mathcal{P}^u,n)\geqslant H_0$. In this case $HR(\mathcal{P},n)=\frac{H(\mathcal{P}^r,n)}{H(\mathcal{P}^u,n)}$ is also a continuous function on A, therefore it is limited and there exist two positive constants $0<HR_0<HR_1$ such that $HR_0<HR(\mathcal{P},n)<HR_1$.

On region B, based on the previous lemmas:

$$HR(\mathcal{P},n)\leqslant\frac{\sum_{i=1}^{n}\frac{4\sigma_r^2}{i^2}\log_2\frac{i^2}{4\sigma_r^2}+(1-8\sigma_r^2)\log_2\frac{1}{1-8\sigma_r^2}}{\frac{1}{2}\left(\frac{\sigma_u^2}{2(n-1)^2}\log_2\frac{2(n-1)^2}{\sigma_u^2}+(1-\frac{\sigma_u^2}{2(n-1)^2})\log_2\frac{2(n-1)^2}{2(n-1)^2-\sigma_u^2}\right)}=M(\mathcal{P},n).\quad(21)$$

The right-hand side estimate $M(\mathcal{P},n)$ is continuous, positive and finite for any $\sigma_u=\sigma(\mathcal{P}^u)>0$. For the region B boundary, $\sigma_u^2\to 0$, substitute the expression for the residual variance $\sigma_r^2=2\sigma_u^2(1-\rho)$ into $M(\mathcal{P},n)$ and consider $\lim_{\sigma_u^2\to 0}M(\mathcal{P},n)$. From the last lemma,

$$\lim_{\sigma_u^2\to 0}M(\mathcal{P},n)=\frac{\sum_{i=1}^{n}\frac{4*2(1-\rho)}{i^2}}{\frac{1}{2}\frac{1}{2(n-1)^2}}=32(1-\rho)(n-1)^2\sum_{i=1}^{n}\frac{1}{i^2}<64(1-\rho)(n-1)^2.$$

Thus $HR(\mathcal{P},n)$ is limited over the region B as well.

■

The theorem states that for any number of intensity levels n there is a lower bound on compression ratio achieved by subtracting from u any image $\rho-$correlated to u. This conclusion is far from trivial because as $H(u) \to 0$, one might expect that $\frac{H(u-v)}{H(u)}$ can become anything. However, the theorem proves that the presence of correlation between u and v prevents this ratio from diverging.

Corollary 6. $\lim_{\rho \to 1} HR(\mathcal{P}, n) = 0$

Proof. The proof follows directly from $0 \leqslant HR(\mathcal{P}, n) < 64(1-\rho)(n-1)^2$ ■

This result demonstrates that the ratio of residual entropy to the original image entropy $HR(\mathcal{P}, n) = \frac{H(u-v)}{H(u)}$ will uniformly vanish for correlation $\rho(u, v) \to 1$ no matter how the intensities in u or v are distributed or how small $H(u)$ is.

Corollary 7. $HR^{\sup}(n, \rho)$ and $HR^{\inf}(n, \rho)$ are continuous on $\rho \geqslant 0$ with

1. $\lim_{\rho \to 1} HR^{\sup}(n, \rho) = 0$
2. $\lim_{\rho \to 1} HR^{\inf}(n, \rho) = \lim_{\rho \to 1} HR^{\sup}(n, \rho)$

Proof.

Continuity follows from the previous theorem. By definition (16)

$HR^{\sup}(n, \rho) = \sup_{\mathcal{P}:\, \sigma(\mathcal{P}^u)=\sigma(\mathcal{P}^v),\ \rho(u,v)=\rho} HR(\mathcal{P}, n)$

$\leqslant \sup_{\mathcal{P}:\, \sigma(\mathcal{P}^u)=\sigma(\mathcal{P}^v),\ \rho(u,v)=\rho} 64(1-\rho)(n-1)^2 \to 0$ as $\rho \to 1$. Since $HR^{\sup}(n, \rho) \geqslant 0$, it proves 1.

Since by definition $0 \leqslant HR^{\inf}(n, \rho) \leqslant HR^{\sup}(n, \rho)$, and $\lim_{\rho \to 1} HR^{\sup}(n, \rho) = 0$, then $\lim_{\rho \to 1} HR^{\inf}(n, \rho) = 0 = \lim_{\rho \to 1} HR^{\sup}(n, \rho)$.

■

MONOTONICITY OF $HR^{\sup}(n,\rho)$

We have proved that $\lim_{\rho \to 1} HR^{\sup}(n,\rho) = 0$. This result demonstrates that increasing correlation between two integer sources will inevitably result in an arbitrarily high worst-case compression ratio $C^{\inf}(n,\rho) = \frac{1}{HR^{\sup}(n,\rho)}$. However, this behavior is asymptotic and occurs only when $\rho \to 1$, which may be hard to achieve in practice. For the practical applications, one would rather know whether increasing ρ will always result in increased worst-case compression ratio $C^{\inf}(n,\rho)$ or not, no matter what integer sources are being compressed. This is the question of proving that for any given n the function $HR^{\sup}(n,\rho) = \frac{1}{C^{\inf}(n,\rho)}$ is a non-increasing function of ρ. This hypothesis has already been supported with our numerical results (Figure 23, Tables 3 and 4), and will be proven in this section.

We will prove the stronger result of the monotonicity of

$$HR_u^{\sup}(n,\rho) = \sup_{v:\, \sigma(v)=\sigma(u),\ \rho(u,v)=\rho} \frac{H(u-v)}{H(u)}. \tag{22}$$

Since

$$HR^{\sup}(n,\rho) = \sup_u HR_u^{\sup}(n,\rho),$$

given a non-increasing $HR_u^{\sup}(n,\rho)$, $HR^{\sup}(n,\rho)$ must be non-decreasing as well.

Since n is considered to be a constant, we will use notations $HR_u^{\sup}(n,\rho)$ and $HR_u^{\sup}(\rho)$ as equivalent.

Lemma 11. Difference model residual $r = u - v$ has uniform probability distribution $\mathcal{P}^r = \{P(r = r_0) = R\}$ if and only if correlation $\rho(u,v) \leqslant 0$.

Proof.

We are still using the assumption $\sigma(u) = \sigma(v) = \sigma$, and in this case the covariance

$$cov(u,v) = \rho(u,v)\sigma^2 = \sum_{-n \leqslant i,j \leqslant n} ijP_{ij},$$

where P_{ij} is the bivariate probability $P(u = i \text{ and } v = j)$. Without loss of generality we assume that images u and v have intensities in the range $[-n, n]$, which can be always achieved with an intensity range translation. Since the largest value of ij for $-n \leqslant i, j \leqslant n$ is achieved on the boundary $\max(|i|, |j|) = n$ and is equal to $n(n - |i - j|) = n(n - |r_0|)$, then

$$\begin{aligned}
&\rho(u,v)\sigma^2 = \sum_{-n \leqslant i,j \leqslant n} ijP_{ij} = \sum_{-2n \leqslant r_0 = i-j \leqslant 2n} \left(\sum_{i-j=r_0,\ -n \leqslant i,j \leqslant n} ijP_{ij} \right) \\
&\leqslant \sum_{-2n \leqslant r_0 = i-j \leqslant 2n} \left(\sum_{i-j=r_0,\ -n \leqslant i,j \leqslant n} n(n - |r_0|)P_{ij} \right) \\
&= n \sum_{-2n \leqslant r_0 = i-j \leqslant 2n} \left((n - |r_0|) \sum_{i-j=r_0,\ -n \leqslant i,j \leqslant n} P_{ij} \right) \\
&= n \sum_{-2n \leqslant r_0 \leqslant 2n} ((n - |r_0|)P(r = r_0)) \\
&= n \sum_{-2n \leqslant r_0 \leqslant 2n} ((n - |r_0|)R) = nR \sum_{-2n \leqslant r_0 \leqslant 2n} (n - |r_0|) \\
&= 2nR \left(2 \sum_{0 \leqslant r_0 \leqslant 2n} (n - |r_0|) - n \right) = 2nR \left(2 \sum_{-n \leqslant k = n - r_0 \leqslant n} k - n \right) = -2n^2R < 0,
\end{aligned}$$

which means $\rho(u,v) < 0$. ■

Theorem 2. For any integer image u function $HR_u^{\sup}(\rho)$ is a decreasing function of $\rho > 0$.

Proof.

Let $\mathcal{P} = \{P(u = i \text{ and } v = j) = P_{ij}\}$ be the bivariate probability distribution for the pairs of intensities (u, v). The proof will be to find some differential transform $\Omega(\alpha, \mathcal{P}) : \mathcal{P} \to \mathcal{P}_\alpha$ with parameter α such that for any α, $0 < \alpha < \alpha_0$,

1.Correlation $\rho(\alpha) = \rho(\mathcal{P}_\alpha)$ decreases with α,

2.Entropy $H(\alpha) = H(\mathcal{P}_\alpha)$ increases with α,

3.$\Omega(0, \mathcal{P}) = \mathcal{P}$.

The existence of $\Omega(\alpha, \mathcal{P})$ will demonstrate that at least for some specific subset of inter-image bivariate probabilities $\mathcal{P}_\alpha$ (parameterized with α) decreasing correlation to the given u leads to increasing entropy $H(u - v)$ (and consequently $\frac{H(u-v)}{H(u)}$). Since $HR_u^{\sup}(\rho(\alpha))$ by definition (22) must be at least as large as $\frac{H(\alpha)}{H(u)}$, and $H(\alpha) \geqslant H(0)$, then $HR_u^{\sup}(\rho(\alpha)) \geqslant HR_u^{\sup}(\rho(0))$ for $\rho(\alpha) < \rho(0)$, which means decreasing monotonicity of $HR_u^{\sup}(\rho)$.

The transform $\Omega(\alpha, v)$ can be constructed in the following way:

$$\Omega(a, b, c, d, \alpha, \mathcal{P}) = \begin{cases} P_{ac} \rightarrow -\alpha + P_{ac} \\ P_{bd} \rightarrow -\alpha + P_{bd} \\ P_{ad} \rightarrow \alpha + P_{ad} \\ P_{bc} \rightarrow \alpha + P_{bc} \\ P_{ij} \rightarrow P_{ij} \qquad otherwise \end{cases} . \tag{23}$$

This transform (23) modifies only four probabilities in $\mathcal{P}$, redistributing their values. If $\mathcal{P}_\alpha = \Omega(a, b, c, d, \alpha, \mathcal{P})$, the total probability (equal to 1), images u and v, and therefore entropies $H(u)$, $H(v)$ and variances $\sigma(u), \sigma(v)$ will remain unchanged after applying $\Omega(a, b, c, d, \alpha, \cdot)$ to $\mathcal{P}$. However $\Omega(a, b, c, d, \alpha, \cdot)$ will affect $\rho(u, v)$ and $H(u - v)$ as follows:

1.$\frac{d\rho}{d\alpha} = \frac{d}{d\alpha}\left(\frac{1}{\sigma^2} \sum_{-n \leqslant i,j \leqslant n} ijP_{ij}\right) = -ac - bd + ad + bc = (a - b)(d - c)$, which is negative if $b > a$ and $d > c$,

2. $\frac{dH(u-v)}{d\alpha} = \frac{d}{d\alpha} \sum_{-2n \leqslant r_0 = i-j \leqslant 2n} \Omega(a,b,c,d,\alpha,\mathcal{P})$

$= \frac{d}{d\alpha}(-(R_{ac}-\alpha)\log_2(R_{ac}-\alpha) - (R_{bd}-\alpha)\log_2(R_{bd}-\alpha)$

$-(R_{ad}+\alpha)\log_2(R_{ad}+\alpha) - (R_{bc}+\alpha)\log_2(R_{bc}+\alpha)),$

where R_{xy} stands for residual probability $\mathcal{P}^r(x-y) = \sum_{-n \leqslant i,j \leqslant n,\ i-j=x-y} P_{ij}$. After differentiating with respect to α

$\frac{dH(u-v)}{d\alpha} = \log_2(R_{ac}-\alpha) + \log_2(R_{bd}-\alpha) - \log_2(R_{ad}+\alpha) - \log_2(R_{bc}+\alpha) =$

$= \log_2 \frac{(R_{ac}-\alpha)(R_{bd}-\alpha)}{(R_{ad}+\alpha)(R_{bc}+\alpha)}.$

This expression becomes positive when

$$\frac{(R_{ac}-\alpha)(R_{bd}-\alpha)}{(R_{ad}+\alpha)(R_{bc}+\alpha)} > 1, \tag{24}$$

or $(R_{ac}-\alpha)(R_{bd}-\alpha) > (R_{ad}+\alpha)(R_{bc}+\alpha)$. Consider only $\alpha > 0$, then all residual probabilities R_{xy} are non-negative and inequality (24) due to its continuity will be true for some α, $0 < \alpha < \alpha_0$, if

$$\frac{R_{ac}R_{bd}}{R_{ad}R_{bc}} > 1. \tag{25}$$

However, one can always satisfy (25) choosing R_{ac} and R_{bd} as the largest residual probabilities (from the previous lemma, all residual probabilities cannot be equal when $\rho > 0$, therefore the strictly largest probabilities exist). With this choice we have a small positive parameter α such that:

1. $\frac{d\rho}{d\alpha} < 0$
2. $\frac{dH(u-v)}{d\alpha} > 0$

which means with respect to this parameter α (along a one-dimensional tangent spanned by α) $\frac{dH(u-v)}{d\rho} = \frac{dHR(\rho)}{d\rho} < 0$, i.e., $HR_u(\rho(\alpha))$ is a decreasing function. ■

The theorem proves that an increase in inter-image correlation will always result in a decrease in the worst-case entropy of the difference between the images. In other words, the worst-case compression ratio produced by the difference compression will decrease with the increasing inter-image correlation.

Discrete Max-Entropy Distributions

In this section we will derive the discrete $(2n-1)$-state distribution which, for given variance σ, has maximal entropy. We assume that distribution states correspond to $(2n-1)$ integer numbers from $-n+1$ to $n-1$, each with probability p_i, $i=-n+1,..,n-1$. Note that neither entropy nor variance depend on the distribution mean $\sum_{i=-n+1}^{n-1} ip_i$, therefore we can also assume without loss of generality that $\sum_{i=-n+1}^{n-1} ip_i = 0$ (otherwise we can shift all state values by $\sum_{i=-n+1}^{n-1} ip_i$).

In this maximize the entropy

$$H = -\sum_{i=-n+1}^{n-1} p_i \log_2 p_i,$$

subject to constraints

$$\sum_{i=-n+1}^{n-1} p_i = 1 \text{ and } \sum_{i=-n+1}^{n-1} i^2 p_i = \sigma^2.$$

To do so, form the Lagrangian:

$$\Phi(p_i,\alpha,\beta) = -\sum_{i=-n+1}^{n-1} p_i \log_2 p_i + \alpha\left(\sum_{i=-n+1}^{n-1} i^2 p_i - \sigma^2\right) + \beta\left(\sum_{i=-n+1}^{n-1} p_i - 1\right),$$

and solve for

$$\frac{\partial}{\partial p_i}\Phi(p_i,\alpha,\beta) = -\log_2 p_i - \log_2 e + \alpha i^2 + \beta = 0.$$

This yields $p_i = \frac{1}{e}2^{\alpha i^2+\beta}$, where the constants α and β must be determined to satisfy the constraints

$$\sum_{i=-n+1}^{n-1} 2^{\alpha i^2+\beta} = e \text{ and } \sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2+\beta} = e\sigma^2.$$

From the first constraint $\beta = \log_2 e - \log_2 \sum_{i=-n+1}^{n-1} 2^{\alpha i^2}$, which after substitution into the second constraint results in:

$$\sigma^2 = \frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}}. \tag{26}$$

Then the entropy of the image $H = -\sum_{i=-n+1}^{n-1} p_i \log_2 p_i = -\sum_{i=-n+1}^{n-1} p_i(-\log_2 e + \alpha i^2 + \beta) =$

$$= \log_2 e \sum_{i=-n+1}^{n-1} p_i - \alpha \sum_{i=-n+1}^{n-1} i^2 p_i - \beta \sum_{i=-n+1}^{n-1} p_i = \log_2 e - \alpha\sigma^2 - \beta$$

$$= \log_2 e - \alpha \frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}} - (\log_2 e - \log_2 \sum_{i=-n+1}^{n-1} 2^{\alpha i^2}),$$

or

$$H = \log_2 \sum_{i=-n+1}^{n-1} 2^{\alpha i^2} - \alpha \frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}}. \tag{27}$$

This is the maximum entropy an image with $(2n-1)$ intensity levels and known variance σ can have, where the parameter α is found from (26). Figure 28 shows the dependency $\sigma^2(\alpha)$ for values of $n = 2$, 3 and 4 (one can prove $0 \leqslant \sigma^2(\alpha) \leqslant (n-1)^2$). Since equation (26) cannot be resolved for α analytically, but still provides a one-to-one continuous and monotone correspondence between σ and α, we will use α rather than σ as a variance characteristic of this distribution.

Modeling $HR^{\sup}(n,\rho)$ Behavior with Extremal Entropy Distributions

As mentioned, computing the exact $HR^{ext}(n,\rho)$ values involves optimization over $n^2 - 2$ independent variables, which is prohibitively complicated even for small n.

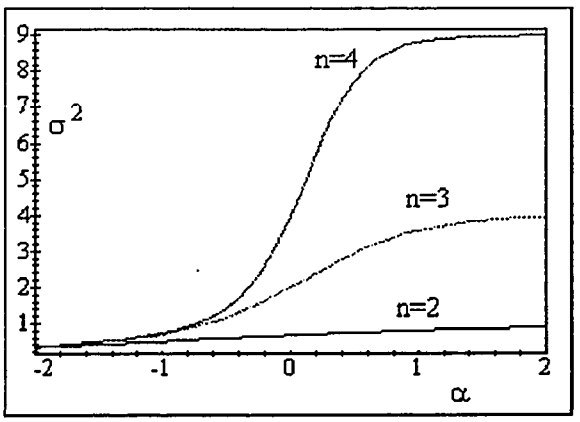

Figure 28: $\sigma^2(\alpha)$.

One way to avoid this difficulty is to develop some approximate estimates for these functions instead of computing the exact values. In this section, we will construct an accurate upper bound on $HR^{\sup}(n,\rho)$, which is also exact for $n = 2$.

Consider any n-ary image u, $n \geqslant 2$. Its entropy $H(u)$ is assumed to always be positive, i.e., the image has at least two distinct intensities m and j, with probabilities p_m and p_j respectively. Since entropy grows with an increasing number of different states, $H(u)$ will be minimized if we assume that u has *exactly* two intensities. Then, if the probability of intensity level m is $p = p_m$, $p_j = 1-p$, and $H(u) = -p\log_2 p-(1-p)\log_2(1-p)$. The variance of this binary image is $\sigma^2(u) = (m-j)^2p(1-p)$; therefore the variance of the residual $r = u - v$ is $\sigma^2(r) = 2\sigma^2(u)(1-\rho) = 2(m-j)^2p(1-p)(1-\rho)$, provided that correlation $\rho(u,v) = \rho$. Given this known residual variance $\sigma(r)$, and known number of residual intensities $2n-1$, the residual entropy $H(r)$ will be maximized by the "discrete normal" distribution found in the previous section and in this case $H(r) = \log_2 \sum_{i=-n+1}^{n-1} 2^{\alpha i^2} - \alpha \frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}}$, where α can be uniquely determined for each $\sigma(r)$ from (26). Finally, $H(r)$ will increase if $\sigma(r)$ increases,

which occurs with m and j, $1 \leqslant m, j \leqslant n$, chosen as far apart as possible, i.e., $|m-j| = n-1$. In this case for the image u, variance $\sigma^2(u) = \sigma^2 = (n-1)^2 p(1-p)$, i.e., $p = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\frac{\sigma^2}{(n-1)^2}}\right)$, and

$$\begin{aligned} H(u) &= -\frac{\left(1-\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}\log_2\frac{\left(1-\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2} \\ &\quad -\frac{\left(1+\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}\log_2\frac{\left(1+\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}, \end{aligned}$$

where $\sigma^2 = \sigma^2(u) = \frac{1}{2(1-\rho)}\sigma^2(r) = \frac{1}{2(1-\rho)}\frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}}$. Thus, using this argument, we minimize $H(u)$ and maximize $H(r)$ at the same time with the proper choice of intensity distributions, which leads to:

$$\begin{aligned} HR(u,v) &= \frac{H(r)}{H(u)} \\ &\leqslant \frac{\log_2 \sum_{i=-n+1}^{n-1} 2^{\alpha i^2} - \alpha\frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}}}{-\frac{\left(1-\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}\log_2\frac{\left(1-\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2} - \frac{\left(1+\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}\log_2\frac{\left(1+\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}} \\ &= M(n,\rho,\alpha), \end{aligned} \tag{28}$$

with

$$\sigma^2 = \frac{1}{2(1-\rho)}\frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}}. \tag{29}$$

Then

$$HR^{\inf}(n,\rho) \leqslant HR^{\sup}(n,\rho) \leqslant \sup_{\alpha} M(n,\rho,\alpha). \tag{30}$$

$M(n, \rho, \alpha)$ must be only maximized with respect to one independent variable α. This maximization can be carried out with great precision if one observes the behavior of $M(n, \rho, \alpha)$, shown on Figures 29 and 30 for two respective values of $n = 2$ and $n = 200$. For each n, the plot on the right represents a magnified region of the left plot for $0.9 \leqslant \rho \leqslant 1$.

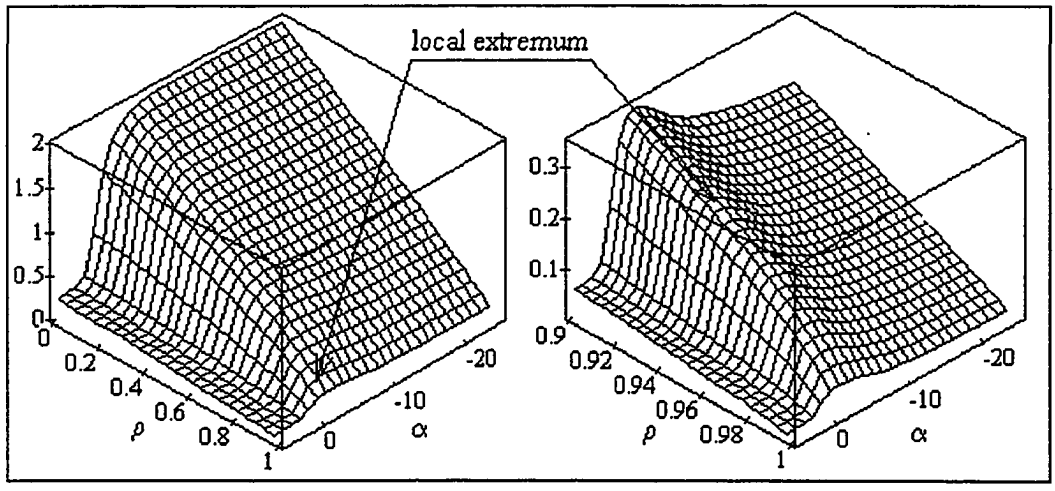

Figure 29: $M(n, \rho, \alpha)$ for $n = 2$.

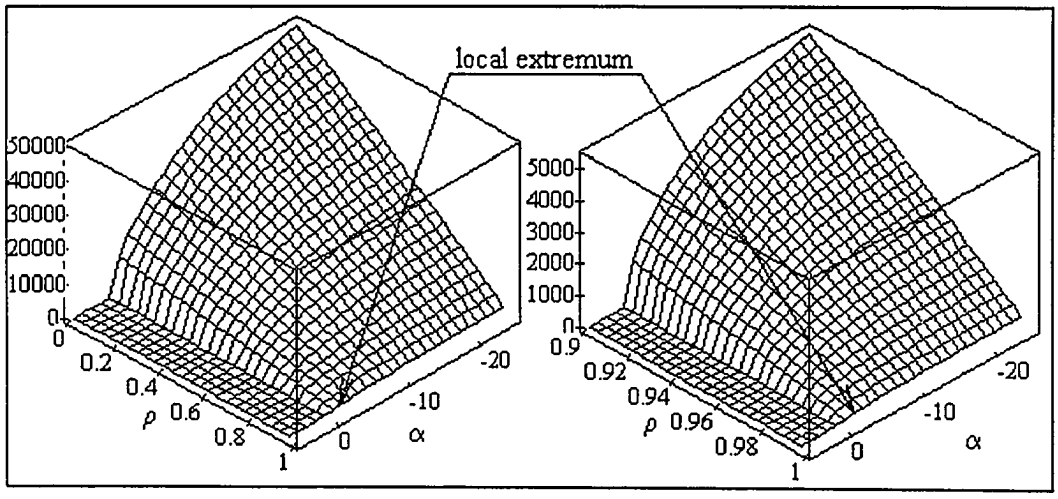

Figure 30: $M(n, \rho, \alpha)$ for $n = 200$.

One can observe that in general $M(n,\rho,\alpha)$ tends to reach its maximized values as $\alpha \to -\infty$. The only exception to this is when n is small and ρ is large; then $M(n,\rho,\alpha)$ has a local wavy extremum shown on the left plot on Figure 29. This extremum becomes negligibly small as n increases (Figure 30). This leads to the conclusion that for large n and $\rho < 1$ approximately

$$\sup_{\alpha} M(n,\rho,\alpha) \approx \lim_{\alpha\to-\infty} M(n,\rho,\alpha).$$

The exact value of $\lim_{\alpha\to-\infty} M(n,\rho,\alpha)$ can be determined as follows:

Lemma 12. $\lim_{\alpha\to-\infty} M(n,\rho,\alpha) = 2(n-1)^2(1-\rho)$

Proof.

$\sum_{i=-n+1}^{n-1} 2^{\alpha i^2} = 1 + 2\sum_{i=1}^{n-1} 2^{\alpha i^2} = 1 + 2^{1+\alpha} + 2^{1+3\alpha}\sum_{i=2}^{n-1} 2^{\alpha(i^2-3)}$. Here $\sum_{i=2}^{n-1} 2^{\alpha(i^2-3)} < \sum_{i=2}^{\infty} 2^{\alpha i} = \frac{1}{1-2^{2\alpha i}}$, which vanishes as $\alpha \to -\infty$. In other words, vanishing tails in $\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}$ and $\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}$ can be truncated because the lower order terms dominate as $\alpha \to -\infty$ or $n \to \infty$. Therefore, truncating all tails to $n = 2$, and applying some well-known limits,

$H(r) = \log_2 \sum_{i=-n+1}^{n-1} 2^{\alpha i^2} - \alpha \frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}} \to \log_2(1+2^{1+\alpha}) - \alpha\frac{2^\alpha}{1+2^{1+\alpha}} \to 2^{1+\alpha} - \alpha 2^\alpha$
$= (2-\alpha)2^\alpha =\to -\alpha 2^\alpha$ as $\alpha \to -\infty$, and

$\sigma^2(r) = \frac{\sum_{i=-n+1}^{n-1} i^2 2^{\alpha i^2}}{\sum_{i=-n+1}^{n-1} 2^{\alpha i^2}} \to 2^\alpha$. Then, since $\sigma^2(u) = \frac{\sigma^2(r)}{2(1-\rho)} \to 0$,

$\frac{\left(1+\sqrt{1-4\frac{\sigma^2(u)}{(n-1)^2}}\right)}{2} \to \frac{\left(1+1-2\frac{\sigma^2(u)}{(n-1)^2}\right)}{2} = 1 - \frac{\sigma^2(u)}{(n-1)^2}$, and $\frac{\left(1-\sqrt{1-4\frac{\sigma^2(u)}{(n-1)^2}}\right)}{2} \to \frac{\sigma^2(u)}{(n-1)^2}$,

resulting in

$H(u) = -\frac{\left(1-\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}\log_2\frac{\left(1-\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2} - \frac{\left(1+\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}\log_2\frac{\left(1+\sqrt{1-4\frac{\sigma^2}{(n-1)^2}}\right)}{2}$

$\rightarrow -\frac{\sigma^2(u)}{(n-1)^2}\log_2\frac{\sigma^2(u)}{(n-1)^2} - \left(1-\frac{\sigma^2(u)}{(n-1)^2}\right)\log_2\left(1-\frac{\sigma^2(u)}{(n-1)^2}\right)$

$\rightarrow -\frac{\sigma^2(u)}{(n-1)^2}\log_2\frac{\sigma^2(u)}{(n-1)^2} + \left(1-\frac{\sigma^2(u)}{(n-1)^2}\right)\frac{\sigma^2(u)}{(n-1)^2}$

$\rightarrow -\frac{\sigma^2(u)}{(n-1)^2}\log_2\frac{\sigma^2(u)}{(n-1)^2} \rightarrow -\frac{2^\alpha}{2(1-\rho)(n-1)^2}\log_2\frac{2^\alpha}{2(1-\rho)(n-1)^2}$. Finally,

$M(n,\rho,\alpha) = \frac{H(r)}{H(u)} \rightarrow \frac{-\alpha 2^\alpha}{-\frac{2^\alpha}{2(1-\rho)(n-1)^2}\log_2\frac{2^\alpha}{2(1-\rho)(n-1)^2}} = \frac{2\alpha(1-\rho)(n-1)^2}{\alpha-\log_2(2(1-\rho)(n-1)^2)}$

$= \frac{2(1-\rho)(n-1)^2}{1-\frac{1}{\alpha}\log_2(2(1-\rho)(n-1)^2)} \rightarrow 2(1-\rho)(n-1)^2.$

■

This leads to the following asymptotic estimate for large n :

$$HR^{\sup}(n,\rho) \approx 2(1-\rho)(n-1)^2 = L^{\sup}(n,\rho).$$

However, this estimate will fail as $\rho \rightarrow 1$, when the local extremum of $M(n,\rho,\alpha)$ becomes important. In this case, we can improve the estimate considerably with the following argument. One can observe on Figures 29 and 30, the local extremum for $M(n,\rho,\alpha)$ occurs for $\alpha < 0$ and, as shown in the lemma, all sums involved in $M(n,\rho,\alpha)$ can be truncated for negative α without noticeable loss of accuracy. In other words, if $M_k(n,\rho,\alpha)$ is $M(n,\rho,\alpha)$ with all summation series truncated to some constant k, then for large n $\sup_\alpha M(n,\rho,\alpha) \approx \sup_\alpha M_k(n,\rho,\alpha)$, with the equality made arbitrarily accurate with appropriate choice of k. The advantage of using truncated $M_k(n,\rho,\alpha)$ is that, unlike $M(n,\rho,\alpha)$, $M_k(n,\rho,\alpha)$ is a function of the product $2(1-\rho)(n-1)^2$, rather than of ρ and n separately (see formulas (28) and (29)). This immediately leads to $M_k(n,\rho_n,\alpha) = M_k(2,\rho_2,\alpha)$ for any n, ρ_n and ρ_2 connected by $2(1-\rho_n)(n-1)^2 = 2(1-\rho_2)(2-1)^2$, or $\rho_2 = 1-(1-\rho_n)(n-1)^2$. Thus, $M_k(n,\rho,\alpha)$ $= M_k(2,1-(1-\rho)(n-1)^2,\alpha)$ and

$$HR^{\sup}(n,\rho) \lesssim \sup_{\alpha} M_k(n,\rho,\alpha) = \sup_{\alpha} M_k(2, 1-(1-\rho)(n-1)^2, \alpha)$$
$$\approx \sup_{\alpha} M(2, 1-(1-\rho)(n-1)^2, \alpha).$$

Finally, $M(n,\rho,\alpha)$ assumes a binary distribution for the image u; therefore for $n = 2$ its optimization must produce exactly the same result as obtained earlier for the binary image model:

$$\sup_{\alpha} M(2,\rho,\alpha) = HR^{\sup}(2,\rho).$$

This leads to

$$HR^{\sup}(n,\rho) \lesssim HR^{\sup}(2, 1-(1-\rho)(n-1)^2) = E^{\sup}(n,\rho).$$

We examined and tabulated the function $HR^{\sup}(2,\rho)$ before; obtaining estimates for $HR^{\sup}(n,\rho)$ is equivalent to changing variables in $HR^{\sup}(2,\rho)$. Figure 31 shows the behavior of both linear $L^{\sup}(n,\rho) = 2(1-\rho)(n-1)^2$ and $E^{\sup}(n,\rho) = HR^{\sup}(2, 1-(1-\rho)(n-1)^2)$ estimates for $HR^{\sup}(n,\rho)$. As n increases, they become indistinguishable.

Correlation Threshold for Difference and Regression-Based Compression

One of the tasks in this research was to estimate the correlation ρ_C between two n-ary images, sufficient to reach the given compression ratio C. As one can see from the previous section, $HR^{\sup}(n,\rho) = \frac{1}{C} \approx 2(1-\rho)(n-1)^2$, which yields a worst-case estimate $\rho_C \approx 1 - \frac{1}{2C(n-1)^2}$. This value rapidly approaches 1 as n increases, making this result quite pessimistic, for example, using this estimate, $C = 2$ compression ratio can be guaranteed for any two 8-bit ($n = 256$) images only if the inter-image correlation

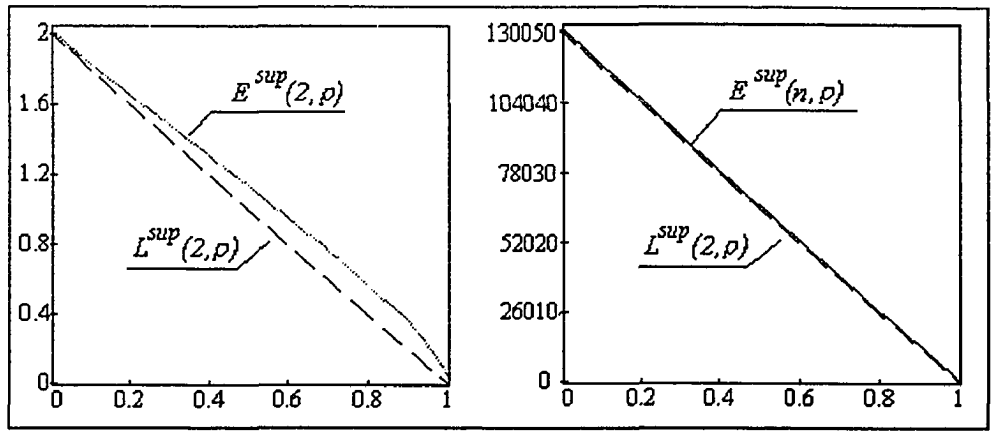

Figure 31: $HR^{\sup}(n,\rho)$ estimates for $n = 2$ and $n = 256$.

approaches $1 - \frac{1}{2*2*(256-1)^2} = .999996$. This high correlation is truly impossible in practice; however, we know that the difference model works for certain classes of images. On the other hand, the best-case estimate for ρ_C for $n > 2$ must lie below that for the binary case, determined earlier to be $\rho_{C=2} = 0.79$. The n-ary images with two distinct intensity clusters (background-foreground) will behave closer to the binary model, producing low correlation thresholds for the given compression ratios.

What happens if the difference model based on compressing $r = u - v$ is replaced by a more general linear regression $r = u - \beta v$? It follows from the linear regression theory that for the β optimally determined with least squares to minimize $||r||$, the residual variance $\sigma^2(r) = (1 - \rho^2)\sigma^2(u)$ (for the difference model $\sigma^2(r) = 2(1 - \rho)\sigma^2(u)$). Therefore one can obtain similar estimates for regression-based compression simply replacing $2(1 - \rho)$ by the smaller[28] $(1 - \rho^2)$ in all previous derivations. In particular, $HR^{\sup}_{reg}(n,\rho) \approx (1 - \rho^2)(n-1)^2$, and $\rho_C \approx \sqrt{1 - \frac{1}{C(n-1)^2}}$. This expression

[28] Since we consider correlation $\rho : 0 \leqslant \rho \leqslant 1$.

yields (for $C = 2$ worst-case) $\rho_C \approx \sqrt{1 - \frac{1}{2*(256-1)^2}} = 999996$, which is still very much the same as the difference model (difference and regression compression models converge as $\rho \rightarrow 1$).

Finally, the same reasoning with extremal probability distributions can estimate $HR^{\text{inf}}(n, \rho)$, which will result in $HR^{\text{inf}}(n, \rho) \gtrsim \frac{2(1-\rho)}{(n-1)^2}$ for the difference model. This estimate will rapidly vanish for increasing n, proving that carefully chosen n-ary images can be compressed with remarkably high compression ratios. At this point, additional study of $HR^{\text{inf}}(n, \rho)$ and $HR^{\text{sup}}(n, \rho)$ is not of interest since these functions converge to their trivial boundaries 0 and ∞ respectively. We will limit our discussion to the more specific and practical models of database compression, valid for certain common classes of images (such as medical images), rather than for general n-ary images.

Visualizing Image Similarity: From Correlation Plots To Probability Surfaces

Our previous discussion shows that the correlation ρ between two signals, $u = u[i]_{i=1}^{n}$ and $v = v[i]_{i=1}^{n}$, may not be sufficient to determine useful boundaries for a compression ratio C obtained by replacing u with $r = u - v$, or the more general expression $r = u - \beta v$. In particular, more information about the inter-image distribution $\mathcal{P}(x, y) = P(u = x \text{ and } v = y)$ is needed since C is a function of $\mathcal{P}$.

Traditionally, correlation plots were used to provide more details about the image-to-predictor correlation. A correlation plot for images u and v presents correlation $\rho(u, v)$ as a set of points with coordinates $(x_i, y_i) = (u[i], v[i])_{i=1}^{n}$. Image similarity means encountering a $u[i] \approx v[i]$, producing a correlation plot concentrated along the line $y = x$. Figure 32 shows correlation plots between two CT and two MR 8-bit images; one can observe that the majority of points lie within the $y = x$ vicinity, which can be viewed as an indication of image similarity.

However, correlation plots have even more information than is expressed with a correlation plot. The probability of a point (x_i, y_i) being on this plot matters much more than its presence. We visualized this missing information by adding a third probabilistic dimension $P = P\{(u[i], v[i]) = (x_i, y_i)\} = P(u = x_i \text{ and } v = y_i)$, which shows the probability of each point (x_i, y_i). This transforms a correlation plot into a three-dimensional plot of the surface $\mathcal{P}(x, y)$, because a correlation plot is an (x, y) projection of the set $\{(x, y, \mathcal{P}(x, y)) : \mathcal{P}(x, y) > 0\}$. This three-dimensional inter-image probability plot for the two CT images used for Figure 32 is shown

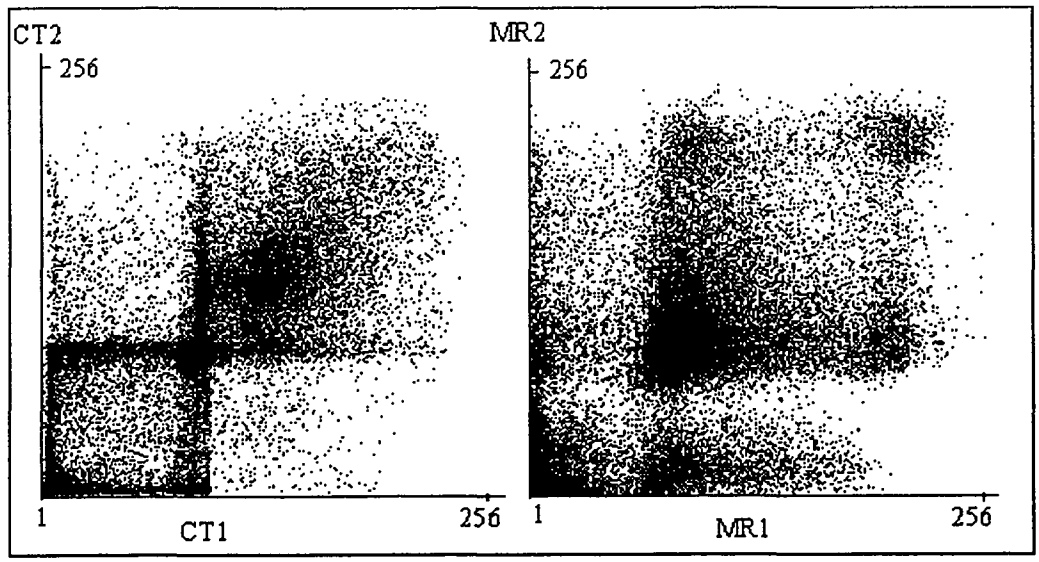

Figure 32: CT and MR correlation plots.

Figure 33 with two different scales for the P axis, and this leads to several important observations.

First, as one can observe, that these do not show a "straight line pattern" correlation. Conversely, the similarity between u=CT1 and v=CT2 is a juxtaposition of the four clusters in $\mathcal{P}(x,y)$ which are:

1. The (1,1) cluster (CT1 background, CT2 background) plus its vicinity accounts for 50% of the total inter-image probability. We will refer to it as the b-b cluster.

2. The (88,88) cluster (CT1 foreground, CT2 foreground), corresponding to the regions where the CT1 foreground with an average intensity close to 88 overlaps with that for CT2 (hereafter referred as f-f cluster). The f-f cluster and the b-b cluster both contribute to a higher inter-image correlation and better compression ratio.

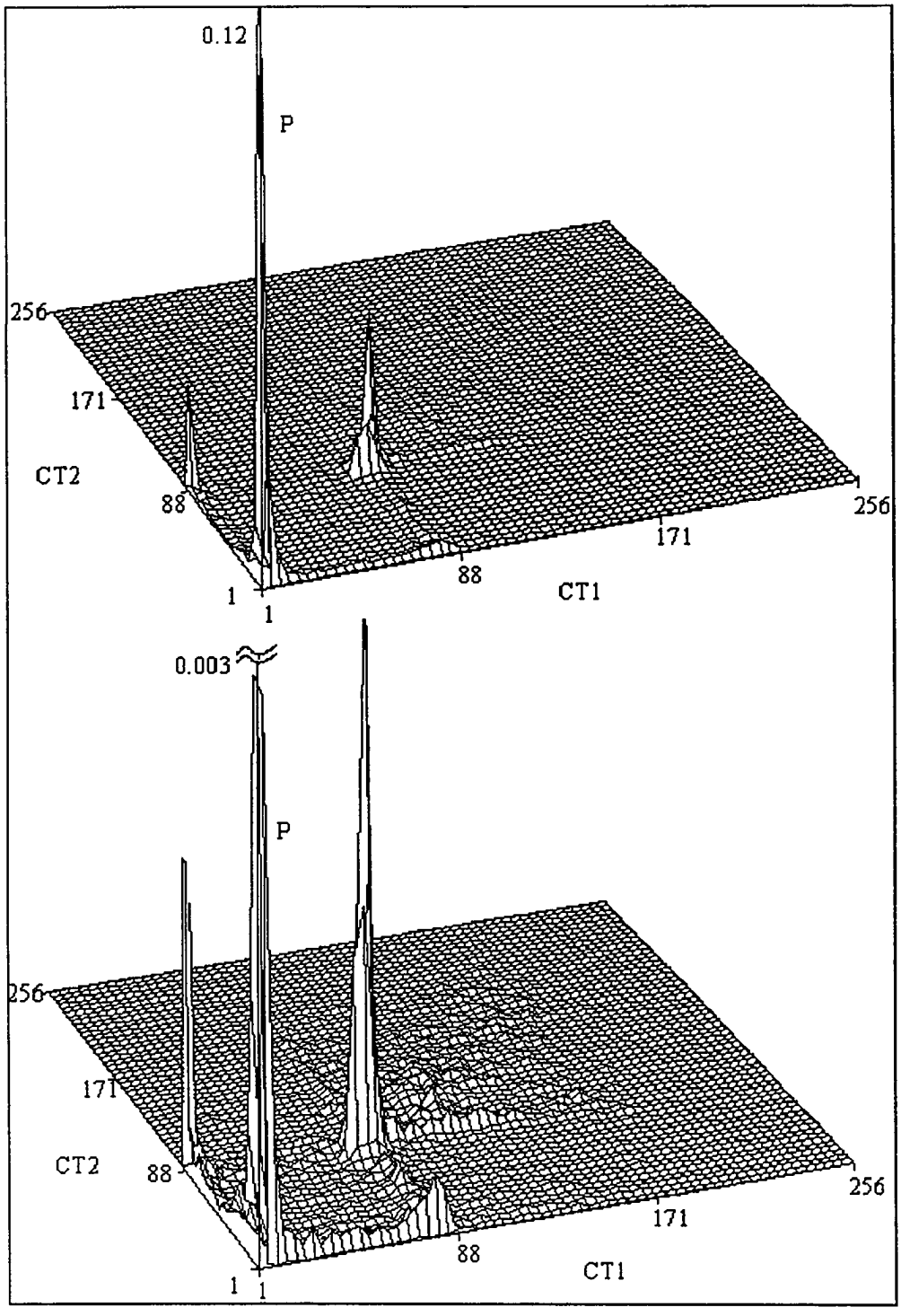

Figure 33: $\mathcal{P}(x, y)$ surface for two similar CT images.

3. The(1,88) (b-f) and the smaller (88,1) (f-b) clusters correspond to the background-foreground overlapping of the two images. These two clusters are responsible for a lower inter-image correlation and a lower compression ratio.

The bottom graph on Figure 33 shows the same inter-image probability surface with 5x magnification along the vertical axis. One can see that the contribution of all other areas (visible on the correlation plot on Figure 32) is negligeable with respect to these four principal clusters. Figure 34 represents inter-image probability plots (original and magnified) for two similar MR images (MR1 and MR2) used for Figure 32. In this case, f-f, b-f and f-b clusters are more spread out and therefore lower, but still contribute most of inter-image bivariate probability.

We also visualized simple autoregressive correlation between each pixel and its left neighbor with the results presented on Figure 35. One can observe that many visually-appealing features of correlation plots, e.g., straight-line clustering or an elliptical area on the CT plot have very little effect on correlation and become almost invisible on the probability surfaces. However, the same 4-cluster model still applies with vanishing b-f and f-b clusters. Since b-f and f-b clusters are responsible for image dissimilarities, the use of autoregressive predictive models on Figure 35 provides much better predictability than the inter-image predictors used for Figures 33 and 34.

This four-cluster interpretation captures virtually all inter-image interaction and can be a valid model for studying compression ratios obtained with predictive models. From this model, marginal image distributions should have two intensity clusters, which explains the presence of two distinct background and foreground clusters in the CT and MR image histograms shown earlier. Using this approach we will introduce

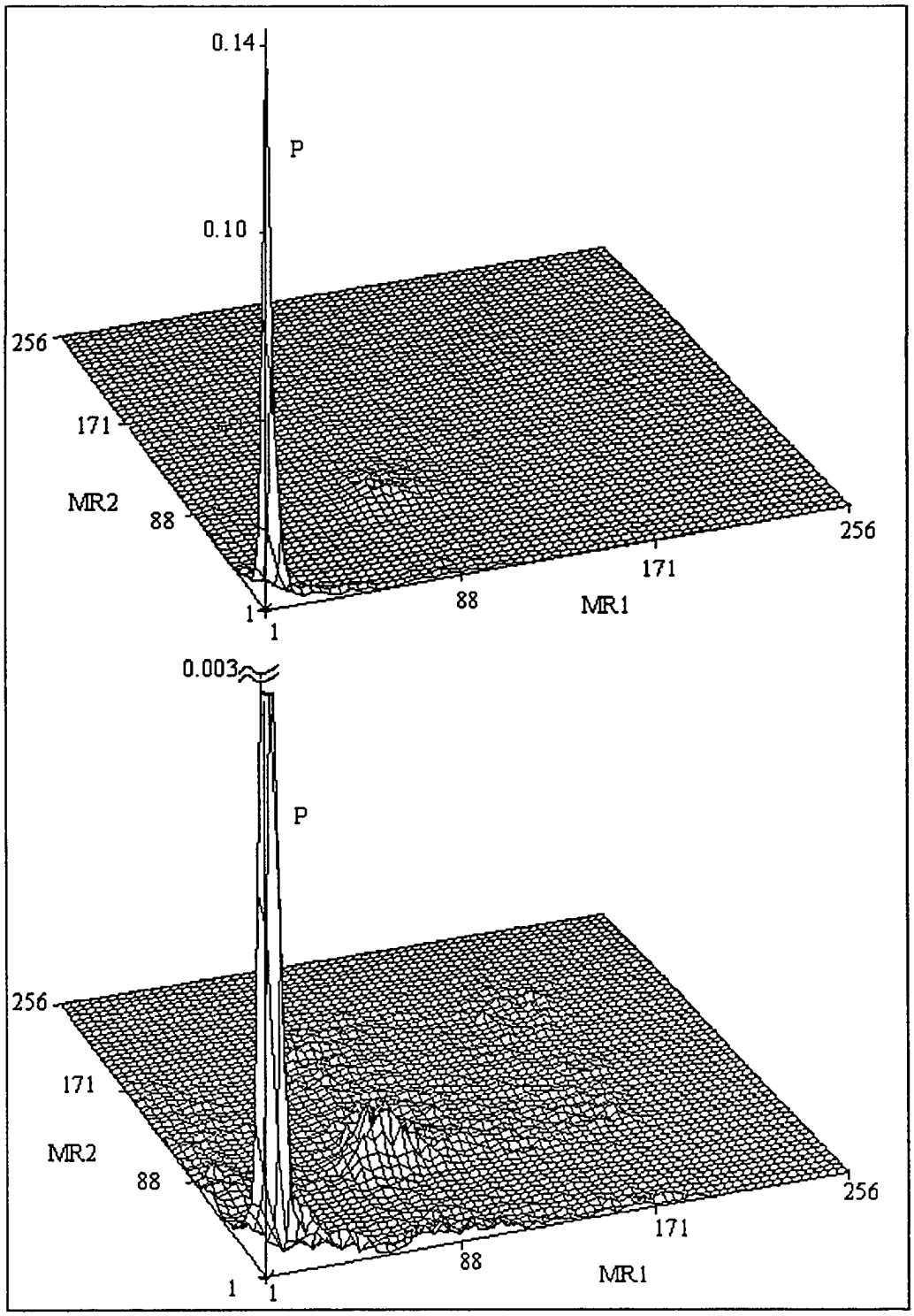

Figure 34: $\mathcal{P}(x, y)$ surface for two similar MR images.

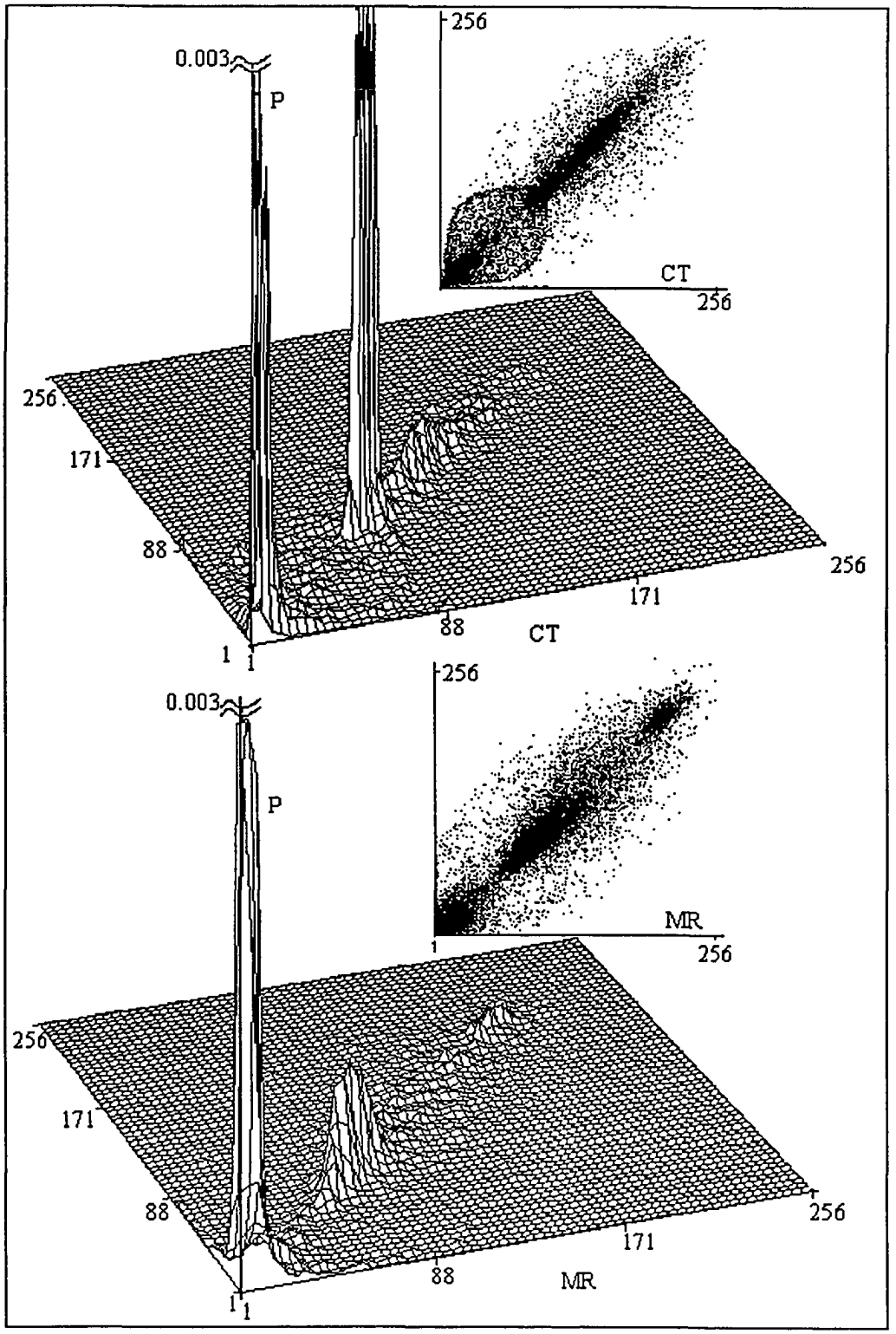

Figure 35: Visualizing autocorrelation.

and develop the simplified four-cluster model shown on Figure 36. Figure 36 presents the clusters with four rectangular probability regions p_1, p_2, p_3 and p_4, that assume the background intensities in each image lie between 1 and a, and the foreground intensities between b and n. The probabilities p_i can be viewed as the average cluster probabilities over the respective rectangular projections as shown on Figure 36. This model is also an extension to the binary model examined earlier.

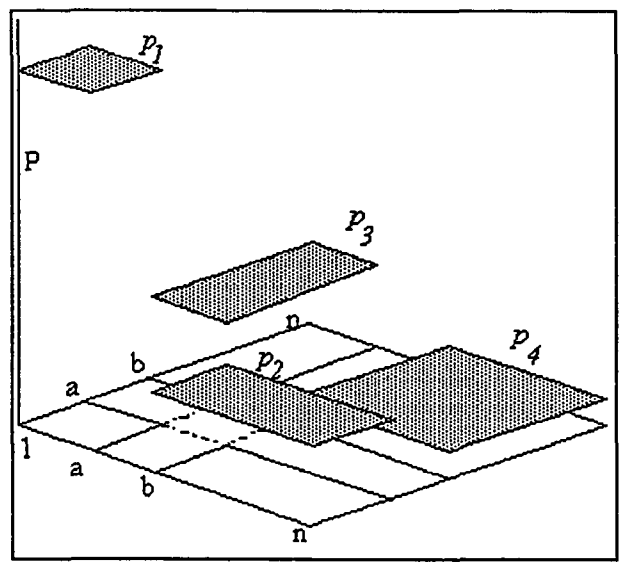

Figure 36: Four-cluster model.

Four-Cluster Model Derivation

In this section the details for the estimates of $HR^{ext}(n, \rho)$ are given assuming the four-cluster model. The derivation is similar to the binary case examined earlier. Figure 36 completely defines the bivariate probability for any integers x, y :

$$\mathcal{P}(x,y) = \begin{cases} p_1, & 1 \leqslant x,y \leqslant a, \\ p_2, & b \leqslant x \leqslant n, 1 \leqslant y \leqslant a, \\ p_3, & 1 \leqslant x \leqslant a, b \leqslant y \leqslant n, \\ p_4, & b \leqslant x,y \leqslant n, \\ 0 & otherwise. \end{cases}$$

We impose two constraints:

$$\sum_{1 \leqslant i,j \leqslant n} \mathcal{P}(x,y) = 1 = p_1 a^2 + a(p_2+p_3)(n-b+1) + p_4(n-b+1)^2, \text{ and } (31)$$

$$\sigma^2(u) = \sigma^2(v). \quad (32)$$

The probability distribution of image u intensities:

$$\mathcal{P}^u(x) = \sum_y \mathcal{P}(x,y) = \begin{cases} ap_1 + (n-b+1)p_3, & 1 \leqslant x \leqslant a, \\ ap_2 + (n-b+1)p_4, & b \leqslant x \leqslant n, \\ 0, & otherwise, \end{cases}$$

and therefore

$$\sigma^2(u) = \sum_{i=1}^{a} i^2(ap_1 + (n-b+1)p_3) + \sum_{i=b}^{n} i^2(ap_2 + (n-b+1)p_4)$$
$$- \left(\sum_{i=1}^{a} i(ap_1 + (n-b+1)p_3) + \sum_{i=b}^{n} i(ap_2 + (n-b+1)p_4) \right)^2.$$

The expression for $\sigma^2(v)$ is obtained from $\sigma^2(u)$ swapping p_2 and p_3 :

$$\sigma^2(v) = \sum_{i=1}^{a} i^2(ap_1 + (n-b+1)p_2) + \sum_{i=b}^{n} i^2(ap_3 + (n-b+1)p_4)$$
$$- \left(\sum_{i=1}^{a} i(ap_1 + (n-b+1)p_2) + \sum_{i=b}^{n} i(ap_3 + (n-b+1)p_4) \right)^2.$$

Then

$$\sigma^2(u) - \sigma^2(v) = (n-b+1)(p_3-p_2) \sum_{i=1}^{a} i^2 + a(p_2-p_3) \sum_{i=b}^{n} i^2$$

$$- \left((n-b+1)(p_3-p_2) \sum_{i=1}^{a} i + a(p_2-p_3) \sum_{i=b}^{n} i \right)$$
$$* \left(\sum_{i=1}^{a} i(2ap_1 + (n-b+1)(p_2+p_3)) + \sum_{i=b}^{n} i(a(p_2+p_3) + 2(n-b+1)p_4) \right)$$
$$= (p_3-p_2)[(n-b+1) \sum_{i=1}^{a} i^2 - a \sum_{i=b}^{n} i^2 - \left((n-b+1) \sum_{i=1}^{a} i - a \sum_{i=b}^{n} i \right)$$
$$* \left(\sum_{i=1}^{a} i(2ap_1 + (n-b+1)(p_2+p_3)) + \sum_{i=b}^{n} i(a(p_2+p_3) + 2(n-b+1)p_4) \right)] = 0$$

This quadratic polynomial equation (with respect to p_i) will have two real roots that satisfy:

1. $p_3 = p_2$, or

2. $(n-b+1) \sum_{i=1}^{a} i^2 - a \sum_{i=b}^{n} i^2 - \left((n-b+1) \sum_{i=1}^{a} i - a \sum_{i=b}^{n} i \right)$
$$* \left(\sum_{i=1}^{a} i(2ap_1 + (n-b+1)(p_2+p_3)) + \sum_{i=b}^{n} i(a(p_2+p_3) + 2(n-b+1)p_4) \right) = 0$$

One can demonstrate that case 2 is unacceptable. Case 2 results in a linear equation in (p_2+p_3), however, the first imposed condition (31) is a linear equation with respect to (p_2+p_3) as well. These two linear equations are different and they will lead to contradiction unless additional assumptions are made about the remaining variables, which we are not willing to do while considering the general case. Thus, satisfying (31) and (32) leads to a unique possible solution only when $p_3 = p_2$ in (32). Note that when $p_3 = p_2$ the images u and v have the same probability distributions (histograms), which is a very natural result for two similar images. Then the system

$$\begin{cases} p_3 = p_2 \\ p_1 a^2 + (p_2+p_3)a(n-b+1) + p_4(n-b+1)^2 = 1 \end{cases}$$

solved for p_2 and p_3 yields

$$p_3 = p_2 = -\frac{1}{2} \frac{a^2 p_1 + 2p_4 n - 2p_4 nb + p_4 n^2 + p4b^2 - 2p_4 b + p4}{a(n-b+1)}. \tag{33}$$

All probabilities are expressed as functions of a, b, n, p_1 and p_4. It is straightforward to express image variances $\sigma^2(u)$ and $\sigma^2(v)$, correlation $\rho(u,v)$, entropies $H(u)$ and $H(u-v)$, and the entropy reduction function $HR(n,\rho) = H(u-v)/H(u) = 1/C(n,\rho)$ as functions of these variables. The derivation of these expressions was similar to our binary case. The final formula for $HR(n,\rho)$ was determined with Maple® and requires about three pages to present. This is believed to not contribute to the clarity of results if reproduced. The result (33) was used primarily to run numerical $HR(n,\rho)$ minimizations to find:

$$\begin{aligned} HR^{\sup}(n,\rho) &= \frac{1}{C^{\inf}(n,\rho)} = \sup_{a,b,n,p_1,p_4,\rho=\rho(u,v)} HR(n,\rho), \\ HR^{\inf}(n,\rho) &= \frac{1}{C^{\sup}(n,\rho)} = \inf_{a,b,n,p_1,p_4,\rho=\rho(u,v)} HR(n,\rho). \end{aligned}$$

Figure 37 presents the results of this numerical optimization carried out for $n = 2, 4, 8, 16, 32, 64$ and 128.

Additional observations about this four-cluster model:

1. Images u and v have the same intensity distributions, which leads to $H(u) = H(v)$ and $HR^{\sup}(n,\rho) \leqslant 2$ (as proved earlier). Then $HR^{\sup}(n,\rho) = 2$ corresponds to independent, i.e., uncorrelated images u and v.

2. When n increases, $HR^{\sup}(n,\rho)$ approximately converges to 1.4 for $\rho > 0.7$, and rapidly vanishes to 0 only when ρ becomes very close to 1. This means that if we consider MR or CT images having a typical $0.7 \leqslant \rho \leqslant 0.85$, and accept the four-cluster model as a good approximation for the inter-image bivariate probability,

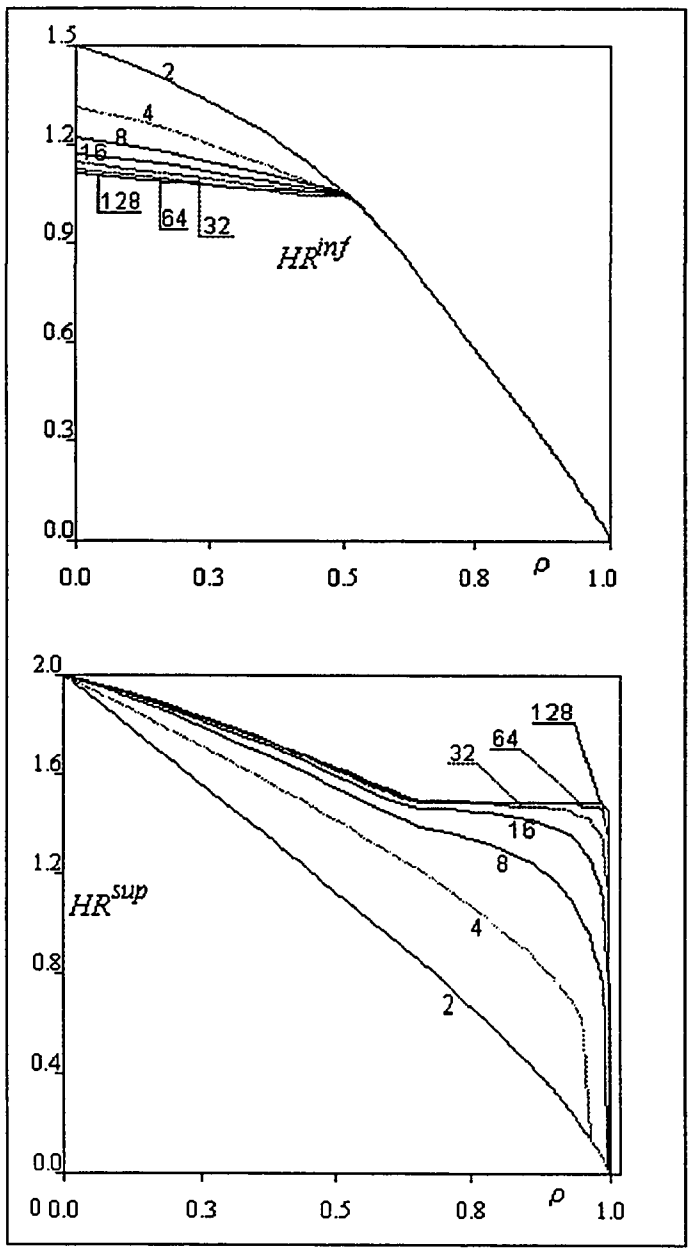

Figure 37: $HR^{\inf}(n,\rho)$ and $HR^{\sup}(n,\rho)$ for four-cluster model.

the application of difference compression to these images can result in a 40% increase in entropy. Clearly, difference models do not guarantee compression.

3. $HR^{\inf}(n,\rho)$ curves for different values of n and $\rho > 0.5$ converge to the straight line $HR^{\inf}(n,\rho) = 2(1-\rho)$ which improves our earlier estimate $HR^{\inf}(n,\rho) \gtrsim \frac{2(1-\rho)}{(n-1)^2}$. It is interesting to observe that as n increases, the highest (best-case) value of $HR^{\inf}(n,0)$ decreases toward 1. Figure 37 suggests for large n :

$$HR^{\inf}(n,\rho) \approx \begin{cases} 1 & 0 \leqslant \rho \leqslant 0.5 \\ 2(1-\rho) & 0.5 \leqslant \rho \leqslant 1 \end{cases}$$

with only 8% error for $n > 128$. This means that for large n and any correlation $\rho_0 \geqslant 0$ there exist at least two images u and v with identical histograms such that a) $\rho(u,v) = \rho_0$, and b) $H(u-v) = H(u)$ $(= H(v))$. For large n and for the best-case images, difference models are guaranteed to not increase the entropy for any inter-image correlation.

The 4–cluster model also permits the behavior of the correlation threshold $\rho(n,C)$ to be analyzed. This function provides a worst and best case estimate on inter-image correlation ρ between two n-ary images sufficient to ensure a compression ratio C for difference compression. This can be determined from the same data used for Figure 37, with the results shown on Figure 38.

Curves on the left plot on Figure 38 show best-case estimates for inter-image correlation ρ providing the respective compression ratios: $C = 1$ (practical threshold), 2 (typical lossless) and 3 (good lossless). They remain almost constant since the best-case compression ratio function $HR^{\inf}(n,\rho)$ behaves as a straight line for $\rho \geqslant 0.5$

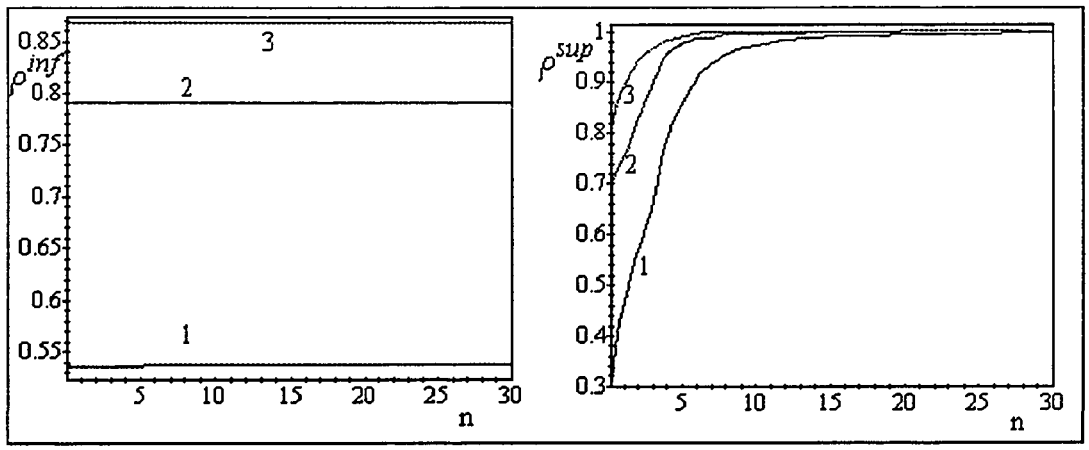

Figure 38: Curves $\rho^{\inf}(n, C)$ (left) and $\rho^{\sup}(n, C)$ for $C = 1, 2$ and 3.

(Figure 37). Note: since $\rho^{\inf}(n, 3) > 0.85$, a 1:3 difference compression is impossible for CT and MR images with average $0.7 \leqslant \rho \leqslant 0.85$.

Worst-case estimates $\rho^{\sup}(n, C)$ on the right plot exhibit very fast convergence to 1 as n increases. This corresponds to the fast decay to 0 for $HR^{\sup}(n, \rho)$ on Figure 37.

In the previous sections we primarily studied difference models, achieving set compression replacing similar images by their differences. We determined numerical estimates for the compression ratio function $HR(\rho)$ (inverse to compression ratio $C(\rho)$) and found that compressing $r = u - v$ instead of u or v can be beneficial only when correlation $\rho = \rho(u, v)$ is very close to 1.

However, one may try to improve these estimates introducing more general linear regression-based set compression when r is determined by

$$r = u - (\alpha + \beta v), \tag{34}$$

and the constants α and β are chosen to minimize $||r|| = \sqrt{r^T r}$. Note that α will shift all image intensities by a constant value which has no effect on the entropy $H(r)$. Therefore, in terms of compression model (34) will perform as efficiently as

$$r = u - \beta v. \tag{35}$$

The value of β can be determined with the following lemma

Lemma 13. $||r||^2 = r^T r$ in (35) is minimized by $\beta = \frac{\sigma(u)}{\sigma(v)}\rho(u, v)$.

Proof.

$||r||^2 = r^T r = (u - \beta v)^T (u - \beta v) = u^T u - \beta(u^T v + v^T u) + \beta^2 v^T v,$

where[29] $u^T u = \sigma^2(u)$, $v^T v = \sigma^2(v)$ and $u^T v = v^T u = \rho\sigma(u)\sigma(v)$. Then optimal

[29]To be more accurate, we should use $u - \overline{u}$ instead of u, and similar centered vectors instead of v and r. However, as stated earlier, shifts in intensity domain do not affect the entropies, and therefore need not to be considered.

$$\beta = \frac{u^T v + v^T u}{2v^T v} = \frac{\sigma(u)}{\sigma(v)}\rho(u,v).$$

■

In particular, the assumption of $\sigma(u) = \sigma(v)$ yields $\beta = \rho(u,v)$. Since βv is not an integer vector any more, instead of (35) one uses

$$r = u - \lfloor \rho v \rfloor \tag{36}$$

with $\lfloor . \rfloor$ representing integer truncation.

Replacing difference model $r = u - v$ by the more general $r = u - \lfloor \rho v \rfloor$ introduces many changes into behavior of the $HR(\rho)$ and $C(\rho)$ functions. First, note that for $\rho = 0$ equation (36) becomes $r = u$, which means that $HR(0) = \frac{H(r)}{H(u)} = 1$ and $C(0) = \frac{1}{HR(0)} = 1$. When $\rho \to 1$, model (36) converges to difference model $r = u - v$ examined earlier, with $HR(1) = 0$ and $C(1) = +\infty$. Finally, our numerical experiments indicate[30] that there always are positively-correlated images u, v with $HR > 1$. This means that functions $HR^{\text{sup}}(n,\rho)$ and $HR^{\text{inf}}(n,\rho)$ must have local maxima (local minima for $C^{\text{sup}}(n,\rho)$ and $C^{\text{inf}}(n,\rho)$) and are not monotone as they were for difference predictors.

The presence of integer truncation in (36) makes it impossible to analyze analytically, but all derivations from the previous section remain valid allowing for convenient numerical study. Therefore, we performed numerical optimization for the

[30] See for example the two MR images used in the introduction. Another numerical example is $u = [0,2,2]$, $v = [2,2,0]$. Entropies $H(u) = H(v) = \frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 \frac{3}{2} = 0.918296$. Correlation $\rho(u,v) = 0.5$ and $r = u - \lfloor \rho v \rfloor = [-1,1,2]$ with $H(r) = \log_2 2 = 1.58496$. Then $HR(\rho = 0.5) = \frac{1.58496}{0.918296} = 1.72598 > 1$.

4-cluster model with regression predictor (36). The resulting plots for $HR^{\inf}(n,\rho)$ and $HR^{\sup}(n,\rho)$ for $n = 4, 8, 16, 32, 64$ and 128 are shown on Figure 39.

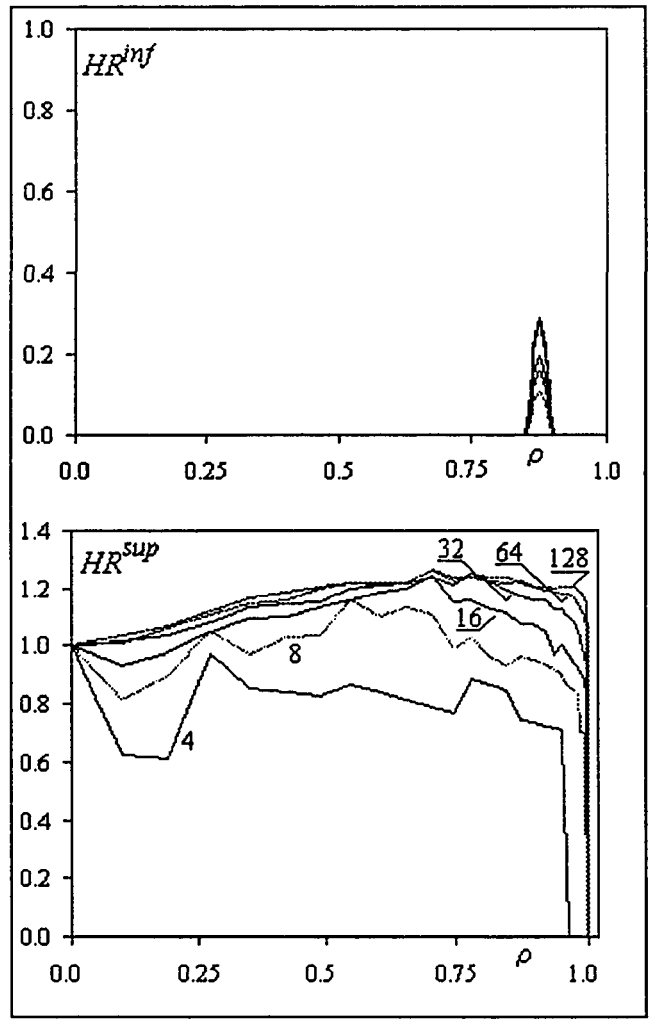

Figure 39: Regression $HR^{\inf}(n,\rho)$ and $HR^{\sup}(n,\rho)$ for four-cluster model.

These results look very different from the difference models on Figure 37. This is due to the presence of integer truncation function in $HR^{\inf}(n,\rho)$ which becomes 0 almost everywhere except one compact maximum at $\rho = 0.875$. It is worth mentioning that, within our acceptable numerical error, different values of intensity levels n in $HR^{\inf}(n,\rho)$ have the same extremal ρ. This ρ corresponds to the best-case

regression-based compression producing the worst results. Plots for $HR^{\sup}(n,\rho)$ are also influenced by truncation, especially for small values of n. As n increases however these plots become more stable and smooth, with local maximum at $\rho = 0.75$.

These results suggest that regression models (including both autoregressive and inter-image predictors) perform better than difference models: for example, all plots $HR^{\sup}(n,\rho)$ for different n on Figure 39 stay below 1.4, while same plots on Figure 37 approach to 2. However, as ρ increases, both regressive and difference models converge and have very similar compression. The local extrema in regressive predictors indicate that values of $0.75 < \rho < 0.87$ are the worst for entropy reduction, since they maximize functions $HR^{\sup}(n,\rho)$ and $HR^{\inf}(n,\rho)$ (minimize compression ratios $C^{\sup}(n,\rho)$ and $C^{\inf}(n,\rho)$). Note that inter-image correlation between CT or MR images falls in this range, which means that inter-image predictors for these similar image classes will perform with their lowest efficiency.

Average Case Study With 4-Cluster Model And Model Validation

The worst and the best-case estimated for $HR^{\sup}(n,\rho)$ and $HR^{\inf}(n,\rho)$, obtained with the 4-cluster bivariate probability model, are better and more applicable compared to the most general case when no assumptions were made about the bivariate probability distribution. However, the difference between the worst $HR^{\sup}(n,\rho)$ and the best $HR^{\inf}(n,\rho)$ increases with n, leaving a wider range for guessing about the most typical values of $HR(\rho)$ between these two extreme cases. Therefore we used the same 4-cluster model to perform a numerical average-case study and determine the values of

$$HR^{avg}(n,\rho) = \frac{1}{C^{avg}(n,\rho)} = average_{a,b,p_1,p_4} HR(n,\rho),$$

assuming that variables a, b, p_1 and p_4 are distributed uniformly within their respective ranges. The resulting values for $HR^{avg}(n,\rho)$ are shown on the Figure 40. The values of $n = 2,4,6,8,10,16,32,64,128$ and 256 were used for the difference $HR^{avg}(n,\rho)$ plot, and $n = 4,6,8,10,16,32,64,128$ and 256 for the regressive $HR^{avg}(n,\rho)$.

These average-case curves lie closer to the worst-case estimates and exhibit the same rate of decay as $\rho \to 1$. Since all $HR(\rho)$ curves were obtained from the theoretical 4-cluster model, and the 4-cluster model was built from studying the shape of bivariate probability surface for CT and MR images, it is interesting to observe how\ actual CT and MR data fit between these numerical curves. Two sets of these images with $n = 256$ intensity levels were used to find HR values. Figure 41, top, shows

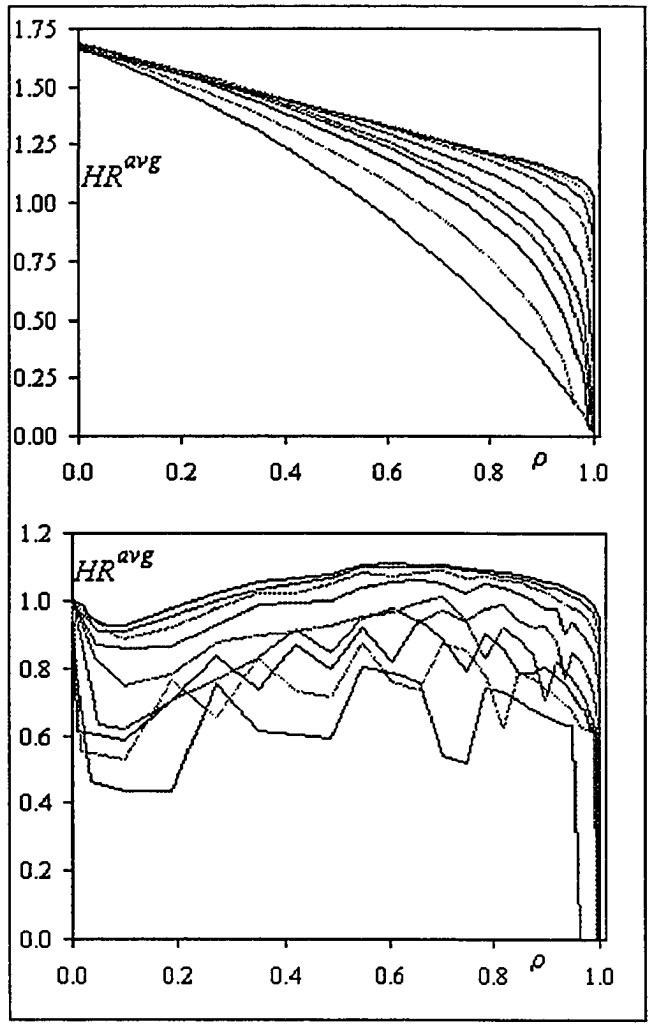

Figure 40: $HR^{avg}(n, \rho)$ for the difference (top) and regression.

three curves, $HR^{\sup}(256,\rho)$, $HR^{avg}(256,\rho)$ and $HR^{\inf}(256,\rho)$, and 200 points with coordinates $(\rho(u,v), HR(u,v))$ computed for CT and MR images. HR values on this plot were determined assuming difference compression: $HR(u,v) = \frac{H(u-v)}{H(u)}$, where images u and v were taken either from the same similar class (CT or MR) for inter-image prediction, or as shifts of the same image, $u[i] = v[i-1]_{i=1}^{N}$, for autoregressive models.

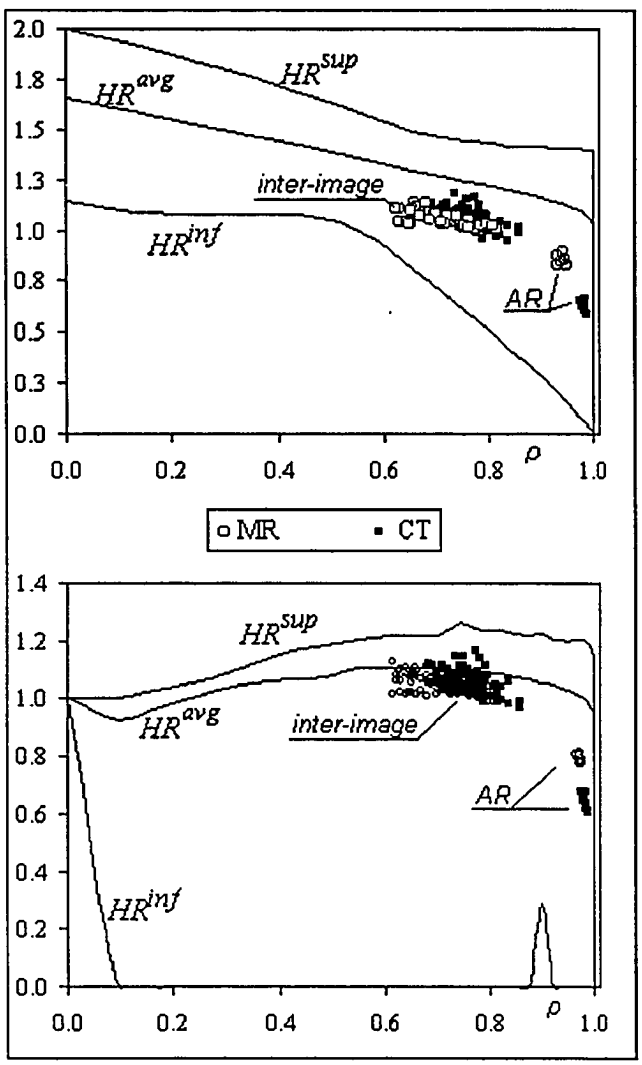

Figure 41: 4-cluster model validation.

Observe that

1. The average case function $HR^{avg}(256, \rho)$ lies closer to the worst case $HR^{sup}(256, \rho)$, than to the best-case $HR^{inf}(256, \rho)$. This means that on average, predictive compression may perform almost as bad as the worst case.

2. Image data is well localized with respect to these curves, with points never exceeding $HR^{sup}(256, \rho)$ nor falling below $HR^{inf}(256, \rho)$. This supports the validity of the 4-cluster model and all performance predictions based on it.

3. Note that for the inter-image predictors, the HR coordinate in the image data $(\rho(u, v),\ HR(u, v))$ stays typically above, but very close to 1. This demonstrates again that inter-image prediction increases image entropy and cannot be used for efficient set compression.

Figure 41, bottom, is similar to the top plot, but shows the same results for the regression-based predictors with $HR(u, v) = \frac{H(u - v\rho(u,v))}{H(u)}$. Note how close image data is clustered with respect to the $HR^{avg}(256, \rho)$ curve, once again proving the validity of our theoretical approach.

Nearly-Lossless Extensions To Lossless CAR Compression

Introduction

During the last decade, several compression methods have emerged as a compromise between traditional lossless (relatively low compression ratios) and lossy (information loss) compression techniques. *Almost-lossless* compression is a probabilistic approach that guarantees that most of the image pixels (for instance, 95%) will retain their intensities, as defined by Karray [37]. *Perceptually-lossless* compression introduces new information measures, explained by Karunasekera, S.A and N.G.Kingsbury [44], based on subjective human sensitivity to different image details, ensuring that certain losses in image information will remain unperceived by the human eye [42], [40], [41], [43]. Finally, *nearly-lossless* (NL) compression, as presented by Chen [38] and Ke and Marcellin [39], allows compression of an image assuming that every pixel value can be changed by some small ε (lossless compression corresponds to $\varepsilon = 0$). Small values of ε can substantially improve the image compression ratio without any visible changes in the image [40]. Straightforward NL compression for an image with N intensity levels is often performed by reducing this number to $N/(2\varepsilon + 1)$ quantization levels [45], and optimally replacing each original pixel intensity by some value from the reduced (quantized) intensity range. Theoretically, this decreases the image entropy by $log_2(2\varepsilon + 1)$. However, each pixel is deemed to lose a certain intensity, and the image to lose most of its colors. This may produce visible artifacts even for small ε. In this research, instead of reducing the image entropy through intensity range reduction, we propose to improve the image compressing properties for the speci-

fied compressing transforms. Autoregressive (AR) model compression was chosen as an example. Our Nearly-Lossless Autoregressive (NLAR) compression preserves the image intensity range and only redistributes some pixel intensities within $\pm\varepsilon$ error interval to optimally decrease the entropy of the AR-compressed image. Since pixels are not forced to a reduced intensity scale, compression improvement becomes more moderate, but the fidelity, accuracy and perceptual quality increase. Moreover, this technique is computationally simple and can be used to implement many previous variations of nearly and perceptually lossless compression.

NLAR Algorithm Derivation

As an example, the 5-th order AR model was used to derive and test our technique; in general, one can use any AR model, of any size, image-dependent or image-independent. With the operators B and L representing 1-pixel bottom and left image shifts respectively, the 5-th order AR model can be represented as:

$$u = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)u + r,$$

where $L^m B^n u[i,j] = u[i-m, j-n]$. Optimal values for $\beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5]^T$ are found with traditional least squares regression as $\beta = (v^T v)^{-1}(v^T u)$, where matrix v has five columns, all are shifts of the vector u:

$$v = (Lu, Bu, LBu, L^2 u, B^2 u).$$

Increasing the $[x,y]$ pixel intensity by e intensity units is equivalent to adding to u the image $e\Delta^{[x,y]}$ where:

$$\Delta^{[x,y]}[i,j] = \begin{cases} 0, & if \quad [i,j] \neq [x,y] \\ 1 & if \quad [i,j] = [x,y] \end{cases}$$

This leads to the following:

$$\begin{cases} u = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)u + r \\ (u + e\Delta^{[x,y]}) = (\widetilde{\beta}_1 L + \widetilde{\beta}_2 B + \widetilde{\beta}_3 LB + \widetilde{\beta}_4 L^2 + \widetilde{\beta}_5 B^2)(u + e\Delta^{[x,y]}) \\ \qquad\qquad +(r + e\delta^{[x,y]}) \end{cases} \tag{37}$$

where $\delta^{[x,y]}$ is the change in image residual caused by altering one pixel value in the image u by one intensity unit. Theoretically, every small change in u will produce a small change in the corresponding optimal AR model, i.e., $\beta \neq \widetilde{\beta}$. This makes the system (37) nonlinear. To overcome this problem, we solve (37) with the following two iterative steps:

Step 1.Assume $\beta \approx \widetilde{\beta}$ and solve

$$\begin{cases} u = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)u + r \\ (u + e\Delta^{[x,y]}) = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)(u + e\Delta^{[x,y]}) \\ \qquad\qquad +r + e\delta^{[x,y]} \end{cases} \tag{38}$$

Step 2. Update the optimal β as the optimal AR model for $\widetilde{u} = u + e\Delta^{[x,y]}$.

We will consider each step separately and demonstrate how it reduces the residual variance. Since a small ε will not affect the range of the residual distribution, variance minimization inevitably results in reduced entropy.

Step 1 : Residual Variance Minimization in sup Norm.

For step 1, using equation (38) and linear property of recurrent equations:

$$\Delta^{[x,y]} - (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)\Delta^{[x,y]} = \delta^{[x,y]}.$$

Since both $\Delta^{[x,y]}$ and the model coefficients β_i are known, for the 5-th order AR we determine

$$\delta^{[x,y]} = \begin{cases} 1 & at \quad [x,y], \\ -\beta_1 & at \quad [x+1,y], \\ -\beta_2 & at \quad [x,y+1], \\ -\beta_3 & at \quad [x+1,y+1], \\ -\beta_4 & at \quad [x+2,y], \\ -\beta_5 & at \quad [x,y+2], \\ 0 & otherwise. \end{cases} \cdot \tag{39}$$

Altering the $[x,y]$ pixel of the u image by some e will result in changing the residual r by $e\delta^{[x,y]}$. This residual change $e\delta^{[x,y]}$ is used to minimize the residual variance. Since only six residual pixels will be affected by (39), we consider only their contribution in the residual variance, that is:

$$\begin{aligned} \sigma^{[x,y]}(e) \quad = \quad & (r[x,y]+e)^2 + (r[x+1,y]-e\beta_1)^2 + (r[x,y+1]-e\beta_2)^2 \\ & +(r[x+1,y+1]-e\beta_3)^2 + (r[x+2,y]-e\beta_4)^2 \\ & +(r[x,y+2]-e\beta_5)^2. \end{aligned} \tag{40}$$

So, the optimal choice for e to minimize $\sigma^{[x,y]}(e)$ is :

$$\begin{aligned} e_{min} &= (-r[x,y] + (r[x+1,y]\beta_1 + r[x,y+1]\beta_2 + r[x+1,y+1]\beta_3 \\ &\quad + r[x+2,y]\beta_4 + r[x,y+2]\beta_5)/(1+\beta_1^2+\beta_2^2+\beta_3^2+\beta_4^2+\beta_5^2). \end{aligned}$$

Since $e = 0$ corresponds to the old variance, and $\sigma^{[x,y]}(e)$ is a quadratic polynomial in e, any $e \in [0, e_{min}]$ will decrease the residual variance (entropy). This leads to the following nearly-lossless residual variance reduction algorithm:

1. Choose $\varepsilon > 0$ and initialize a "lost" image du as a 0-intensity image. The du image will keep the part of u we are going to sacrifice, with constraint $sup_{[i,j]} |du[i,j]| \leqslant \varepsilon$.

2. **For** each pixel $[i,j]$ in u **do:**

2.1 Compute e_{min} , */* optimal error for residual variance reduction */*

2.2 Find highest k, $k \leqslant 1$, such that $|du[i,j] + ke_{min}| \leqslant \varepsilon$

2.3 **If** $k > 0$ **then:** */* residual variance for $[i,j]$ can be minimized */*

a.compute $e = ke_{min}$ */* acceptable image error */*

b update:

$$du[i,j] \rightarrow du[i,j] + e;$$

$$r[i,j] \rightarrow r[i,j] + e;$$

$$r[i+1,j] \rightarrow r[i+1,j] - e\beta_1;$$

$$r[i,j+1] \rightarrow r[i,j+1] - e\beta_2;$$

$$r[i+1,j+1] \rightarrow r[i+1,j+1] - e\beta_3;$$

$$r[i+2,j] \rightarrow r[i+2,j] - e\beta_4;$$

$$r[i,j+2] \rightarrow r[i,j+2] - e\beta_5;$$

2.4 **Endif**

3. **Enddo**

The loop 2, minimizing the total residual variance for the 5-point vicinity of each $u[i,j]$, may be repeated until no more pixels are updated (we found one or two iterations sufficient to closely reach convergence). Only if necessary, it will change the $u[i,j]$ intensity ensuring that the cumulative change in it does not exceed ε. This will produce the "lost" image du such that:

1. $sup\ |du| \leqslant \varepsilon$

2. image $\widetilde{u} = u + du$, for the given AR model with coefficients β, has smaller residual variance than u.

Replacing u with $\widetilde{u}$, du is lost, but image autocorrelation properties are improved, which leads to better image compression.

Step 2 : Optimizing Model Parameters

Step two is only needed when image-dependent AR models are considered. In this case, optimal values of β coefficients are always determined. In the previous section we replaced :

$$u = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)u + r, \tag{41}$$

by

$$u + du = \widetilde{u} = (\beta_1 L + \beta_2 B + \beta_3 LB + \beta_4 L^2 + \beta_5 B^2)\widetilde{u} + r', \tag{42}$$

such that $sup|du| = sup|u - \widetilde{u}| \leqslant \varepsilon$. Note that the β coefficients that were optimal for the u image in (41) do not have to remain optimal for $\widetilde{u}$ in (42). So we can further reduce the residual r' variance by replacing suboptimal values of β with their optimal values. Then in step 2 the AR model for $\widetilde{u}$ is updated as:

$$\widetilde{u} = (\widetilde{\beta}_1 L + \widetilde{\beta}_2 B + \widetilde{\beta}_3 LB + \widetilde{\beta}_4 L^2 + \widetilde{\beta}_5 B^2)\widetilde{u} + \widetilde{r},$$

by recomputing the optimal coefficient vector $\widetilde{\beta}$ as $\widetilde{\beta} = (v^T v)^{-1}(v^T \widetilde{u})$, where the matrix v has columns $v = (L\widetilde{u}, B\widetilde{u}, LB\widetilde{u}, L^2\widetilde{u}, B^2\widetilde{u})$, and residual

$$\widetilde{r} = \widetilde{u} - (\widetilde{\beta}_1 L + \widetilde{\beta}_2 B + \widetilde{\beta}_3 LB + \widetilde{\beta}_4 L^2 + \widetilde{\beta}_5 B^2)\widetilde{u}.$$

For the optimal $\widetilde{\beta}$, we always have $||\widetilde{r}|| < ||r'||$ (equality within chosen tolerance means convergence of the iterative process). That is, the second iterative step (updating β) will also always result in reduced residual variance.

Combining two iterative steps in one, yields the following nearly-lossless image compression algorithm:

1. Choose nearly-lossless compression error $\varepsilon > 0$
2. Initialize du as a nil image
3. **Do**

3.1 Compute $\widetilde{u} = u + du$ and its optimal β-coefficients **or** choose an image-independent AR model

3.2 Scan image $\widetilde{u}$ as in the previous section, reducing residual variance pixel by pixel

while (the decrease in residual variance does not reach the convergence threshold)

4. Replace u by $\widetilde{u} = u + du$ and compress it.

Numerical Results

Trends

Our NLAR technique was applied to ten CT images and the Lena image (256×256×8) (Figures 42 and 43). The correlation between residual entropy and variance was found as high as 0.9994 on average, which demonstrates that fixed-range residual variance reduction is equivalent to the entropy reduction. Table 5 summarizes NLAR compression of the Lena image for different values of the error ε (Fig. 1-3).

Table 5: NLAR Lena compression.

ε	r_ε entropy	r_ε variance	r_ε intensity range	Lena_ε intensity range
0	4.77	116.61	200	188
1	4.36	101.99	193	190
2	4.09	90.48	185	192
3	3.87	80.71	184	194
4	3.68	72.67	178	194
5	3.54	65.42	176	194

One can observe that increased ε reduces the entropy, variance and intensity range of the AR residual r, used to encode and store the Lena image (original entropy 7.28). However, it does not reduce the intensity range of the Lena image. Figure 44 shows how the entropy of the NLAR-compressed Lena image changes with respect to ε. We

also compared the performance of our technique to lossy JPEG compression. The amount of loss in JPEG was chosen to produce small values of ε, and the Lena image was compressed with NLAR and JPEG. The results are summarized in the Table 6. One can observe that for small ε, when nearly-lossless compression is needed, NLAR compression greatly outperforms JPEG.

Table 6: NLAR vs. JPEG.

ε	NLAR entropy	JPEG entropy	NLAR improvement
1	4.36	5.20	20 %
2	4.09	4.53	11 %
3	3.87	4.05	5 %
4	3.68	3.65	-1 %

Speed of Convergence

The number of iterations needed for convergence of the NLAR image refinement algorithm was determined. This determination used values of $\varepsilon = 1$, 3 and 9, and CT images of human brains. The results are shown on Figure 45. One can see that for small ε only 2-3 iterative steps are required to reach convergence. Moreover, our proposed technique only prepares the image of interest for more efficient lossless AR compression, i.e., the image has to be modified only once to improve its compression properties.

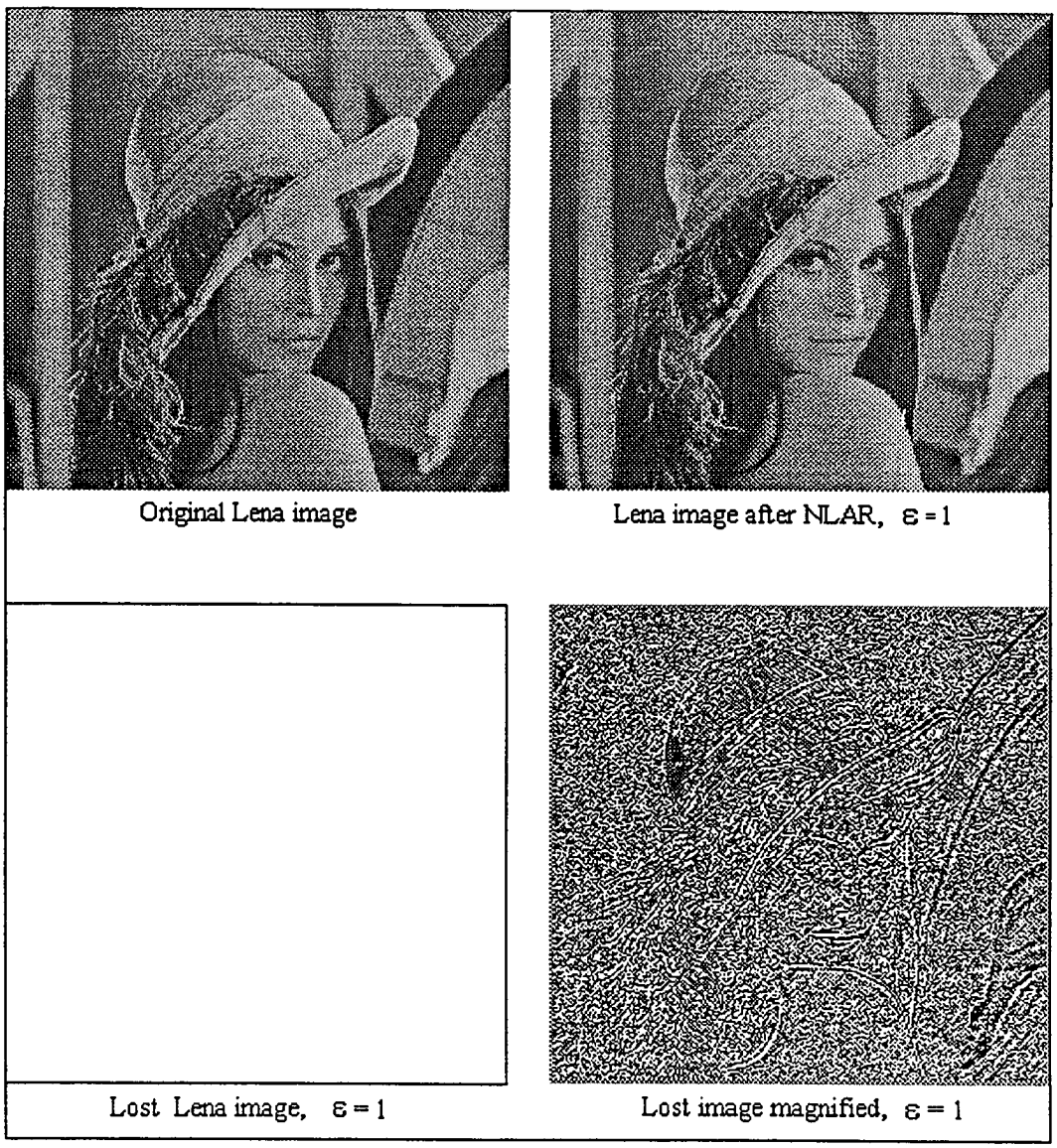

Figure 42: NLAR compression, $\varepsilon = 1$.

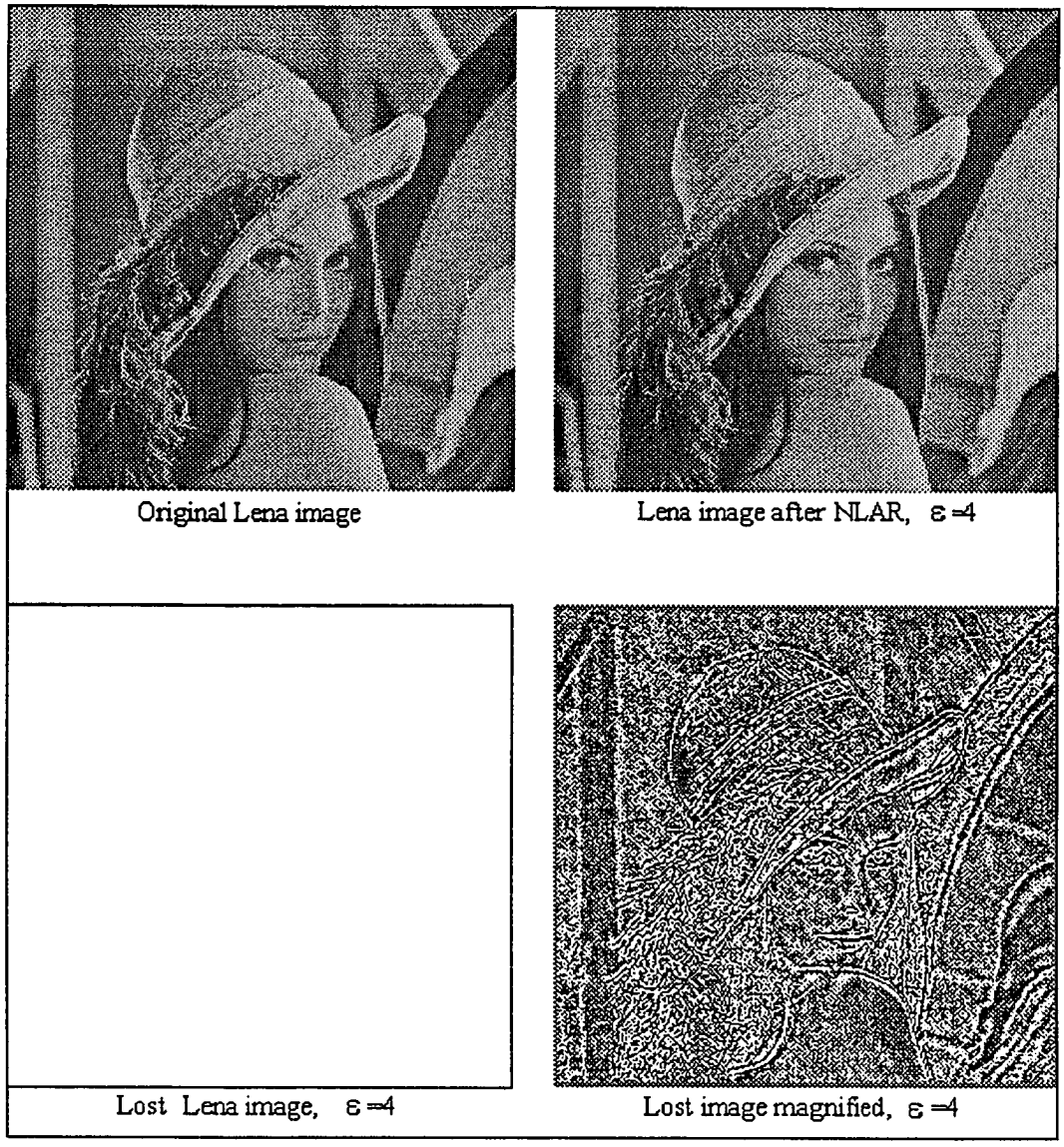

Figure 43: NLAR compression, $\varepsilon = 4$.

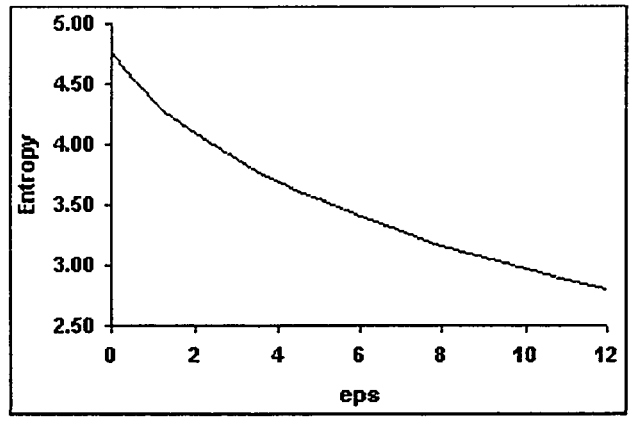

Figure 44: Nearly-lossless Lena compression.

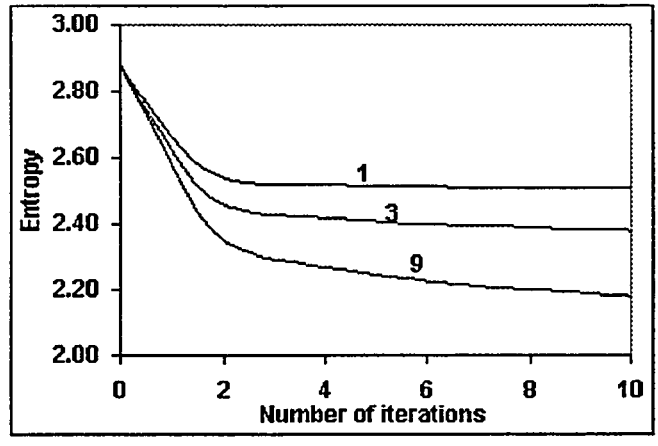

Figure 45: Convergence of NLAR.

Conclusion

The principal advantages of the proposed nearly-lossless compression technique are :

1. Preserving the original intensity range of the image. Instead of decreasing the image entropy, in general it makes the image more "compressible" with respect to the chosen compression transform.

2. More accurate control on lost information. For relatively small ε, it ensures that the changes made in the image cannot be perceived.

3. Flexible control on lost information. The parameter ε can easily be made region or intensity dependent as $\varepsilon = \varepsilon[i, j]$. ε may equal 0 for the most important regions of the image where nothing can be lost, and be greater than 0 for less important regions, e.g., background.

Thus, the augmented accuracy and flexibility of this technique makes it superior with respect to traditional nearly lossless compression. Finally, this algorithm can be extended to virtually any image compressing transform such as FFT, DCT and wavelet-based.

Scaling (resizing) images can influence set compression strategies. In particular, uniform image scaling to smaller sizes will typically result in both increased inter-image and decreased intra-image similarities. This is illustrated in Figures 46 and 47, with CT and MR images at different scales from the original (256 × 256) pixels to the smallest (4 × 4) pixels.

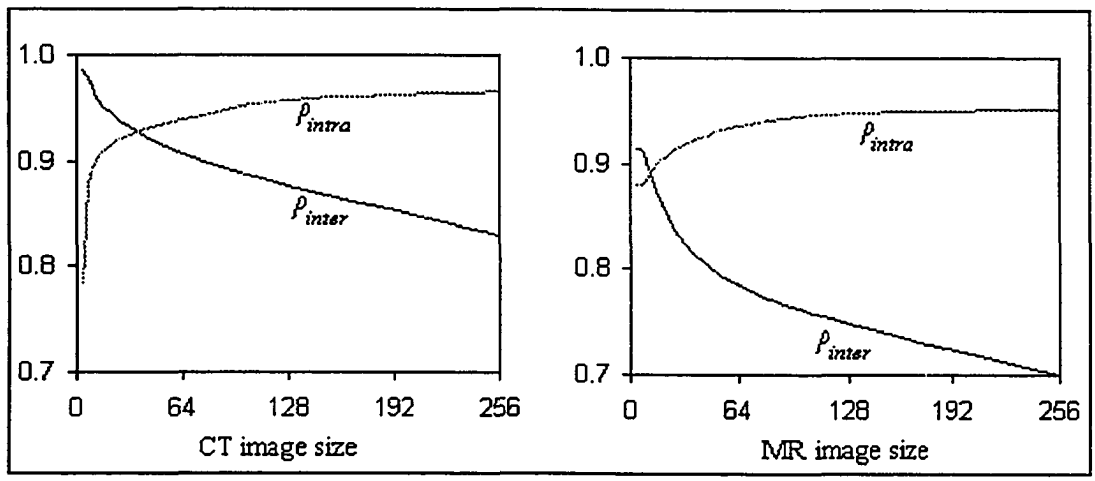

Figure 46: Changes in ρ for different image sizes.

One can observe that resizing the images to smaller uniform scales leads to the inter-image similarity becoming dominate. The reduction to smaller sizes, especially with the interpolation smoothing of intensity reduces the noise and local details, which are responsible for low inter-image correlation and high image difference entropy. Smaller scaled images preserve only the most general features of the similar image set and the inter-image similarity increases.

Figure 47 shows how the inter-image similarity becomes more important than intra-image similarity for CT images starting from approximately 32 × 32, and for

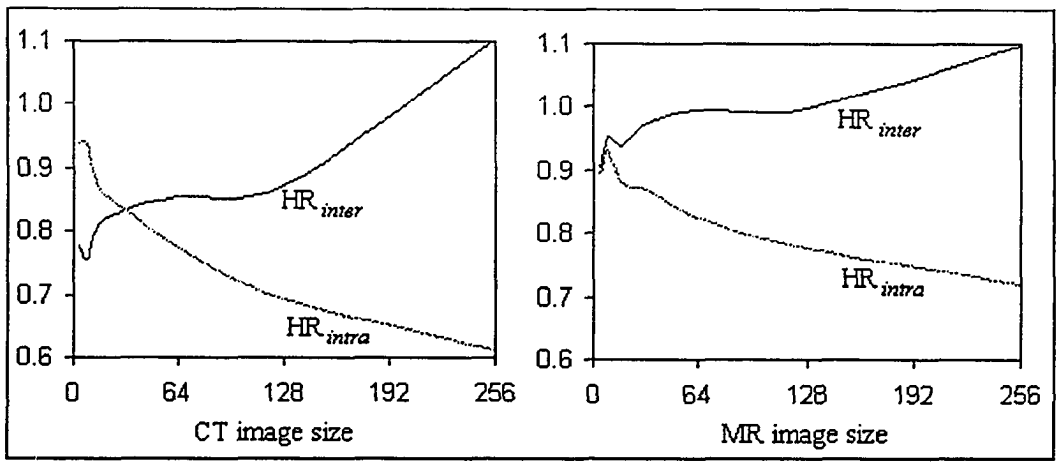

Figure 47: Changes in HR for different image sizes.

4×4 MR images. This implies that with small images inter-image prediction provides better compression when compared to common AR models. Conversely, there are multi-resolution compression techniques (e.g., wavelet compression) which use copies of the original image on reduced scales. In particular, for an image u, one pass of a 2-dimensional wavelet transform will produce four images $\{u_{ll}, u_{lh}, u_{hl}, u_{hh}\}$, each half of the original size of u, where u_{ll} is obtained from u with low-frequency filters in the x and y, u_{lh} - with low-frequency in x and high frequency in y directions, etc. For example, using the simplest 2-dimensional Haar wavelet to reversibly decompose[31] u yields:

[31] Despite integer truncation, u can still be uniquely recovered from $\{u_{ll}, u_{lh}, u_{hl}, u_{hh}\}$.

$$u \longrightarrow \begin{cases} u_{ll}[x,y] = \frac{1}{4}\lfloor u[2x,2y] + u[2x,2y+1] + u[2x+1,2y] + u[2x+1,2y+1] \rfloor \\ u_{lh}[x,y] = \frac{1}{2}\lfloor u[2x,2y] + u[2x,2y+1] \rfloor - \frac{1}{2}\lfloor u[2x+1,2y] + u[2x+1,2y+1] \rfloor \\ u_{hl}[x,y] = \frac{1}{2}\lfloor u[2x,2y] + u[2x+1,2y] \rfloor - \frac{1}{2}\lfloor u[2x,2y+1] + u[2x+1,2y+1] \rfloor \\ u_{hh}[x,y] = u[2x,2y] - u[2x,2y+1] - u[2x+1,2y] + u[2x+1,2y+1] \end{cases} \tag{43}$$

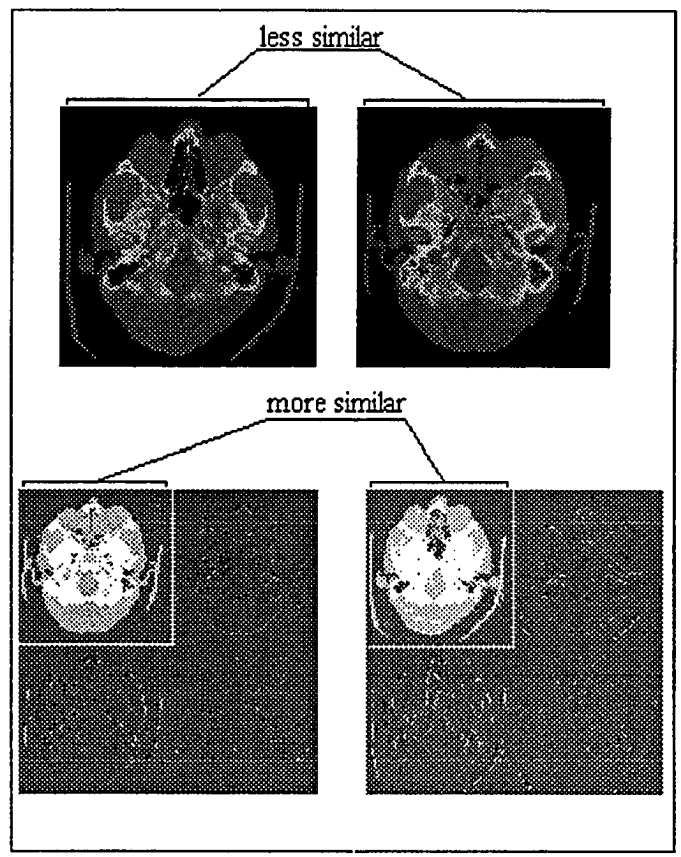

Figure 48: Increasing similarity in wavelet transform.

Subimages u_{lh}, u_{hl} and u_{hh} represent image-specific edges along the y, x and $x = y$ axes respectively. These images are very hard to forecast or compress with any inter-image predictive model. Conversely, u_{ll} is the "smoothed" u on a reduced scale, and

contains the most general features of u. As Figure 46 shows, this smoothed scale reduction will tend to make similar images even more similar, i.e., $\rho(u_{ll}, v_{ll}) > \rho(u, v)$ (Figure 48). On the other hand, intra-image correlation in u_{ll} will be smaller when compared to u, making u_{ll} often harder to compress with wavelet compression[32] (43). Therefore, it becomes natural to introduce difference set compression into low-frequency, low-scale passes of the wavelet compression. Such hybrid compression algorithm for predictive wavelet set compression (PWSC) for two similar images u and v can be outlined as follows:

1. **Apply** one pass of wavelet transform $W()$ to each similar image.
2. **Compute** entropies $H(u - v)$ and $H(\{u_{ll}, u_{lh}, u_{hl}, u_{hh}\})$
3. **If** $\min(H(u - v), H(\{u_{ll}, u_{lh}, u_{hl}, u_{hh}\})) > H(u)$ **Then Stop**
4. **If** $(H(u - v) < H(\{u_{ll}, u_{lh}, u_{hl}, u_{hh}\}))$

   a.**Then** $u = u - v$; /* *difference compression* */

   b.**Else** $u = u_{ll}$; $v = v_{ll}$; /* *wavelet compression* */
5. **Goto** 1.

At each step, this adaptive algorithm (line 4.) will choose between wavelet and difference compression which one provides the smallest compressed image entropy. Then (line 5) it will iteratively reapply this procedure until the condition in line 3 is satisfied. This condition occurs when neither wavelet nor difference compression can produce further entropy reduction. Before this condition is met, the algorithm will

[32]Note that (43) is essentially an intra-image (autoregressive) predictor, since it attempts to approximate each pixel value with the values of its three neighbors.

continue to reduce each similar (sub)image into four, trying to improve wavelet compression of the low-frequency (sub)images with difference inter-image compression.

We tested this technique with CT images. Simple difference compression does not work for this data and increases the total set entropy. This proposed hybrid approach improved the wavelet compression ratio for a pair of similar images by 5%, which is a good result for the difficult to compress CT data (Table 7).

Table 7: PWSC compression.

Entropy	u=CT1	v=CT2	Total entropy
Original images	5.94	5.52	11.46
After wavelet compression	3.98	3.81	7.79 (100%)
After PWSC	3.78	3.66	7.44 (95%)

To store a similar database in PWSC compressed form, one have to:

1. Store a few low-scale predictor (reference) images.

2. Store the low-scale copy of any other similar image as its predicted (from the reference set) residual.

3. Store high-frequency $\{u_{lh}, u_{hl}, u_{hh}\}$ images for each image u.

To restore an image u, its low-frequency component u_{ll} is recovered from the residual and the reference images first. Then the full u is recovered with inverse wavelet transform from $\{u_{ll}, u_{lh}, u_{hl}, u_{hh}\}$. Since this is lossless in both wavelet and set difference parts, the resulting compression scheme is lossless.

The lossless predictive set compression was analyzed numerically and theoretically. The numerical analysis was based on two classes of similar images: MR and CT human brain scans. These images are difficult to compress with any of the previously suggested approaches. The origins of this difficulty were determined, and this lead to a better and more reliable CAR set compression.

The theoretical analysis of the predictive set compression lead to the study of the binary and 4-cluster models. These models were introduced as good approximations to the observed inter-image bivariate intensity distributions, and lead to mathematically accurate best, worst and average case estimates for the set compression ratio $C(\rho)$ as a function of inter-image correlation ρ. These estimates can be used to evaluate any existing set compression technique or algorithm, as well as for further theoretical analysis.

Finally, some extensions to the lossless compression were introduced. NLAR models allow improvement in image and set compression properties for the given AR model. PWSC compression extends the predictive compression to the sets where originally it did not work well, and naturally links it with the wavelet compression. Both techniques become beneficial for the sets of images where the high level of noise or local details prohibit the straightforward application of the predictive set compression.

Detailed conclusive remarks were also given in the end of each chapter.

References

[1] C.E. Shannon, A Mathematical Theory of Communication, Bell Systems Technical Journal, Vol. 27, 1048, pp.379-423, 623-656.

[2] R.J. McEliece, The Theory of Information and Coding, Addison-Wesley Publishing Company, 1977, pp. 15-69.

[3] R.E.Blahut, Principles and Practice of Information Theory, Addison-Wesley Publishing Company, 1987, pp. 47-95.

[4] J.Aczel, Z.Daroczy, On Measures of Information and Their Characterizations, Academic Press, New York, 1975, pp.26-45

[5] V.K. Madisetti, D.B. Williams, The Digital Signal Processing Handbook, IEEE Press, 1998, pp.51-1 - 53-25

[6] N. Moayeri, A Near-Lossless Trellis-Searched Predictive Image Compression System, Proceedings of the 8th Annual Meeting of the IEEE Engineering in Medicine and Biology Society, 1986, pp.93-95.

[7] D. A. Ortendahl, The Application of Principal Component Analysis to Multivariate MRI Data, IEEE 8th Annual Conference of the Engineering in Medicine and Biology Society, pp.1065-1067.

[8] H. Grahn, N.M.Szeverenyi, M.W. Roggenbuck, F. Delaglio and P.Geladi, Data Analysis of Multivariate Magnetic Resonance Images: A Principal Component Approach, Chemometrics and Intelligent Laboratory Systems, 5(1989) 311-322.

[9] Y. W. Nijim, S. D. Stearns and W. B. Michael, Differentiation Applied to Lossless Compression of Medical Images, IEEE Transactions on Medical Imaging, Vol. 15, No. 4, August 1996, pp.555-559.

[10] J. B. Farison, Y. Park, Q. Yu and H. Lu, KL transformation of spatially-invariant image sequences, SPIE conference February'97, contact jfarison@eng.utoledo.edu.

[11] J.S. Taur and C.W. Tao, Medical Image Compression Using Principal Component Analysis, IEEE 8th Annual Meeting of the Engineering in Medicine and Biology Society, 1986, pp.903-905.

[12] K. Chen and T.V. Ramabadran, Near-Lossless Compression of Medical Images Through Entropy-Coded DPCM, IEEE Transactions on Medical Imaging, Vol.13, No.3, September 1994.

[13] H. Soltanian-Zadeh, Optimal Linear Transformation for MRI Feature Extraction, IEEE Transactions on Medical Imaging, Vol. 15, No. 6, December 1996, pp.749-767.

[14] H. Lee, Y. Kim, A.H. Rowberg and E.A. Riskin, Statistical Distributions of DCT Coefficients and Their Application to an Interframe Compression Algorithm for 3D Medical Images, IEEE Transactions on Medical Imaging, Vol.12, No.3, September 1993.

[15] R. Sharman, J. M. Tyler and O. Pianykh, Wavelet-based registration and compression of sets of images, SPIE Conference on AeroSense, Orlando, Florida,20-25 April 1997

[16] R. A. Johnson, D. W. Wichern, "Applied Multivariate Statistical Analysis", Prentice Hall, 1995

[17] H. Lee, Y. Kim, A.H. Rowberg and E.A. Riskin, Statistical Distributions of DCT Coefficients and Their Application to an Interframe Compression Algorithm for 3D Medical Images, IEEE Transactions on Medical Imaging, Vol.12, No. 3, September 1993, pp.478-485

[18] G.R. Kuduvalii and R.M. Rangayyan, Performance Analysis of Reversible Image Compression Techniques for High-Resolution Digital Teleradiology, IEEE Transactions on Medical Imaging, Vol.11, No. 3, September 1992, pp.430-445

[19] T.V. Ramabadran and K. Chen, The Use of Contextual Information in the Reversible Compression of Medical Images, IEEE Transactions on Medical Imaging, Vol.11, No. 2, June 1992, pp.185-195.

[20] A. Ramaswamy and W.B. Mikhael, A Mixed Transform Approach for Efficient compression of Medical Images, IEEE Transactions on Medical Imaging, Vol.15, No. 3, June 1996, pp. 343-352.

[21] S. Wong, L. Zaremba and H.K. Huang, Radiologic Image Compression - A Review, Proceedings of the IEEE, Vol. 83, No. 2, February 1995, pp. 194-218.

[22] H.G. Musmann, Predictive Image Coding, W.K.Pratt, Ed., in Image Transmission Techniques, Advanced Electronics electron Physics. Orlando, FL: Academic, 1979, pp.73-112.

[23] H. Murase and M. Lindenbaum, Partial Eigenvalue Decomposition of Large Images Using Spatial Temporal Adaptive Method, IEEE Transactions on Medical Imaging, Vol.4, No. 5, May 1995, pp.620-629.

[24] E. A. Gifford, B. R. Hunt and M. W. Marcellin, Image Coding Using Adaptive Recursive Interpolative DPCM, IEEE Transactions on Medical Imaging, Vol.4, No. 8, August 1995, pp. 1061-1069.

[25] R. Sharman, J. M. Tyler and O. Pianykh, Wavelet-based registration and compression of sets of images, SPIE Conference on AeroSense, Orlando, Florida,20-25 April 1997

[26] G.A.Korn, T.M.Korn, Mathematical Handbook For Scientists And Engineers, Section 18, McGraw-Hill Book Company, 1968

[27] M. Rabbani and P.W. Jones, "Digital Image Compression Techniques", *SPIE Optical Engineering Press,* 1991.

[28] A. Nosratinia, N. Mohsenian, M.T.Orchard and B. Liu, "Interframe Coding of Magnetic Resonance Images", *IEEE Transactions on Medical Imaging*, Vol.15, No.5, October 1996.

[29] M. Das and S. Burgett, "Lossless Compression of Medical Images Using Two-Dimensional Multiplicative Autoregressive Models", *IEEE Transactions on Medical Imaging*, Vol.12, No.4, December 1993.

[30] M. B.Priestley, "Non-linear and Non-stationary Time Series Analysis", Academic Press, 1987.

[31] S. Wang, L. Zaremba, D. Gooden and H.K. Huahg, "Radiologic Image Compression - A Review", *Proceedings of the IEEE*, Vol. 83, No. 2, February 1995.

[32] P.Roos and M.A. Viergever, "Reversible Interframe Compression of Medical Images: A Comparison of Decorrelation Methods", *IEEE Transactions on Medical Imaging*, Vol.10, No. 4, Dec. 1991.

[33] P.Roos, M.A. Viergever, M.C. Van Dijke and J.H. Peters, "Reversible Interframe Compression of Medical Images", *IEEE Transactions on Medical Imaging*, Vol.7, No. 4, Dec. 1988.

[34] J. D. Hamilton, "Time Series Analysis", Princeton University Press, 1994.

[35] O. S. Pianykh, J. M. Tyler and R. Sharman, "Autoregressive Models for Compressing Similar Data", *SPIE Conference on AeroSense*, Orlando, Florida,13-17 April 1998.

[36] John Neter, Michael H. Kutner, Christopher J. Nachtsheim and William Wasserman, Applied Linear Statistical Models, Irwin, 1996, pp.217-353

[37] Karray, L., Rioul, O. and P.Duhamel (1994). L^{∞}-Coding of Images: A Confidence Interval Criterion, Proceedings ICIP, Vol. 2, pp. 888-892, Nov. 1994

[38] Chen, K. and T.V. Ramabadran (1994). Near-Lossless Compression of Medical Images Through Entropy-Coded DPCM, IEEE Transactions on Medical Imaging, Vol.13, No. 3, September 1994.

[39] Ke, L. and M.W.Marcellin (1995). Near-Lossless Image Compression: Minimum-Entropy, Constrained-Error DPCM, IEEE International Conference on Image Processing, Washington, D.C, October 1995 (complete paper is in preparation for publication in IEEE Transactions on Image Processing)

[40] P.C.Cosman, H.C. Davidson, C.J.Bergin, C-W. Tseng, L.E.Moses, E.A.Riskin, R.A.Olshen and R.M.Gray (1994). Thoracic CT Images: Effect of Lossy Image Compression on Diagnostic Accuracy, Radiology, Vol. 190, pp.517-524, February 1994

[41] Mathews, V.J. and P.J. Hahn (1995). Vector Quantization Using the L_∞ Distortion Measure, IEEE International Conference on Image Processing, Washington, D.C., October 23-26, 1995.

[42] H. Lee, Y. Kim, E.A. Riskin, A.H. Rowberg and M.S. Frank (1995). A Predictive Classified Vector Quantizer and Its Subjective Quality Evaluation for X-Ray CT Images, IEEE Transactions on Medical Imaging, Vol. 14, No. 2, pp. 397-406, June 1995.

[43] Ran, X. and N. Farvardin (1995). A Perceptually Motivated Three-Component Image Model (Parts I and II), IEEE Transactions on Image Processing, Vol. 4, No. 4, pp.401-415 and pp. 430-447, April 1995.

[44] Karunasekera, S.A and N.G.Kingsbury (1995), A Distortion Measure for Blocking Artifacts in Images Based on Human Visual Sensitivity, IEEE Transactions on Image Processing, Vol. 4, No. 6, pp.713-724, June 1995.

[45] Langdon, G.G. and C.A.Haidinyak (1995). Experiments with Lossless and Virtually Lossless Image Compression Algorithms, Proceedings of SPIE, Still Image Compression, Vol. 2418, pp. 21-27, Feb. 1995.

VITA

Oleg S. Pianykh was born in Gorky, Russia, on November 8, 1968. He received his equivalent of a bachelor's degree and a master of science degree in applied mathematics and physics from Moscow State University, Moscow, Russia, in 1994. Currently he is finishing his doctoral research in the Computer Science Department, Louisiana State University. His major research interests include realistic image synthesis, compression and analysis of similar images, simulation and artificial intelligence. He will receive the degree of Doctor of Philosophy in August, 1998.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Oleg S. Pianykh

Major Field: Computer Science

Title of Dissertation: Lossless Set Compression of Correlated Information

Approved:

John M. Tyler
Major Professor and Chairman

Dean of the Graduate School

EXAMINING COMMITTEE:

Warren N Waggenspack

J. Hürrelbrink

Date of Examination:

May 25, 1998

IMAGE EVALUATION TEST TARGET (QA-3)

150mm

6″

APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989