Performance of the Kullback-Leibler information gain for predicting image fidelity

J.A. García, J. Fdez-Valdivia, R. Rodríguez-Sánchez  
Depto. Ciencias de la Computación e I.A.  
E.T.S. de Ingeniería Informática.  
Univ. de Granada. 18071 Granada. Spain  
{J.Fdez-Valdivia, jags, rosa}@decsai.ugr.es

Xose R. Fdez-Vidal  
Depto. Física Aplicada. Facultad de Física  
Univ. de. Santiago de Compostela.  
15706 Santiago de Compostela. Spain  
faxose@usc.es

Abstract

This paper presents a new method for characterizing information of a compressed image relative to the original one. We show how the Kullback-Leibler information gain is based on three basic postulates which are natural for image processing and thus desirable. As an example of the proposed measure, we analyze the effects of lossy compression on the identification of breast cancer microcalcifications. We also show the comparative results of the Kullback-Leibler information gain and various quantitative measures for predicting image fidelity in the sense of diagnostic usefulness.

1 Introduction

The problem of evaluating the quality of a compressed image in specific applications may be solved using a suitable rate-distortion measure. In practice, however, this procedure may be difficult to apply for two reasons: 1) The lack of knowledge of what distortion measure is more suitable for images [3, 10], and 2) the characteristics of the human visual system are not well understood, yet [1, 13], so the problem of finding a general enough measure of perceptual quality has proven to be an elusive goal. As an example, consider compression of digitized mammograms (most of the authors digitize them with a spatial resolution of 0.1 or 0.05 mm producing a huge amount of data). Radiologists look for certain signs and characteristics indicative of cancer when evaluating a mammogram. Among these signs is the presence of clustered microcalcifications. Individual breast cancer microcalcifications appear as small objects of variable shape with a roughly circular average form, and therefore compression artifacts as blurring or blocking could affect to the number or extension of these microcalcifications. In such scenario, the important point is how to find an optimal compression level at which to trade-off between compression ratio and source fidelity in the sense of diagnostic usefulness. Optimization of this trade-off between image fidelity and coding rate requires an ability to measure distortion [9, 11]. However, as we above stated, the perceived distortion in visual content is a very difficult quantity to measure, as the characteristics of the human visual system are complex and not well understood. In practice, highly imperfect distortion models such as the sum of squared differences or its equivalents, known as mean squared error (MSE) or peak signal-to-noise ratio (PSNR) are used in most actual evaluations.

To circumvent the lack of knowledge of what distortion measures are more suitable for images, here we propose to use a Kullback-Leibler information gain [8]. This measure of information gain is to be characterized in Section 2 with a minimal number of properties which are natural for image processing and thus desirable. This section also determines the form of all error functions that satisfy these properties which we have stated to be desirable.

As an example of the proposed measure, several experiments are performed to investigate the comparative results of the Kullback-Leibler information gain and various quantitative distortion measures for predicting image fidelity in the sense of diagnostic usefulness in digital mammograms (Section 3). The main conclusions of this paper are summarized in Section 4.

2 Axiomatic Characterization of Relative Information

Let $X_1, X_2, \ldots, X_N$ be a sequence of $N$ symbols from the set of gray levels $G = \{1, 2, \ldots, |G|\}$. A 2D digital image can be interpreted as a sequence $X_1, X_2, \ldots, X_N$ of $N$ symbols, with $X_i$ being the gray level at pixel $i$. We use the notation $\mathbf{x}$ to denote a particular sequence of gray levels $x_1, \ldots, x_N$.  

1051-4651/02 $17.00 (c) 2002$ IEEE
Let \( I \) be the original image; \( I_{q(i)} \) be the reconstruction of the original image at compression ratio \( q(i) \); \( P \) and \( P_{q(i)} \) be the discrete probability distributions, with \( P = (p(l/I))_l \) and \( P_{q(i)} = (p(l/I_{q(i)}))_l \), that characterize the probability of occurrence of each gray level \( l \) in the original image \( I \) and the reconstruction \( I_{q(i)} \), respectively.

Then \( P_{q(1)}, P_{q(2)}, \ldots P_{q(i)}, \ldots P_{q(K)} \) denote the discrete probability distributions for the reconstructions of the original image at compression ratios \( q(1), \ldots q(i), \ldots q(K) \), respectively; where if \( i \leq j \) then, \( q(i) \leq q(j) \); that is, \( q(1) \) is the lowest compression ratio and \( q(K) \) is the highest compression ratio.

As stated in [4], the Kullback-Leibler information gain plays a key role in obtaining the best achievable level of compression and thus desirable.

The first postulate states a property of how unexpected a single event of a digital image was.

**Postulate 1.** A measure \( \mathcal{U} \) of how unexpected the single event "gray level \( l \) occurs" was, depends only upon its probability \( p \).

This means that there exists a function \( h \) defined in \([0, 1]\) such that

\[
\mathcal{U}(\text{"gray level } l \text{ occurs"}) = h(p). \tag{1}
\]

This is a natural property because it often happens that the structure of a certain scene cannot be determined exactly due to various reasons (e.g., it is possible that some of the details may not be observable or the observer who makes an attempt to investigate the structure may no take all the relevant factors governing the structure into consideration). Under such circumstances, the structure of an image can be characterized statistically by a discrete probability distribution.

Our second postulate is formulated to obtain a reasonable estimate of how unexpected a digital image was from some probability distribution by means of the mathematical expectation of how unexpected its single events were from this distribution.

Let \( p(l/R) \) and \( p(l/I) \) be the probability of occurrence of gray level \( l \) in a reference image \( R \) and another of input \( I \), respectively. Suppose that every possible observation from \( p(l/R) \) is also a possible observation from \( p(l/I) \). This is to avoid the contingency that \( p(l/R) \neq 0 \) and \( p(l/I) = 0 \).

As stated in Postulate 1, if the single events of the reference image \( R \) are characterized by an "estimated" distribution \( Q = \{ p(l/I) \mid l = 0, 1, \ldots, n \} \), then the function \( h(p(l/I)) \), with \( l = 0, 1, \ldots, n \), returns a measure of how unexpected each single event "gray level \( l \) occurs" was from \( Q \). Thus, assuming that \( P = \{ p(l/R) \mid l = 0, 1, \ldots n \} \) is the "true" probability distribution of the reference image \( R \), we have that:

**Postulate 2.** The mathematical expectation of the discrete random variable \( h(Q) \), which can assume the values

\[
h(p(l/I)) \quad h(p(\bar{l}/I)) \ldots h(p(\bar{n}/I))
\]

with respective probabilities

\[
p(l/R) \cdot p(n/R)
\]

is an estimate \( \mathcal{U}_P(Q) \) of how unexpected the reference image \( R \) was from \( Q = \{ p(l/I) \} \), i.e.,

\[
\mathcal{U}_P(Q) = E_P[ h(Q) ] = \sum_{l} p(l/R) h(p(l/I)) \tag{2}
\]

with \( E_P \) denoting the mathematical expectation in \( P \).

The following postulate relates the estimate of how unexpected the reference image was from an "estimated" distribution \( Q \) and the estimate from the "true" distribution \( P \).

**Postulate 3.** The reference image \( R \) with "true" probability distribution \( P \) is more unexpected from an "estimated" distribution \( Q \) than from the "true" distribution \( P \).

The following inequality expresses how the reference image is more unexpected when it is characterized by \( Q \) than when is characterized by \( P \):

\[
\mathcal{U}_P(Q) \geq \mathcal{U}_P(P). \tag{3}
\]

with \( \mathcal{U}_P(Q) \) and \( \mathcal{U}_P(P) \) being estimates of how unexpected the reference image was from the "estimated" distribution \( Q \) and from the "true" distribution \( P \), respectively.

The true distribution \( Q \) of the input image \( I \) may be interpreted as an estimated distribution of the reference image \( R \) (with "true" distribution \( P \)). Thus, we can define a measure of information gain of the reference image from the input image by the difference between the estimate of how unexpected the reference image was from \( Q \) and that from \( P \).

**Definition 4: A measure of information gain.** Given the reference image \( R \) with "true" probability distribution \( P = \{ p(l/R) \} \), a measure of the information gain of the reference image \( R \) from the input image \( I \) with "true" distribution \( Q = \{ p(l/I) \} \) is:

\[
\mathcal{E}(P,Q) = \mathcal{U}_P(Q) - \mathcal{U}_P(P) \tag{4}
\]

with \( \mathcal{U}_P(Q) \) and \( \mathcal{U}_P(P) \) being estimates of how unexpected the reference image was from \( Q \) and \( P \), respectively. \( \mathcal{U}_P(Q) \) and \( \mathcal{U}_P(P) \) are defined as given in Postulate 2, and such that satisfy the inequality (3) in Postulate 3.

The following result serves to determine the form of the measure \( \mathcal{E}(P,Q) \).
Theorem 2. Let $\mathcal{E}(PQ)$ be a measure of information gain for the discrimination between two images as given in Definition 4, i.e.,

$$\mathcal{E}(PQ) = \mathcal{U}(Q) - \mathcal{U}(P)$$

with $P = \{ p(\ell|I) \}$ and $Q = \{ q(\ell|I) \}$. Then, the measure of relative information $\mathcal{E}$ is equal to the Kullback-Leibler’s information gain\(^\text{[8]}\) between $P$ and $Q$ up to a nonnegative multiplicative constant, i.e.,

$$\mathcal{E}(PQ) = a \sum_\ell p(\ell|I) \log \frac{p(\ell|I)}{q(\ell|I)} . \quad (5)$$

with $a \geq 0$ constant; the base of the logarithm is 2.

Proof. See [4].

In conclusion, any error function such that satisfies Postulates 1, 2, and 3, has to be of the form of the Kullback-Leibler information function up to a nonnegative multiplicative constant.

3 Performance of the Kullback-Leibler Information Gain for Predicting Image Fidelity

This section analyzes the effects of lossy compression on a specific application of image feature extraction. The coding scheme will be the international standard JPEG [6]. And the particular problem of feature extraction here discussed is the identification of breast cancer microcalcifications. Thus the point is to evaluate whether lossy JPEG coding at the different compression ratios \(q(i)\) can be used for the specific problem of detection of individual microcalcifications.

Digitized mammograms from the Nijmegen database [7] and the MIAS database [12] were used to investigate this point. Figure 1 shows the 10 digitized mammographic images used in this experiment. These ten images contain some clustered microcalcifications. From each selected image we obtained the region 256 x 256 in size where the microcalcifications are present. Clusters of fine, granular microcalcifications in mammograms may be an early sign of disease. Individual grains are difficult to detect and segment due to size and shape variability and because the background mammogram texture is typically inhomogeneous. The visibility of the clusters is highly variable and is often degraded by the high frequency texture of breast tissue (e.g., the fine vasculature of the breast). Consultant radiologists review the original images to detect microcalcifications.

All mammograms in the MIAS database were digitized with a scanning microdensitometer (Scandig3) to 50x50 micron resolution. These images were reduced to 200 micron pixel edge and clipped/padded, so that every image is 8 bits, 1024x1024 pixels. Images in the Nijmegen database are in raw format and are 2048x2048 in size. In this database, images were digitized from film with an Eikonix 1412 12 bits CCD camera. A sampling aperture of 0.05 mm in diameter and a 0.1 mm sampling distance were used. No preprocessing step was carried out on the images.

The method for characterizing compression losses using images reconstructed under various degrees of lossy compression is the false positive to true positive ratio for microcalcification detection. In order to derive the false positive to true positive detection ratio, the images reconstructed after compression were also reviewed to detect individual microcalcifications. Since it is rated diagnostic usefulness rather than general appearance or simply line or edge patterns, this study relates diagnostic accuracy to compression level.

Fig. 2 shows one example illustrative of this method for characterizing compression losses. The original image is given in this figure. Various images reconstructed under different degrees of lossy compression are also illustrated in Fig. 3. Fig. 2 shows both the curve of true positive (TP) and the curve of false positive (FP) for individual calcification detection. This same figure illustrates the false positive
Figure 2. Method for characterizing compression losses: the false positive to true positive (FP/TP) ratio for the sequence of images at compression ratios \( q(1), \ldots, q(99) \).

We can observe that when the diagnostic usefulness of the reconstructed images significantly decreases (see Figure 3), the false positive to true positive ratio also shows a significant increase, as illustrated in Fig. 2. This means that the false positive to true positive ratio evaluation of reconstructed images is sufficient for evaluations that take into account diagnostic usefulness.

Now we show the performance of the Kullback-Leibler information gain for predicting image fidelity in terms of diagnostic usefulness. This will be evaluated by making use of the degree of correlation between the false positive to true positive (FP/TP) ratio and the Kullback-Leibler gain. The information gain is applied to quantify the image distortion by means of the difference between the original mammographic image and the images reconstructed after compression.

The correlation coefficient is given by:

\[
\text{corr}(y, z) = \frac{\sum_{i=1}^{99} (y_i - \bar{y})(z_i - \bar{z})}{\left[ \sum_{i=1}^{99} (y_i - \bar{y})^2 \sum_{i=1}^{99} (z_i - \bar{z})^2 \right]^{1/2}},
\]

where \( y_i \) is the Kullback-Leibler information gain.

Figure 3. Images reconstructed under various degrees of lossy compression

\[
KLI(q(i)) \text{ between the discrete probability distribution } P \text{ of the original image and the distribution } P_{q(i)} \text{ of the image reconstructed at level of compression } q(i); \ z_i \text{ is the value of the FP/TP ratio for the image reconstructed at } q(i); \text{ and } \mu_y = \sum_{i=1}^{99} y_i/99 \text{, } \mu_z = \sum_{i=1}^{99} z_i/99.
\]

In this section we also explore the comparative results of the Kullback-Leibler gain and two quantitative measures: the root mean square error (RMS), and the peak-signal-to-noise ratio (PSNR) (the distortion measures often used in R-D optimization) defined as:

\[
RMS(R, I) = \sqrt{\frac{1}{n \times m} \sum_{x=1}^{n} \sum_{y=1}^{m} (R(x, y) - I(x, y))^2}.
\]

\[
PSNR(R, I) = 10 \log_{10} \left( \frac{\max[R(x, y)] - \min[R(x, y)]}{\frac{1}{n \times m} \sum_{x=1}^{n} \sum_{y=1}^{m} (R(x, y) - I(x, y))^2} \right).
\]

where \( R(x, y) \) represents the original mammographic image and \( I(x, y) \) represents the image reconstructed after lossy compression.

The comparative results of the Kullback-Leibler information gain and those of RMS, and PSNR are illustrated in Table 1. For each mammographic image, the correlation coefficient \( \text{corr}(y, z) \) between a particular distortion measure and the FP/TP ratio is used to study the efficacy of
distribution is the best estimator of the population distribution, i.e. if the sample contains $N$ data $x_1, x_2, \cdots, x_N$ then the maximum likelihood estimator of the population distribution is found by assigning to each $x_j$ the probability $1/N$. Given the original sample $S$ of size 10 shown in Fig. 1, to get a sampling distribution for $\rho(y, z)$ we simply run the following procedure: [2]

**Procedure: Bootstrap sampling.**

Repeat $i = 1, 2, \ldots, 100$ times:

- Draw a bootstrap pseudosample $S_i$ of size 10 from $S$ by sampling with replacement as follows:
  - Repeat 10 times: Select a member of $S$ at random and add it to $S_i$.
  - Calculate and record the value of $\rho(y, z)$ for $S_i$.

Once we have the bootstrap sampling distribution of $\rho(y, z)$, the important point is to produce good confidence intervals for $\rho(y, z)$ automatically. “Good” means that the bootstrap intervals should closely match exact confidence intervals in those special situations where statistical theory yields an exact answer, and should give dependably accurate coverage probabilities in all situations. The most advanced bootstrap intervals are the $BC_a$ intervals and they come close to this criteria of goodness. The construction of $BC_a$ intervals using the bootstrap sampling is described in [2].

Table 2 shows the 93% $BC_a$ confidence intervals for $\rho(y, z)$.

### Table 2. 93% $BC_a$ confidence intervals for $\rho(y, z)$

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho(KL, \frac{P}{TP})$</td>
<td>(0.823, 0.944)</td>
</tr>
<tr>
<td>$\rho(RMSE, \frac{P}{TP})$</td>
<td>(0.792, 0.855)</td>
</tr>
<tr>
<td>$\rho(PSNR, \frac{P}{TP})$</td>
<td>(0.701, 0.731)</td>
</tr>
</tbody>
</table>

From these confidence intervals, we also claim that the hypothesis $\rho(y, z) > 0.9$ is only accepted at 95% for the KL information gain.

Table 1. Comparative results of the Kullback-Leibler information gain and those of $RMSE$ and $PSNR$ (correlation values).

<table>
<thead>
<tr>
<th>Image</th>
<th>$KL, \frac{P}{TP}$</th>
<th>$RMSE, \frac{P}{TP}$</th>
<th>$PSNR, \frac{P}{TP}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>0.843</td>
<td>0.731</td>
<td>0.630</td>
</tr>
<tr>
<td>#2</td>
<td>0.951</td>
<td>0.922</td>
<td>0.854</td>
</tr>
<tr>
<td>#3</td>
<td>0.971</td>
<td>0.896</td>
<td>0.821</td>
</tr>
<tr>
<td>#4</td>
<td>0.886</td>
<td>0.826</td>
<td>0.736</td>
</tr>
<tr>
<td>#5</td>
<td>0.789</td>
<td>0.835</td>
<td>0.737</td>
</tr>
<tr>
<td>#6</td>
<td>0.818</td>
<td>0.750</td>
<td>0.652</td>
</tr>
<tr>
<td>#7</td>
<td>0.824</td>
<td>0.775</td>
<td>0.700</td>
</tr>
<tr>
<td>#8</td>
<td>0.848</td>
<td>0.723</td>
<td>0.623</td>
</tr>
<tr>
<td>#9</td>
<td>0.944</td>
<td>0.829</td>
<td>0.727</td>
</tr>
<tr>
<td>#10</td>
<td>0.823</td>
<td>0.792</td>
<td>0.701</td>
</tr>
<tr>
<td><strong>MEAN</strong></td>
<td><strong>0.870</strong></td>
<td><strong>0.808</strong></td>
<td><strong>0.718</strong></td>
</tr>
</tbody>
</table>

Each measure for predicting image distortion (in terms of diagnostic usefulness). At the bottom of each of the columns is given the mean $\rho(y, z)$ of the correlation coefficients in that column. The Kullback-Leibler (KL) gain yields the highest value of the mean $\rho(y, z) = 0.87)$. Therefore, the KL gain shows the best overall performance.

We now attempt to answer a basic question: How accurate are these data summaries?

The point is that we need some kind of statistical microscope to predict the accuracy of the estimated value of the mean of the correlation coefficients for a particular measure. What we would like to know is the true value of $\rho(y, z)$ for the particular measure, that is the value of $\rho(y, z)$ we would see if we inspect any set of mammographic images from the population, and not just a sample of ten images (Fig. 1). In other cases, how do we know that $\rho(y, z)$ might not come out much less favorably if the experiment were run again?

Thus the point may be rewritten as follows: Given a particular measure, how to generate knowledge about the distribution of all values of $\rho(y, z)$ corresponding to any possible sample $S_i$ drawn from the same population of the original sample $S$ of ten mammographic images?

Here we use bootstrap sampling for making statistical inference. The bootstrap is a data-based simulation method for assessing statistical accuracy, [2]. It is only recently developed because it requires modern computer power to simplify the calculations of traditional statistical theory.

The only assumption is that the original sample $S$ of ten mammographic images given in Fig. 1 is representative of the population. Then we can treat our sample $S$ of size 10 as if it is the population. And we can simulate the sampling by resampling from $S$, just as we would sample from a population. This method works well because the sample
Table 3. 95% BCa confidence intervals for the differences of means

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{corr} \left( KL, \frac{M}{T} \right) - \text{corr} \left( \frac{M}{T}, \frac{R}{P} \right) )</td>
<td>(0.0170, 0.074)</td>
</tr>
<tr>
<td>( \text{corr} \left( KL, \frac{M}{T} \right) - \text{corr} \left( \frac{M}{T}, \frac{S}{N} \right) )</td>
<td>(0.0270, 0.087)</td>
</tr>
<tr>
<td>( \text{corr} \left( KL, \frac{M}{T} \right) - \text{corr} \left( \frac{M}{T}, \frac{PSN}{M} \right) )</td>
<td>(0.086, 0.156)</td>
</tr>
<tr>
<td>( \text{corr} \left( KL, \frac{M}{T} \right) - \text{corr} \left( \frac{M}{T}, \frac{PSNR}{M} \right) )</td>
<td>(0.118, 0.184)</td>
</tr>
</tbody>
</table>

4 Conclusions

The Kullback-Leibler information gain can be characterized with a minimal number of properties which are natural for image processing and thus desirable. We have shown that any error function that verifies the three postulates defined in Section 2 has to be of the form of the Kullback-Leibler information function up to a nonnegative multiplicative constant.

But the Kullback-Leibler gain is equally important due to a different reason: it can adequately capture statistical differences between a pair of digital images for situations of practical relevance. As an example of this fact, we have shown that the information gain may be used to quantify the image fidelity in the sense of diagnostic usefulness by means of the difference between the original mammographic image and the decoded images. Moreover, the Kullback-Leibler information gain was found to be significantly better for predicting image distortion (in the sense of diagnostic usefulness) than \( R_{M} \) and \( P_{SNR} \). Bootstrap methods for assessing statistical accuracy were used to produce this inference. Additional discussions of approaches for quantifying visual distinctness by using relative information can be found in [5].

Acknowledgments

The digitized mammographic images were provided by courtesy of the Mammographic Image Analysis Society as well as the National Expert and training centre for breast cancer screening and the Department of radiology at the University of Nijmegen, The Netherlands. This research was sponsored by the Spanish Board for Science and Technology (DGES) under grant PB98-1374.