Comparison Of Lossy Coders For Still Images.

J.A. Garcia,* J. Fdez-Valdivia, Rosa Rodriguez-Sánchez, and, Xose R. Fdez-Vidal

Abstract

A criterion for selecting a lossy coder for still images is proposed. The "optimum" coder using this criterion is selected to maximize the minimum Chernoff information between pairs of distributions which are the most difficult to distinguish from achievable and non-achievable regions using this coder. The resulting coder has the property that the average Bayesian error probability, for the problems of hypothesis testing between distributions from achievable and non-achievable regions, is less than the average Bayesian error probability for any other choice of coder.

The coder selection procedure may be applied without a knowledge of what distortion measure is more suitable for images and does not require an explicit evaluation of error probabilities. Several examples of coder selection are included to illustrate the procedure.

I. Introduction

The problem of selecting a lossy image coder from a family being tested for 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 [1], [2], [3], [4], [5], [6], and 2) the characteristics of the human visual system are too complex and not well understood, yet [7], [8], [9], [10].

In this paper, an alternative criterion for coder selection is proposed which does not require an ability to measure distortion or a knowledge of the properties of the human visual perception. The approach used is based on a necessary and sufficient condition for the probability of error between pairs of distributions from achievable and non-achievable regions, which is satisfied when the optimum coder is compared with any other choice of possible coders. This approach consists of two steps. Given a lossy coder in the class being compared, we first find the best achievable compression ratio which determines a boundary between achievable and non-achievable regions using the coder. The second step consists of characterizing an optimum coder in the sense of the occurrence of a particularly unfavorable pair of hypotheses which are the most difficult to distinguish from achievable and non-achievable regions.

J.A. Garcia, J. Fdez-Valdivia and Rosa Rodriguez-Sánchez are with the *Departamento de Ciencias de la Computación e I.A., E.T.S. de Ingeniería Informática, Universidad de Granada, 18071 Granada, Spain.* **E-mail:** jags@decsai.ugr.es. **Fax:** +34.958.243317. **Phone:** +34.958.244235.

Xose R. Fdez-Vidal is with the *Departamento de Fisica Aplicada, Facultad de Fisica, Universidad de Santiago de Compostela, 15706 Santiago de Compostela, Spain*

The trade-off between image fidelity and coding rate is reached with several techniques, but all of them require an ability to measure distortion. The problem is that finding a general enough measure of perceptual quality has proven to be an elusive goal. Section (II) proposes a novel technique for deriving an optimal compression ratio for lossy coding based on the relationship between information theory and the problem of testing hypotheses. Given a set of discrete probability distributions, $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(K)}\}$, characterizing statistically the reconstructions of the original image at various compression ratios $q(1), q(2), \cdots, q(K)$, we have that each level of compression $q(i)$ divides the set of distributions in two subsets: $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$ and $\{P_{q(i+1)}, P_{q(i+2)}, \cdots, P_{q(K)}\}$. The best achievable compression ratio for lossy coding is then defined as the highest level of compression that more clearly separates the two subsets, where the clarity of separation is formulated in terms of the probabilities of error over all the problems of hypothesis testing to decide between two distributions from different subsets. The resultant level of compression determines a boundary between achievable and non-achievable regions in the trade-off between source fidelity and coding rate. As demonstrated in [11], the resultant bounds are tight for situations of practical relevance (i.e., correspond to high coding rate). And these performance bounds are directly achievable by a constructive procedure.

After the preliminary definitions and results in Section (II), two different procedures for coder selection are obtained in Section (III). In this section, we first prove that a lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$ on an error probability basis for classical hypothesis testing (i.e., one of the individual error probabilities goes to zero and the other is minimized), if and only if $\mathcal{L}_\alpha$ is the optimum coder in the sense of the occurrence of a particularly unfavorable pair of hypotheses which are the most difficult to distinguish by means of the Kullback-Leibler information gain[12]. This result allows us to outline a maximin Kullback-Leibler procedure for coder selection.

In classical hypothesis testing the two individual probabilities of error are treated separately, and hence this approach lacks symmetry. Instead, we can follow a Bayesian approach in which the overall probability of error for hypothesis testing is given by the weighted sum of the individual probabilities. In Section (III), we also prove that a lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$ on a Bayesian error probability basis, if and only if $\mathcal{L}_\alpha$ is the optimum coder in the sense of the occurrence of a particularly unfavorable pair of hypotheses which are the most difficult to distinguish by means of the Chernoff information[13]. This result implies a maximin Chernoff procedure for coder selection on a Bayesian error probability basis.

Section (IV) gives examples where the procedure is applied to compare several coders in the encoding of still images. The examples have been selected to illustrate the relative simplicity of the calculations involved in applying the procedure. The main conclusions of this paper are summarized in Section (V).

II. The Best Achievable Compression Ratio

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

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 $P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}, \cdots, P_{q(K)}$ denote the discrete probability distributions for the reconstructions of the original image at compression ratios $q(1), \cdots, q(i), \cdots, q(K)$, 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.

In this setting, a classical statistical hypothesis testing to decide between any two distributions $P_{q(j)}$ and $P_{q(n)}$ can be stated as follows:

$$H_1: \quad P = P_{q(j)}$$
$$H_2: \quad P = P_{q(n)}$$

Consider the general decision function $g(x_1, \cdots, x_N)$ where $g(x_1, \cdots, x_N) = 1$ implies that H_1 is accepted and $g(x_1, \cdots, x_N) = 2$ implies that H_2 is accepted. The two probabilities of error are defined as:

$$\begin{aligned} \alpha_{j,n}^N &= Pr\{g(x_1, \cdots, x_N) = 2/H_1 \text{ true}\} \\ \beta_{j,n}^N &= Pr\{g(x_1, \cdots, x_N) = 1/H_2 \text{ true}\}\,. \end{aligned} \tag{1}$$

In general, we wish to minimize both probabilities, but there is a trade-off. Thus we minimize the probability of error $\beta_{j,n}^N$ subject to a constraint on the other probability of error: $\alpha_{j,n}^N < \varepsilon$.

Let $\beta_{j,n}^{\varepsilon,N}$ be the lowest achievable probability of error $Pr\{g(x_1, \cdots, x_N) = 1/H_2 \text{ true}\}$ subject to $\alpha_{j,n}^N < \varepsilon$ for this problem. Note that more similar distributions $P_{q(j)}$ and $P_{q(n)}$ will produce higher values of the best achievable probability of error $\beta_{j,n}^{\varepsilon,N}$, and therefore, $\beta_{j,n}^{\varepsilon,N}$ may be interpreted as a measure of similarity between $P_{q(j)}$ and $P_{q(n)}$.

We have that, for each compression ratio $q(i)$, we can divide the set of distributions corresponding to the reconstructions of the original in two subsets: $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$ and $\{P_{q(i+1)}, \cdots, P_{q(K)}\}$.

In the following, for each level of compression $q(i)$, we will define the probability of error over all the decision problems of hypothesis testing to decide between a distribution in $\{P_{q(i+1)}, \cdots, P_{q(K)}\}$ and any other in $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$.

Definition 1. For a given level $q(i)$, let $q(j)$ be any compression ratio such that $i < j$. Consider all the decision problems of hypothesis testing between two alternatives, $H_1 : P = P_{q(j)}$ and $H_2 : P = P_{q(n)}$, with n such that $1 \leq n \leq i$. Let $\beta_{j,n}^{\varepsilon,N}$, with $1 \leq n \leq i$, be the lowest achievable probability of error corresponding to each hypothesis test when the other individual probability of error $\alpha_{j,n}^{N} < \varepsilon$. Then the overall probability of error for these i decision problems is defined as the average sum of the individual probabilities of error $\beta_{j,n}^{\varepsilon,N}$, i.e.,

$$\beta_{\leftarrow}^{\varepsilon,N}(i,j) = \frac{1}{i} \sum_{1 \leq n \leq i} \beta_{j,n}^{\varepsilon,N} \,. \tag{2}$$

The overall probability of error $\beta_{\leftarrow}^{\varepsilon,N}(i,j)$ is essentially determined by the individual probability of error $\beta_{j,n}^{\varepsilon,N}$ for the hypothesis test between $P_{q(j)}$ and the distribution in $\{P_{q(1)}, \cdots, P_{q(n)}, \cdots, P_{q(i)}\}$ that is closest to $P_{q(j)}$ (for further details see Appendix A).

For a given level of compression $q(i)$, the best value in the overall probability of error $\beta_{\leftarrow}^{\varepsilon,N}(i,j)$, with j such that $i < j$, is simply given by:

$$\beta_{\leftarrow}^{\varepsilon,N}(i) = \min_{i < j \leq K} \{\beta_{\leftarrow}^{\varepsilon,N}(i,j)\} \,. \tag{3}$$

Thus, the best value $\beta_{\leftarrow}^{\varepsilon,N}(i)$ is given by the overall probability of error $\beta_{\leftarrow}^{\varepsilon,N}(i,j)$ for the distribution $P_{q(j)}$ in $\{P_{q(i+1)}, \cdots, P_{q(K)}\}$ that is furthest away from the subset $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$.

Similarly we can define, for each compression ratio $q(i)$, the probability of error over all the decision problems of hypothesis testing to decide between a distribution in $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$ and any other in $\{P_{q(i+1)}, P_{q(i+2)}, \cdots, P_{q(K)}\}$ as follows.

For a given level $q(i)$, let $q(l)$ be a compression ratio such that $l \leq i$; that is, $q(l) \leq q(i)$. Consider now all the decision problems of hypothesis testing between two alternatives:

$$H_1 : \quad P = P_{q(m)}$$

$$H_2 : \quad P = P_{q(l)}$$

with m such that $i < m \leq K$. Let $\beta_{m,l}^{\varepsilon,N}$ be the lowest achievable probability of error corresponding to each hypothesis test when the other individual probability of error $\alpha_{m,l}^{N} < \varepsilon$.

Definition 2. The overall probability of error for these new $K - i$ decision problems is given by:

$$\beta_{\rightarrow}^{\varepsilon,N}(l,i) = \frac{1}{K-i} \sum_{i < m \leq K} \beta_{m,l}^{\varepsilon,N} \,. \tag{4}$$

In this case, the overall probability of error $\beta^{\varepsilon,N}_{\rightarrow}(l,i)$ is essentially determined by the individual probability of error $\beta^{\varepsilon,N}_{m,l}$ for the hypothesis test between $P_{q(l)}$ and the distribution in $\{P_{q(i+1)},\cdots,P_{q(m)},\cdots,P_{q(K)}\}$ that is closest to $P_{q(l)}$ (see proof of Theorem 1 in Appendix A).

For a given compression ratio $q(i)$, the best value in the overall probability of error $\beta^{\varepsilon,N}_{\rightarrow}(l,i)$, with l such that $l \leq i$, is simply given by:

$$\beta^{\varepsilon,N}_{\rightarrow}(i) = \min_{1\leq l\leq i}\{\beta^{\varepsilon,N}_{\rightarrow}(l,i)\}\,. \tag{5}$$

Thus, the best value $\beta^{\varepsilon,N}_{\rightarrow}(i)$ is given by the overall probability of error $\beta^{\varepsilon,N}_{\rightarrow}(l,i)$ that corresponds to the distribution $P_{q(l)}$ in $\{P_{q(1)},P_{q(2)},\cdots,P_{q(i)}\}$ that is furthest away from the subset $\{P_{q(i+1)},P_{q(i+2)},\cdots,P_{q(K)}\}$.

For a given compression ratio $q(i)$ with $1 \leq i < K$, let $\beta^{\varepsilon,N}(i)$ be the average sum of the best overall probability of error $\beta^{\varepsilon,N}_{\rightarrow}(i)$ and the best overall probability of error $\beta^{\varepsilon,N}_{\leftarrow}(i)$, i.e.,

$$\beta^{\varepsilon,N}(i) = \frac{1}{2}\left[\beta^{\varepsilon,N}_{\rightarrow}(i) + \beta^{\varepsilon,N}_{\leftarrow}(i)\right]\,. \tag{6}$$

Levels of compression $q(i)$ that more clearly separate the set of distributions in two subsets $\{P_{q(1)},P_{q(2)},\cdots,P_{q(i)}\}$ and $\{P_{q(i+1)},P_{q(i+2)},\cdots,P_{q(K)}\}$, will produce lower values of the average sum $\beta^{\varepsilon,N}(i)$ of the best overall probabilities of error $\beta^{\varepsilon,N}_{\rightarrow}(i)$ and $\beta^{\varepsilon,N}_{\leftarrow}(i)$.

We will now define some new notation to express equality to first order in the exponent, as suggested in [13].

Definition 3. The notation $a_N \doteq b_N$ means

$$\lim_{N\to\infty}\frac{1}{N}\log\frac{a_N}{b_N} = 0$$

and therefore $a_N \doteq b_N$ implies that a_N and b_N are equal to the first order in the exponent. The logarithmic base 2 is used throughout this paper unless otherwise stated.

The best achievable compression ratio can now be defined as follows.

Definition 4: Best achievable compression ratio. The best achievable compression ratio $q(i^*)$ is the highest level of compression that attains the asymptotically best value in $\beta^{\varepsilon,N}(i)$, i.e.,

$$\min_{1\leq i<K}\beta^{\varepsilon,N}(i) \doteq \beta^{\varepsilon,N}(i^*)\,, \tag{7}$$

and also verifies a condition of consistency

$$\beta^{\varepsilon,N}_{\rightarrow}(i^*) \doteq \beta^{\varepsilon,N}_{\leftarrow}(i^*)\,. \tag{8}$$

The following result states the relationship between the best achievable compression ratio $q(i^*)$ and the Kullback-Leibler information gain.

Theorem 1. Let $q(i^*)$ be the highest compression ratio such that:

$$d^B(i^*) = d^F(i^*) \tag{9}$$

where

$$d^B(i) = \max_{i<j\leq K} \left\{ \min_{1\leq n\leq i} KL(P_{q(j)}, P_{q(n)}) \right\} \tag{10}$$

and

$$d^F(i) = \max_{1\leq l\leq i} \left\{ \min_{i<m\leq K} KL(P_{q(m)}, P_{q(l)}) \right\} \tag{11}$$

with $KL(P_{q(\dagger)}, P_{q(\ddagger)})$ being the Kullback-Leibler information gain[12] between $P_{q(\dagger)}$ and $P_{q(\ddagger)}$:

$$KL(P_{q(\dagger)}, P_{q(\ddagger)}) = \sum_l p(l/I_{q(\dagger)}) \log \frac{p(l/I_{q(\dagger)})}{p(l/I_{q(\ddagger)})},$$

where $P_{q(\dagger)} = (p(l/I_{q(\dagger)}))_l$ characterizes the probability of occurrence of each gray level l in the decoded image $I_{q(\dagger)}$ at compression ratio $q(\dagger)$.

Then, $q(i^*)$ is the best achievable compression ratio for the sequence of levels $q(1), q(2), \cdots, q(K)$.

Proof. See Appendix A.

This result implies that the performance bound given by the best achievable compression ratio is operational in that it is directly achievable by a constructive procedure. This optimal level will allow us to determine a boundary between achievable and non-achievable regions, and consequently, it provides useful information to benchmark specific applications. It is worth noting that the Kullback-Leibler information gain appears from asymptotic results as we look for a practical algorithm that comes close to the optimal performance dictated by the bound.

III. The comparison of lossy coders.

A. *Coder Selection Based On Error Probability For Classical Hypothesis Testing.*

Given a lossy coder $\mathcal{L}_\alpha$, let $q(i^*_\alpha)$ be the best achievable compression ratio for coding the original image I by using $\mathcal{L}_\alpha$.

If $\{P^\alpha_{q(1)}, \cdots, P^\alpha_{q(K)}\}$ is the set of discrete probability distributions for the decoded images at compression ratios $q(1), \cdots, q(K)$ by using coder $\mathcal{L}_\alpha$, then we have that the best achievable compression ratio $q(i^*_\alpha)$ is such that divides the set $\{P^\alpha_{q(1)}, \cdots, P^\alpha_{q(K)}\}$ in two subsets:

$$A_\alpha = \{P^\alpha_{q(1)}, \cdots, P^\alpha_{q(i^*_\alpha)}\}$$

and

$$B_\alpha = \{P^\alpha_{q(i^*_\alpha+1)}, \cdots, P^\alpha_{q(K)}\}$$

where A_α determines the achievable region for coding I by $\mathcal{L}_\alpha$, and B_α is the non-achievable region.

Consider all the decision problems of hypothesis testing between any two distributions $P^{\alpha}_{q(m)}$ and $P^{\alpha}_{q(l)}$, with $P^{\alpha}_{q(m)} \in B_\alpha$ and $P^{\alpha}_{q(l)} \in A_\alpha$, i.e.,

$$\begin{aligned} H^{\alpha}_{q(m)} &: \; P = P^{\alpha}_{q(m)} \\ H^{\alpha}_{q(l)} &: \; P = P^{\alpha}_{q(l)} \end{aligned}$$

Let N be the size in pixels of the original image I; and $Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)} \text{ or } H^{\alpha}_{q(l)})$ be the best achievable overall probability of error for hypotheses $H^{\alpha}_{q(m)}$ vs. $H^{\alpha}_{q(l)}$ when one of the individual probabilities of error, $\alpha^{N}_{m,l}$, is fixed to ε and the other, $\beta^{\varepsilon,N}_{m,l}$, is minimized (i.e., $\alpha^{N}_{m,l}$ and $\beta^{\varepsilon,N}_{m,l}$ as given in equation (21) and equation (22), in Appendix B).

Then we have that the average error probability over all pairs of hypotheses $H^{\alpha}_{q(m)}$ vs. $H^{\alpha}_{q(l)}$ with the a priori probability $Pr(H^{\alpha}_{q(m)}) \times Pr(H^{\alpha}_{q(l)})$ is:

$$Pr^{\varepsilon,N}_{\alpha}(e) = \sum_{P^{\alpha}_{q(m)} \in B_\alpha} \sum_{P^{\alpha}_{q(l)} \in A_\alpha} Pr(H^{\alpha}_{q(m)}) \times Pr(H^{\alpha}_{q(l)}) \times Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)} \text{ or } H^{\alpha}_{q(l)})\,. \tag{12}$$

For each $\mathcal{L}_\alpha$ in the class of lossy coders being compared, let D_α be the best achievable exponent in the average probability of error $Pr^{\varepsilon,N}_{\alpha}(e)$ when, for each hypothesis testing, one of the individual probabilities of error goes to zero and the other is minimized, i.e.,

$$D_\alpha = \lim_{\varepsilon \to 0} \lim_{N \to \infty} -\frac{1}{N} \log Pr^{\varepsilon,N}_{\alpha}(e)\,. \tag{13}$$

We can now define when a lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$.

Definition 5. A lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$ on an error probability basis for classical hypothesis testing (i.e., one of the individual probabilities of error goes to zero and the other is minimized) if D_α is the best achievable error exponent for any lossy coder in the class being compared, i.e.,

$$D_\alpha \geq D_\beta\,, \quad \text{for all } \beta\,. \tag{14}$$

where D_α and D_β denotes the best achievable exponent in the average probability of error for $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$, respectively.

The following result states that the relation of preference given in Definition 5 guards against the occurrence of a particularly unfavorable pair of hypotheses which are the most difficult to distinguish in the sense of the Kullback-Leibler information gain.

Theorem 2. A lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$ on an error probability basis for classical hypothesis testing, if and only if $\mathcal{L}_\alpha$ is the optimum coder in the sense of the occurrence of a particularly unfavorable pair of hypotheses which are the most difficult to distinguish by means of the Kullback-Leibler information gain; that is, for any other $\mathcal{L}_\beta$ in the class being compared,

$$\min\left\{KL(P^{\alpha}_{q(m)}, P^{\alpha}_{q(l)}) \;\mid\; P^{\alpha}_{q(m)} \in B_\alpha,\, P^{\alpha}_{q(l)} \in A_\alpha\right\} \geq \min\left\{KL(P^{\beta}_{q(m)}, P^{\beta}_{q(l)}) \;\mid\; P^{\beta}_{q(m)} \in B_\beta,\, P^{\beta}_{q(l)} \in A_\beta\right\}\,,$$

with $KL(P_{q(\dagger)}, P_{q(\ddagger)})$ being the Kullback-Leibler information gain between $P_{q(\dagger)}$ and $P_{q(\ddagger)}$.

Proof. See Appendix C.

This result allows us to outline a maximin Kullback-Leibler procedure for coder selection: Select a lossy coder $\mathcal{L}_\dagger$ which achieves the maximum value of the minimum of

$$\left\{KL(P^\dagger_{q(m)}, P^\dagger_{q(l)}) \;\mid\; P^\dagger_{q(m)} \in B_\dagger,\, P^\dagger_{q(l)} \in A_\dagger\right\}$$

as the optimum coder in the class of lossy coders being compared.

B. Coder Selection On A Bayesian Error Probability Basis.

Since we treat the two individual probabilities of error separately, the approach in the last section lacks symmetry. Instead, we can follow a Bayesian approach in which the overall probability of error for hypothesis testing is given by the weighted sum of the individual probabilities $\alpha^N_{m,l}$ and $\beta^N_{m,l}$.

Consider all the problems of hypothesis testing between $P^\alpha_{q(m)}$ and $P^\alpha_{q(l)}$, with $P^\alpha_{q(m)} \in B_\alpha$ and $P^\alpha_{q(l)} \in A_\alpha$, i.e.,

$$\begin{array}{lcl} H^\alpha_{q(m)} & : & P = P^\alpha_{q(m)} \\ H^\alpha_{q(l)} & : & P = P^\alpha_{q(l)} \end{array}$$

Let $Pr^N(e/H^\alpha_{q(m)} \text{ or } H^\alpha_{q(l)})$ be the best achievable Bayesian probability of error for testing $H^\alpha_{q(m)}$ vs. $H^\alpha_{q(l)}$, where the Bayesian probability of error is the weighted sum of the individual probabilities of error $\beta^N_{m,l} = Pr(P^\alpha_{q(m)}/H^\alpha_{q(l)} \text{ is true})$ and $\alpha^N_{m,l} = Pr(P^\alpha_{q(l)}/H^\alpha_{q(m)} \text{ is true})$.

By Chernoff's theorem (see Theorem 12.9.1 in [13]), we have that the best achievable Bayesian probability of error is

$$Pr^N(e/H^\alpha_{q(m)} \text{ or } H^\alpha_{q(l)}) \doteq 2^{-N\times C(P^\alpha_{q(m)}, P^\alpha_{q(l)})}\,, \tag{15}$$

where $C(P^\alpha_{q(m)}, P^\alpha_{q(l)})$ is the Chernoff information[13]:

$$C(P^\alpha_{q(m)}, P^\alpha_{q(l)}) = -\min_{0\le\lambda\le 1} \log \sum_l \left[\left(p(l/I^\alpha_{q(m)})\right)^\lambda \times \left(p(l/I^\alpha_{q(l)})\right)^{1-\lambda}\right]\,, \tag{16}$$

with $P^\alpha_{q(\dagger)} = (p(l/I^\alpha_{q(\dagger)}))_l$ characterizing the probability of occurrence of each gray level l in the decoded image $I^\alpha_{q(\dagger)}$ at compression ratio $q(\dagger)$ by using lossy coder $\mathcal{L}_\alpha$.

Then we have that the average Bayesian probability of error over all pairs of hypotheses $H^\alpha_{q(m)}$ vs. $H^\alpha_{q(l)}$ with the a priori probability $Pr(H^\alpha_{q(m)}) \times Pr(H^\alpha_{q(l)})$ is

$$Pr^N_\alpha(e) = \sum_{P^\alpha_{q(m)}\in B_\alpha} \sum_{P^\alpha_{q(l)}\in A_\alpha} Pr(H^\alpha_{q(m)}) \times Pr(H^\alpha_{q(l)}) \times Pr^N(e/H^\alpha_{q(m)} \text{ or } H^\alpha_{q(l)})\,, \tag{17}$$

with $Pr^N(e/H^\alpha_{q(m)} \text{ or } H^\alpha_{q(l)})$ being the best achievable Bayesian probability of error for hypotheses $H^\alpha_{q(m)}$ vs. $H^\alpha_{q(l)}$, as given in equation (15).

For each $\mathcal{L}_\alpha$ in the class of lossy coders being compared, let E_α be the best achievable exponent in the average Bayesian probability of error $Pr_\alpha^N(e)$, i.e.,

$$E_\alpha = \lim_{N\to\infty} -\frac{1}{N}\log Pr_\alpha^N(e)\,. \tag{18}$$

We can now define when a lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$ based on Bayesian error probabilities as follows.

Definition 6. A lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$ on a Bayesian error probability basis if E_α is the best achievable error exponent for any lossy coder in the class being compared, i.e.,

$$E_\alpha \ge E_\beta\,, \quad \text{for all } \beta \tag{19}$$

where E_α and E_β denotes the best achievable exponent in the average Bayesian error probability for $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$, respectively.

The following result states that the relation of preference given in Definition 6 guards against the occurrence of a particularly unfavorable pair of hypotheses that are the most difficult to distinguish in the sense of the Chernoff information.

Theorem 3. A lossy coder $\mathcal{L}_\alpha$ is preferred to any other $\mathcal{L}_\beta$ on a Bayesian error probability basis, if and only if

$$\min\left\{C(P^\alpha_{q(m)},P^\alpha_{q(l)})\ \mid\ P^\alpha_{q(m)}\in B_\alpha,\, P^\alpha_{q(l)}\in A_\alpha\right\} \ge \min\left\{C(P^\beta_{q(m)},P^\beta_{q(l)})\ \mid\ P^\beta_{q(m)}\in B_\beta,\, P^\beta_{q(l)}\in A_\beta\right\} \tag{20}$$

with $C(P_1,P_2)$ being the Chernoff information between P_1 and P_2.

Proof. See Appendix D.

This result implies a maximin Chernoff procedure for coder selection on a Bayesian error probability basis: Select a lossy coder $\mathcal{L}_\dagger$ which achieves the maximum value of the minimum of

$$\left\{C(P^\dagger_{q(m)},P^\dagger_{q(l)})\ \mid\ P^\dagger_{q(m)}\in B_\dagger,\, P^\dagger_{q(l)}\in A_\dagger\right\}$$

as the optimum coder in the class of lossy coders being compared.

IV. Comparative Study

The purpose of this section is to compare various image compression methods. Our comparative study is confined to still images and is based on the coder selection criteria as given in Section (III). Both coders and images have been selected to illustrate the applicability of the procedure and the relative simplicity of the calculations involved in applying it. Given the abundant literature in image coding methods and for the purpose of this paper, the compression methods tested are not meant to offer a cross section of mainstream methods at or near the state of art.

A. The Compression Methods

In this comparative study we consider the following image coders:

(EPWIC). Embedded predictive wavelet image coding[14] uses an explicit model of the conditional statistical relationships between coefficients in different wavelet subbands. Subband coefficients are encoded one bitplane at a time using a non-adaptive arithmetic encoder that utilizes conditional probabilities calculated from the model. It takes advantage of the joint statistics between wavelet subbands at adjacent scales and orientations. The coder is based on a separable orthonormal wavelet decomposition using 9-tap symmetric (linear-phase) filters designed in [15]. Orthonormal wavelet decompositions have proven to be extremely effective for image compression.

(BTPC). Binary tree predictive coding[16] uses a non-causal, shape-adaptive predictor to decompose an image into a binary tree of prediction errors and zero blocks. Fast compression performance is comparable with joint Photographers Expert Group (JPEG) for photographs, with GIF for graphics, and superior to the state of the art for composite images.

(SPIHT). An algorithm that operates through set partitioning in hierarchical trees[17] and accomplishes completely embedded coding. This algorithm uses the principles of partial ordering by magnitude, set partitioning by significance of magnitudes with respect to a sequence of octavely decreasing thresholds, ordered bit plane transmission, and self-similarity across scale in an image wavelet transform. The results of this coding algorithm with its embedded code and fast execution are so impressive that it is a serious candidate for standardization in future image compression systems.

B. The Test Images

Four natural test images have been used in this study. These images are presented in Figure 2:

(Micro). A digitized mammogram was selected from the Nijmegen database[18] containing some clustered microcalcifications. 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. From the selected mammogram we obtained the region 256×256 in size where the microcalcifications are present. The resultant grayscale image, 8 bits deep, is illustrated in Figure 2. 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 test image to detect microcalcifications.

(Bird). This is also a grayscale image, 8 bits deep, 256 pixels square. There are several features of fine detail including the crest feathers, small white feathers cast against black, the band on the beak, and the nostril. The bird's toes, beak, and eyes are medium resolution structures crucial to accurate reproduction. The body of the bird exhibits a visually pleasing texture surrounded by strong edges. The background gradually increases in luminance from the bottom right corner to the top left corner but it is not entirely smooth.

(Lena). This is a classic of the image processing literature. It is an image of detailed content, as might be found in typical photographs.

(Goldhill). This is another classic image. Its main characteristic is the lack of large uniform areas which makes it the most challenging of the photorealistic images.

C. Results

The four test images were coded at various compression ratios using the three coding methods. The lossy compression comparisons were made by setting all coders to achieve similar information gain between the original image and the image reconstructed under the highest degree of lossy compression using the coder. The information gain was measured by the Kullback-Leibler information function.

The first step of the coder selection procedure consists of finding the best achievable compression ratio which determines a boundary between achievable and non-achievable regions using each coder in the class being compared.

Figures 3, 4, and 5 show one example illustrative of the first step in which the best achievable compression ratio is obtained for each coder (EPWIC, BTPC, and SPIHT) by using the constructive procedure given in Section (II). The test image in this example is Micro.

Figure 3 illustrates the plot of $d^B(i)$ and $d^F(i)$ for EPWIC. As suggested in Theorem 1 the best achievable compression ratio $q(i^*)$ is attained if $d^B(i^*) = d^F(i^*)$. For the test image $Micro$, it happens at a high compression ratio of $59:1$ (see Figure 3).

Figure 3 also shows a plot of the false positive to true positive ratio (FP/TP) for microcalcification detection upon images reconstructed under various degrees of lossy compression using EPWIC. In order to derive the false positive to true positive detection ratio, the images reconstructed after compression were also reviewed to detect individual microcalcifications. This method can be used for characterizing compression losses using decoded images. Since it is rated diagnostic usefulness rather than general appearance or simply line or edge patterns, this method relates diagnostic accuracy to compression level. In fact, when the diagnostic usefulness of the reconstructed images significantly decreases, the false positive to true positive ratio also shows a significant increase.

Best Achievable Compression Ratio			
IMAGE	*SPIHT*	*EPWIC*	*BTPC*
Micro	80 : 1	59 : 1	68 : 1
Lena	50 : 1	44 : 1	21 : 1
Goldhill	44 : 1	23 : 1	16 : 1
Bird	57 : 1	51 : 1	31 : 1

TABLE I

BEST ACHIEVABLE COMPRESSION RATIO FOR EACH IMAGE AND LOSSY CODER.

Figure 3 illustrates the fact that the best achievable compression ratio corresponds to a relatively low FP/TP ratio. As can be seen from this same figure, the image reconstructed at the best achievable compression ratio can still be used for detection of individual microcalcifications.

Figures 4 and 5 illustrate the same analysis for BTPC and SPIHT, respectively. In each case, we show the plot of $d^B(i)$ and $d^F(i)$, the point at which $d^B(i^*) = d^F(i^*)$, the relationship between this point and the FP/TP ratio, as well as a reconstruction after compression at the best achievable ratio $q(i^*)$.

As can be seen from Figures 4 and 5, the false positive to true positive ratio for the reconstruction at level $q(i^*)$ indicates that the imperfect source representation given by the image compressed at $q(i^*)$ is sufficient for evaluations that take into account diagnostic usefulness. Conversely, the diagnostic usefulness of images reconstructed at levels $q(i)$, with $i > i^*$, significantly decreases, i.e., the corresponding false positive to true positive rate shows a significant increase. This means that the best achievable compression ratio $q(i^*)$ determines a performance bound for the mammographic image Micro; that is, it will allow us to distinguish between the best achievable operating points and those that are suboptimal or unachievable for the given image.

As illustrated in Table I, the best achievable compression ratio is tight for situations of practical relevance (i.e., it corresponds to high coding rate). Thus, for the test image Micro, the best achievable compression ratio using SPIHT is $80 : 1$ and using BTPC is $68 : 1$. Note that these performance bounds are operational in that they were directly achievable by a constructive procedure, as suggested in Theorem 1. Table I also shows the best achievable ratio that was obtained for each image and lossy coder in this study. Additional discussions of the effects of lossy compression at the best achievable level on image feature applications can be found in [11].

The second step of the coder selection procedure consists of characterizing an optimum coder in the sense of the occurrence of a particularly unfavorable pair of hypotheses which are the most

Maximin Kullback-Leibler procedure			
IMAGE	*SPIHT*	*EPWIC*	*BTPC*
Micro	$\boxed{0.007312}$	0.001295	0.002864
Lena	$\boxed{0.007433}$	0.002329	0.000533
Goldhill	$\boxed{0.006167}$	0.005376	0.001179
Bird	$\boxed{0.016430}$	0.003692	0.000029

TABLE II

Maximin Kullback-Leibler criterion for coder selection.

Maximin Chernoff procedure			
IMAGE	*SPIHT*	*EPWIC*	*BTPC*
Micro	$\boxed{0.001945}$	0.000328	0.000717
Lena	$\boxed{0.001835}$	0.000582	0.000132
Goldhill	$\boxed{0.001558}$	0.001353	0.000295
Bird	$\boxed{0.004158}$	0.000883	0.000007

TABLE III

Maximin Chernoff criterion for coder selection.

difficult to distinguish from achievable and non-achievable regions.

Table II illustrates the results of the maximin Kullback-Leibler procedure for coder selection (Section (III-A)). From these results we conclude that, for the test image Micro, the optimum coder on an error probability basis for classical hypothesis testing is SPIHT. This means that SPIHT achieves the maximum value of the Kullback-Leibler information gain between the pair of hypotheses which are the most difficult to distinguish from achievable and non-achievable regions. Both achievable and non-achievable regions using SPIHT are illustrated in Figure 6.

Of course, the proposed technique for coder selection can be applied on different data sets. This fact is illustrated in Table II. This table shows, for each test image, the optimum coder on an error probability basis using the maximin Kullback-Leibler procedure. The maximum value of the minimum of the Kullback-Leibler information gain between pairs of hypotheses from achievable and non-achievable regions is displayed in a box. In each case, the optimum coder is SPIHT. The respective achievable and non-achievable regions using SPIHT are illustrated in Figures 7, 8, and 9.

Table III illustrates the results of the maximin Chernoff procedure for coder selection (Section (III-B)). For the test image Micro, the optimum coder on a Bayesian error probability basis is SPIHT. This means that SPIHT also achieves the maximum value of the Chernoff information between the pair of hypotheses which are the most difficult to distinguish from achievable and non-achievable regions. Table III also illustrates, for Lena, Goldhill, and Bird, the optimum coder on a Bayesian error probability basis. In each case, the maximum value of the minimum of the Chernoff information gain between pairs of hypotheses from achievable and non-achievable regions is displayed in a box. Again, for each test image, the optimum coder is SPIHT.

V. Conclusion

The previous examples illustrate the technique for obtaining "optimum coders" using both the maximin Kullback-Leibler procedure and the maximin Chernoff procedure. Both criteria optimize coder performance for the pair of hypotheses which are the most difficult to distinguish from achievable and non-achievable regions in the sense of the Kullback-Leibler gain and the Chernoff information, respectively. For this reason the coders obtained by these procedures tend to have a low probability of error for the problems of hypothesis testing between distributions from achievable and non-achievable regions. Thus if the rate-distortion measure is unknown or cannot be conveniently evaluated for a specific application, the maximin Kullback-Leibler procedure and the maximin Chernoff procedure provide useful criteria for coder selection. In the maximin Kullback-Leibler criterion we treat the two individual probabilities of error for hypothesis testing separately, and therefore this approach lacks symmetry. Instead, the maximin Chernoff criterion follows a Bayesian approach in which the overall probability of error for hypothesis testing is given by the weighted sum of the individual probabilities. In the previous examples, however, the optimum coder obtained by both procedures is the same.

Although the examples are selected to illustrate the use of these techniques in coder selection problems, the procedures may be applied to the comparison of any collection of lossy coders by setting all coders to achieve similar information gain between the original image and the image reconstructed under the highest level of compression.

VI. Appendix A

Proof of Theorem 1.[11] To prove the theorem, firstly, we have to prove that the probabilities of error $\beta^{\varepsilon,N}_{\leftarrow}(i)$ and $\beta^{\varepsilon,N}_{\rightarrow}(i)$ behave exponentially with exponents given by the functions $d^B(i)$ and $d^F(i)$, respectively. Secondly, we show that the best achievable compression ratio corresponds to the argument i that produces the maximum value of the minimum of $d^B(i)$ and $d^F(i)$. We then show that $d^F(i)$ is non-decreasing and $d^B(i)$ is non-increasing, and thus, the maximum value of the minimum of $\{d^F(i), d^B(i)\}$ is attained if $d^F(i) = d^B(i)$.

By Stein's lemma (see Appendix B), when the type of error $\alpha_{\dagger,\ddagger}^{N} < \varepsilon$ we have that the lowest probability of error $\beta_{\dagger,\ddagger}^{\varepsilon,N}$ for the hypothesis test between $P_{q(\dagger)}$ and $P_{q(\ddagger)}$ is asymptotically:

$$\beta_{\dagger,\ddagger}^{\varepsilon,N} \doteq 2^{-N\times KL(P_{q(\dagger)},P_{q(\ddagger)})}$$

with N being the image size in pixels; $KL(P_{q(\dagger)}, P_{q(\ddagger)})$ being the Kullback-Leibler information gain between $P_{q(\dagger)}$ and $P_{q(\ddagger)}$.

Then

$$\begin{aligned}
\beta_{\leftarrow}^{\varepsilon,N}(i,j) &= \frac{1}{i}\sum_{1\leq n\leq i}\beta_{j,n}^{\varepsilon,N} \\
&\doteq \frac{1}{i}\sum_{1\leq n\leq i} 2^{-N\times KL(P_{q(j)},P_{q(n)})} \\
&\doteq 2^{-N\min_{1\leq n\leq i} KL(P_{q(j)},P_{q(n)})}
\end{aligned}$$

since the exponential rate is determined by the worst exponent. And it follows that

$$\begin{aligned}
\beta_{\leftarrow}^{\varepsilon,N}(i) &= \min_{i<j\leq K}\{\beta_{\leftarrow}^{\varepsilon,N}(i,j)\} \\
&\doteq \min_{i<j\leq K} 2^{-N\min_{1\leq n\leq i} KL(P_{q(j)},P_{q(n)})} \\
&= 2^{-N\max_{i<j\leq K}\left\{\min_{1\leq n\leq i} KL(P_{q(j)},P_{q(n)})\right\}} \\
&= 2^{-N\times d^{B}(i)}\,.
\end{aligned}$$

Similarly, we can prove that

$$\begin{aligned}
\beta_{\rightarrow}^{\varepsilon,N}(l,i) &= \frac{1}{K-i}\sum_{i<m\leq K}\beta_{m,l}^{\varepsilon,N} \\
&\doteq \frac{1}{K-i}\sum_{i<m\leq K} 2^{-N\times KL(P_{q(m)},P_{q(l)})} \\
&\doteq 2^{-N\min_{i<m\leq K} KL(P_{q(m)},P_{q(l)})}
\end{aligned}$$

and then

$$\begin{aligned}
\beta_{\rightarrow}^{\varepsilon,N}(i) &= \min_{1\leq l\leq i}\{\beta_{\rightarrow}^{\varepsilon,N}(l,i)\} \\
&\doteq \min_{1\leq l\leq i} 2^{-N\min_{i<m\leq K} KL(P_{q(m)},P_{q(l)})} \\
&= 2^{-N\max_{1\leq l\leq i}\left\{\min_{i<m\leq K} KL(P_{q(m)},P_{q(l)})\right\}} \\
&= 2^{-N\times d^{F}(i)}\,.
\end{aligned}$$

This proves that $\beta_{\leftarrow}^{\varepsilon,N}(i)$ and $\beta_{\rightarrow}^{\varepsilon,N}(i)$ behave exponentially with exponents given by $d^{B}(i)$ and $d^{F}(i)$, respectively.

It follows that, for all i with $1 \leq i < K$,

$$\begin{aligned} \beta^{\varepsilon,N}(i) &= \frac{1}{2}\left[\beta^{\varepsilon,N}_{\rightarrow}(i) + \beta^{\varepsilon,N}_{\leftarrow}(i)\right] \\ &\doteq \frac{1}{2}\left[2^{-N\times d^F(i)} + 2^{-N\times d^B(i)}\right] \\ &\doteq 2^{-N\min\{d^F(i),d^B(i)\}} \end{aligned}$$

since the exponential rate is determined by the worst exponent. Thus, we have that

$$\begin{aligned} \min_{1\leq i<K} \beta^{\varepsilon,N}(i) &\doteq \min_{1\leq i<K}\left\{2^{-N\min\{d^F(i),d^B(i)\}}\right\} \\ &= 2^{-N\max_{1\leq i<K}\left\{\min\{d^F(i),d^B(i)\}\right\}}. \end{aligned}$$

We will now show that $d^F(i)$ is non-decreasing:

$$\begin{aligned} d^F(i+1) &= \max_{1\leq l\leq i+1}\left\{\min_{i+1<m\leq K} KL(P_{q(m)},P_{q(l)})\right\} \\ &\geq \max_{1\leq l\leq i}\left\{\min_{i+1<m\leq K} KL(P_{q(m)},P_{q(l)})\right\} \\ &\geq \max_{1\leq l\leq i}\left\{\min_{i<m\leq K} KL(P_{q(m)},P_{q(l)})\right\} \\ &= d^F(i) \end{aligned}$$

and $d^B(i)$ is non-increasing:

$$\begin{aligned} d^B(i+1) &= \max_{i+1<j\leq K}\left\{\min_{1\leq n\leq i+1} KL(P_{q(j)},P_{q(n)})\right\} \\ &\leq \max_{i+1<j\leq K}\left\{\min_{1\leq n\leq i} KL(P_{q(j)},P_{q(n)})\right\} \\ &\leq \max_{i<j\leq K}\left\{\min_{1\leq n\leq i} KL(P_{q(j)},P_{q(n)})\right\} \\ &= d^B(i). \end{aligned}$$

Since $d^F(i)$ is non-decreasing and $d^B(i)$ is non-increasing, the maximum value of the minimum of $\{d^F(i), d^B(i)\}$ is attained if $d^F(i) = d^B(i)$. This is illustrated in Figure 1.

Hence, if $q(i^*)$ is so that

$$d^F(i^*) = d^B(i^*)$$

then we have that

$$\min\{d^F(i^*), d^B(i^*)\} = \max_{1\leq i<K}\left\{\min\{d^F(i), d^B(i)\}\right\}$$

and, consequently

$$\begin{aligned} \min_{1\leq i<K} \beta^{\varepsilon,N}(i) &\doteq 2^{-N\max_{1\leq i<K}\left\{\min\{d^F(i),d^B(i)\}\right\}} \\ &= 2^{-N\min\{d^F(i^*),d^B(i^*)\}} \\ &\doteq \beta^{\varepsilon,N}(i^*). \end{aligned}$$

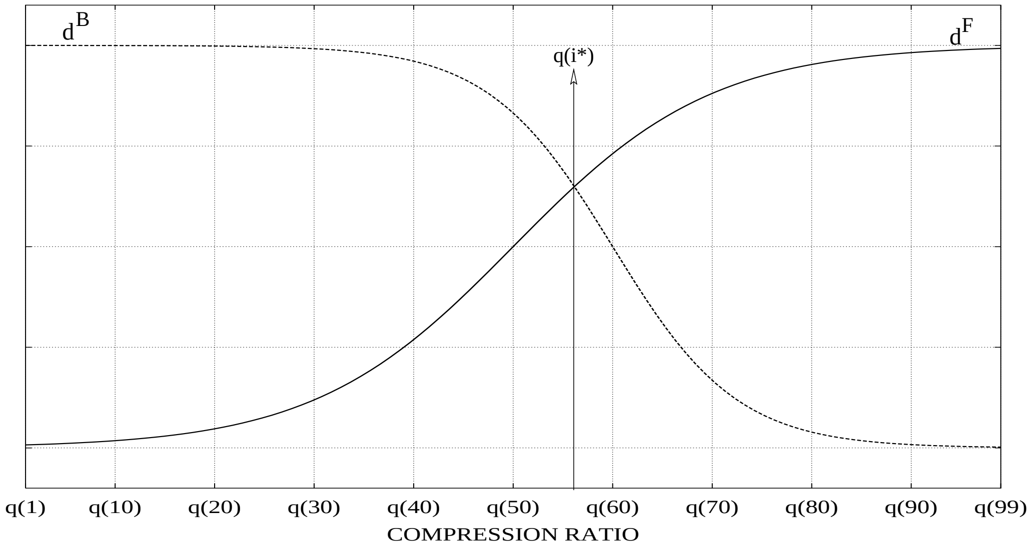

Fig. 1. $d^F(i)$ is non-decreasing and $d^B(i)$ is non-increasing; the maximum value of the minimum of $\{d^F(i), d^B(i)\}$ is attained if they are equal.

And, given the equality $d^F(i^*) = d^B(i^*)$, it follows that

$$\begin{aligned} \beta_{\rightarrow}^{\varepsilon,N}(i^*) &\doteq 2^{-N\times d^F(i^*)} \\ &= 2^{-N\times d^B(i^*)} \\ &\doteq \beta_{\leftarrow}^{\varepsilon,N}(i^*)\,. \end{aligned}$$

This proves the theorem.

VII. Appendix B

Stein's lemma: Let $X_1, X_2, \cdots, X_N$ be i.i.d. $\sim Q$. Consider the hypothesis test between two alternatives, $H_1 : Q = P_1$ and $H_2 : Q = P_2$, where $KL(P_1, P_2) < \infty$. Let A_N be an acceptance region for hypothesis 1. Let the individual probabilities of error be

$$\alpha^N = Pr\{A_N^C / H_1 \text{ true}\}\,,\ \beta^N = Pr\{A_N / H_2 \text{ true}\} \tag{21}$$

and for $0 < \varepsilon < \frac{1}{2}$, define

$$\beta^{\varepsilon,N} = \min_{\alpha^N < \varepsilon} \beta^N\,. \tag{22}$$

Then

$$\lim_{\varepsilon\to 0} \lim_{N\to\infty} \frac{1}{N} \log \beta^{\varepsilon,N} = -KL(P_1, P_2).$$

Proof. See Theorem 12.8.1 in reference [13].

VIII. Appendix C

Proof of Theorem 2. The overall probability of error for hypotheses

$$\begin{aligned} H^{\alpha}_{q(m)} &: P = P^{\alpha}_{q(m)} \\ H^{\alpha}_{q(l)} &: P = P^{\alpha}_{q(l)} \end{aligned}$$

when one of the individual probabilities of error, $\alpha^N_{m,l}$, is fixed to ε and we minimize the other individual probability of error is

$$\begin{aligned} Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)} \text{ or } H^{\alpha}_{q(l)}) &= \frac{Pr(H^{\alpha}_{q(l)})}{Pr(H^{\alpha}_{q(m)}) + Pr(H^{\alpha}_{q(l)})} \times Pr(P^{\alpha}_{q(m)}/H^{\alpha}_{q(l)} \text{ is true}) \\ &\quad + \frac{Pr(H^{\alpha}_{q(m)})}{Pr(H^{\alpha}_{q(m)}) + Pr(H^{\alpha}_{q(l)})} \times Pr(P^{\alpha}_{q(l)}/H^{\alpha}_{q(m)} \text{ is true}) \\ &\overset{(a)}{=} \lambda_{m,l} \times \beta^{\varepsilon,N}_{m,l} + (1-\lambda_{m,l}) \times \alpha^N_{m,l} \\ &\overset{(b)}{=} \lambda_{m,l} \times \beta^{\varepsilon,N}_{m,l} + (1-\lambda_{m,l}) \times \varepsilon\,, \end{aligned} \tag{23}$$

where

(a) follows by defining $\lambda_{m,l} = \frac{Pr(H^{\alpha}_{q(l)})}{Pr(H^{\alpha}_{q(m)})+Pr(H^{\alpha}_{q(l)})}$ and by the definition of $\beta^{\varepsilon,N}_{m,l}$ and $\alpha^N_{m,l}$ given in equations (21) and (22), and

(b) follows from the fact that $\alpha^{\varepsilon,N}_{m,l}$ is fixed to ε.

Then, we have that

$$\begin{aligned} \lim_{\varepsilon\to 0} Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)} \text{ or } H^{\alpha}_{q(l)}) &= \lambda_{m,l} \times \lim_{\varepsilon\to 0} \beta^{\varepsilon,N}_{m,l} + (1-\lambda_{m,l}) \times \lim_{\varepsilon\to 0} \varepsilon \\ &= \lambda_{m,l} \times \lim_{\varepsilon\to 0} \beta^{\varepsilon,N}_{m,l} + 0 \\ &\overset{(a)}{\doteq} \lambda_{m,l} \times 2^{-N\times KL(P^{\alpha}_{q(m)},P^{\alpha}_{q(l)})}\,, \end{aligned} \tag{24}$$

where

(a) follows from Stein's lemma (see Appendix B), with $KL(P^{\alpha}_{q(m)}, P^{\alpha}_{q(l)})$ being the Kullback-Leibler information gain between $P^{\alpha}_{q(m)}$ and $P^{\alpha}_{q(l)}$.

If we average over all pairs of hypotheses $H^{\alpha}_{q(m)}$ and $H^{\alpha}_{q(l)}$ with the a priori probability $Pr(H^{\alpha}_{q(m)}) \times Pr(H^{\alpha}_{q(l)})$ we obtain the average probability of error

$$Pr^{\varepsilon,N}_{\alpha}(e) = \sum_{P^{\alpha}_{q(m)}\in B_\alpha} \sum_{P^{\alpha}_{q(l)}\in A_\alpha} Pr(H^{\alpha}_{q(m)}) \times Pr(H^{\alpha}_{q(l)}) \times Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)} \text{ or } H^{\alpha}_{q(l)})\,,$$

where $Pr^{\varepsilon,N}_{\alpha}(e) > 0$ which follows from the fact that

$$Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)} \text{ or } H^{\alpha}_{q(l)}) = \lambda_{m,l} \times \beta^{\varepsilon,N}_{m,l} + (1-\lambda_{m,l}) \times \varepsilon$$

and by the definition of $\varepsilon > 0$ in equation (22). Then we have that

$$\begin{aligned}
\lim_{\varepsilon\to 0} Pr_{\alpha}^{\varepsilon,N}(e) &= \lim_{\varepsilon\to 0} \sum_{P^{\alpha}_{q(m)}\in B_\alpha} \sum_{P^{\alpha}_{q(l)}\in A_\alpha} Pr(H^{\alpha}_{q(m)}) \times Pr(H^{\alpha}_{q(l)}) \times Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)} \text{ or } H^{\alpha}_{q(l)}) \\
&\overset{(a)}{\doteq} \sum_{P^{\alpha}_{q(m)}\in B_\alpha} \sum_{P^{\alpha}_{q(l)}\in A_\alpha} Pr(H^{\alpha}_{q(m)}) \times Pr(H^{\alpha}_{q(l)}) \times \lambda_{m,l} \times 2^{-N\times KL(P^{\alpha}_{q(m)},P^{\alpha}_{q(l)})} \\
&\overset{(b)}{\doteq} 2^{-N\min\left\{KL(P^{\alpha}_{q(m)},P^{\alpha}_{q(l)})\ \mid\ P^{\alpha}_{q(m)}\in B_\alpha,\, P^{\alpha}_{q(l)}\in A_\alpha\right\}}
\end{aligned} \tag{25}$$

where

(a) follows from equation (24) and from the fact that since A_α and B_α are finite and $Pr(H^{\alpha}_{q(m)}) \times Pr(H^{\alpha}_{q(l)}) \times Pr^{\varepsilon,N}(e/H^{\alpha}_{q(m)}$ or $H^{\alpha}_{q(l)})$ is bounded for all $q(m)$ and $q(l)$, we can interchange sums and limit,

(b) follows from the fact that the exponential rate is determined by the worst exponent.

Thus

$$\begin{aligned}
D_\alpha &= \lim_{\varepsilon\to 0}\lim_{N\to\infty} -\frac{1}{N}\log Pr_{\alpha}^{\varepsilon,N}(e) \\
&\overset{(a)}{=} \lim_{N\to\infty} -\frac{1}{N}\log\left(\lim_{\varepsilon\to 0} Pr_{\alpha}^{\varepsilon,N}(e)\right) \\
&\overset{(b)}{=} \lim_{N\to\infty} -\frac{1}{N}\log\left(2^{-N\min\left\{KL(P^{\alpha}_{q(m)},P^{\alpha}_{q(l)})\ \mid\ P^{\alpha}_{q(m)}\in B_\alpha,\, P^{\alpha}_{q(l)}\in A_\alpha\right\}}\right) \\
&= \min\left\{KL(P^{\alpha}_{q(m)},P^{\alpha}_{q(l)})\ \mid\ P^{\alpha}_{q(m)}\in B_\alpha,\ P^{\alpha}_{q(l)}\in A_\alpha\right\}\ .
\end{aligned} \tag{26}$$

where

(a) follows from the fact that since $Pr_{\alpha}^{\varepsilon,N}(e) > 0$ and $-\frac{1}{N}\log p$ is bounded and continuous in p for all $0 < p \leq 1$, and therefore we can interchange $-\frac{1}{N}\log$ and limit,

(b) follows from equation (25).

Similarly, it follows that

$$D_\beta = \min\left\{KL(P^{\beta}_{q(m)},P^{\beta}_{q(l)})\ \mid\ P^{\beta}_{q(m)}\in B_\beta,\ P^{\beta}_{q(l)}\in A_\beta\right\}\ . \tag{27}$$

Given two lossy coders $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$ we say that $\mathcal{L}_\alpha$ is preferred to $\mathcal{L}_\beta$ on an error probability basis for classical hypothesis testing (one of the individual error probabilities goes to zero and the other is minimized) if we have

$$D_\alpha \geq D_\beta\,,$$

or equivalently, following equations (26) and (27) we have that

$$\min\left\{KL(P^{\alpha}_{q(m)},P^{\alpha}_{q(l)})\ \mid\ P^{\alpha}_{q(m)}\in B_\alpha,\ P^{\alpha}_{q(l)}\in A_\alpha\right\} \geq \min\left\{KL(P^{\beta}_{q(m)},P^{\beta}_{q(l)})\ \mid\ P^{\beta}_{q(m)}\in B_\beta,\ P^{\beta}_{q(l)}\in A_\beta\right\}$$

which can be interpreted as $\mathcal{L}_\alpha$ being the optimum coder in the sense of the occurrence of a particularly unfavorable pair of hypotheses $H^\alpha_{q(m)}$ vs. $H^\alpha_{q(l)}$ which are the most difficult to distinguish by means of the Kullback-Leibler information gain.

This proves the theorem.

IX. Appendix D

Proof of Theorem 3. If we average over all pairs of hypotheses $H^\alpha_{q(m)}$ and $H^\alpha_{q(l)}$ with the a priori probability $Pr(H^\alpha_{q(m)}) \times Pr(H^\alpha_{q(l)})$ we obtain the average Bayesian probability of error

$$\begin{aligned} Pr^N_\alpha(e) &= \sum_{P^\alpha_{q(m)} \in B_\alpha} \sum_{P^\alpha_{q(l)} \in A_\alpha} Pr(H^\alpha_{q(m)}) \times Pr(H^\alpha_{q(l)}) \times Pr^N(e/H^\alpha_{q(m)} \text{ or } H^\alpha_{q(l)}) \\ &\overset{(a)}{\doteq} \sum_{P^\alpha_{q(m)} \in B_\alpha} \sum_{P^\alpha_{q(l)} \in A_\alpha} Pr(H^\alpha_{q(m)}) \times Pr(H^\alpha_{q(l)}) \times 2^{-N \times C(P^\alpha_{q(m)}, P^\alpha_{q(l)})} \\ &\overset{(b)}{\doteq} 2^{-N \min\left\{ C(P^\alpha_{q(m)}, P^\alpha_{q(l)}) \mid P^\alpha_{q(m)} \in B_\alpha,\, P^\alpha_{q(l)} \in A_\alpha \right\}}, \end{aligned} \tag{28}$$

where

(a) follows from equation (15), and

(b) follows from the fact that the exponential rate is determined by the worst exponent.

Thus

$$\begin{aligned} E_\alpha &= \lim_{N\to\infty} -\frac{1}{N} \log Pr^N_\alpha(e) \\ &\overset{(a)}{=} \lim_{N\to\infty} -\frac{1}{N} \log \left(2^{-N \min\left\{ C(P^\alpha_{q(m)}, P^\alpha_{q(l)}) \mid P^\alpha_{q(m)} \in B_\alpha,\, P^\alpha_{q(l)} \in A_\alpha \right\}} \right) \\ &= \min\left\{ C(P^\alpha_{q(m)}, P^\alpha_{q(l)}) \;\mid\; P^\alpha_{q(m)} \in B_\alpha,\, P^\alpha_{q(l)} \in A_\alpha \right\}, \end{aligned} \tag{29}$$

where (a) follows from equation (28).

Similarly, it follows that

$$E_\beta = \min\left\{ C(P^\beta_{q(m)}, P^\beta_{q(l)}) \;\mid\; P^\beta_{q(m)} \in B_\beta,\, P^\beta_{q(l)} \in A_\beta \right\}. \tag{30}$$

Given two lossy coders $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$ we say that $\mathcal{L}_\alpha$ is preferred to $\mathcal{L}_\beta$ on a Bayesian error probability basis if we have

$$E_\alpha \geq E_\beta ,$$

or equivalently, following equations (29) and (30) we have that

$$\min\left\{ C(P^\alpha_{q(m)}, P^\alpha_{q(l)}) \;\mid\; P^\alpha_{q(m)} \in B_\alpha,\, P^\alpha_{q(l)} \in A_\alpha \right\} \geq \min\left\{ C(P^\beta_{q(m)}, P^\beta_{q(l)}) \;\mid\; P^\beta_{q(m)} \in B_\beta,\, P^\beta_{q(l)} \in A_\beta \right\}$$

which can be interpreted as $\mathcal{L}_\alpha$ being the optimum coder in the sense of the occurrence of a particularly unfavorable pair of hypotheses $H^\alpha_{q(m)}$ vs. $H^\alpha_{q(l)}$ which are the most difficult to distinguish by means of the Chernoff information.

This proves the theorem.

Acknowledgments

The digitized mammographic image was provided by courtesy of the Department of Radiology at the University of Nijmegen, The Netherlands. This research was sponsored by the Spanish Board for Science and Technology (CICYT) under grant TIC97-1150.

References

[1] Fdez-Vidal, X.R., Garcia, J.A., Fdez-Valdivia, J., and Rodriguez-Sanchez, R. "The role of integral features for perceiving image discriminability", *Pattern Recognition Letters*, Vol. 18, pp. 733-740, (1997).

[2] Fdez-Vidal, X.R., Garcia, J.A., Fdez-Valdivia, J. "Using models of feature perception in distortion measure guidance", *Pattern Recognition Letters*, Vol. 19, pp. 77-88, (1998).

[3] Martinez-Baena, J., Fdez-Valdivia, J., Garcia, J.A., and Fdez-Vidal, X.R.. "A new distortion measure based on a data-driven multisensor organization", *Pattern Recognition*, Vol. 31, No. 8, pp. 1099-1116, (1998).

[4] Garcia, J.A., Fdez-Valdivia, J., Fdez-Vidal, X.R., and Rodriguez-Sanchez, R. "Visual distinctness based on partially invariant features", *NATO SCI-12 Workshop on Search and Target Acquisition*, Utrecht, The Netherlands, June 21-23, (1999).

[5] Rodriguez-Sanchez. R., Garcia, J.A., Fdez-Valdivia, J., and Fdez-Vidal, X.R. "The RGFF representational model: a system for the automatically learned partitioning of visual patterns in digital images", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, pp. 1044-1073, October (1999).

[6] Fdez-Vidal, X.R., Toet, A., Garcia, J.A., Fdez-Valdivia, J."Computing visual target distinctness through selective filtering, statistical features, and visual patterns", *Optical Engineering*, Vol. 39, No. 1, pp. 267-281, January (2000).

[7] Uttal, William R. *On seeing forms.* Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey, (1988).

[8] Levine, M. W. and Shefner, J. M. *Fundamentals of sensation and perception.* Brooks/Cole Publishing Company, Pacific Grove, California, (1991).

[9] Wandell, Brian A. *Foundations of vision.* Sinauer Associates Inc. Publishers, Sunderland, Massachusetts, (1995).

[10] Bruce, V., Green, P.R., and Georgeson, M.A. *Visual perception.* Psychology Press Publishers, East Sussex, UK, (1996).

[11] Garcia, J.A., Fdez-Valdivia, J, Fdez-Vidal, X.R., and Rodriguez-Sanchez, Rosa."Best achievable compression ratio for lossy image coding", Technical report 99-03-24, Department of Computer Science and Artificial Intelligence, University of Granada, Spain, (1999).

[12] Kullback, S. *Information theory and statistics.* Gloucester, Mass. Peter Smith, (1978).

[13] Cover, T., and Thomas, J. *Elements of Information Theory*, Wiley Series in Telecommunications, John Wiley and Sons, Inc., (1991).

[14] Buccigrossi R.W., and Simoncelli, E.P. "Progressive wavelet image coding based on a conditional probability model." *Proceedings ICASSP-97*, April 21-24, Munich, Germany, (1997).

[15] Simoncelli, E.P., and Adelson, E.H. "Subband transforms", In John W. Woods, editor, *Subband image coding,* Chapter 4, pp. 143-192, (1990).

[16] Robinson, J.A., "Efficient general purpose image compression with binary tree predictive coding." *IEEE Trans. on Image Processing*, Vol. 6, No. 4, pp. 601-608, (1997).

[17] Said, A., and Pearlman, W.A., "A new fast and efficient image coder based on set partitioning in hierarchical trees." *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 6, pp. 243-250, (1996).

[18] Karssemeijer, N. "Adaptive noise equalization and image analysis in mammography." in *13th Int. Conf. Inform. Processing Med. Imag.*, Flagstaff, pp. 472-486, (1992).

Fig. 2. Four natural test images used in the comparative study.

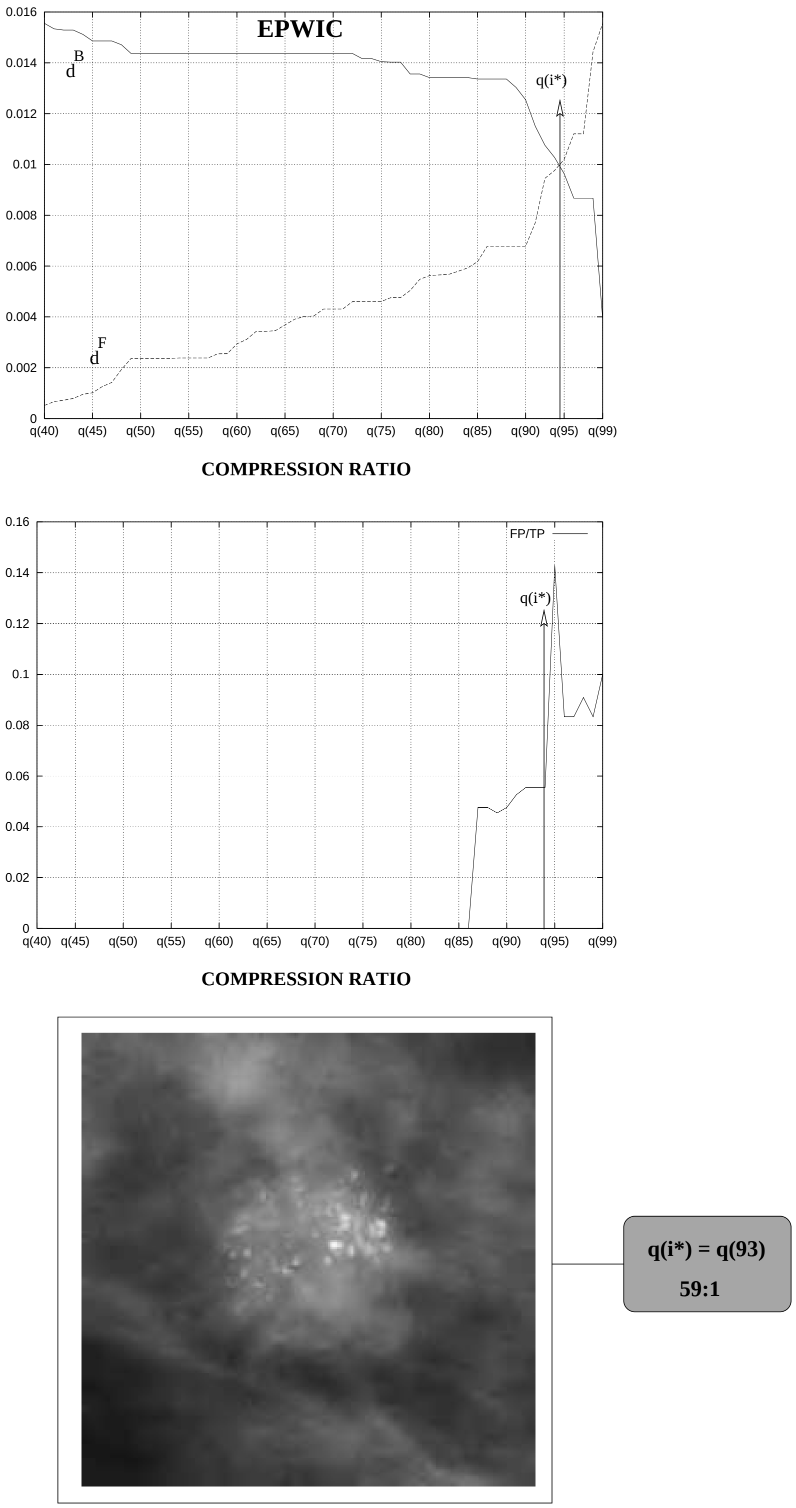

Fig. 3. Plot of $d^B(i)$ and $d^F(i)$ for EPWIC; Best achievable compression ratio $q(i^*)$; the relationship between this point and the false positive to true positive (FP/TP) microcalcification detection ratio upon the images reconstructed under different degrees of lossy compression using EPWIC.

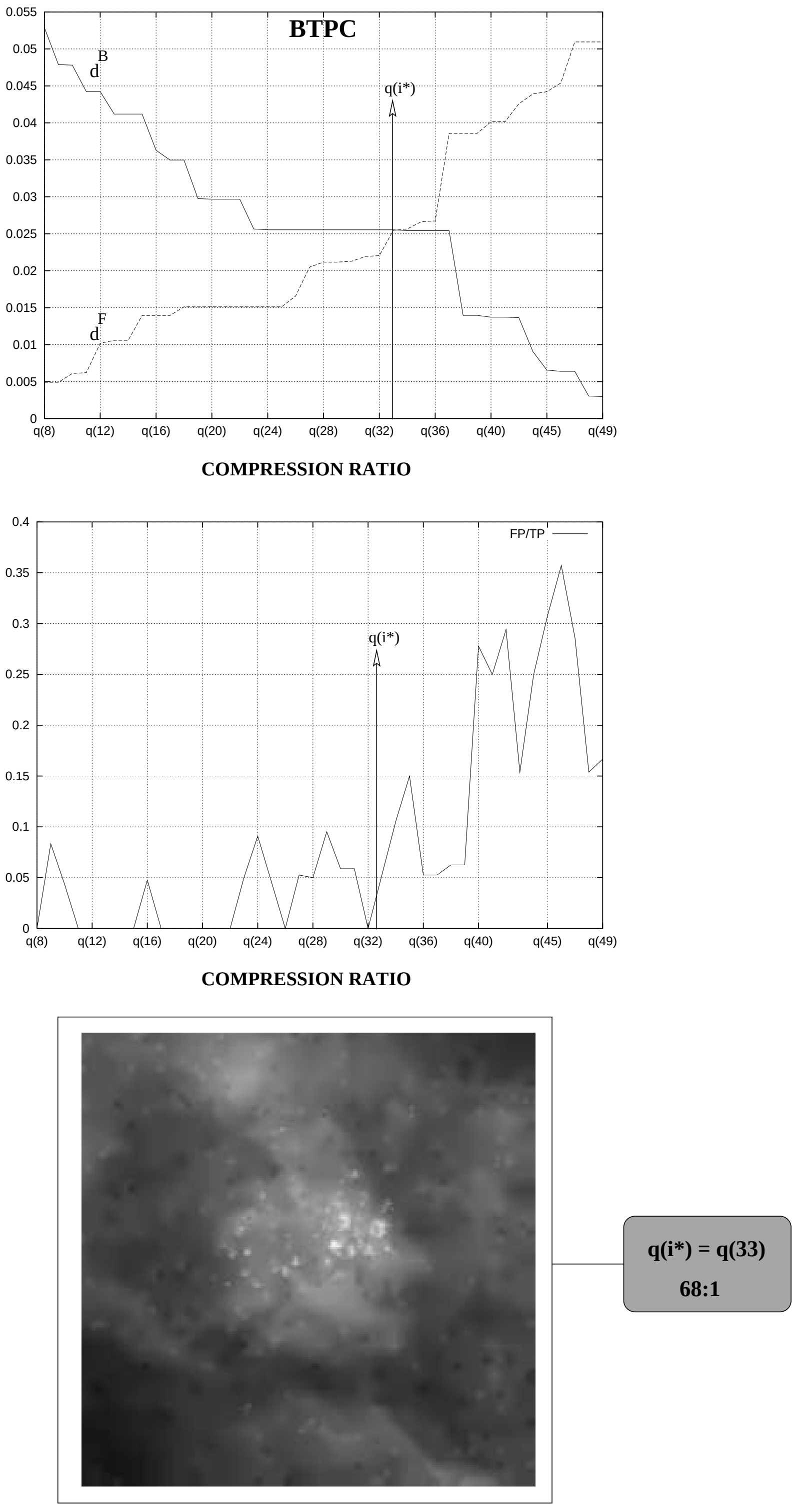

Fig. 4. Plot of $d^B(i)$ and $d^F(i)$ for BTPC; Best achievable compression ratio $q(i^*)$; the relationship between this point and the false positive to true positive (FP/TP) microcalcification detection ratio upon the images reconstructed under different degrees of lossy compression using BTPC.

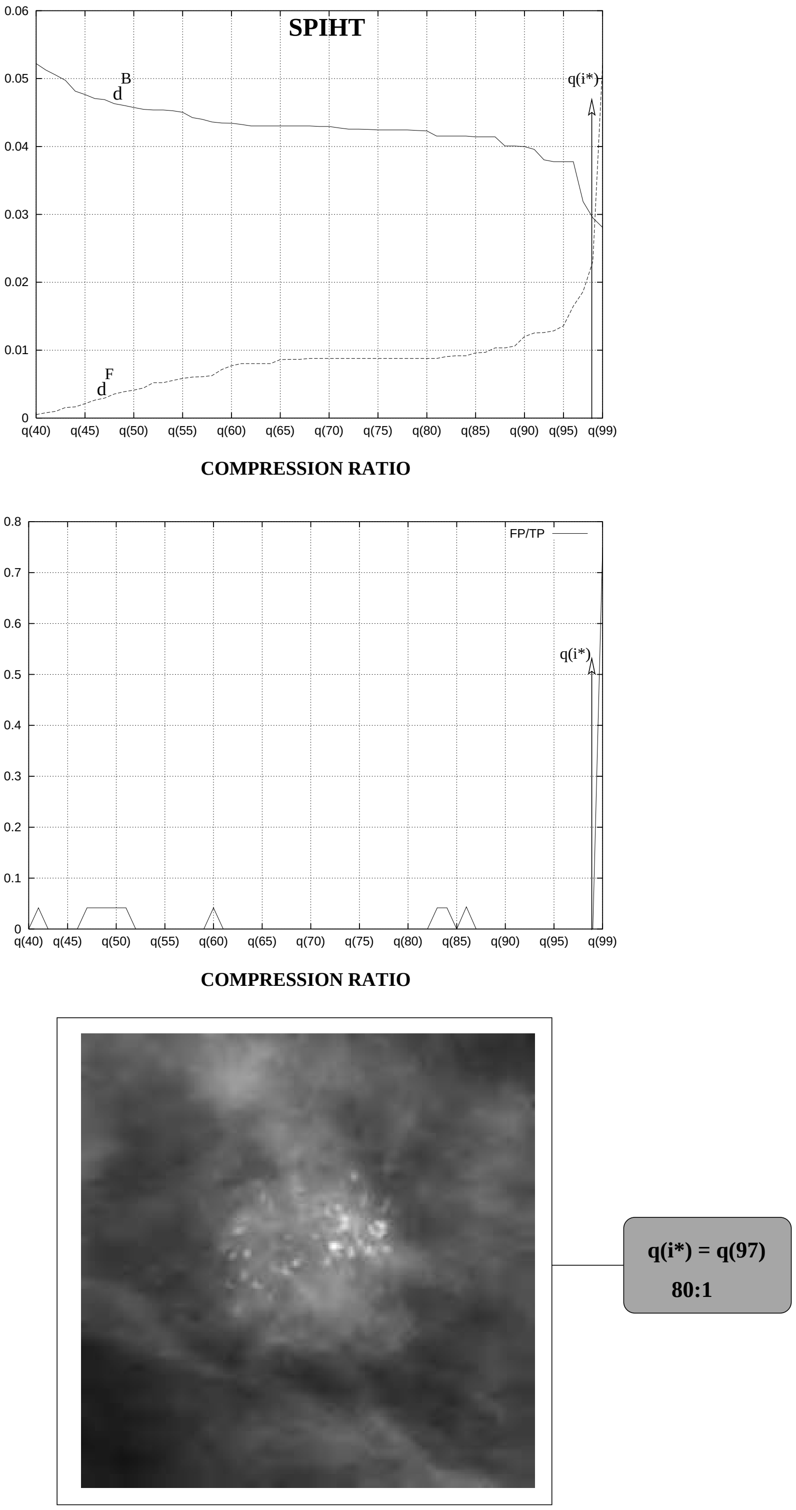

Fig. 5. Plot of $d^B(i)$ and $d^F(i)$ for SPIHT; Best achievable compression ratio $q(i^*)$; the relationship between this point and the false positive to true positive (FP/TP) microcalcification detection ratio upon the images reconstructed under different degrees of lossy compression using SPIHT.

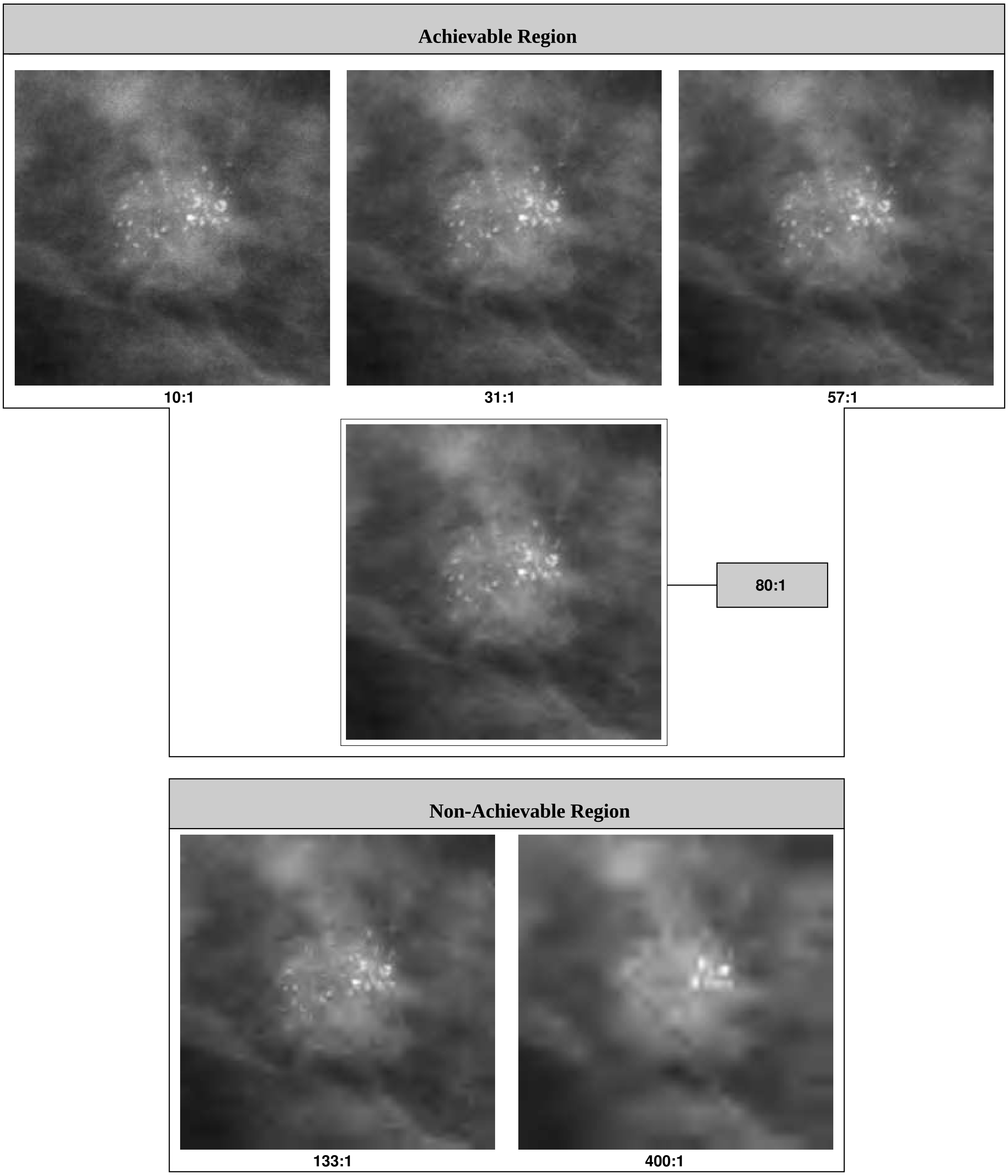

Fig. 6. Achievable and non-achievable regions using the optimum coder SPIHT.

Fig. 7. Achievable and non-achievable regions using the optimum coder SPIHT.

Fig. 8. Achievable and non-achievable regions using the optimum coder SPIHT.

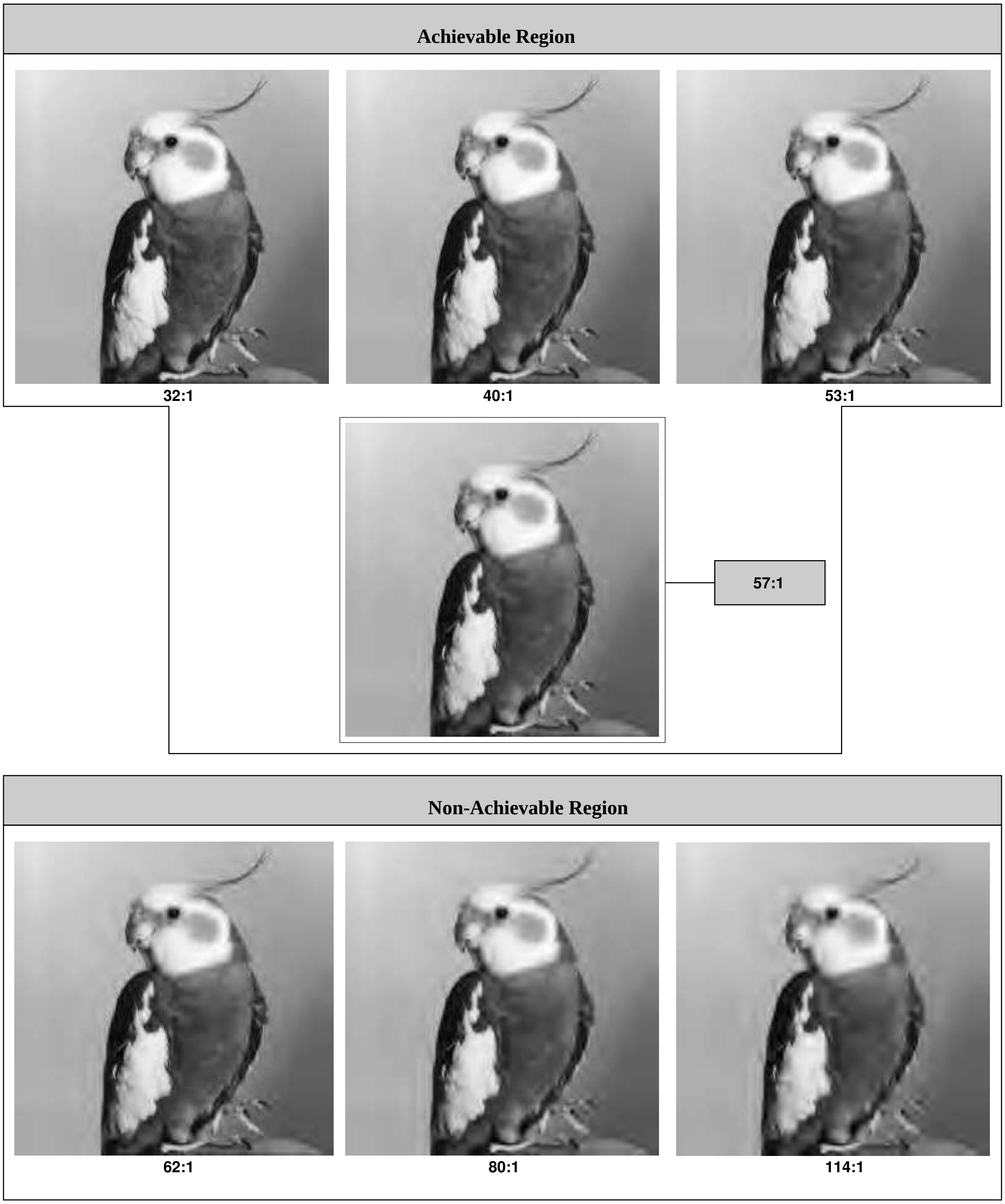

Fig. 9. Achievable and non-achievable regions using the optimum coder SPIHT.