Coder Selection For Lossy Compression Of Biomedical Images.

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

Abstract

A criterion for selecting lossy coders deemed good enough to submit to clinical evaluation is proposed. This approach does not require an ability to measure distortion or a knowledge of the properties of the human visual perception. The coder selection procedure consists of three steps. For each lossy coder in the class being compared, we first find those levels of compression which are to be the only natural solutions for the trade-off between source fidelity and coding rate. The second step of the coder selection procedure consists of finding the "optimum" achievable region for each coder, where the optimum achievable region has the property that the average Bayesian error probability for the problems of hypothesis testing between probability distributions associated with decoded images at compression ratios from achievable and non-achievable regions, is less than the average Bayesian error probability for any other choice of achievable region. In the third step of the procedure, a lossy coder is selected on a Bayesian error probability basis as outlined in the following.

A problem of statistical hypothesis testing to decide between two distributions which are difficult to distinguish tends to have a high probability of error. Thus, a coder that is optimal in this sense of error probabilities tends to produce a high value in the best achievable error probability over the problems of hypothesis testing between the probability distribution associated with the original image and any probability distribution from the optimal achievable region. And clearly, we should not seriously consider any criterion for selecting a lossy coder which does not meet this condition. The "optimum" coder using the proposed criterion is selected to minimize the maximum Chernoff information between the distribution for the original and any distribution from the optimal achievable region. The resulting coder has the property that the best achievable probability of error over the problems of hypothesis testing between these distributions, is greater than the best Bayesian error probability for any other choice of coder.

Here we use the minimax Chernoff criterion to characterize the coder which is optimum for lossy compression of a biomedical image. Several examples of coder selection are included to illustrate the procedure. EDICS: 1-OTHA, 1-STIL.

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*

I. Introduction

The loss of information renders lossy compression of biomedical images controversial because of the potential loss of quality and consequent problems regarding liability. How does one decide if a lossy coding technique is good enough for a specific application? A traditional approach for evaluating the comparative performance of lossy compression methods in biomedical applications is to simulate the application in a carefully designed experiment, gather necessary data in a way that interferes with the simulation as little as possible, and analyze the resulting data to assess the validity of the hypothesis in a specific diagnostic application, [1]. Given the abundant literature in image coding methods, this kind of evaluation is usually impractical both because of the funding constraints and because of the time required for thorough clinical experiments and subsequent statistical analysis. It implies that a protocol for clinical evaluation of lossy compression methods needs an acceptable technology for selecting coders deemed good enough to submit to clinical simulation and statistical analysis. And the intent of this paper is exactly to develop such a technology.

The problem of selecting a coder for potential use in lossy compression of biomedical images 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 digital images, [2], [3], [4], [5], [6], [7]; and 2) the behavior of image coding is not well understood for so-called non-difference distortion measures where the distortion is not measured by a function of the difference between the source and the decoded images, however non-difference distortion measures occur naturally in image coding, [8], [9]. In this paper, an alternative criterion for coder selection is proposed as follows.

It often happens that the structure of the images 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 the original image and the decoded images can be characterized statistically by discrete probability distributions. Let us assume the probability distributions associated with the reconstructions of the original image at compression ratios $q(1), \cdots, q(i), \cdots, q(K)$ as those given by $P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}, \cdots, P_{q(K)}$; where $q(1)$ is the lowest compression ratio and $q(K)$ is the highest compression ratio.

Given a fixed coder, for each level of compression $q(i)$, the set of distributions corresponding to the reconstructions of the original image can be divided into two subsets:

$$\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\} \text{ and } \{P_{q(i+1)}, \cdots, P_{q(K)}\}.$$

The first step of the coder selection procedure consists of finding those levels of compression $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$ which are to be the only natural solutions, in terms of clarity of separation of the entire set of distributions into two subsets, for the trade-off between source fidelity and coding rate across the sequence of compression ratios (see Section II). To this aim, the concept of best achievable compression ratio for lossy coding is defined as the highest level of compression that more clearly separates the two subsets $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$ and $\{P_{q(i+1)}, \cdots, P_{q(K)}\}$, 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 these subsets. And therefore, the best achievable compression ratio so defined determines a boundary between achievable and non-achievable regions in the trade-off between source fidelity and coding rate. The performance bound given by the best achievable compression ratio is operational in that it is directly achievable by a constructive procedure.

The second step of the coder selection procedure consists of finding, for each coder in the class being compared, an optimal achievable region $A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$, with $q(i_j^*)$ in the subset of levels $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$ that more clearly separates the entire set of distributions into two subsets (see Section III). The "optimum" achievable region A_j is selected to maximize the minimum Chernoff information between pairs of distributions which are the most difficult to distinguish from the corresponding achievable and non-achievable regions. The resulting region A_j 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 $A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$ with $q(i_j^*)$ in the subset $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$.

Let us assume that P is the probability distribution associated with the original image, and $P^{\alpha}_{q(l)}$ is a probability distribution in the "optimum" achievable region $A_\alpha = \{P^{\alpha}_{q(1)}, \cdots, P^{\alpha}_{q(i_j^*)}\}$ using the coder $\mathcal{L}_\alpha$. The decision problem of statistical hypothesis testing to decide between two distributions $P^{\alpha}_{q(l)}$ and P, which are difficult to distinguish, tends to have a high probability of error. Hence, optimal coders $\mathcal{L}_\alpha$ should tend to produce a high value of the best achievable probability of error $Pr^N_\alpha(e)$ over the pairs of hypotheses

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

And clearly, we should not seriously consider any criterion for selecting a lossy coder which might select $\mathcal{L}_\beta$ when $Pr^N_\alpha(e) > Pr^N_\beta(e)$. Our criterion for coder selection meets this condition. In the third step of the coder selection procedure, the "optimum" coder is selected to minimize the maximum Chernoff information between the probability distribution P associated with the original image and any probability distribution $P^{\alpha}_{q(l)}$ from the optimal achievable region using the coder $\mathcal{L}_\alpha$

(see Section IV). The resulting coder has the property that the best Bayesian error probability over the problems of hypothesis testing between these distributions is greater than the best Bayesian error probability for any other choice of coder.

Section V gives examples where the procedure is applied to compare several coders in the encoding of biomedical 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 VI.

II. Natural Solutions For The Trade-off Between Source Fidelity And Coding Rate

To select the levels of compression which are to be the only natural solutions for the trade-off between source fidelity and coding rate, firstly, we have to define the concept of best achievable compression ratio. Secondly, we prove that the performance bound given by the best achievable compression ratio is operational in that it is directly achievable by a constructive procedure. We then apply a hierarchical method for selecting those ratios which are "locally optimal" in the sense of the definition of best achievable compression ratio. It allows us to analyze the behavior of image coders at different levels of performance across compression ratios.

A. 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 statistical hypothesis testing to decide between any two distributions $P_{q(j)}$ and $P_{q(n)}$ can be stated as follows:

$$H_{q(j)}: \quad Q = P_{q(j)}$$

$$H_{q(n)}: \quad Q = P_{q(n)}$$

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

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

Let $Pr^N(e/H_{q(j)} \text{ or } H_{q(n)})$ be the best achievable Bayesian probability of error for testing

$$H_{q(j)} : Q = P_{q(j)} \text{ vs. } H_{q(n)} : Q = P_{q(n)}$$

where the Bayesian probability of error is the weighted sum of the individual probabilities of error $\beta_{j,n}^N$ and $\alpha_{j,n}^N$. Note that distributions $P_{q(j)}$ and $P_{q(n)}$ which are more difficult to distinguish tend to produce higher values of the best achievable Bayesian probability of error $Pr^N(e/H_{q(j)} \text{ or } H_{q(n)})$, and therefore, $Pr^N(e/H_{q(j)} \text{ or } H_{q(n)})$ may be interpreted as a measure of similarity between $P_{q(j)}$ and $P_{q(n)}$.

For each level of compression $q(i)$, the set of distributions corresponding to the reconstructions of the original image can be divided into two subsets: $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$ and $\{P_{q(i+1)}, \cdots, P_{q(K)}\}$. The best achievable compression ratio for lossy coding at various levels $q(1), q(2), \cdots, q(K)$ will be defined in the following 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 these subsets.

Given a fixed level of compression $q(i)$, let $q(j)$ be any compression ratio such that $i < j$. Consider all the decision problems of hypothesis testing between two alternatives, $H_{q(j)} : Q = P_{q(j)}$ and $H_{q(n)} : Q = P_{q(n)}$, with n such that $1 \leq n \leq i$. Then the average Bayesian probability of error over all pairs of hypotheses $H_{q(j)}$ vs. $H_{q(n)}$ with the a priori probability $Pr(H_{q(j)}) \times Pr(H_{q(n)})$ is

$$Pr_{\leftarrow}^N(e; i, j) = \sum_{1 \leq n \leq i} Pr(H_{q(j)}) \times Pr(H_{q(n)}) \times Pr^N(e/H_{q(j)} \text{ or } H_{q(n)}), \tag{2}$$

with $Pr^N(e/H_{q(j)} \text{ or } H_{q(n)})$ being the best achievable Bayesian probability of error for hypotheses $H_{q(j)}$ vs. $H_{q(n)}$.

The average Bayesian probability of error $Pr_{\leftarrow}^N(e; i, j)$ is essentially determined by the probability of error $Pr^N(e/H_{q(j)} \text{ or } H_{q(n)})$ 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 average Bayesian probability of error

$Pr^N_{\leftarrow}(e;i,j)$, with j such that $i<j$, is simply given by:

$$Pr^N_{\leftarrow}(e;i) = \min_{i<j\leq K}\{Pr^N_{\leftarrow}(e;i,j)\}\,. \tag{3}$$

Thus, $Pr^N_{\leftarrow}(e;i)$ is given by the average Bayesian probability of error $Pr^N_{\leftarrow}(e;i,j)$ associated with 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 best value in the average Bayesian 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:

$$\begin{aligned} H_{q(m)}:&\quad Q = P_{q(m)} \\ H_{q(l)}:&\quad Q = P_{q(l)} \end{aligned}$$

with m such that $i<m\leq K$. Let $Pr^N(e/H_{q(m)} \text{ or } H_{q(l)})$ be the best achievable Bayesian probability of error for testing $H_{q(m)}$ vs. $H_{q(l)}$.

Then, the average Bayesian probability of error for these new $K-i$ decision problems is given by:

$$Pr^N_{\rightarrow}(e;l,i) = \sum_{i<m\leq K} Pr(H_{q(m)})\times Pr(H_{q(l)})\times Pr^N(e/H_{q(m)} \text{ or } H_{q(l)})\,. \tag{4}$$

In this case, the average probability of error $Pr^N_{\rightarrow}(e;l,i)$ is essentially determined by the Bayesian probability of error $Pr^N(e/H_{q(m)} \text{ or } H_{q(l)})$ 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 average Bayesian probability of error $Pr^N_{\rightarrow}(e;l,i)$, with l such that $l\leq i$, is simply given by:

$$Pr^N_{\rightarrow}(e;i) = \min_{1\leq l\leq i}\{Pr^N_{\rightarrow}(e;l,i)\}\,. \tag{5}$$

Thus, $Pr^N_{\rightarrow}(e;i)$ is given by the overall probability of error $Pr^N_{\rightarrow}(e;l,i)$ 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 $Pr^N(e;i)$ be the average sum of $Pr^N_{\rightarrow}(e;i)$ and $Pr^N_{\leftarrow}(e;i)$, i.e.,

$$Pr^N(e;i) = \frac{1}{2}\left[Pr^N_{\rightarrow}(e;i) + Pr^N_{\leftarrow}(e;i)\right]\,. \tag{6}$$

Recall that distributions $P_{q(j)}$ and $P_{q(n)}$ which are more difficult to distinguish tend to produce higher values of the best achievable Bayesian probability of error $Pr^N(e/H_{q(j)} \text{ or } H_{q(n)})$. Hence, optimal levels of compression $q(i)$ in the sense that more clearly separate the set of distributions into two subsets $\{P_{q(1)}, P_{q(2)}, \cdots, P_{q(i)}\}$ and $\{P_{q(i+1)}, P_{q(i+2)}, \cdots, P_{q(K)}\}$, tend to produce lower values of the average probability of error $Pr^N(e;i)$.

We will now define some new notation to express equality to first order in the exponent, as suggested in [10].

Definition 1. 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 2: Best achievable compression ratio. The best achievable compression ratio $q(i^*)$ for the ordered sequence of levels $q(1), q(2), \cdots, q(K)$, is the highest level of compression that attains the asymptotically best value in $Pr^N(e;i)$, i.e.,

$$\min_{1\leq i<K} Pr^N(e;i) \doteq Pr^N(e;i^*)\,, \tag{7}$$

and also verifies a condition of consistency

$$Pr^N_{\rightarrow}(e;i^*) \doteq Pr^N_{\leftarrow}(e;i^*)\,. \tag{8}$$

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

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

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

where

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

and

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

with $C(P_{q(\dagger)}, P_{q(\ddagger)})$ being the Chernoff information[10] between $P_{q(\dagger)}$ and $P_{q(\ddagger)}$.

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 level of compression will allow us to determine a natural boundary between achievable and non-achievable regions, and consequently, it provides useful information to benchmark specific applications. It is worth noting that the Chernoff information appears from asymptotic results as we look for a practical algorithm that comes close to the optimal performance dictated by the bound.

B. A Hierarchical Method For Selecting Natural Solutions For The Rate-Distortion Trade-off

Given a fixed coder in the class being compared, the compression ratio $q(i^*)$ as given by Theorem 1 is such that divides the ordered sequence of levels $q(1), q(2), \cdots, q(K)$, into two subsets to satisfy Definition 2: $\{q(1), q(2), \cdots, q(i^*)\}$ and $\{q(i^*+1), q(i^*+2), \cdots, q(K)\}$. Thus, $q(i^*)$ is selected as the best achievable compression ratio, in the sense of Definition 2, when the behavior of such a lossy coder was analyzed at the coding performance level associated with the sequence of compression ratios $q(1), q(2), \cdots, q(K)$. But, how do we select those ratios which are optimal at different levels of performance for the lossy coder? Theorem 1 is next applied upon the shorter sequence of ratios $q(1), q(2), \cdots, q(i^*)$ which is therefore divided into two new subsets to satisfy the definition of best achievable compression ratio at this new performance level. And this hierarchical method is sequentially applied up to reach the lowest compression ratio $q(1)$. At any individual stage this method selects that ratio which is "locally optimal" in the sense of Definition 2, which allows us to analyze the behavior of the image coder at different levels of performance across compression ratios.

For a given coder, let $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$ be the resultant sequence of best achievable compression ratios; where $q(i_1^*)$ is the best achievable compression ratio for the sequence $q(1), q(2), \cdots, q(K)$; $q(i_2^*)$ is the best achievable compression ratio for the sequence $q(1), q(2), \cdots, q(i_1^*)$, and so forth. The decoded images at these few levels $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$ are to be the only natural solutions (in terms of clarity of separation of the respective sets of distributions into two subsets) for the trade-off between source fidelity and coding rate, across the entire sequence of levels of compression.

III. A Criterion for Selecting The "Optimum" Achievable Region For Lossy Coding.

If $\{P_{q(1)}, \cdots, P_{q(K)}\}$ is the set of discrete probability distributions for the decoded images at compression ratios $q(1), \cdots, q(K)$, then we have that any achievable ratio $q(i_j^*)$ in the subset $\{q(i_1^*), q(i_2^*), \cdots, q(i_s^*)\}$ is such that divides the set of distributions $\{P_{q(1)}, \cdots, P_{q(K)}\}$ into two subsets:

$$A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$$

and

$$B_j = \{P_{q(i_j^*+1)}, \cdots, P_{q(K)}\}$$

where A_j determines the corresponding achievable region, and B_j is the non-achievable region.

In the following a criterion for selecting an optimal achievable region $A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$, with $q(i_j^*)$ in the subset $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$, is proposed. The "optimum" achievable region A_j using this criterion is selected, for each image coder in the class being compared, to maximize the minimum Chernoff information between pairs of distributions which are the most difficult to distinguish from the corresponding achievable and non-achievable regions A_j and B_j. The resulting region A_j has the property that the average Bayesian error probability, for the problems of hypothesis testing between distributions from achievable and non-achievable regions A_j and B_j, is less than the average Bayesian error probability for any other choice of $A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$ with $q(i_j^*)$ in the subset $\{q(i_1^*), q(i_2^*), \cdots, q(i_s^*)\}$.

Consider all the problems of hypothesis testing between $P_{q(m)}$ and $P_{q(l)}$, with $P_{q(m)} \in B_j$ and $P_{q(l)} \in A_j$, i.e.,

$$\begin{aligned} H_{q(m)} &: \quad Q = P_{q(m)} \\ H_{q(l)} &: \quad Q = P_{q(l)} \end{aligned}$$

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

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

$$Pr^N(e; A_j) = \sum_{P_{q(m)} \in B_j} \sum_{P_{q(l)} \in A_j} Pr(H_{q(m)} \times Pr(H_{q(l)}) \times Pr^N(e/H_{q(m)} \text{ or } H_{q(l)}) \,, \tag{12}$$

with $Pr^N(e/H_{q(m)} \text{ or } H_{q(l)})$ being the best achievable Bayesian probability of error for hypotheses $H_{q(m)}$ vs. $H_{q(l)}$.

For each achievable region $A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$ with $q(i_j^*)$ in the subset $\{q(i_1^*), q(i_2^*), \cdots, q(i_s^*)\}$, let $E(A_j)$ be the best achievable exponent in the average Bayesian probability of error $Pr^N(e; A_j)$, i.e.,

$$E(A_j) = \lim_{N \to \infty} -\frac{1}{N} \log Pr^N(e; A_j) \,. \tag{13}$$

We can now define when the achievable region A_j is preferred to any other A_k based on Bayesian error probabilities as follows.

Definition 3. The achievable region A_j is preferred to any other on a Bayesian error probability basis if $E(A_j)$ is the best achievable error exponent for any $A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$ with $q(i_j^*)$ in the subset $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$, i.e.,

$$E(A_j) \geq E(A_k) \,, \quad \text{for all } k \text{ in } \{1, 2, \cdots, s\} \tag{14}$$

where $E(A_j)$ and $E(A_k)$ denotes the best achievable exponent in the average Bayesian error probability associated with A_j and A_k, respectively.

The following result states that the relation of preference given in Definition 3 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 2. The achievable region $A_j = \{P_{q(1)}, \cdots, P_{q(i_j^*)}\}$ is preferred to any other $A_k = \{P_{q(1)}, \cdots, P_{q(i_k^*)}\}$ on a Bayesian error probability basis, if and only if

$$\min\left\{C(P_{q(m)}, P_{q(l)}) \ \mid \ P_{q(m)} \in B_j,\ P_{q(l)} \in A_j\right\} \geq \min\left\{C(P_{q(m)}, P_{q(l)}) \ \mid \ P_{q(m)} \in B_k,\ P_{q(l)} \in A_k\right\} \tag{15}$$

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

Proof. See Appendix B.

This result implies a maximin Chernoff procedure for selection of an achievable region on a Bayesian error probability basis: Select the achievable region A_j which produces the maximum value of the minimum of

$$\left\{C(P_{q(m)}, P_{q(l)}) \ \mid \ P_{q(m)} \in B_j,\ P_{q(l)} \in A_j\right\}$$

as the optimum achievable region in the class $A_1, \cdots, A_j, \cdots, A_s$ being compared.

IV. Coder Selection On A Bayesian Error Probability Basis.

Given a lossy coder $\mathcal{L}_\alpha$, let $\{P^\alpha_{q(1)}, \cdots, P^\alpha_{q(K)}\}$ be the set of discrete probability distributions associated with the decoded images at compression ratios $q(1), \cdots, q(K)$ using coder $\mathcal{L}_\alpha$; A_α be the "optimum" achievable region for coding the original image I using $\mathcal{L}_\alpha$:

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

where $q(i^*_\alpha)$ is the corresponding best achievable ratio.

Consider all the problems of hypothesis testing between two distributions $P^\alpha_{q(l)}$ and P, with $P^\alpha_{q(l)} \in A_\alpha$ and $P = (p(l/I))_l$ characterizing the probability of occurrence of each gray level l in the original image I, i.e.,

$$\begin{aligned} H^\alpha_{q(l)} \quad &: \quad Q = P^\alpha_{q(l)} \\ H_P \quad &: \quad Q = P \end{aligned}$$

The best value in the Bayesian probability of error over all the pair of hypotheses $H^\alpha_{q(l)}$ vs. H_P is

$$Pr^N_\alpha(e) = \min_{P^\alpha_{q(l)} \in A_\alpha} Pr^N(e/H^\alpha_{q(l)} \text{ or } H_P), \tag{16}$$

with $Pr^N(e/H^\alpha_{q(l)}$ or $H_P)$ being the Bayesian probability of error for hypotheses $H^\alpha_{q(l)}$ vs. H_P.

For each $\mathcal{L}_\alpha$ in the class of lossy coders being compared, let E_α be the exponent in the best probability of error $Pr_\alpha^N(e)$ over the problems of hypothesis testing $H_{q(l)}^\alpha$ vs. H_P, i.e.,

$$E_\alpha = \lim_{N\to\infty} -\frac{1}{N}\log Pr_\alpha^N(e)\,. \tag{17}$$

Recall that a statistical hypothesis testing to decide between two distributions $P_{q(l)}^\alpha$ and P which are difficult to distinguish tends to have a high probability of error. Hence, optimal coders $\mathcal{L}_\alpha$ tend to produce a high value of the best probability of error $Pr_\alpha^N(e)$ over the pairs of hypotheses $H_{q(l)}^\alpha$ vs. H_P or, equivalently, a low exponent E_α in the best probability of error over the pair of hypotheses $H_{q(l)}^\alpha$ vs. H_P. And clearly, we should not seriously consider any criterion for selecting a lossy coder which might select $\mathcal{L}_\beta$ when $E_\alpha < E_\beta$. In the following we propose a criterion for coder selection which meets this condition. Thus, the problem of selecting an "optimum" coder for biomedical images is solved by minimizing the exponent in the best probability of error obtained for a lossy coder over all the possible coders.

Definition 4. 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 worst achievable error exponent for any lossy coder in the class being compared, i.e.,

$$E_\alpha \le E_\beta\,, \quad \text{for all } \beta \tag{18}$$

with E_α and E_β being the exponent in the best Bayesian error probability for $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$, respectively.

The following result states that the relation of preference given in Definition 4 guards against the occurrence of a particularly unfavorable pair of hypotheses $H_{q(l)}^\alpha : Q = P_{q(l)}^\alpha$ vs. $H_P : Q = P$ which are the most easy 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

$$\max\left\{C(P_{q(l)}^\alpha, P)\ \mid\ P_{q(l)}^\alpha \in A_\alpha\right\} \le \max\left\{C(P_{q(l)}^\beta, P)\ \mid\ P_{q(l)}^\beta \in A_\beta\right\} \tag{19}$$

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

Proof. See Appendix C.

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

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

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

V. 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 (IV). 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.

A. The Compression Methods

In this comparative study we consider the following image coders:

(EPWIC). Embedded predictive wavelet image coding[11] 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 [12]. Orthonormal wavelet decompositions have proven to be extremely effective for image compression.

(BTPC). Binary tree predictive coding[13] 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[14] 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.

(FRACTAL). A technique to store images in a small amount of space based on the fact that most natural scenes contain affine redundancy, or self-similarity, and that this redundancy can be exploited to compress images. Here we use an implementation given in [15]. The main advantage of fractal compression is that it can be used to create very small files that decompress very quickly. This quality is becoming more and more important in applications such as multimedia and animation. Its main disadvantage is that compressing images are of unpredictable quality.

(JPEG). JPEG's proposed standard aims to be generic, to support a wide variety of applications for continuous-tone images [16]. The JPEG standard includes a DCT-based method that is specified for lossy compression. JPEG features a simple lossy technique known as the Baseline method, a subset of the other DCT-based modes of operation. The Baseline method has been by far the most widely implemented JPEG method to date, and is sufficient in its own right for a large number of applications.

B. The Test Images

Fourteen test images have been used in this study, as showed in Figure 2. They are examples of specific applications: (i) Fluorescency, #1; (ii) cytological cells, #2, #3, and, #4; (iii) DNA, #5, and #6; (iv) chromosomes, #7; (v) ecography, #8; (vi) radiology, #9; (vii) mammography, #10; (viii) magnetic resonance, #11; (ix) muscle cells, #12; and (x) nematodes, #13, and, #14.

C. Results

The test images were coded at various compression ratios using the five coding methods. The lossy compression comparisons were made by setting all the coders to achieve a constant energy constraint on the decoded image at the highest degree of compression as well as at the lowest compression ratio.

The *first step* of the coder selection procedure consists of finding, for each lossy coder in the class being compared, those levels of compression $q(i_1^*), q(i_2^*), \cdots, q(i_s^*)$ which are to be the only natural solutions, in terms of the clarity of separation of the entire set of distributions into two subsets, for the trade-off between source fidelity and coding rate across the sequence of compression ratios. To this aim we use a hierarchical method for selecting those ratios which are "locally optimal" in the sense of the definition of best achievable compression ratio. It allows us to analyze the behavior of image coders at different levels of performance across compression ratios.

Figure 3 shows one example illustrative of the first step in which the best achievable compression ratio, that divides the entire sequence of levels $q(1), q(2), \cdots, q(K)$ into two subsets to satisfy Definition 2, is obtained for SPIHT by using the constructive procedure given in Section (II). The test image in this example can be explained as follows. A digitized mammogram was selected from the Nijmegen database[17] 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.

Figure 3 illustrates the plot of $d^B(i)$ and $d^F(i)$ for SPIHT. As suggested in Theorem 1 the best achievable compression ratio $q(i^*)$ is attained if $d^B(i^*) = d^F(i^*)$. For the entire sequence of levels $q(1), q(2), \cdots, q(K)$, with $K = 69$, it happens at a high compression ratio of $80:1$.

Figure 3 also shows a plot of the false positive to true positive ratio (FP/TP) for microcalcification detection on images reconstructed under various degrees of lossy compression using SPIHT. 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. Figure 3 illustrates the fact that the best achievable compression ratio $q(i^*)$ obtained for the entire sequence of levels $q(1), q(2), \cdots, q(K)$, corresponds to a low FP/TP ratio. 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; that is, it will allow us to distinguish between achievable operating points and those that are suboptimal or unachievable for the given image. Additional discussions of the effects of lossy compression at the best achievable compression ratio on image feature applications can be found in [18].

The *second step* of the coder selection procedure consists of finding the "optimum" achievable region for each coder, where the optimum achievable region has the property that the average Bayesian error probability for the problems of hypothesis testing between probability distributions associated with decoded images at compression ratios from achievable and non-achievable regions, is less than the average Bayesian error probability for any other choice of achievable region. The "optimum" achievable region is so selected to maximize the minimum Chernoff information between pairs of distributions which are the most difficult to distinguish from the corresponding achievable and non-achievable regions, as suggested by Theorem 2.

Figures 4–6 show examples illustrative of this second step in which the optimum achievable region is obtained for EPWIC, BTPC, and, FRACTAL, respectively. In each figure, the left column shows the images reconstructed under the lowest degree of lossy compression using the respective coder.

Optimum Achievable Region					
Image	*SPIHT*	*JPEG*	*EPWIC*	*BTPC*	*FRACTAL*
#1	38 : 1	79 : 1	159 : 1	10 : 1	8 : 1
#2	114 : 1	53 : 1	80 : 1	40 : 1	11 : 1
#3	38 : 1	84 : 1	35 : 1	24 : 1	25 : 1
#4	44 : 1	55 : 1	40 : 1	15 : 1	10 : 1
#5	19 : 1	15 : 1	20 : 1	13 : 1	8 : 1
#6	13 : 1	16 : 1	11 : 1	8 : 1	8 : 1
#7	38 : 1	40 : 1	34 : 1	24 : 1	8 : 1
#8	114 : 1	60 : 1	80 : 1	14 : 1	8 : 1
#9	50 : 1	42 : 1	15 : 1	37 : 1	8 : 1
#10	80 : 1	76 : 1	48 : 1	12 : 1	23 : 1
#11	53 : 1	34 : 1	12 : 1	12 : 1	12 : 1
#12	33 : 1	25 : 1	9 : 1	18 : 1	8 : 1
#13	133 : 1	79 : 1	68 : 1	15 : 1	8 : 1
#14	18 : 1	37 : 1	10 : 1	8 : 1	8 : 1

TABLE I

Best achievable compression ratio associated with the optimum achievable region for each image and lossy coder.

The right column shows the images reconstructed under the highest compression ratio using the coder. And the middle column presents, for each test image, the images reconstructed under the best achievable compression ratio associated with the optimum achievable region. Note that these performance bounds are operational in that they were achievable by a constructive procedure, as suggested in Theorem 2.

Table I illustrates the results of the maximin Chernoff procedure for selection of the optimum achievable region. Thus this table shows the best achievable ratio associated with the optimum achievable region that was obtained for each image and lossy coder in this study. As illustrated in Table I, the best achievable compression ratio is tight for situations of practical relevance (i.e., it corresponds to relatively high coding rate).

The *third step* of the coder selection procedure consists of characterizing an optimum coder on a Bayesian error probability basis, as suggested by Theorem 3.

Table II illustrates the results of the minimax Chernoff procedure for coder selection (Section (IV)). From these results we conclude that, for the mammographic image #10, the optimum coder on an error probability basis is BTPC. This means that BTPC achieves the minimum value of the maximum of the Chernoff information between the probability distribution associated with the original image and any distribution from the optimum achievable region. Figure 8 illustrates, for each coder, the decoded image at the best achievable degree of compression (as dictated by the optimum achievable region) for the mammographic image.

Minimax Chernoff Procedure					
Image	*SPIHT*	*JPEG*	*EPWIC*	*BTPC*	*FRACTAL*
#1	0.040706	0.808833	0.046055	$\boxed{0.021710}$	0.034475
#2	1.154643	1.480808	1.142317	$\boxed{1.138290}$	1.139715
#3	0.117405	0.878438	0.107947	$\boxed{0.060841}$	0.229942
#4	0.137720	0.558733	0.155796	$\boxed{0.066088}$	0.207428
#5	0.604459	0.601902	0.608105	0.602191	$\boxed{0.599515}$
#6	$\boxed{0.202519}$	0.300387	0.206484	0.207002	0.204421
#7	0.524747	1.711144	0.567529	$\boxed{0.087676}$	1.026819
#8	0.011638	0.307328	0.009730	$\boxed{0.007419}$	0.008732
#9	0.003962	0.293141	$\boxed{0.002380}$	0.010348	0.012247
#10	0.346777	0.830912	0.344553	$\boxed{0.338408}$	0.347731
#11	0.389048	1.603840	0.326301	0.297035	$\boxed{0.295902}$
#12	0.059834	0.576816	$\boxed{0.032994}$	0.038129	0.043834
#13	0.126980	0.439066	0.127569	$\boxed{0.119677}$	0.122278
#14	0.787098	1.505670	0.791218	$\boxed{0.550484}$	0.908783

TABLE II

Minimax Chernoff criterion for coder selection.

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 minimax Chernoff procedure. The minimum value of the maximum of the Chernoff information between the distribution of the original image and any distribution from the optimum achievable region, is displayed in a box. It follows that: (i) BTPC is the optimum coder for test images #1, #2, #3, #4, #7, #8, #10, #13, and #14; (ii) FRACTAL is the optimum coder for test images #5, and #11; (iii) EPWIC is the optimum coder for test images #9, and #12; and (iv) SPIHT is the optimum coder for test image #6. From these results we claim that BTPC shows the best overall performance. It can be visually evaluated in Figures 7–9, where the decoded images at the best achievable degree of compression as dictated by the optimum achievable region are illustrated for some test images of example.

VI. Conclusion

The previous examples illustrate the technique for obtaining "optimum coders" using the minimax Chernoff procedure. It optimizes coder performance for pairs of hypotheses, between the distribution of the original image and any distribution from the optimum achievable region, which are the most easy to distinguish in the sense of the Chernoff information. For this reason the coders obtained by this procedure tend to have a high probability of error for the problems of hypothesis testing between the distribution of the original image and any distribution from the optimum achievable region. Thus if the rate-distortion measure is unknown or cannot be conveniently evaluated for a biomedical application, the minimax Chernoff procedure provides a useful criterion for coder selection.

VII. Appendix A

Proof of Theorem 1. To prove the theorem, firstly, we have to prove that the probabilities of error $Pr^N_{\leftarrow}(e;i)$ and $Pr^N_{\rightarrow}(e;i)$ 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 Chernoff's theorem (see Theorem 12.9.1 in [10]), we have that the best achievable Bayesian probability of error $Pr^N(e/H_{q(\dagger)} \text{ or } H_{q(\ddagger)})$ for the hypothesis test between $P_{q(\dagger)}$ and $P_{q(\ddagger)}$ is:

$$Pr^N(e/H_{q(\dagger)} \text{ or } H_{q(\ddagger)}) \doteq 2^{-N\times C(P_{q(\dagger)},P_{q(\ddagger)})}, \tag{20}$$

with N being the image size in pixels; $C(P_{q(\dagger)}, P_{q(\ddagger)})$ being the Chernoff information[10]:

$$C(P_{q(\dagger)}, P_{q(\ddagger)}) = -\min_{0\le\lambda\le 1} \log \sum_l \left[\left(p(l/I_{q(\dagger)})\right)^{\lambda} \times \left(p(l/I_{q(\ddagger)})\right)^{1-\lambda} \right], \tag{21}$$

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

Then

$$\begin{aligned} Pr^N_{\leftarrow}(e;i,j) &= \sum_{1\le n\le i} Pr(H_{q(j)}) \times Pr(H_{q(n)}) \times Pr^N(e/H_{q(j)} \text{ or } H_{q(n)}) \\ &\overset{(a)}{\doteq} \sum_{1\le n\le i} Pr(H_{q(j)}) \times Pr(H_{q(n)}) \times 2^{-N\times C(P_{q(j)},P_{q(n)})} \\ &\overset{(b)}{\doteq} 2^{-N \min_{1\le n\le i} C(P_{q(j)},P_{q(n)})} \end{aligned}$$

where

(a) follows from Chernoff's theorem and equation (20), and

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

And it follows that

$$\begin{aligned} Pr^N_{\leftarrow}(e;i) &= \min_{i<j\le K} \{Pr^N_{\leftarrow}(e;i,j)\} \\ &\doteq \min_{i<j\le K} 2^{-N \min_{1\le n\le i} C(P_{q(j)},P_{q(n)})} \end{aligned}$$

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

Similarly, we can prove that

$$
\begin{aligned}
Pr^N_{\rightarrow}(e;l,i) &= \sum_{i<m\le K} Pr(H_{q(m)}) \times Pr(H_{q(l)}) \times Pr^N(e/H_{q(m)} \text{ or } H_{q(l)}) \\
&\doteq \sum_{i<m\le K} Pr(H_{q(m)}) \times Pr(H_{q(l)}) \times 2^{-N\times C(P_{q(m)},P_{q(l)})} \\
&\doteq 2^{-N \min_{i<m\le K} C(P_{q(m)},P_{q(l)})}
\end{aligned}
$$

since the exponential rate is determined by the worst exponent, and then

$$
\begin{aligned}
Pr^N_{\rightarrow}(e;i) &= \min_{1\le l\le i}\{Pr^N_{\rightarrow}(e;l,i)\} \\
&\doteq \min_{1\le l\le i} 2^{-N \min_{i<m\le K} C(P_{q(m)},P_{q(l)})} \\
&= 2^{-N \max_{1\le l\le i}\left\{\min_{i<m\le K} C(P_{q(m)},P_{q(l)})\right\}} \\
&= 2^{-N\times d^F(i)} .
\end{aligned}
$$

This proves that $Pr^N_{\leftarrow}(e;i)$ and $Pr^N_{\rightarrow}(e;i)$ behave exponentially with exponents given by $d^B(i)$ and $d^F(i)$, respectively.

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

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

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

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

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

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

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

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

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.

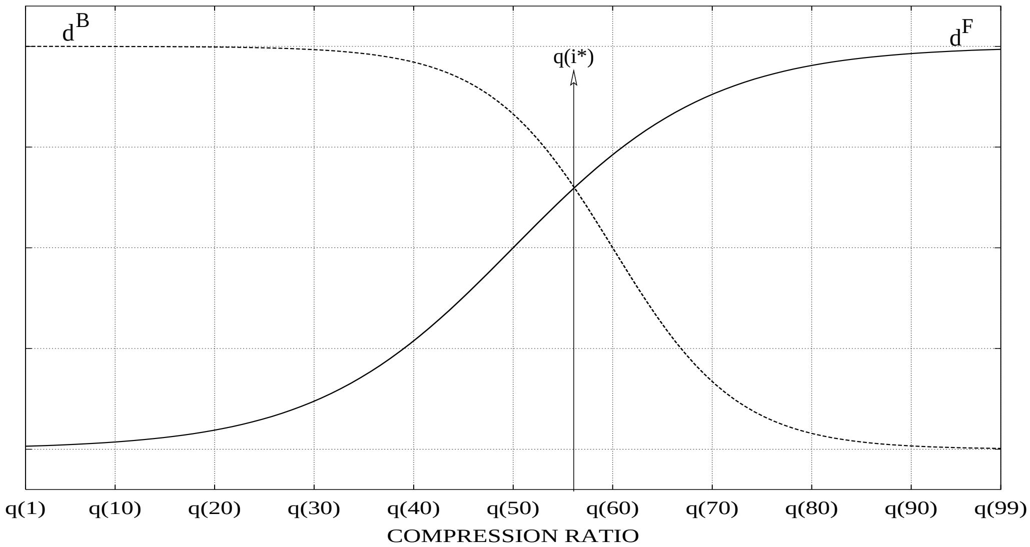

Fig. 1. $d^F(i)$ is non-decreasing and $d^B(i)$ is non-increasing; the maximum value of the minimum of $\{d^F(i), d^B(i)\}$ is attained if they are equal.

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

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

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

$$Pr^N_{\rightarrow}(e;i^*) \doteq 2^{-N\times d^F(i^*)}$$

$$
\begin{aligned}
&= 2^{-N\times d^B(i^*)} \\
&\doteq Pr^N_{\leftarrow}(e;i^*)\,.
\end{aligned}
$$

This proves the theorem.

VIII. Appendix B

Proof of Theorem 2. If we average over all pairs of hypotheses $H_{q(m)}$ and $H_{q(l)}$ with the a priori probability $Pr(H_{q(m)}) \times Pr(H_{q(l)})$ we obtain the average Bayesian probability of error

$$
\begin{aligned}
Pr^N(e;A_j) &= \sum_{P_{q(m)}\in B_j}\sum_{P_{q(l)}\in A_j} Pr(H_{q(m)} \times Pr(H_{q(l)}) \times Pr^N(e/H_{q(m)} \text{ or } H_{q(l)}) \\
&\overset{(a)}{\doteq} \sum_{P_{q(m)}\in B_j}\sum_{P_{q(l)}\in A_j} Pr(H_{q(m)}) \times Pr(H_{q(l)}) \times 2^{-N\times C(P_{q(m)},P_{q(l)})} \\
&\overset{(b)}{\doteq} 2^{-N\min\left\{C(P_{q(m)},P_{q(l)})\ |\ P_{q(m)}\in B_j,\, P_{q(l)}\in A_j\right\}}\,,
\end{aligned} \tag{22}
$$

where

(a) follows from equation (20), and

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

Thus

$$
\begin{aligned}
E(A_j) &= \lim_{N\to\infty} -\frac{1}{N}\log Pr^N(e;A_j) \\
&\overset{(a)}{=} \lim_{N\to\infty} -\frac{1}{N}\log\left(2^{-N\min\left\{C(P_{q(m)},P_{q(l)})\ |\ P_{q(m)}\in B_j,\, P_{q(l)}\in A_j\right\}}\right) \\
&= \min\left\{C(P_{q(m)},P_{q(l)})\ \mid\ P_{q(m)}\in B_j,\, P_{q(l)}\in A_j\right\}\,,
\end{aligned} \tag{23}
$$

where (a) follows from equation (22).

Similarly, it follows that

$$
E(A_k) = \min\left\{C(P_{q(m)},P_{q(l)})\ \mid\ P_{q(m)}\in B_k,\, P_{q(l)}\in A_k\right\}\,. \tag{24}
$$

Given two achievable regions A_j and A_k we say that A_j is preferred to A_k on a Bayesian error probability basis if we have

$$
E(A_j) \geq E(A_k)\,,
$$

or equivalently, following equations (23) and (24) we have that

$$
\min\left\{C(P_{q(m)},P_{q(l)})\ \mid\ P_{q(m)}\in B_j,\, P_{q(l)}\in A_j\right\} \geq \min\left\{C(P_{q(m)},P_{q(l)})\ \mid\ P_{q(m)}\in B_k,\, P_{q(l)}\in A_k\right\}
$$

which can be interpreted as A_j being the optimum achievable region in the sense of the occurrence of a particularly unfavorable pair of hypotheses $H_{q(m)} : Q = P_{q(m)}$ vs. $H_{q(l)} : Q = P_{q(l)}$, with $P_{q(m)} \in B_j$ and $P_{q(l)} \in A_j$, which are the most difficult to distinguish by means of the Chernoff information. This proves the theorem.

IX. Appendix C

Proof of Theorem 3. The best value in the Bayesian probability of error over the pairs of hypotheses $H^{\alpha}_{q(l)}$ vs. H_P is

$$\begin{aligned} Pr^N_{\alpha}(e) &= \min_{P^{\alpha}_{q(l)} \in A_{\alpha}} Pr^N(e/H^{\alpha}_{q(l)} \text{ or } H_P) \\ &\overset{(a)}{\doteq} \min_{P^{\alpha}_{q(l)} \in A_{\alpha}} 2^{-N \times C(P^{\alpha}_{q(l)}, P)} \\ &= 2^{-N \max\left\{C(P^{\alpha}_{q(l)}, P) \;|\; P^{\alpha}_{q(l)} \in A_{\alpha}\right\}}, \end{aligned} \tag{25}$$

where (a) follows from equation (20).

Thus

$$\begin{aligned} E_{\alpha} &= \lim_{N \to \infty} -\frac{1}{N} \log Pr^N_{\alpha}(e) \\ &\overset{(a)}{=} \lim_{N \to \infty} -\frac{1}{N} \log \left(2^{-N \max\left\{C(P^{\alpha}_{q(l)}, P) \;|\; P^{\alpha}_{q(l)} \in A_{\alpha}\right\}} \right) \\ &= \max\left\{ C(P^{\alpha}_{q(l)}, P) \;\;|\;\; P^{\alpha}_{q(l)} \in A_{\alpha} \right\}, \end{aligned} \tag{26}$$

where (a) follows from equation (25).

Similarly, it follows that

$$E_{\beta} = \max\left\{ C(P^{\beta}_{q(l)}, P) \;\;|\;\; P^{\beta}_{q(l)} \in A_{\beta} \right\}. \tag{27}$$

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

$$E_{\alpha} \leq E_{\beta},$$

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

$$\max\left\{ C(P^{\alpha}_{q(l)}, P) \;\;|\;\; P^{\alpha}_{q(l)} \in A_{\alpha} \right\} \leq \max\left\{ C(P^{\beta}_{q(l)}, P) \;\;|\;\; P^{\beta}_{q(l)} \in A_{\beta} \right\}$$

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 $H^{\alpha}_{q(l)}$ vs. H_P which are the most easy 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] Cosman, P.C., Tseng, C., Gray, R.M., Olshen, R.A., Moses, L.E., Davidson, H.C., Bergin, C.J., Riskin, E.A., "Tree-structured vector quantization of CT chest scans: Image quality and diagnostic-accuracy,", *IEEE Trans. on Medical Imaging,* Vol. 12, No. 4, pp. 727-739, (1993).

[2] 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).

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 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).

[8] Linder, T., Zamir, R., Zeger, K. "High-resolution source coding for non-difference distortion measures: multidimensional companding," *IEEE Trans. on Information Theory,* Vol. 45, No. 2, pp. 548-561, (1999).

[9] Li, J., Chaddha, N., and Gray, R.M. "Asymptotic performance of vector quantizers with the perceptual distortion measure," Proc. IEEE Int. Symp. Information Theory, pp. 55, Germany, June, (1997).

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

[11] 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).

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

[13] 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).

[14] 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).

[15] Paynter, G.W., "Fractal Image Compression," *0657.420 Report of an Investigation.* Department of Computer Science. The University of Waikato, Hamilton, New Zealand, (1995).

[16] Wallace, G., Vivian, R., and Polusen, H., "Subjective testing results for still picture compression algorithms for international standardization." In *Proceedings of the IEEE Global Telecommunications Conference,* pp. 1022-1027, IEEE Communications Society, November (1988).

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

[18] 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, $ftp://decsai.ugr.es/pub/diata/tech\_rep/TR990324.ps.Z$, Department of Computer Science and Artificial Intelligence, University of Granada, Spain, (1999).

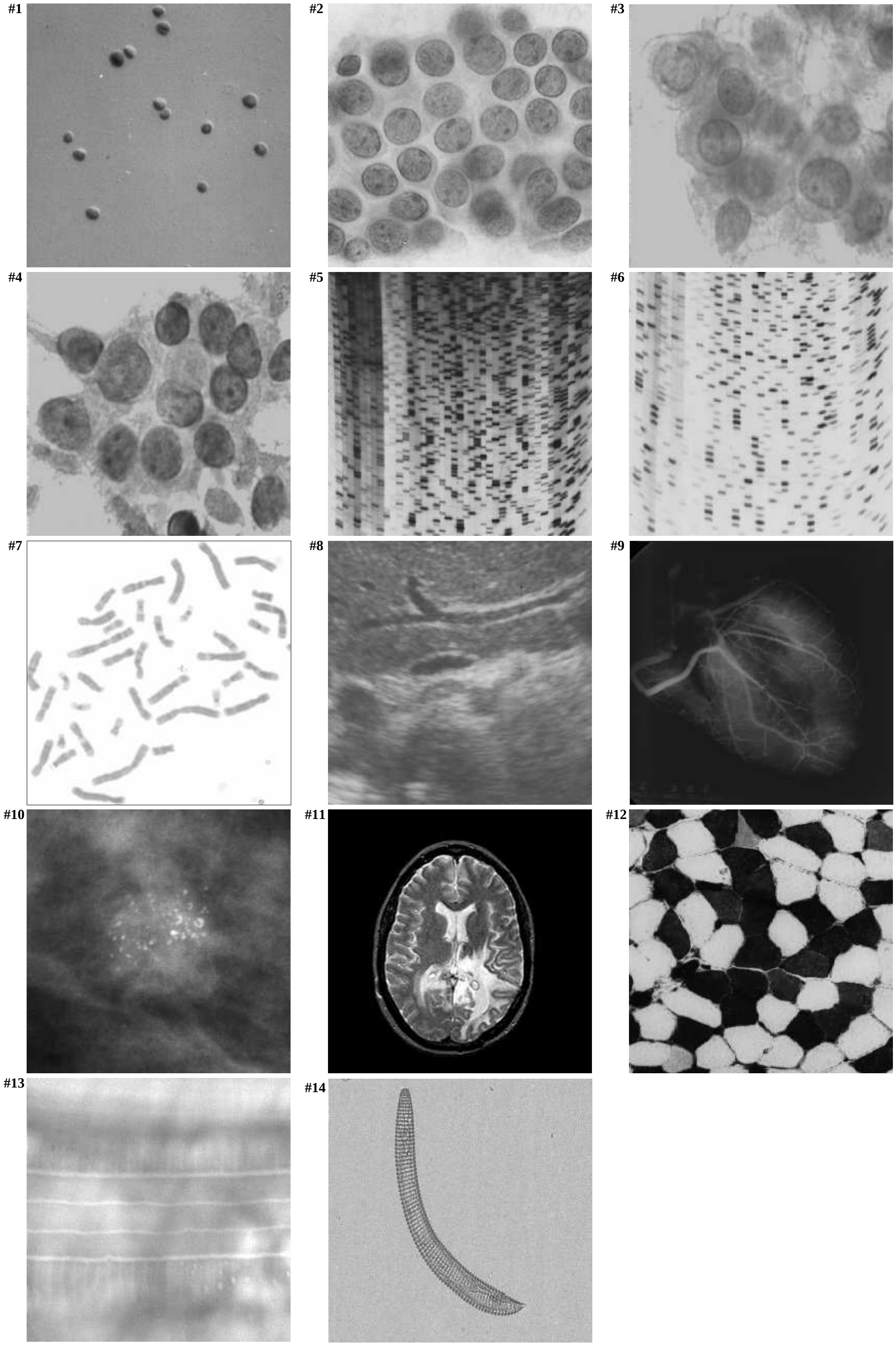

Fig. 2. Fourteen biomedical images used in the comparative study.

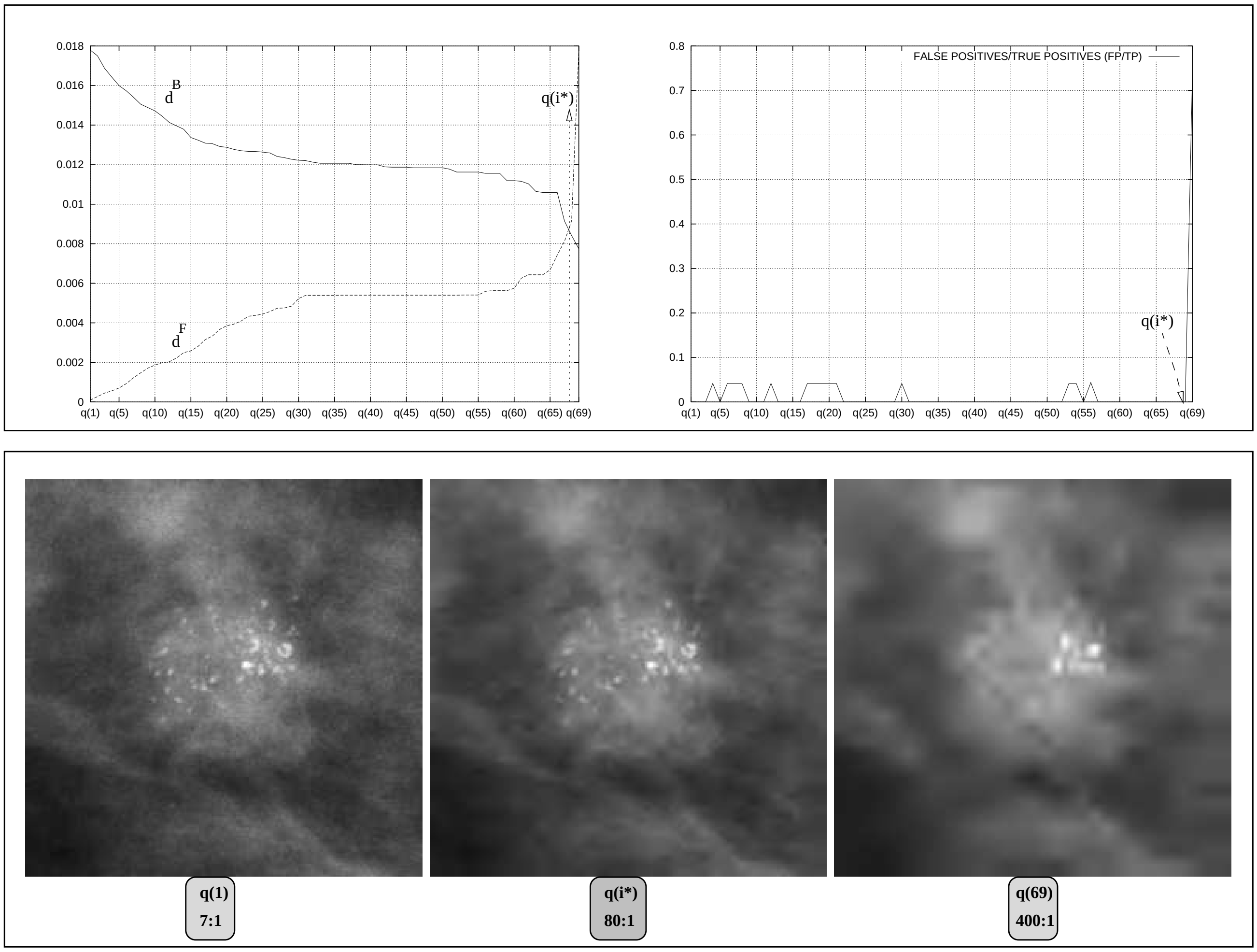

Fig. 3. Plot of $d^B(i)$ and $d^F(i)$ for SPIHT; Best achievable compression ratio $q(i^*)$; the relationship between this point and the false positive to true positive (FP/TP) microcalcification detection ratio for the images reconstructed under different degrees of lossy compression using SPIHT.

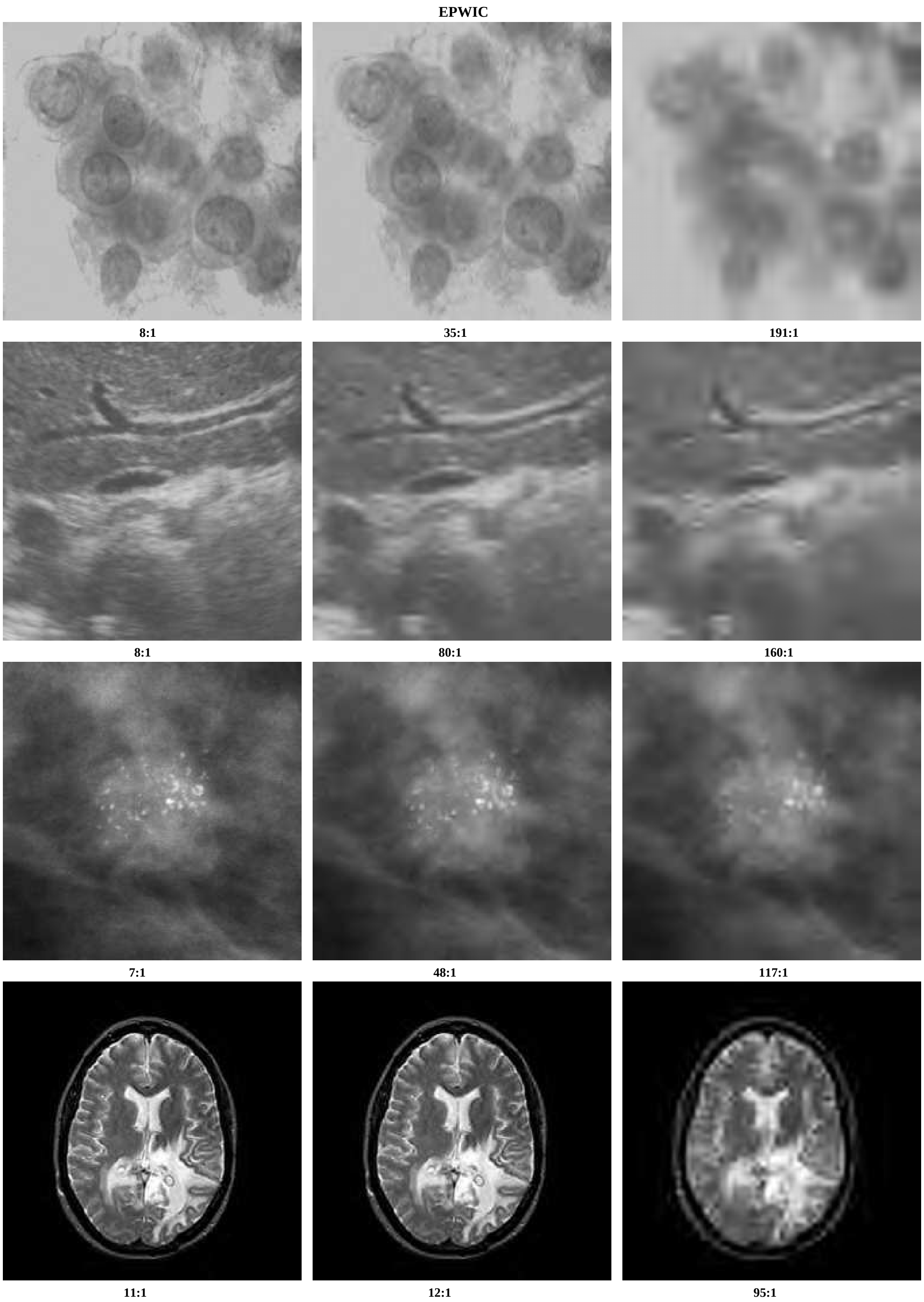

Fig. 4. Optimum achievable regions obtained for EPWIC: On the left column, decoded images at the lowest degree of lossy compression; on the middle column, decoded images at the best achievable compression ratio associated with the optimum achievable region; on the right column, decoded images at the highest degree of lossy compression.

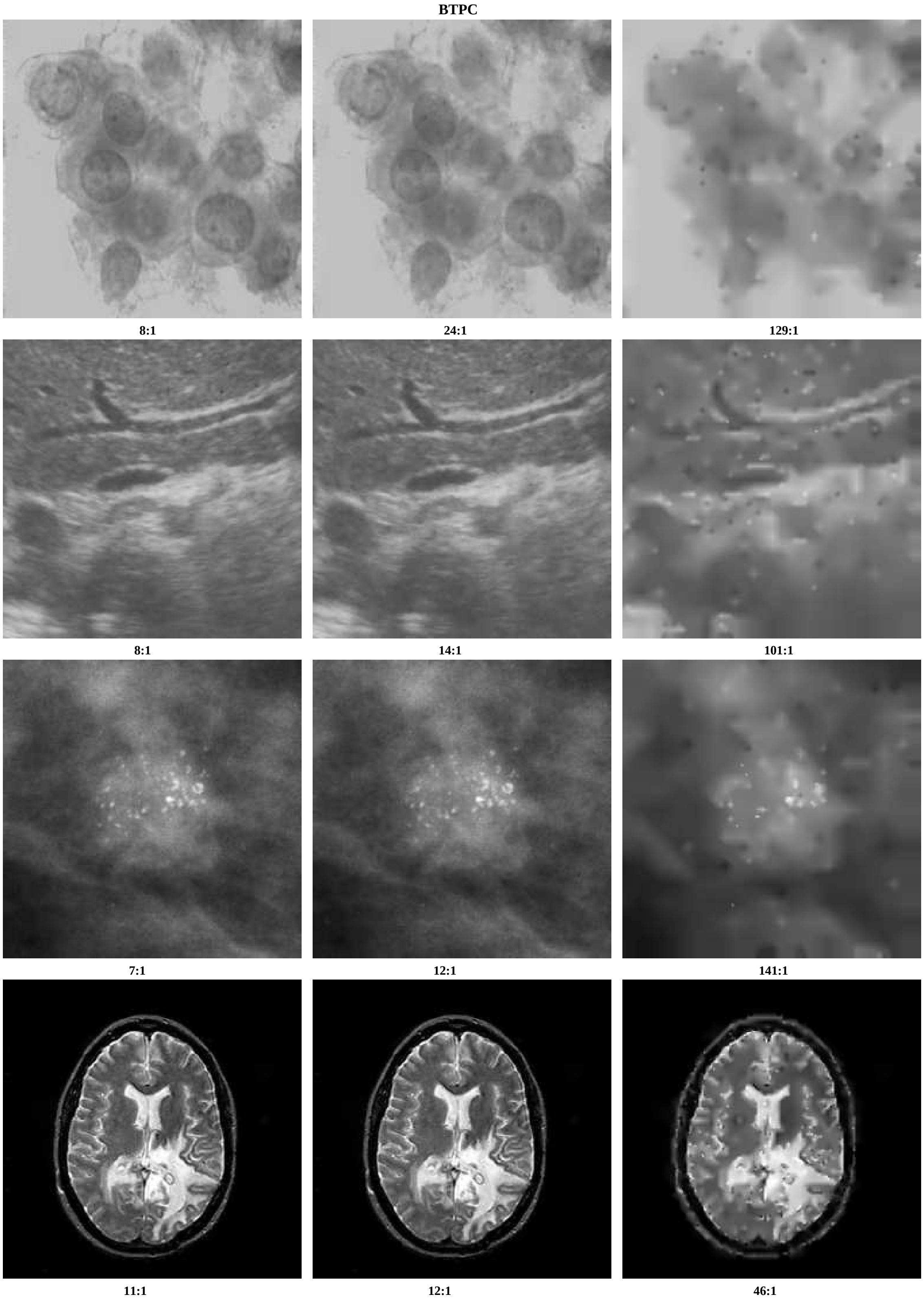

Fig. 5. Optimum achievable regions obtained for BTPC: On the left column, decoded images at the lowest degree of lossy compression; on the middle column, decoded images at the best achievable compression ratio associated with the optimum achievable region; on the right column, decoded images at the highest degree of lossy compression.

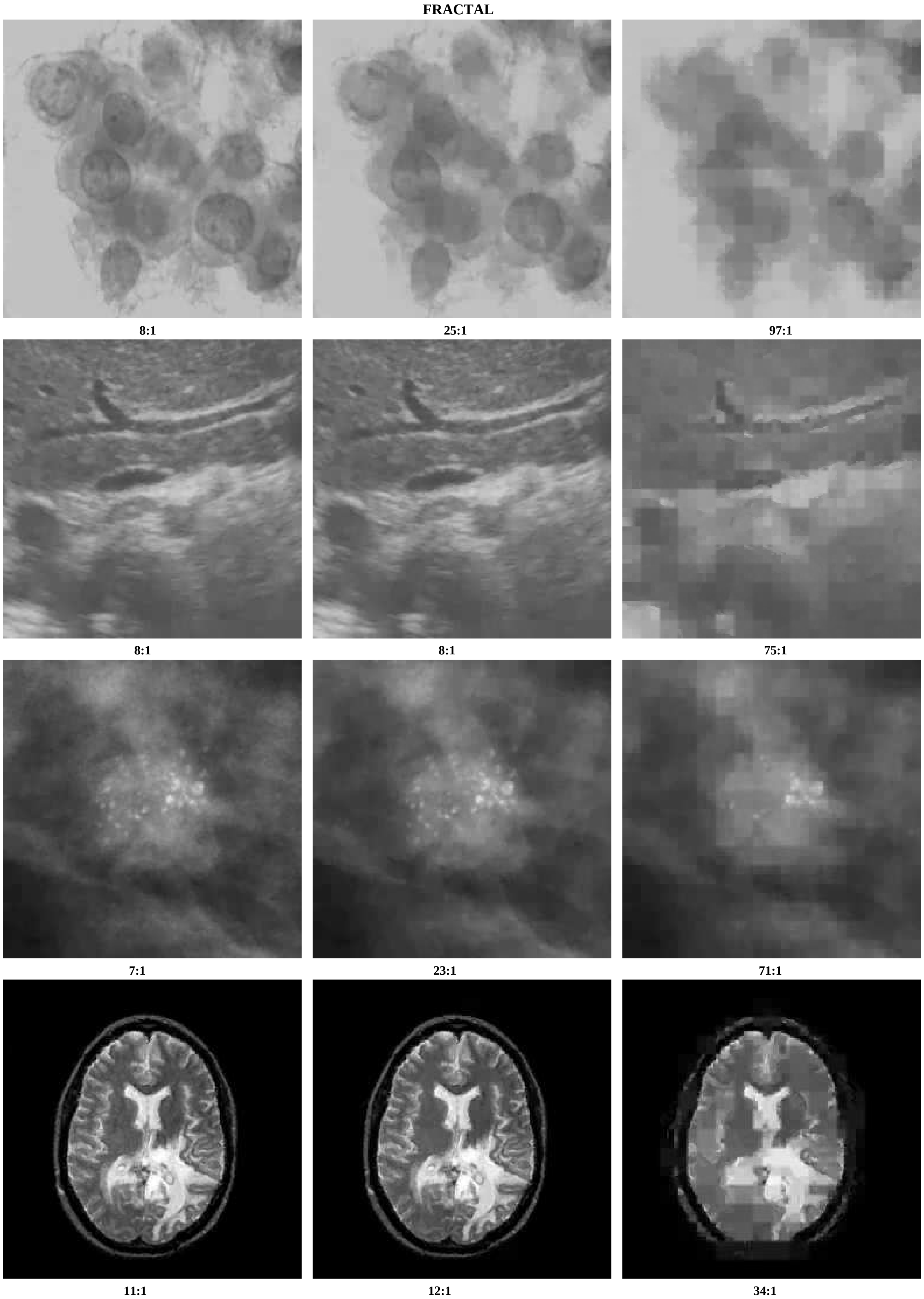

Fig. 6. Optimum achievable regions obtained for FRACTAL: On the left column, decoded images at the lowest degree of lossy compression; on the middle column, decoded images at the best achievable compression ratio associated with the optimum achievable region; on the right column, decoded images at the highest degree of lossy compression.

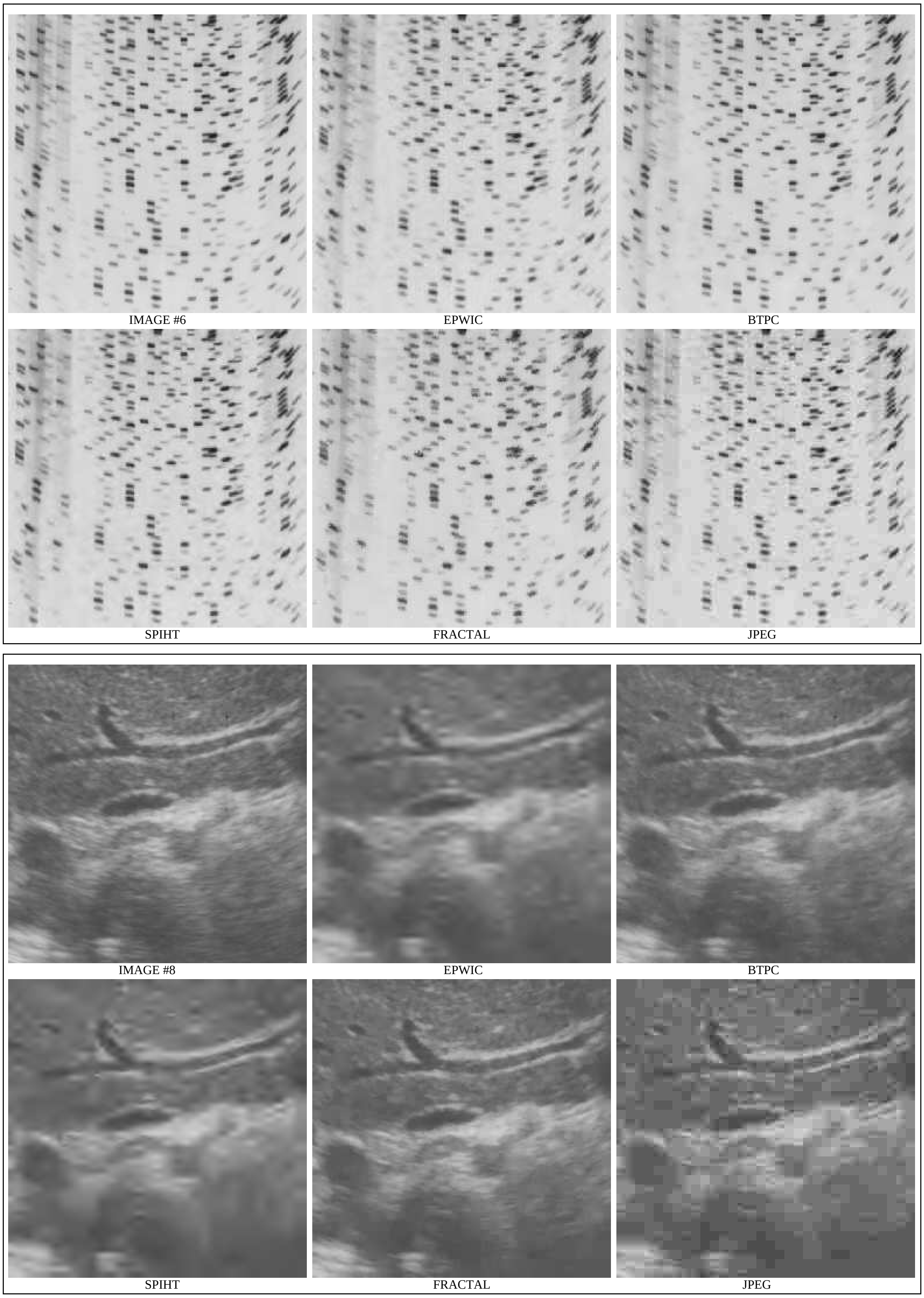

Fig. 7. For each lossy coder, decoded image at the best achievable degree of compression as dictated by the optimum achievable region.

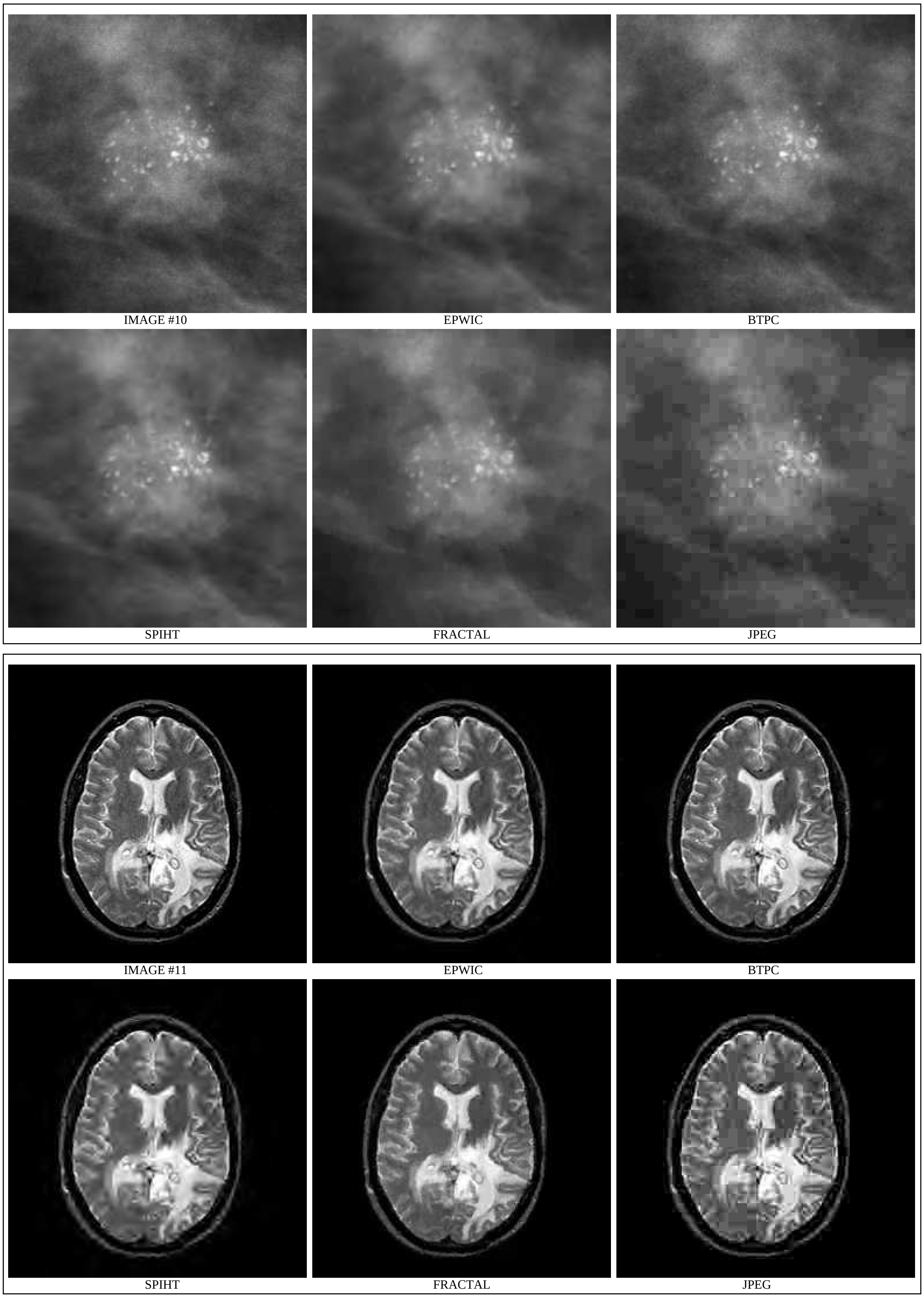

Fig. 8. For each lossy coder, decoded image at the best achievable degree of compression as dictated by the optimum achievable region.

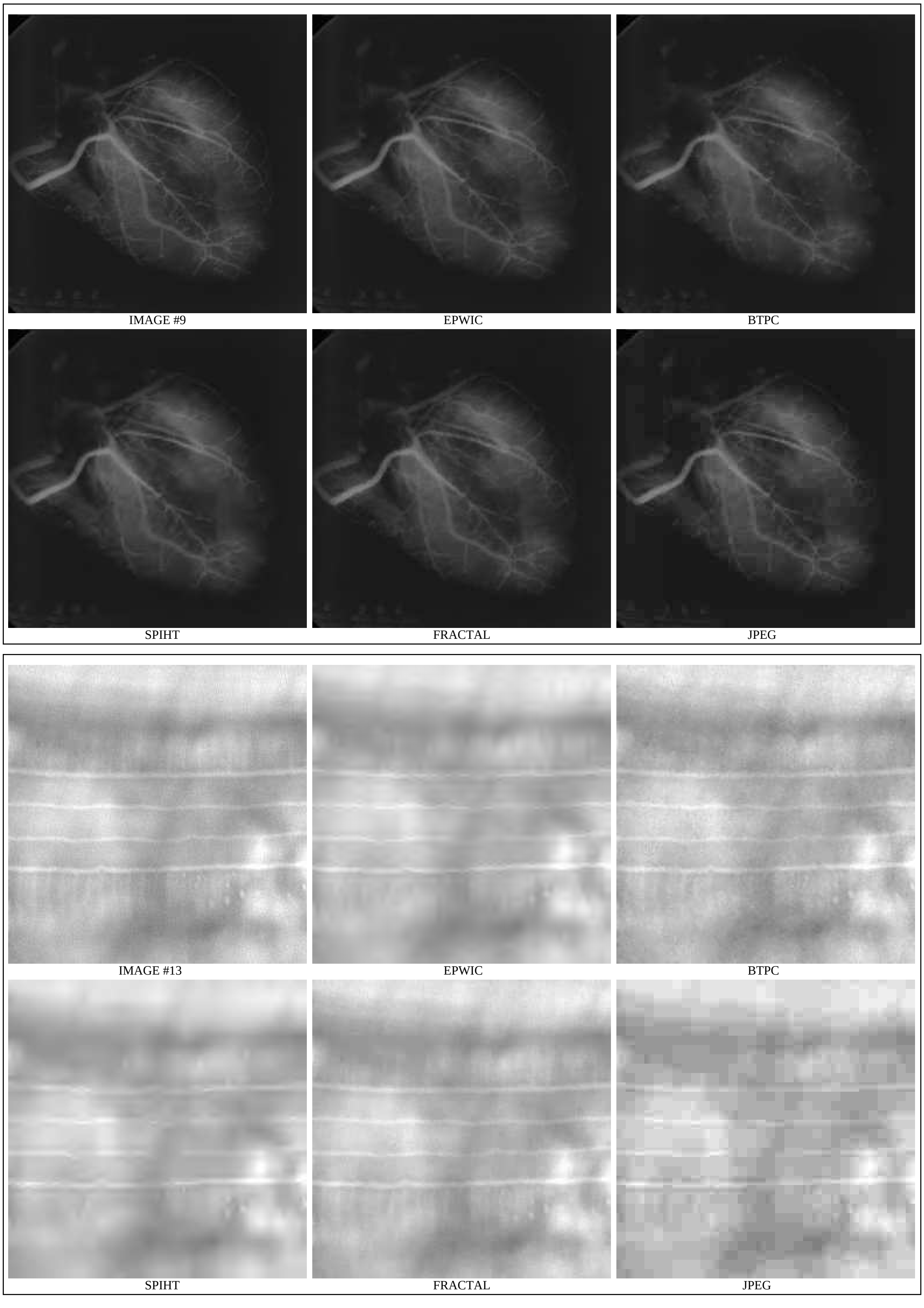

Fig. 9. For each lossy coder, decoded image at the best achievable degree of compression as dictated by the optimum achievable region.