Vector quality measure of lossy compressed medical images

Artur Przelaskowski

Institute of Radioelectronics, Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warszawa, Poland

Tel: +48 22 6607917; fax: +48 22 8251363; e-mail: arturp@ire.pw.edu.pl

Vector quality measure of lossy compressed medical images

Artur Przelaskowski

Abstract

A numerical measure, which is able to predict diagnostic accuracy rather than subjective quality is required for compressed medical image assessment. The objective of this study is to present a proposition for a new vector measure of image quality reflecting diagnostic accuracy. Construction of such measure includes the formation of a diagnostic quality pattern based on the subjective ratings of local features of an image, playing an essential role in the detection and classification of any lesion. Experimental results contain the opinions of 9 radiologists: 2 test designers and 7 observers who rated digital mammograms. The correlation coefficient between the numerical equivalent of the vector measure and subjective rates is over 0.9.

Keywords: Quality measures; Subjective rating; Diagnostic accuracy; Lossy image compression

1. Introduction

An important issue of using irreversible compression in order to meet storage and image transfer requirements is the proper characterization and measurement of diversified distortions of processed medical images. Lossy image coders are used to increase image management effectiveness but the acceptable distortion level is often an open question in medical applications. Because the diagnostic accuracy of originals should be preserved in reconstructed images reliable measures of diagnostic accuracy are sought. The most common means of measuring diagnostic accuracy for computer-processed medical images is based on receiver operating characteristic (ROC) analysis which has its origins in the theory of signal detection [1]. Nevertheless, ROC analysis has no natural extension to the evaluation of measurement accuracy in compressed medical images [2]. Erickson [3] suggested that ROC studies evaluating both low- and high- frequency features as well as textures are likely to be most valuable. Results of ROC analysis have a statistical nature referred to many diagnostic decisions, large set of test images and studied processing methods. However, the nature of individual diagnosis is not statistical and pathology detection concerns concrete cases. Single image information is interpreted in appropriate terms of diagnosis on the grounds of radiologists' knowledge and experience. Thus measuring of a single, concrete image in diagnostic terms is really important because it reflects radiological practice and is useful for compression effects acceptance. Further, the results of subjective assessment, i.e. pathology detection, quality rating etc., may vary depending on test conditions, radiologists' experience, tiredness and other compression-independent factors. Subjective methods do not provide any constructive methodology for

performance improvement. It is hard to include these methods into the optimisation process of coder efficiency. The purpose of our research is to design an image quality measure based on assessment of local image features important for diagnosis. This measure is intended to predict diagnostic accuracy of a single compressed image related to the original and could be useful in lossy coder construction.

A key attribute of simple and often used numerical distortion measures, such as mean squared error (MSE) or signal-to-noise-ratio (SNR) is objectivity. Besides, these measures feature ease of computation but they do not provide for the precise characteristics of the complex nature of possible distortions. For example, MSE is a good distortion indicator for random errors but rather unsuitable for structured, correlated or other compression-dependent errors. An objective distortion measure that reflects perceptual quality perception or usefulness in rating of diagnostic information degradation is desired. Such an objective prediction of coder performance both with respect to bit rate and image quality or diagnostic value would lead to systematic design of compression procedures.

Better distortion characteristics in a qualitative and quantitative sense are available by applying objective vector measures, which can extend separate error specification. Hosaka plots [4] are a typical example. This graphical method is used to characterise a number of reconstructed image features and to compare them with the corresponding features of the original. A difference between two feature vectors is a vector measure presented in a graphical form. Other vector measures, e.g. Eskicioglu charts [5], are univariate. It means that error vectors are extracted independently from the original and the reconstructed images. Generally, vector measures have an accepted complexity, an increased degree of correlation with subjective quality evaluation but they are inconvenient for the comparison of coded images because of difficulty in the interpretation of multidimensional graphical forms (a criterion is not unequivocal). The solution is a scalar equivalent of vector measure as final quality score for comparisons and acceptance fixing. This equivalent should be constructed with a criterion of the highest correlation with diagnostic accuracy of analysed images for medical applications. Miyahara [6] presented the determination method of several distortion factors combined by regression into a single number representative of the quality of a given image. Picture Quality Scale (PQS) is a methodology for the determination of objective quality metrics to evaluate the quality of coded still images. This approach is based on

the perceptual properties of human vision and extensive engineering experience with the observation of actual image disturbances resulting from image coding. Two basic disadvantages of PQS are as follows: it was not provided for medical applications (PQS approach was based on the perceptual properties of human vision), and there was no proposed graphical form to extend error characteristics for an intensive image quality evaluation. Additionally, it is mainly adjusted to the distortions caused by block transform coders.

For medical images, it is really important that a computable objective measure is able to predict diagnostic accuracy rather than subjective quality. We propose a new method that incorporates diagnostic quality estimation into the construction process of a vector distortion measure to make it appropriate for diagnostic accuracy assessment. The assumptions of measure designing and construction are as follows:

- different kinds of errors are captured by a set of factors - elements of the vector measure;
- diagnostic quality pattern (DQP) is approximated for each image; it is made by subjective ratings with protocols grouping basic 'diagnostic features' of the image, i.e. elements which are crucial in lesion detection and definition; reviewer opinion about the ability to detect lesions and classification are notified; a numerical scale with a corresponding diagnostic description for each number is used;
- perceptual quality evaluation of diagnostically important image features is incorporated into the optimisation process of the vector measure; DQP is used for more reliable calculation of a single number - the equivalent of diagnostically related image quality; this vector measure equivalent is a final numerical factor assigned to the encoded image, useful in compression technique optimisation and for an estimation of acceptable-in-diagnosis ratios;
- graphical presentation of the multidimensional vector measure of diagnostic image quality can make error analysis more penetrating and the comparison of compression algorithms more comprehensive; the complex distortion characteristics include global and local reconstruction accuracy, data correlation, random and structured errors, and the distribution of error energy.

A hybrid image quality measure verified in a pilot study is the main contribution of this paper. 9 radiologists took part in the procedure of original and reconstructed mammograms assessment to estimate a 'quality of

abnormalities features' perceptually. Close to one hundred images were rated with a scale describing the contrast, sharpness, shape and outlines of abnormalities, i.e. the quality of diagnostically important features of lesion detection process. More than 200 digital mammograms were acquired and used in test design.

2. Methods

A large set of quality measures was considered to select the most effective factors in the construction of hybrid vector measure of reconstruction accuracy. We tested scalar numerical measures of global and local distortions, well known from research concerning image quality evaluation [5][7]. Next, vector quality measure called Hosaka plots were tested [8] because Hosaka plots were often used as a good compressed image quality measure [9]. Moreover, five factors of PQS were considered as a good measure of various distortions appearing in lossy compression. Initially, we approximated an accuracy pattern (partly in diagnostic terms) for images of different modalities in *ad hoc* organised subjective tests [8], selected and optimised the most appropriate combination of the scalar factors. Hybrid measure for compressed image accuracy was constructed from this background. This measure is called hybrid because of the combining of vector (plots) and scalar (equivalent) distortion representation, and because of the subjective (accuracy pattern in initial stage) and objective (in computable final stage) character of given results.

2.1. Image quality measures

Scalar and vector quality measures influencing the process of hybrid measure design were described briefly in this subsection.

2.1.1. Objective scalars

Given an original image $f(x,y)$ and a distorted, compressed-reconstructed image $\hat{f}(x,y)$, we tested several objective scalar measures as listed below:

- Average Difference: $$AD = \frac{1}{MN}\sum_{x,y}\left|f(x,y) - \hat{f}(x,y)\right| \quad (1)$$
- Correlation Quality: $$CQ = \frac{\sum_{x,y} f(x,y)\cdot \hat{f}(x,y)}{\sum_{x,y} f(x,y)} \quad (2)$$
- Maximum Difference: $$MD = \max_{x,y}\{|f(x,y) - \hat{f}(x,y)|\} \quad (3)$$

- Image Fidelity: $$IF = 1 - \frac{\sum_{x,y}[f(x,y) - \hat{f}(x,y)]^2}{\sum_{x,y}[f(x,y)]^2} \quad (4)$$

- Mean Square Error: $$MSE = \frac{1}{MN}\sum_{x,y}[f(x,y) - \hat{f}(x,y)]^2 \quad (5)$$

- Peak Signal to Noise Ratio: $$PSNR = 10\log_{10}\frac{MN \cdot [\max_{x,y}\{f(x,y)\}]^2}{\sum_{x,y}[f(x,y) - \hat{f}(x,y)]^2} \quad (6)$$

- Chi-Square Measure: $$\chi^2 = \frac{1}{MN}\sum_{x,y}\frac{\left[f(x,y) - \hat{f}(x,y)\right]^2}{f(x,y)} \quad (7)$$

These measures can characterize different kinds of distortions in lossy compression and thus could be considered as potential elements of vector quality measure. The majority of the scalar measures are based on point differences between the original and the reconstructed images. A better reconstruction of original means lower point differences. Hence lower value of *AD*, *CQ*, *MD*, *MSE* and χ^2, and higher value of *PSNR* and *IF* means reconstruction done in better way.

2.1.2. Hosaka plots

Hosaka plots are a measure of a number of features of reconstructed image in relation to the corresponding features of the original. The difference between two feature vectors generates a vector error measure (differences in the corresponding features of the original and reconstruction) plotted in a graphical form. The procedure of plots calculation consists of the following operations: quad-tree segmentation of the original image (the same block classification is also applied to the reconstructed image), estimation of mean intensity values and mean standard deviation in each class of blocks (depending on the size of the blocks), and computation and presentation of two error vectors (containing differences in mean intensity and mean standard deviation of the corresponding original and reconstructed classes). Vector errors are plotted in polar coordinates where the radius is the feature error value, the left half plan contains vectors of differences in standard deviation with vectors of mean intensity in the right half plan each equally spaced for the different classes (blocks of size 1×1 - only for mean intensity, and 2×2, 4×4, 8×8, 16×16) so as to fill the plan, as shown in Fig. 1. The area of the plots is related to the image quality (i.e. is inversely proportional to reconstruction fidelity to original) while the structure

of the diagram depends on the type of distortion. The area of plots in the left half plan (called the noise area) is more sensitive to an additive noise, while the right side area (called the fidelity of reconstruction area) more precisely reflects the fidelity of intensity reconstruction, e.g. blurring effects. These left and right plot areas were calculated as the areas of corresponding polygons (Fig. 1) and used as the proposition of Hosaka plots scalar equivalent in our initial study [8].

[Figure 1]

2.1.2. Picture Quality Scale

PQS was designed for the evaluation of coded achromatic still images on the basis of two factors of global random disturbances and three factors of structured and localised errors. But firstly the image signal is transformed into one which is proportional to the visual perception of luminance (most medical images are greyscale) using Weber-Fechner's Law and the contrast sensitivity. Secondly, spatial frequency weighting to the errors is applied. Next, objective quality factors, which quantify the majority of image degradation, are computed. Finally, the space of distortion factors is reduced by principal component analysis, and a linear combination of the reduced space components approximates the results of subjective tests (mean opinion score). PQS was used for coder comparisons and efficiency evaluation [10][11]. We tested the degree of correlation between PQS factors and diagnostic quality pattern (DQP). Further, the combination of five factors optimised by linear regression to increase this correlation was tested in our experiments.

2.2. Diagnostic quality pattern

Subjective rating tests based on subjective observer perception of compressed image features are used for verification and optimisation of numerical measures. We proposed diagnostic features evaluation, rating the perceptibility of reconstructed elements of local image structures which influence the ability of lesion detection and differentiation. The procedure of evaluation concerns the observing of lesion symptoms, all initially pointed out abnormalities and important structures, and rating of selected image local features which are diagnostically important on a multilevel scale. Perception of those elements is conclusive in the final decision of radiologists relating to lesion detection and classification.

A diagnostic quality pattern was estimated for mammograms as an example of the modality of medical imaging which is particularly difficult to diagnose. Many studies have shown that agreement among radiologists interpreting a test set of mammograms is relatively low [12]. Four elements were considered in designed mammogram rating procedure: contrast, sharpness, lesion shape and outline of chosen subtle structures including direct symptoms tumours (diversified morphology), spiculated lesions, circumscribed masses and microcalcifications. Two experts from two medical centres selected test mammograms containing reliable representatives of these symptoms. Trained observers rated these structures perceptibility on a scale of 1 (weak, indistinct, scarcely perceptible, distorted) to 3 (distinct, clearly perceptible, regular, beyond a doubt). A sum of four scores was a general score of image quality. The *DQP* value for each encoded image is average of the scores given by all observers to the image.

2.3. Vector measure of medical image quality

Diagnostic quality measure of lossy compressed medical images was defined. This hybrid vector measure could be used for estimating an acceptable compression ratio (high correlation with subjective scores where the diagnostic unacceptability of feature quality may be the description of certain scale number), comparing images compressed by different coders (scalar equivalent), and analysing the character of image disturbances (graphs). Moreover, it may be applied in a construction process for optimisation of medical image coders (e.g. quantization procedure according to rate-distortion definition). The vector measure contains six selected factors, which are divided into three groups: point accuracy errors, local structured errors and random errors. The definitions of factors belonging to each of these groups are presented in the following subsections.

2.3.1. Measures of point accuracy errors

Point accuracy errors are measured by the twin factors of global and local error characteristics. They are as follows:

- V_1 (average pixel error)

According to the equation (1):

$$V_1 = AD = \frac{1}{MN}\sum_{x,y}\left|f(x,y) - \hat{f}(x,y)\right| \tag{8}$$

This factor characterizes mean point error and hence reconstruction accuracy. Because it is the mean difference between the values of the original image $f(x,y)$ and the reconstructed image $\hat{f}(x,y)$, V_1 as an integral-manner measure does not capture individual picks of image quality but shows general level of pixel reconstruction accuracy.

- V_2 (maximum pixel error)

According to the equation (3):

$$V_2 = 10 \cdot MD = 10 \cdot \max_{x,y}\{|\, f(x,y) - \hat{f}(x,y)\,|\} \tag{9}$$

The maximum difference of corresponding pixel values is an important factor for preserving small, diagnostically important structures which cannot be changed in an archiving process. This differential-manner measure is a good supplement of V_1 and both are constructed in the original and reconstructed image data domain.

2.3.2. Measures of local structured errors

Two factors of local structured errors were taken from PQS because of their high correlation with DQP. To provide a more uniform perceptual scale, the images are transformed using $g(x,y) = k \cdot f(x,y)^{1/2.2}$, which closely approximates Weber-Fechner's Law for contrast sensitivity. The frequency weighted error $e_w(x,y)$ is just contrast adjusted error $e_g(x,y) = g(x,y) - \hat{g}(x,y)$ filtered with $S_a(u,v) = s(\omega)O(\omega,\theta)$, where $s(\omega) = 1.5e^{-\sigma^2\omega^2/2} - e^{-2\sigma^2\omega^2}$, $\sigma = 2$, $\omega = \frac{2\pi f}{60}$, $f = \sqrt{u^2 + v^2}$ and $O(\omega,\theta) = \frac{1 + e^{\beta(\omega - \omega_0)}\cos^4 2\theta}{1 + e^{\beta(\omega - \omega_0)}}$, $\theta = \tan^{-1}(u/v)$, $\beta = 8$, $f_0 = 11.13$ cycle/degree.

Frequency weighted errors are used in definitions of V_3 and V_4 factors in the following way:

- V_3 (correlated errors in 5×5 window)

$$V_3 = \frac{1}{MN}\sum_{x,y} v_3(x,y), \tag{10}$$

where

$$v_3(x,y) = \sum_{(k,l)\in W} |r(x,y,k,l)|^{0.25} \tag{11}$$

and

$$r(x,y,k,l) = \frac{1}{y-1}\left[\sum e_w(i,j)e_w(i+k,j+l) - \frac{1}{y}\sum e_w(i,j)\sum e_w(i+k,j+l)\right] \tag{12}$$

This factor characterizes local spatial correlation and is defined as the summation over the entire image of local error correlation. The sums are computed over the set of pixels where (i,j) and $(i+k,j+l)$ both lie in the 5×5 window centred at (x,y) and W is the set of lags to include in the computation.

- V_4 (preserving high contrast edges)

$$V_4 = \frac{1}{N_K}\sum_{x,y} v_4(x,y), \tag{13}$$

where

$$v_4(x,y) = I_M(x,y)\,|e_w(x,y)|\,(S_h(x,y) + S_v(x,y)), \tag{14}$$

horizontal masking factor: $S_h(x,y) = e^{\{-0.04 A_h(x,y)\}}$, $A_h(x,y) = \dfrac{|f(x,y-1) - f(x,y+1)|}{2}$, and vertical masking factor $S_v(x,y)$ defined similarly. $I_M(x,y)$ is an indicator function which selects pixels close to high intensity transitions. N_K is the number of pixels whose 3×3 Kirsch edge response is greater or equal to threshold value $K = 400$.

Factor V_4 deals with psycho-physical effects which affect the perception of errors in the vicinity of high contrast transitions. It is visual masking which refers to the reduced visibility of disturbances in activity areas.

2.3.3. Measures of random errors

The measures of random errors are constructed in the following way:

- V_5 (integral square with frequency weighting defined by CCIR)

$$V_5 = 1000 \cdot \frac{\sum_{x,y} v_5(x,y)}{\sum_{x,y} f^2(x,y)} \tag{15}$$

where $v_5(x,y) = [e_f(x,y) * w_{TV}(x,y)]^2$ and $e_f(x,y) = f(x,y) - \hat{f}(x,y)$. This factor is defined similarly to normalised mean square error with frequency weighting defined by CCIR 567-1, where $W_{TV}(f) = \frac{1}{1+(f/f_c)^2}$, $f = \sqrt{u^2+v^2}$, $f_c = 5.56$ cycles/degree. Factor V_5 is also defined in PQS. Together with the next factor, they characterize the energy of the difference between original and reconstructed images. Random disturbances introduced by coder or random errors of the original image reduced by coder can be well described by these factors.

- V_6 (integral square normalised by pixel values)

According to the equation (7):

$$V_6 = 10 \cdot \chi^2 = \frac{10}{MN} \sum_{x,y} \frac{\left[f(x,y) - \hat{f}(x,y)\right]^2}{f(x,y)}. \tag{16}$$

This metric without frequency weighting gives additional information about random errors.

2.3.4. Forms of vector quality measure

The hybrid vector measure (HVM) of compressed images is defined as the scalar equivalent of diagnostic quality and a graphical form which is useful for more insensitive distortion analysis and detailed efficiency comparison of compression methods.

- Scalar equivalent

The scalar equivalent, i.e. a linear combination of selected factors with the coefficients calculated by linear regression to increase the degree of correlation with DQP, was optimised during the initial stage. Medical test images of chosen modality were compressed, evaluated in a sense of computable scalar factors and rated in

subjective tests. The equivalent values are intended to approximate DQP. Hence, we verified the following formula of *HVM* definition:

$$HVM = \sum_{i=1}^{6} \alpha_i V_i \,, \tag{17}$$

where α_i are fitted to *DQP* by linear regression.

- Graphs

Subsets of error factors as the separated fields are included in a graphical form of HVM. It characterizes the different distortions of categories mentioned above. It is simply three rectangles growing down because of a negative meaning of the distortions defined by three sets of factors. HVM plots are presented in Fig. 2.

[Figure 2]

The initial stage is complex and time consuming because of DQP estimation and the necessity of fitting *HVM* to the pattern for the chosen class of medical images. But next, a computable stage of quality evaluation of a single reconstructed image is not time consuming because it is a fully numerical procedure according to equations (8-17). A fast, objective estimation of diagnostic image quality could be useful in practice.

3. Experiments and results

A pilot study was arranged to present the usefulness and potential reliability of HVM. Initially, a set of over 200 digitised exams (film scanned to 12 bpp) was overviewed, interpreted, compressed and analyzed by judges: 2 experts-radiologists and 1 expert in medical image processing during the process of test design. Next, 11 mammograms used for HVM evaluation were selected according to the rules stated in p.2.2. All of theses 11 test images contain one or more lesions, subtle abnormalities and unclear tissue structures which could be interpreted as lesions in worse conditions of observation (e.g. caused by lossy compression). Those appearances were pointed out for each observer to evaluate their diagnostic meaning. Images were compressed by JPEG2000 coder [14] to the following bit rates: 1 bpp, 0.6 bpp, 0.1 bpp and 0.04 bpp. Those bit rates were chosen by judges as 'steps of reconstruction quality notified changes'. Moreover, 2 mammograms were removed from the test set prepared for DQP estimation because of interpretation ambiguity. Subjective scores of originals were so different

that they cannot be used for reliable DQP estimation and compression impact evaluation. Originals and 4 compressed image versions were rated according to the following rules: original and its reconstructions were displayed simultaneously, evaluation was comparative: any feature comparison, ordering, classification was acceptable, viewing conditions (i.e. zooming, viewing distance, brightness, contrast in the range limited by the judges) were fitted according to demands of radiologists, time was unlimited. The mammograms were rated by 7 observers (experts in radiology) from 3 different Warsaw radiology centres.

Furthermore, 6 images chosen from 11 test mammograms were used in the tests of compression efficiency comparison because of their susceptibility to the wavelet compression method used in the experiment (once again selection was made by judges). It means that these images were additionally compressed to the mentioned bit rate values by second wavelet coder MBWT [15]. The numerical notes and subjective rates constituting DQP given for compressed images were compared – see the plots in Fig. 3 and the results given in Tab. 1. All collected results were used to calculate the degree of correlation between numerical measures and *DQP*. Adequate correlation coefficients are given in Tab. 2.

[Figure 3]

[Table 1]

An unequivocal statement as to which of the used wavelet coders is more efficient is impossible. Differences in scores are often less distinguishable than an error of a subjective estimation method. Any regular tendency is difficult to notice – see the results given in Tab. 1 and plots in Fig. 3. However, MBWT reconstruction was rated better than JPEG2000 reconstruction in a range of lower bit rates (0.04 – 0.6 bpp). For bit rate 1 bpp JPEG2000 reconstruction achieved higher scores in more cases. According to numerical quality measures (local and global, vector and scalar) MBWT is slightly more effective than JPEG2000. This tendency was confirmed even by HVM plots (Fig. 2b).

[Table 2]

Popular image quality measures: *MSE* and *PSNR*, often applied in compression applications, reflect the perceptual quality of diagnostic symptoms rather unsatisfactorily (the correlation coefficient is about 0.6). A similar correlation degree was achieved for *AD* (Eq. 1) and *IF* (Eq. 4). Moreover, poor performance was noticed for *CQ* (Eq. 2). The significance of local measures was signalled by a high value of the correlation coefficient for *MD* (Eq. 3). Among global measures the performance of Chi-Square is the best. PQS was optimised in the experiments by fitting weights of a linear operator to increase the degree of correlation between PQS and DQP. Furthermore, *AD-MD*-Chi-Square and other scalar measure combinations were tested. But the highest value of the correlation coefficient was noted for *HVM*. It was over 0.9.

The correlation coefficient between *HVM* and *DQP* values is high enough to state that HVM could be useful for quality evaluation of compressed mammograms. Considering similar results of the initial tests with the images of other modalities (the correlation coefficient between *HVM* and *ad hoc* approximated diagnostic quality pattern of compressed MR images was over 0.98 [8]) one can state that HVM could be useful for assessment of lossy compression effects in management systems of medical image data sets. Because agreement between radiologists interpreting the same set of mammograms is known to be low, the achieved consensus of suitable test images selection, abnormalities interpretations and convergence of radiologists comments noted during the subjective tests seems to be valuable as well as presented results of subjective ratings referred to wavelet encoded mammograms.

4. Conclusions

The purpose of this research was to construct a computable measure of image quality correlated to diagnostic accuracy of compressed images. The reported experiments are only an example of applying this vector measure for mammography applications. Optimisation of presented HVM was based on subjective rating of 'diagnostic local image features and lesion symptoms' perception. Certainly, more training images of concrete modality, observers (radiologists) and arranged tests should be provided to establish more reliable diagnostic patterns intended for testing medical image compression tools. Nevertheless, a growing complexity of the initial stage of HVM design could make this idea impractical. A diagnostically related quality estimate seems to be sufficient in order to test presented conception of hybrid quality measure.

It was clearly shown that the degree of correlation between several objective quality measures and subjectively established DQP is various. Thus, the construction of a numerical objective distortion measure, as a good approximation of diagnostic accuracy by nature psychophysical is very difficult. Nevertheless, design of such measure is very important for archiving and transmission of medical images in databases and complex hospital information systems. HVM provides extended information about error characteristics and is able to predict subjective image quality reflecting diagnostic accuracy. Presented vector measure approximates medical image quality better than PQS and other tested measures. The HVM plots contain more important information about errors and disturbances in compressed medical images than Hosaka plots. At the same time, HVM is not complex, and is useful for compression optimisation in medical applications.

The proposed hybrid measure seems to be a more reliable diagnostic accuracy approximation than any other known numerical measure. It could be potentially accepted by radiologists and applied in practice. The HVM is hybrid in the sense of both: psychovisual quality rating and perception of diagnostically important information. Additionally, it is vector (plot) and scalar (equivalent), subjective and objective. More tests for different medical image modalities and coders with more reliable diagnostic pattern estimation should be arranged to confirm the usefulness of HVM.

Both the convergence of image rates and radiologists comments and remarks noted during subjective tests suggest that lossy wavelet compression (in an acceptable bit rates range: 1 bpp, 0.6 bpp and even 0.1 bpp for several cases) does not reduce the diagnostic quality of original images. The comparison between original and reconstructed images did not demonstrate diagnostically important differences in radiologists' conforming opinions.

References

[1] J.A. Swets, ROC Analysis applied to the evaluation of medical imaging techniques, Investigave Radiology 14 (1979) 109-121.

[2] P.C. Cosman, R.M. Gray, R.A. Olshen, Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy, Proc. IEEE 82 (6) (1994).

[3] B. Erickson, Irreversible compression of medical images, J. Digital Imaging 15 (1) (2002) 5-14.

[4] K. Hosaka, A new picture quality evaluation method, Proc. International Picture Coding Symposium, Tokyo, Japan (1986).

[5] A.M. Eskicioglu, P.S. Fisher, S.C. Siyuan, Image quality measures and their performance, Proc. of the 1994 Space and Earth Science Data Compression Workshop, NASA Conference Publication 3255 (1994) 55-67.
[6] M. Miyahara, K. Kotani, V.R. Algazi, Objective picture quality scale (PQS) for image coding, IEEE Trans. Comm. 46 (9) (1998) 1215-1226.
[7] A.J. Ahumada, Computational Image-Quality Metrics: A Review, SID 93 Digest (1993) 305-307.
[8] A. Przelaskowski, Hybrid vector measures of compressed medical images, SPIE Intern. Symposium on Medical Imaging: Image Performance and Perception, San Diego, USA, http://www.ire.pw.edu.pl/~arturp/Publikacje/ap_HVM.pdf.
[9] P.M. Farrelle, Recursive block coding for image data compression, Springer-Verlag New York (1990).
[10] J. Lu, V.R. Algazi, and R.R. Estes Jr., A comparitive study of wavelet image coders, Optical Engineering 35 (9) (1996).
[11] V.R. Algazi, R.R. Estes Jr., Comparative performance of wavelet and JPEG coders at high quality in very high resolution and quality imaging, Proc. SPIE 3025 (1997) 71-82.
[12] J.G. Elmore, D.L. Miglioretti, L.M. Reisch, M.B. Barton, W. Kreuter, C.L. Christiansen, S.W. Fletcher, Screening Mammograms by Community Radiologists: Variability in False-Positive Rates, J Natl Cancer Inst 94 (2002) 1373-1380.
[13] ISO/IEC 15444-1: JPEG2000 image coding system (2000).
[14] A. Przelaskowski, Details preserved wavelet-based compression with adaptive context-based quantisation, Fundamenta Informaticae 34 (4) (1998) 369-388.

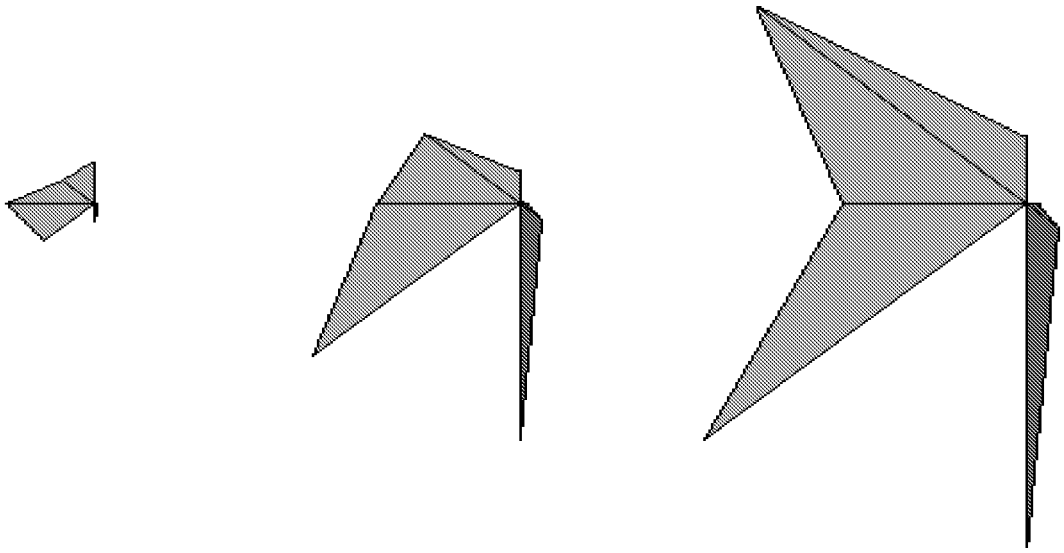

Fig. 1. An example of weighted Hosaka plots. Values of an error vector were weighted according to the number of pixels belonging to each class of blocks. Mammogram was compressed by wavelet coder to 0.6 bits per pixel (bpp), 0.1 bpp and 0.04 bpp. Corresponding Hosaka plots are presented from left to right, respectively.

Tab. 2. Correlation coefficients between *DQP* and selected numerical computable measures. *PQS*(*i*) means *i*-th factor of PQS. The measures proportional to the reconstruction fidelity (i.e. *PSNR*, *IF*, *PQS*) were correlated to *DQP*, and measures inversely proportional to the reconstruction fidelity (i.e. *PQS*(*i*), Chi-Square, *MD*, *MSE*, *AD*, *CQ*, *HVM*) were correlated to reversed pattern (12-*DQP*).

Measures	Correlation with *DQP*	Measures	Correlation with *DQP*
PQS(*1*)	0.7815	*MD*	0.8543
PQS(*2*)	0.6115	*MSE*	0.6162
PQS(*3*)	0.8112	*AD*	0.5903
PQS(*4*)	0.8060	*CQ*	0.1644
PQS(*5*)	0.6374	*IF*	0.6079
PQS	0.7537	*PQS* (optimised)	0.8459
Chi-Square	0.7266	*AD*+*MD*+Chi-Square	0.8625
PSNR	0.5825	***HVM***	**0.9028**

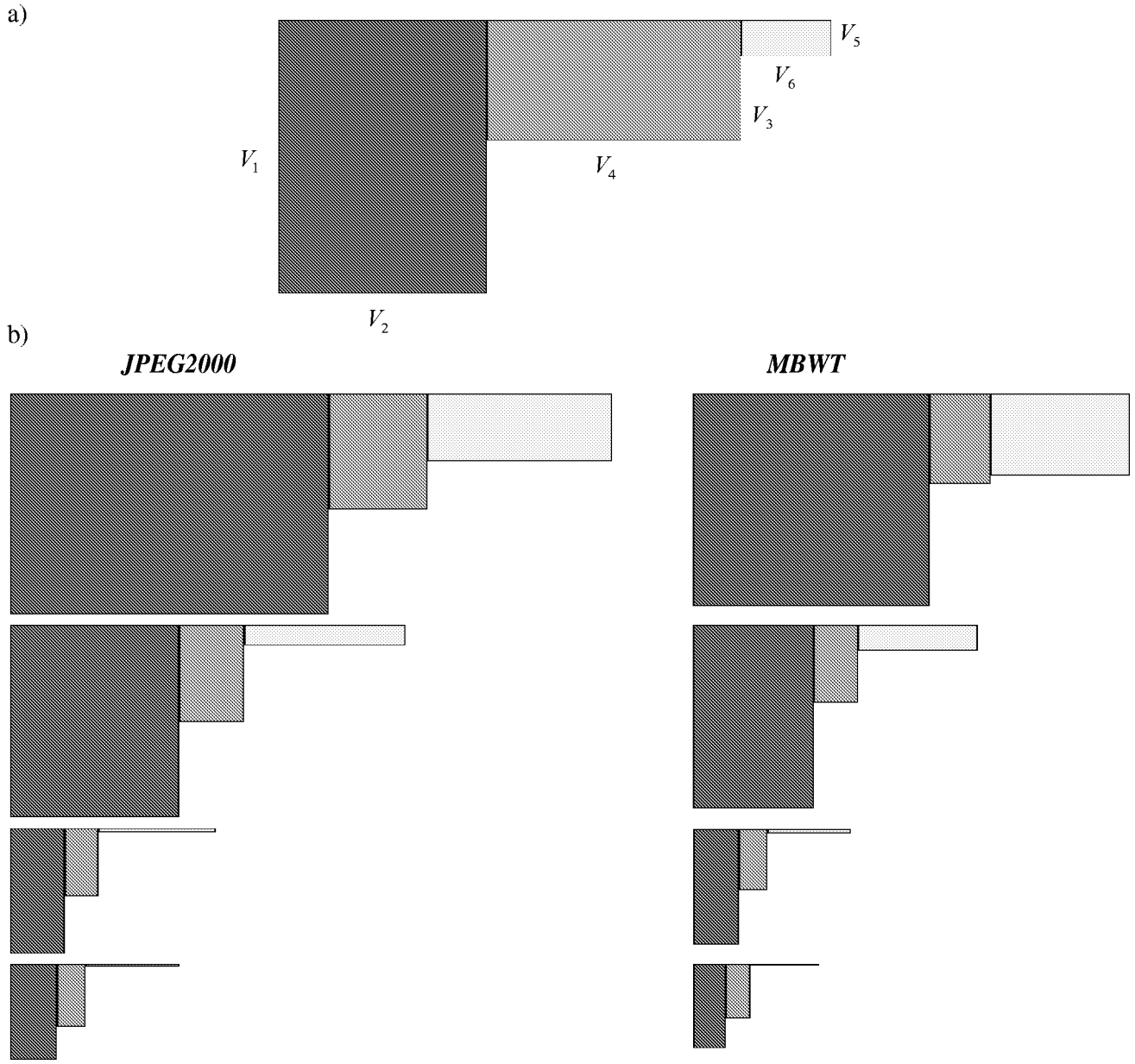

Fig. 2. A graphical form of HVM; a) six factors are included into three groups: a red one informs about point errors, a green field represents structured errors and a yellow rectangle is a sign of random errors; b) a comparison of HVM plots for mammogram compressed by JPEG2000 (left) and MBWT (right); the plots were drawn for four bit rates: 0.04 bpp, 0.1 bpp, 0.6 bpp and 1 bpp (top to bottom, respectively).

Tab. 1. Image quality evaluation: comparison of different objective measures with subjective pattern. Acronyms: A,B,C, ... are successive test images, j – means JPEG2000 reconstructed image, m - means MBWT reconstructed image, 10, 6, 1, 04 - reconstruction bit rates: 1 bpp, 0.6 bpp, 0.1 bpp, 0.04 bpp, respectively. An arrow down means better quality (accuracy) for lower value of the measure, the meaning of an arrow up is opposite. *PQS*(*i*) means *i*-th factor of PQS.

No.	Image	↓ *MSE*	↓ *MD*	↑ *PSNR*	↓ *AD*	↑ *IF*	↓ *CQ*	↓ χ^2	↓ *PQS(1)* [$*10^{-4}$]	↓ *PQS(2)* [$*10^{-3}$]	↓ *PQS(3)*	↓ *PQS(4)*	↓ *PQS(5)*	↑ *PQS*	↓ *HVM*	↑ *DQP*
1.	Aj10	1626.6	361	52.18	31.14	1	4869.4	9.12	.03	.027	.8471	8.804	3.164	4.216	2.5000	**8,57**
2.	Am10	1279.5	254	53.22	27.37	1	4868.3	6.67	.01	.020	.3985	7.660	2.721	4.423	2.5041	**10,14**
3.	Aj6	3086.1	423	49.39	40.97	1	4868.2	11.3	.03	.041	.9279	9.571	3.848	4.024	2.1674	**8,29**
4.	Am6	2725.5	359	49.93	37.58	1	4867.6	8.02	.03	.028	.5269	8.562	3.202	4.240	2.6332	**8,57**
5.	Aj1	8704.9	1306	44.89	62.76	.999	4865.1	15.6	.15	.141	1.783	13.598	7.161	3.053	4.0647	**9,14**
6.	Am1	8499.5	934	44.99	59.78	.999	4864.3	11.6	.18	.083	1.355	10.864	4.936	3.710	3.4565	**9,29**
7.	Aj04	11811.5	2451	43.57	71.90	.999	4863.0	17.9	.47	.332	3.723	16.284	10.999	2.200	6.5220	**4,43**
8.	Am04	11906.0	1821	43.53	69.33	.999	4861.9	13.4	.57	.186	2.838	12.707	6.842	3.224	5.4876	**7,14**
9.	Bj10	3217.7	464	49.21	43.47	1	4900.6	5.98	.02	.009	.2082	6.297	1.546	4.759	1.9627	**11,29**
10.	Bm10	2896.7	343	49.67	40.92	1	4900.0	4.40	.01	.007	.1425	5.765	1.386	4.850	2.0440	**9,43**
11.	Bj6	5989.4	556	46.51	57.15	1	4899.7	7.22	.02	.010	.2209	6.488	1.618	4.724	1.5998	**10,43**
12.	Bm6	5353.0	522	47.00	53.84	1	4899.0	5.51	.02	.010	.1982	6.470	1.636	4.724	2.1641	**8,86**
13.	Bj1	13117.7	1568	43.11	81.25	.999	4897.1	9.98	.08	.030	.4296	9.087	2.790	4.225	3.1672	**7,43**
14.	Bm1	12886.7	994	43.19	79.50	.999	4896.5	8.31	.12	.031	.5406	8.725	2.737	4.281	2.9949	**8,71**
15.	Bj04	16475.4	2960	42.12	90.44	.999	4895.9	11.4	.26	.079	1.261	11.657	4.591	3.651	5.9249	**5,57**
16.	Bm04	16478.9	2922	42.12	89.19	.999	4895.4	9.81	.33	.071	1.177	10.563	3.940	3.883	5.9038	**4,43**
17.	Cj10	10270.7	784	44.17	75.37	1	6781.1	6.69	.02	.008	.3199	7.245	1.498	4.642	2.2887	**10,86**
18.	Cm10	8544.2	551	44.97	68.45	1	6780.6	5.19	.02	.009	.3053	7.131	1.528	4.652	2.5561	**10,00**
19.	Cj6	19975.2	1292	41.28	101.6	.999	6779.2	9.02	.05	.020	.5200	9.165	2.342	4.276	2.9731	**9,57**
20.	Cm6	18626.0	885	41.59	98.20	.999	6778.3	7.35	.05	.016	.5071	8.351	1.988	4.431	2.6439	**10,71**
21.	Cj1	47877.4	2725	37.49	151.6	.998	6772.2	13.8	.19	.051	1.452	11.857	3.673	3.750	4.2010	**6,57**
22.	Cm1	48071.5	2017	37.47	151.3	.998	6771.2	12.7	.30	.057	1.570	11.618	3.468	3.811	3.9119	**8,57**
23.	Cj04	60206.4	5172	36.49	169.5	.998	6768.9	16.7	.60	.116	3.483	14.782	5.266	3.170	8.1224	**4,71**
24.	Cm04	61776.0	3166	36.38	170.9	.998	6768.0	15.8	.79	.125	3.486	14.238	4.960	3.283	6.3556	**5,43**
25.	Dj10	9660.9	688	44.44	73.77	1	8252.3	6.42	.01	.006	.5083	6.918	1.381	4.702	2.0753	**9,71**
26.	Dm10	8615.4	559	44.94	69.31	1	8251.6	5.24	.02	.008	.4543	7.179	1.507	4.650	2.5692	**9,43**
27.	Dj6	17302.3	1320	41.91	95.91	1	8250.7	8.43	.03	.017	.8839	8.991	2.267	4.313	3.1221	**8,43**
28.	Dm6	16120.0	890	42.21	92.67	1	8250.2	6.82	.03	.012	.6416	8.110	1.774	4.493	2.7102	**10,00**
29.	Dj1	41611.0	2695	38.10	144.0	.999	8245.9	12.5	.13	.046	1.978	12.005	3.712	3.731	4.6654	**7,14**
30.	Dm1	40706.5	2169	38.19	141.7	.999	8245.7	10.8	.17	.030	1.387	10.979	2.611	4.009	4.1934	**6,14**
31.	Dj04	52533.3	5523	37.08	160.7	.999	8243.5	14.8	.38	.080	3.430	14.401	4.808	3.281	8.6018	**4,00**
32.	Dm04	54437.6	3437	36.93	162.4	.999	8242.7	13.3	.50	.059	2.585	13.395	3.553	3.575	6.5526	**4,57**
33.	Eaj10	2267.3	434	50.73	35.80	1	5475.6	7.95	.03	.022	.5623	8.200	2.625	4.366	2.5506	**9,86**
34.	Eaj6	4048.7	475	48.22	46.05	1	5474.0	10.5	.03	.037	.6453	9.037	3.379	4.154	2.1121	**9,29**
35.	Eaj1	10949.1	1293	43.89	69.36	.999	5469.8	15.1	.14	.108	1.479	12.841	5.826	3.330	3.6743	**9,43**
36.	Eaj04	14342.4	2774	42.72	78.14	.998	5467.1	17.4	.43	.279	3.403	15.643	9.294	2.512	6.5510	**5,71**
37.	Fj10	2475.7	633	50.35	37.72	1	6993.1	6.94	.02	.016	.4973	8.156	2.317	4.413	3.0406	**8,43**
38.	Fj6	4309.9	515	47.94	48.08	1	6992.2	8.61	.02	.020	.5727	8.713	2.638	4.296	2.5771	**8,86**
39.	Fj1	11410.2	1413	43.72	72.10	.999	6989.4	11.5	.08	.059	1.166	12.218	4.529	3.584	4.5202	**8,00**
40.	Fj04	14996.9	3026	42.53	81.26	.999	6987.9	13.5	.26	.153	2.582	14.835	7.268	2.883	7.3083	**4,71**
41.	Gj10	5382.7	563	46.98	55.13	1	5317.9	6.02	.02	.007	.2086	6.302	1.219	4.803	1.9322	**10,14**
42.	Gj6	9043.1	602	44.73	70.57	.999	5316.7	7.40	.03	.008	.2677	6.692	1.299	4.741	1.5920	**10,71**
43.	Gj1	22114.5	2500	40.84	104.9	.999	5312.1	11.7	.21	.057	1.021	11.119	3.454	3.873	4.8068	**9,43**
44.	Gj04	29262.2	3184	39.63	119.0	.998	5309.8	13.3	.48	.080	1.634	13.070	4.066	3.539	6.5323	**5,71**

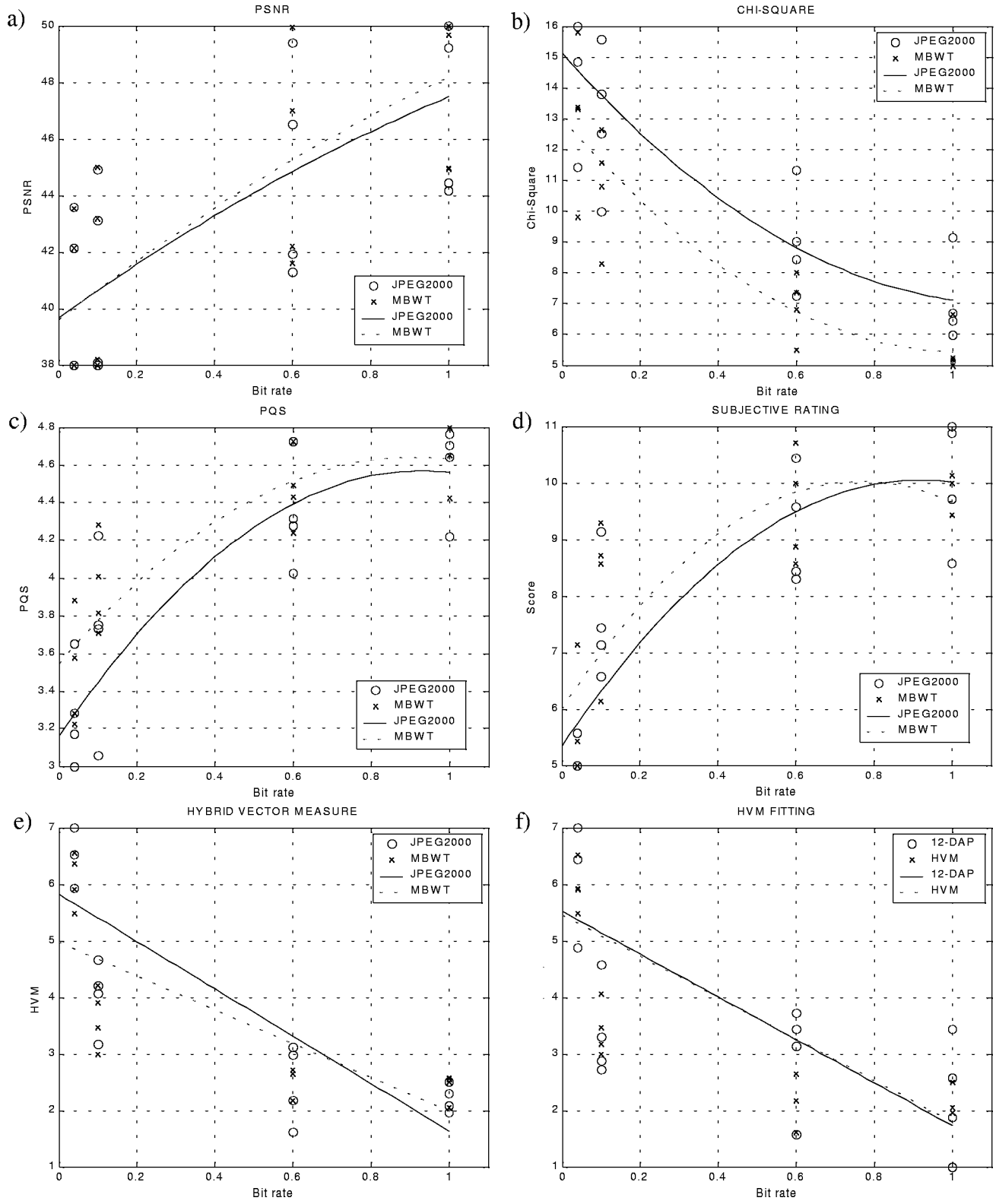

Fig. 3. Results of experiments. Compression efficiency comparison of two wavelet coders: JPEG2000 and MBWT. The results of using scalar quality measures: PSNR(a), Chi-Square (b), PQS (c), and the scores of subjective rating (d), and *HVM* – based evaluation of diagnostic accuracy (e) are presented. Moreover, fitting of HVM to DQP for a set of used test images is shown in f).

SUMMARY

Irreversible compression techniques are required for effective archiving and transmission of large image data sets in hospital information systems (HIS), picture archiving and communications systems (PACS) and teleradiology. But the principal issue is how to characterize and measure different distortions of imperfectly reproduced medical images. Reliable evaluation of compression performance is necessary for efficient coder selection and most of all for acceptable compression ratio calculation.

The primary objective of this study is a proposal for a new vector measure of image quality which could be implemented for lossy compressed medical images. Construction of such a measure includes the formation of a diagnostic quality pattern based on the subjective ratings of mammogram local features playing an essential role in the detection and classification of any lesion. Reliable estimation of this pattern makes vector measure appropriate for assessment of medical image quality. Thus, images with any abnormalities and structures difficult to diagnose could be evaluated in diagnostic accuracy terms by rating quality (state of visibility, perception) of these 'diagnostic features'.

Design of hybrid vector measure and estimation of subjective diagnostic quality pattern is the main contribution of this paper. The vector measure contains of six selected factors, which are divided into three groups: point accuracy errors, local structured errors and random errors. A linear combination of these factors, where coefficients are calculated by linear regression to increase the degree of correlation with diagnostic pattern, forms a scalar equivalent of image quality. The estimation of coefficients made during the initial stage is complex and time consuming because of diagnostic pattern assessment and the necessity of fitting an equivalent to the pattern for the specified class of medical images. But next, a computable stage of vector measure application (quality evaluation of a single reconstructed image) is not time consuming because it is fully numerical procedure.

The demonstrated results of the experiments designed by 3 judges, experts in radiology and image processing, are only an example of practical optimisation of this vector measure. Diagnostic quality pattern was formatted basing on subjective ratings of mammograms done by 7 viewers (radiologists). Next, hybrid measure was optimised according to this pattern. The correlation coefficient between scalar equivalent and subjective

pattern was over 0.9. Graphical presentation of the multidimensional vector measure of image quality can make error analysis more penetrating and precise, e.g. in analysis of wavelet compression effects. Concluding, the proposed vector measure is suggested as a more reliable medical image quality measure than any other numerical measure used in compression efficiency tests. It could be useful in design of modern information systems managing medical image data sets.