XVII IMEKO World Congress
Metrology in the 3rd Millennium
June 22–27, 2003, Dubrovnik, Croatia

NUMERICAL EQUIVALENT OF ACCURACY AS A MEASURE OF MEDICAL IMAGE QUALITY

Artur Przelaskowski

Institute of Radioelectronics, Warsaw University of Technology, Warszawa, Poland

Abstract – The objective of this study is a proposition for a new vector measure of image quality reflecting diagnostic accuracy. The formation of a diagnostic quality pattern (DQP) was based on the subjective ratings of image local features, playing an essential role in the detection and classification of any lesion. The pattern was used for more reliable calculation of a single number - the equivalent of diagnostically related image quality. The equivalent was constructed with a criterion of the highest correlation with approximated diagnostic accuracy of compressed images. 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: Diagnostic image quality, lossy compression

1. INTRODUCTION

Efficient tools of image quality measurement are crucial in many applications of image processing procedures, i.e. quality enhancement, perception quality preserving, irreversible compression and others. Among different conceptions, design of vector measures combining (correlating) extended characteristics of processed image features (by a set of numerical factors) with subjective pattern of image quality seems to be promising approach. Generally, vector measures of image quality have an accepted complexity, a greater degree of correlation with subjective quality evaluation but they are inconvenient for the comparison of coded images because of the difficulty in the interpretation of multidimensional graphical forms. The solution is a scalar equivalent of vector measure as final quality score for comparisons and acceptance fixing. This equivalent should meet the criterion of the highest possible correlation with diagnostic accuracy of analysed images for medical applications. Miyahara [1] presented the determination method of several distortion factors combined by regression into a single number representative of the quality of a given image. This methodology called Picture Quality Scale (PQS) brings together the perceptual properties of human vision and extensive engineering experience as well as the observation of actual image disturbances resulting from image coding. For medical images, an important advantage of a computable objective measure is ability to predict diagnostic accuracy rather than subjective quality.

We selected the most suitable factors and defined a new method that incorporates diagnostic quality estimation into the construction process of a vector distortion measure to make it appropriate for assessment of medical image accuracy.

Mammography, which is difficult to diagnose medical imaging modality, was used as a source of diagnostic image information for the experimental verification of presented conception. Many studies have shown that agreement among radiologists interpreting a test set of mammograms is relatively low [2]. Because of large quantities of data in mammogram databases increased compression efficacy (diagnostically acceptable) is required and could be useful. Wavelet-based image compression algorithms were recognized as a superior method to compress, archive, and electronically disseminate medical imagery [3][4][5].

2. METHODS

Diagnostic quality measure of lossy compressed medical images was defined. The numerical vector consists of 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.

2.1. 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)

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

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)

$$V_2 = 10 \cdot \max_{x,y} \{ | f(x,y) - \hat{f}(x,y) | \} \tag{2}$$

It is an important factor for preserving small, diagnostically important structures which must not be changed in an irreversible compression.

2.2. 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 $S_a(u,\upsilon) = = s(\omega)O(\omega,\theta)$, with $s(\omega) = 1.5e^{-\sigma^2\omega^2/2} - e^{-2\sigma^2\omega^2}$, $\sigma = 2$, $\omega = \frac{2\pi\phi}{60}, \phi = \sqrt{u^2 + \upsilon^2}$ and $O(\omega,\theta) = \frac{1 + e^{\beta(\omega-\omega_0)}\cos^4 2\theta}{1 + e^{\beta(\omega-\omega_0)}}$, $\theta = \tan^{-1}(u/\upsilon)$, $\beta = 8$, $\phi_0 = 11.13$ cycle/degree. Frequency weighted errors are used in the following definitions:

- V_3 (correlated errors in 5×5 window)

$$V_3 = \frac{1}{MN}\sum_{x,y} v_3(x,y), \tag{3}$$

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

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

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)

$$V_4 = \frac{1}{\mathcal{N}_K}\sum_{x,y} v_4(x,y), \tag{4}$$

where $v_4(x,y) = I_{\mathcal{M}}(x,y)\,|e_w(x,y)|\,[S_{ho}(x,y) + S_{ve}(x,y)]$, horizontal masking factor: $S_{ho}(x,y) = e^{\{-0.04A_{ho}(x,y)\}}$, $A_{ho}(x,y) = \frac{|f(x,y-1) - f(x,y+1)|}{2}$, and vertical masking factor $S_{ve}(x,y)$ defined analogously. $I_{\mathcal{M}}(x,y)$ is an indicator function which selects pixels close to high intensity transitions. $\mathcal{N}_K$ 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.

2.3. Measures of random errors

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

- V_5 (integral square with frequency weighting)

$$V_5 = 1000 \cdot \frac{\sum_{x,y} v_5(x,y)}{\sum_{x,y} f^2(x,y)} \tag{5}$$

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 $W_{TV}(\phi) = \frac{1}{1 + (\phi/\phi_c)^2}$, $\phi_c = 5.56$ 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.

- V_6 (integral square normalised by pixel values)

$$V_6 = \frac{10}{MN}\sum_{x,y}\frac{\left\lfloor f(x,y) - \hat{f}(x,y)\right\rfloor^2}{f(x,y)} \tag{6}$$

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

2.4. Diagnostic quality pattern (DQP)

Subjective rating tests based on subjective observer perception of compressed image features are used for verification and optimisation of numerical measures. Diagnostic features evaluation is proposed. Radiologists rate the perceptibility of reconstructed elements of local image structures which influence the ability of lesion detection and differentiation. The procedure of evaluation consists of observing lesion symptoms, all initially pointed out abnormalities and important structures, and rating those selected image local features which are diagnostically important. Perception of those elements is conclusive in the final decision of radiologists relating to lesion detection and classification. Reviewer opinion concerning the ability to detect lesions and classification are notified and a numerical scale with a corresponding diagnostic description for each number is used.

DQP is approximated for each image. Perceptual quality evaluation of diagnostically important image features is incorporated into the optimisation process of the vector measure.

2.5. Forms of vector quality measure

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

DQP is used for more reliable calculation of this single number - the equivalent of *HVM* defined as follows:

$$HVM = \sum_{i=1}^{6}\alpha_i V_i \tag{7}$$

Coefficients α_i are fitted by linear regression (minimizing an error between *HVM* and *DQP*).

HVM is a final numerical factor assigned to the encoded image, useful in compression technique optimization and for an estimation of acceptable-in-diagnosis ratios.

Subsets of error factors as the separated fields are included in a graphical form of HVM which characterizes the different distortions of categories mentioned above. Its shape will assume the simple form of three rectangles growing down because of a negative meaning of the distortions defined by three sets of factors. HVM plots are presented in Fig. 1.

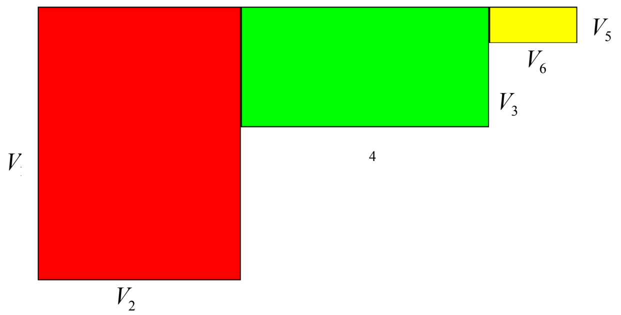


Fig. 1. A graphic form of HVM. HVM consists of six factors $V_1,\ldots,V_6$ defined by (1)-(6) which are included into three groups: a red one (left) informs about point errors (V_1 and V_2), a green field (middle) represents structured errors (V_3 and V_4) and a yellow rectangle (right) is a sign of random errors (V_5 and V_6).

3. EXPERIMENTS

DQP was estimated experimentally for mammograms. Four elements were considered in rating procedure: contrast (related to density), interpretation clarity (visibility, noticeability of lesions, mostly related to detection ability), shape, margin (outline, contour distinction) of chosen subtle structures including direct symptoms tumours (diversified morphology), spiculated lesions, circumscribed masses and microcalcifications. Two experts from 2 medical centres selected test mammograms containing reliable representatives of these symptoms. 7 trained observers rated perceptibility of these structures 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. Estimated *DQP* contains final quality scores for 75 test images (originals and 1 bpp, 0.6 bpp, 0.1 bpp, 0.04 bpp reconstructions of wavelet compression) (more details were given in [6]).

Additionally, quality of images from the same test set was estimated with help of typical scalar measures, PQS and HVM - see the results in Table I, where mean values of numerical rates for four bit rates of wavelet-based mammogram representation were compared to *DQP*. Correlation coefficients between *DQP* and selected numerical computable measures were given in Table II. An extended analysis of the ability of *HVM* to follow *DQP* is presented in Table III. More complex analysis of image quality assessment with use of different measures was done in Fig. 2.

The correlation coefficient between *HVM* and *DQP* (for compressed mammograms) achieved in the experiments equals 0.903. It is high enough to state that HVM could be useful for quality evaluation of compressed mammograms. One can state that HVM could be useful for assessment of lossy compression effects in management systems of medical image data sets.

TABLE I. Image quality evaluation: comparison of different objective measures: $MSE = \frac{1}{MN}\sum_{x,y}[f(x,y)-\hat{f}(x,y)]^2$, $MD = \max_{x,y}\{|f(x,y)-\hat{f}(x,y)|\}$, $PSNR = 10\log_{10}\frac{MN\cdot[\max_{x,y}\{f(x,y)\}]^2}{\sum_{x,y}[f(x,y)-\hat{f}(x,y)]^2}$, *PQS* and *HVM* with subjective pattern *DQP*. An arrow down means better quality (accuracy) for lower value of the measure, the meaning of an arrow up is opposite.

	Bit rate [bpp]	Mean values of quality measures for 120 compressed mammograms					
		↓ *MSE*	↓ *MD*	↑ *PSNR*	↑ *PQS*	↓ *HVM*	↑ *DQP*
Wavelet coder	1	5764	501	47.9	4.61	2.31	9.93
	0.6	11147	781	45.0	4.40	2.51	9.36
	0.1	27685	1802	40.9	3.83	3.79	7.97
	0.04	35704	3432	39.8	3.28	6.68	5.04

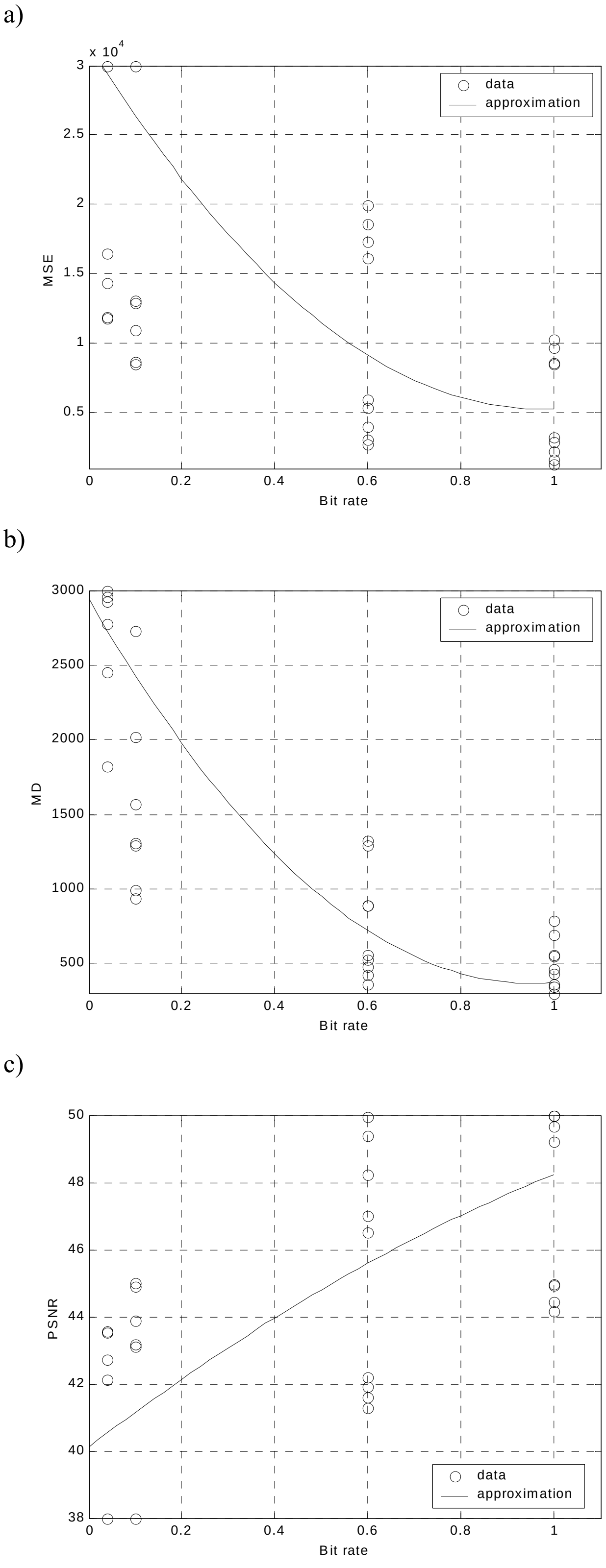
TABLE II. Correlation coefficients between *DQP* and selected numerical computable measures. The measures proportional to the reconstruction fidelity (i.e. *PSNR*, *PQS*) were correlated to *DQP*, and measures inversely proportional to the reconstruction fidelity (i.e. *MD*, *MSE*, *HVM*) were correlated to reversed pattern (12-*DQP*).

Measures	Correlation with *DQP*
MSE	0.616
MD	0.854
PSNR	0.583
PQS	0.754
HVM	**0.903**

TABLE III. Correlation coefficients of successive factors of HVM, three fields of plots and *HVM* with diagnostic local features (four elements of subjective ratings) constituting DQP.

Numerical factors	Correlation with				
	contrast	Interpret. clarity	shape	margin	*DQP*
V_1	0.547	0.590	0.509	0.483	0.590
V_2	0.786	0.795	0.751	0.734	0.854
V_3	0.750	0.786	0.705	0.679	0.806
V_4	0.611	0.645	0.591	0.561	0.637
V_5	0.688	0.780	0.678	0.595	0.782
V_6	0.685	0.728	0.630	0.594	0.727
V_1+V_2	0.713	0.707	0.669	0.651	0.753
V_3+V_4	0.634	0.654	0.618	0.590	0.661
V_5+V_6	0.678	0.755	0.675	0.597	0.759
HVM	0.819	0.840	0.795	0.772	0.903

Popular integral 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). The significance of local measures was signalled by a high value of the correlation coefficient for *MD*.



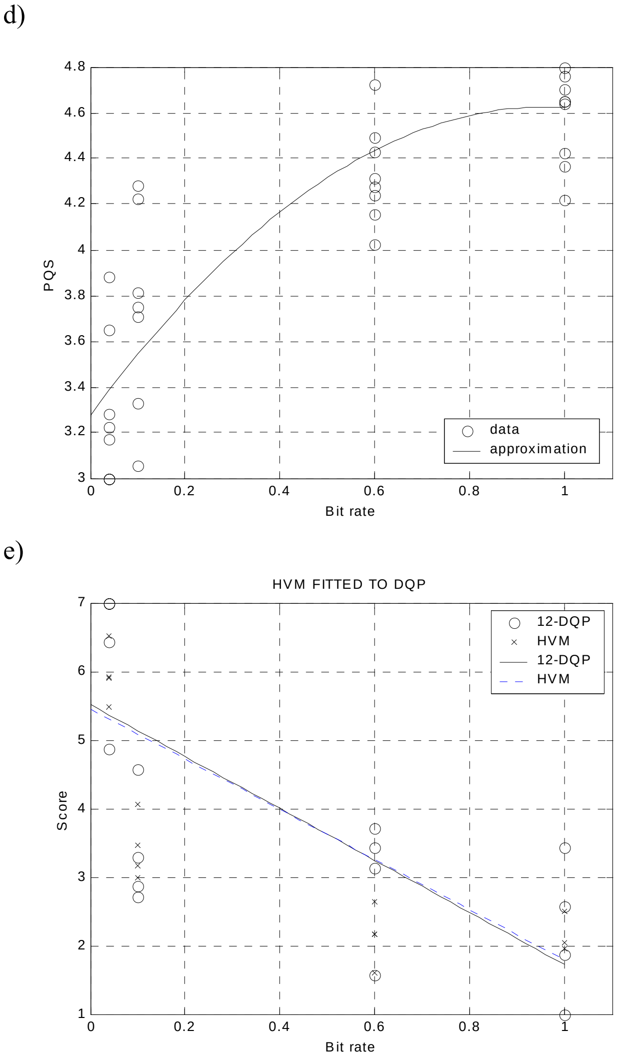


Fig. 2. Experimental results of compression effects evaluation. The results of using scalar quality measures: MSE (a), MD (b), PSNR (c), vector measures PQS (d) and *HVM* fitted to *DQP* (e) for a set of used test images are shown.

The experimental HVM plots for lossy compressed mammograms are presented in Fig. 3 and Fig. 4. Additional information and more penetrating analysis enable reliable characteristics of effects of irreversible compression on images (Fig. 3). Increased point accuracy errors allow predict decreased diagnostic accuracy in experimental assessment and comparison of compression efficiency (Fig. 4).

Generally, the initial stage of HVM design 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 (1)-(6). A fast, objective estimation of diagnostic image quality could be useful in practice.

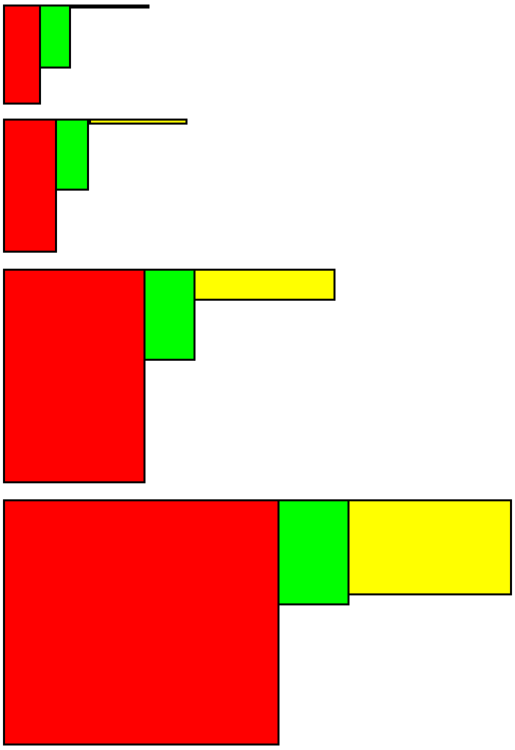


Fig. 3. An example of HVM plots for mammogram compressed by MBWT coder [8]. The plots were drawn for four bit rates: 1 bpp, 0.6 bpp, 0.1 bpp and 0.04 bpp (top to bottom, respectively).

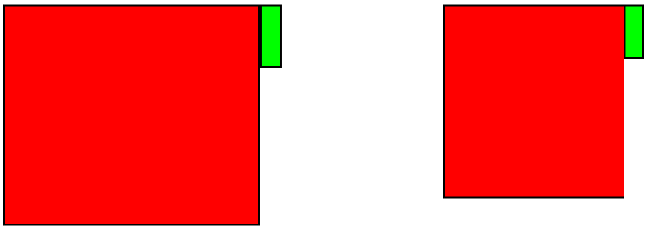


Fig. 4. Regard plots of HVM analysis to *DQP*. The plots were estimated for mammogram compressed by two tested wavelet coders (JPEG2000 [7] left and MBWT right) to 1 bpp. DQP notes for these images were 8.57 and 10.14, respectively what means that significantly better diagnostic quality correlates with clearly decreased distortions (mostly point accuracy errors).

4. CONCLUSION

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 the development of digital archives of medical images, clinical PACS (picture archiving and communication systems), as well as telemedicine networks and RIS (radiology information systems).

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 to 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.

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.

Presented vector measure approximates medical image quality better than PQS and other tested measures. HVM provides extended information about error characteristics of irreversible compressed medical images. At the same time, HVM is not complex, and is useful for compression optimisation in medical applications because it effectively predicts subjective image quality reflecting diagnostic accuracy.

Proposed vector 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.

REFERENCES

[1] M. Miyahara, K. Kotani, V.R. Algazi, "Objective picture quality scale (PQS) for image coding", *IEEE Trans. Comm.*, vol. 46, no. 9, pp. 1215-1226, 1998.

[2] J.G. Elmore, D.L. Miglioretti, L.M. Reisch *et al*., "Screening mammograms by community radiologists: variability in false-positive rates", *J Natl Cancer Inst*, vol. 94, pp. 1373-1380, 2002.

[3] B.J. Erickson, "Irreversible compression of medical images", *J. Digital Imaging*, vol. 15, no. 1, pp.5-14, 2002.

[4] M.-M. Sung, H.-J. Kim *et al*., "Clinical evaluation of compression ratios using JPEG2000 on computed radiography chest images", *J. Digital Imaging*, on line, Sept 2002.

[5] D.H. Foes, E. Muka, R.M. Slone, *et al*., "JPEG 2000 compression of medical imagery", *Proc SPIE*, *PACS Design and Evaluation: Engineering and Clinical Issues,* vol. 3980, pp. 85-96, Feb 2000.

[6] A. Przelaskowski, "Estimation of diagnostic ability by diagnostic features assessment", *presented at* XVII IMEKO World Congress, 2003.

[7] ISO/IEC 15444-1,2: JPEG2000 image coding system (2000) – VM 8.6.

[8] A. Przelaskowski, "Details preserved wavelet-based compression with adaptive context-based quantisation", *Fundamenta Informaticae*, vol. 34, no. 4, pp. 369-388, 1998.

Author: Artur Przelaskowski, PhD, Institute of Radioelectronics, Warsaw University of Technology, ul. Nowowiejska 15/19, 00-665 Warszawa, Poland, phone: +48 22 6607917, fax: +48 22 8251363, e-mail: arturp@ire.pw.edu.pl