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

ESTIMATION OF DIAGNOSTIC ABILITY BY DIAGNOSTIC FEATURES ASSESSMENT TESTS

Artur Przelaskowski[*], *Anna Kukuła*[**]

[*]Institute of Radioelectronics, Warsaw University of Technology, Warszawa, Poland
[**]Radiology Centre, Wolski Hospital, Warszawa, Poland

Abstract – This paper presents a method of simplifying the diagnostic accuracy estimation for compressed mammograms. The proposed method consists of three main stages: a) the selection of abnormal structure features susceptible to wavelet compression method, b) subjective ratings of diagnostically important features and c) estimation of diagnostic pattern for processed mammograms. This pattern reflects diagnostic ability of radiologists using encoded images in clinical practice. Experimental pattern estimation confirmed even higher diagnostic quality of compressed to 1 bpp mammograms than their 12 bpp originals.

Keywords: diagnostic accuracy, subjective rating, irreversible image compression

1. INTRODUCTION

Because the diagnostic accuracy of originals should be preserved in processed images, reliable measures of diagnostic accuracy are required. 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 theory of signal detection. Nevertheless, ROC analysis has no natural extension to the evaluation of measurement accuracy in compressed medical images [1]. Erickson [2] 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 statistical nature referred to many grounded in perception diagnostic decisions, large set of test images and studied processing methods. However, a nature of an individual diagnosis is not statistical and a detection of pathology concerns the concrete case.

The purpose of our research is to design a method of diagnostic accuracy approximation by diagnostic image quality rating based on assessment of local image features important for diagnosis. This approximation is intended to investigate diagnostic accuracy of a single compressed image related to the original. ROC-based analysis was replaced by diagnostic quality pattern (DQP) estimation from subjective ratings of each tested image. The presented procedure of diagnostic accuracy approximation is much less complex and time-consuming than ROC-based analyses, and thus could have larger practical importance.

Mammography is currently the best technique for reliable detection of early, non-palpable, potentially curable breast cancer. The sensitivity of mammography is limited because of poor distinction in radiographic appearance between cancerous tissue and normal breast parenchyma, invisibility of breast lesion in the mammogram received from the concrete imaging system, and observational factors such as fatigue, distraction, and the satisfaction-of-search effect. The efficacy of diagnostic mammography depends on: the optimal radiographic demonstration of the breast, the perception of abnormal features and the correct interpretation.

Generally symptomatology of breast cancer in mammography is based on direct and indirect signs. The direct signs are: mass with low or high density with ill-defined margin, spiculated mass, circumscribed mass. The indirect signs are: microcalcifications, architectural distortion, asymmetric density, focal or diffuse skin thickening, asymmetric ducts, asymmetric veins, nipple retraction. Two experts from different medical centres selected test mammograms containing solid representatives of these symptoms in order to estimate DQP reliably. Moreover, prepared ROIs (Regions of Interests) represent areas that are somewhat important like converging lines and shadows that can be a spiculated mass, densities with irregular margins that can be the start of a subtle cancer, and clusters of bright spots that can be malignant calcium.

2. METHODS

An important approach in radiologists' or CAD-based diagnosis is to extract information (features) from mammograms and use them to make the malignant versus benign assessment of detected lesions. Radiologists are trained to recognize the normal or abnormal appearance of mammograms. Therefore, it is reasonable to detect and classify breast lesions based on their perceptual descriptions completed (correlated) with image features analysis concerning textures, edges, masking noise etc. The components of these descriptions seem to be fundamental in more objective characteristics of diagnostic process.

The following 'medical features' of lesions were considered: the density of the masses, their shape, margin, size, spiculation, calcification and visibility of fine structures, i.e. microcalcifications (appearing as small spots in the picture). Transitions, deformations and local changes of these features caused by compression could be recognized as additional symptoms of pathology or may conceal true lesions. Their perception forms radiologists' decision concerning conditions of lesion detection.

Lossy wavelet compression introduces perceptible changes like removal of "salt-and-pepper" noise and appearing the blurring effects. Moreover, subtle findings (e.g. faint microcalcifications) would be lost but this is not always the case. If they have a significant spatial extent characterised by low frequencies well preserved by wavelet compression subtle findings may be even better visible by higher contrast obtained through removal of high-frequency noise. Therefore, diagnostic accuracy depends on such local image properties as: edge gradient (sharpness) and smoothness, structure blurring, contrast and shape, outline of certain particulars, texture clearness. These 'technical features' were taken into account during characteristics of observed structures because they play an essential role in determining an ability of detection and classification of any abnormalities in processed mammograms.

An approximation of diagnostic accuracy may be performed by rating a quality (state of visibility, perception) of selected 'diagnostic features' that is local image features being a contact between medical properties and technical features. These diagnostic features should be well understandable by radiologists and clear as a subject of subjective rating. Proposed four diagnostic features are as follows: contrast (related to density), interpretation clarity (visibility of lesions, mostly related to detection ability, dependent on the majority of mentioned technical features), shape, margin (outline, contour distinction) of chosen structures including lesions and other abnormalities.

Radiologists observed and analysed mammography exams. The procedure of estimating the diagnostic pattern was based on the following assumptions:

- 'gold standard' is consensus and separate, i.e. determined by the consensus of two judges on the original, reconstructions, other projections and all available exams and results;
- independent observations and decision making by each radiologist in similar, comfortable conditions of clinical practice are provided;
- one test set contains original with lesions (or other pointed out abnormalities) and *N*-1 its reconstructions from different bit rate representation; *N* is limited by presentation conditions and application requirements (e.g. established range of bit rates); several test image sets could be formed (e.g. for different wavelet coders) for one original exam;
- images from one test set are grouped and displayed together; evaluation is comparative, viewing conditions could be fitted according to demands of radiologists; time is unlimited, and evaluation process is divided into one-hour sessions; observers (radiologists) rates perception of diagnostic features in indicated ROIs 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 diagnostic pattern value for each encoded image is average of the scores given by all observers to the image.

3. EXPERIMENTS

Initially, a set of over 200 digitised exams (film scanned to 12 bpp) was overviewed, interpreted, compressed and analyzed in the process of test design prepared by the judges: 2 radiology experts and 1 expert in medical image processing. Following this, 9 mammograms used for diagnostic accuracy approximation were selected (Fig. 1). Test images contained one or more lesions, abnormalities or unclear tissue structures, which could be interpreted as the cases of lesions in worse visual conditions (e.g. caused by lossy compression). Those appearances were pointed out for each observer who was asked to evaluate their diagnostic meaning.

The images were compressed by two wavelet-based coders: JPEG2000[3] and MBWT[4] (with different algorithms of quantization and information ordering) to the following bit rates: 1 bpp, 0.6 bpp, 0.1 bpp and 0.04 bpp. 15 test sets of mammograms were prepared for subjective ratings including original and its 4 reconstructed versions. 7 radiologists from 3 medical centres rated these images in accordance with mentioned rules.

According to the results given in Table 1 (explained more clearly in Table 2) only 6 originals were rated with the highest scores. For other 9 cases reconstructed images (1 bpp and 0.6 bpp) were scored higher than originals. Average score of all original images is 9.83 while it is 9.96 for 1 bpp reconstruction, 9.22 for 0.6 bpp, 7.82 for 0.1 bpp and 4.92 for 0.04 bpp. It may be a proof that in some cases lossy compression can even improve diagnostic quality of mammograms.

Experimental results of compression efficiency evaluation for two wavelet coders were given in Fig. 2. Presented tendency of diagnostic quality degradation for low bit rates of compressed mammograms could be useful for an estimation of diagnostically acceptable compression ratios. MBWT coder seems to be more effective than JPEG2000 coder in a range of lower bit rates (up to 0.6 bpp because of locally fitted adaptive quantization scheme).

Reported experiments were not detection tests but subjective rating of 'diagnostic local image features and lesion symptoms' perception. Estimated pattern may be used for optimisation of numerical measure of diagnostic image quality [5]. Moreover, quantitative assessment of enhancement or degradation of single image quality is possible. Approximated DQP may be useful in design of compression or other image processing procedures. More training images, observers (radiologists) and conducted tests may be assured to establish diagnostic patterns more reliably. Nevertheless, growing complexity of the pattern estimation could make this idea impractical.

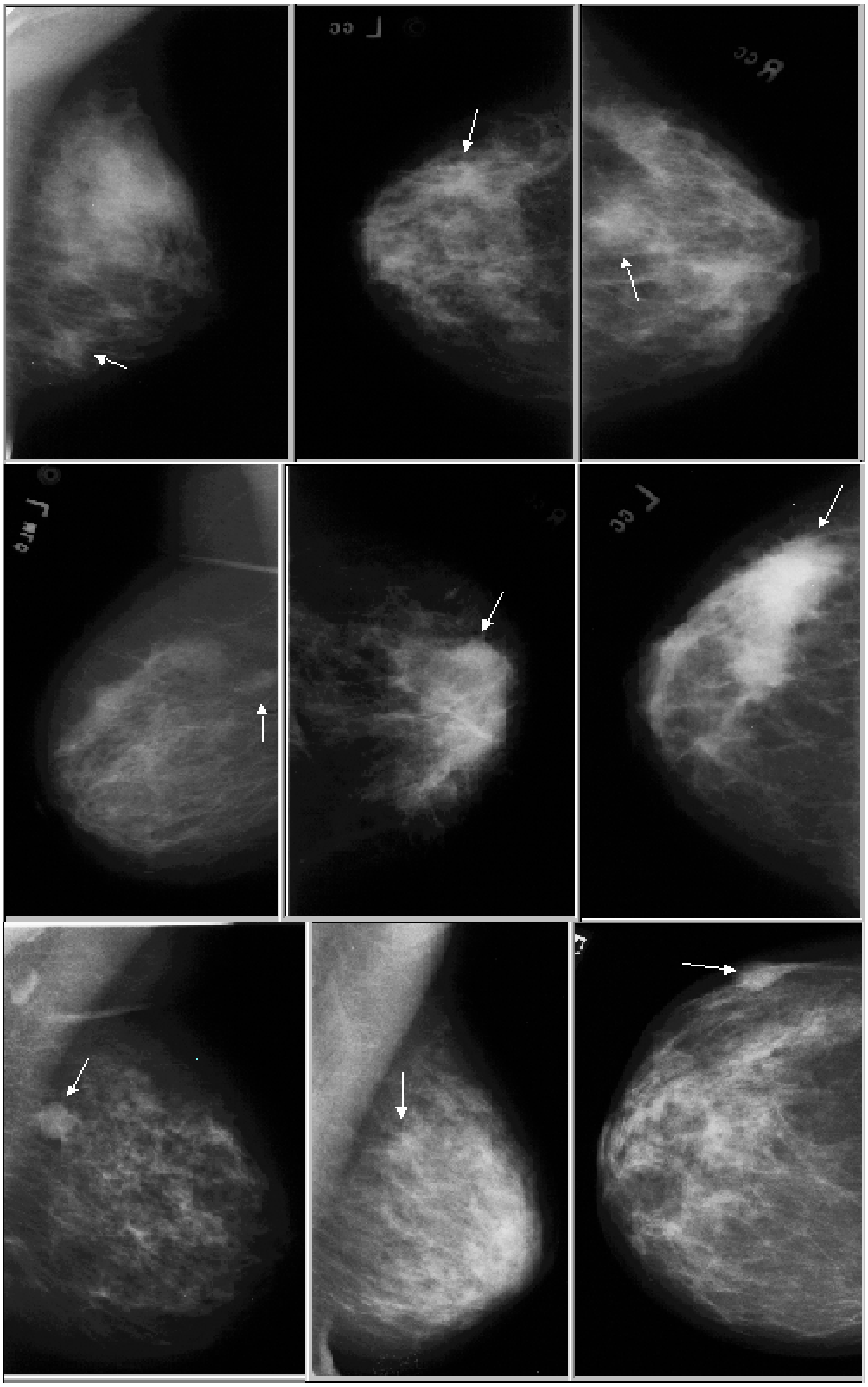


Fig. 1. Test mammograms used in diagnostic pattern estimation. Final test set of 9 images used for pattern estimation. White arrows denote suspected regions. In order from left to right and from top to bottom there are: 1) mass of high-density with ill-defined margin; 2) spiculated mass; 3) spiculated mass – original 6 in Tab. 3, which was particularly difficult to interpret; 4) spiculated lesion; 5) circumscribed mass of high-density to differentiate between malignant and benign; 6,7) circumscribed mass of high-density; 8) spiculated mass of high density; 9) mass of high-density with well defined margin

TABLE I. DQP of mammograms. Each rated separately test set included original (O) and its 4 reconstructions. Sum of scores given for particular mammogram in categories of contrast, interpretation, shape and margin is a general score of image diagnostic accuracy. All general scores obtained by test mammograms form *DQP*.

	Image	Score		Image	Score		Image	Score
Test set 1	O	9.00	**Test set 6**	O	9.43	**Test set 11**	O	11.29
	O_1.0	10.14		O_1.0	10.86		O_1.0	10.43
	O_0.6	8.57		O_0.6	9.57		O_0.6	7.29
	O_0.1	9.29		O_0.1	8.57		O_0.1	6.43
	O_0.04	7.14		O_0.04	5.43		O_0.04	4.29
Test set 2	O	9.57	**Test set 7**	O	9.43	**Test set 12**	O	10.00
	O_1.0	8.43		O_1.0	9.71		O_1.0	9.86
	O_0.6	8.86		O_0.6	8.43		O_0.6	7.71
	O_0.1	8.00		O_0.1	6.14		O_0.1	4.71
	O_0.04	4.71		O_0.04	4.57		O_0.04	4.14
Test set 3	O	10.14	**Test set 8**	O	10.14	**Test set 13**	O	10.71
	O_1.0	10.43		O_1.0	10.14		O_1.0	9.43
	O_0.6	9.57		O_0.6	10.71		O_0.6	10.43
	O_0.1	6.86		O_0.1	9.43		O_0.1	7.43
	O_0.04	4.14		O_0.04	5.71		O_0.04	4.43
Test set 4	O	8.14	**Test set 9**	O	10.43	**Test set 14**	O	10.57
	O_1.0	10.86		O_1.0	8.57		O_1.0	10.00
	O_0.6	10.00		O_0.6	8.29		O_0.6	10.71
	O_0.1	4.29		O_0.1	9.14		O_0.1	6.57
	O_0.04	4.29		O_0.04	4.43		O_0.04	4.71
Test set 5	O	9.00	**Test set 10**	O	10.29	**Test set 15**	O	9.29
	O_1.0	11.29		O_1.0	9.86		O_1.0	9.43
	O_0.6	8.86		O_0.6	9.29		O_0.6	10.00
	O_0.1	8.71		O_0.1	9.43		O_0.1	7.14
	O_0.04	5.57		O_0.04	5.71		O_0.04	4.00

TABLE II. An example of the results of subjective ratings - all scores collected for test set 1 (appendix for the results from Table 1).

Image	Contrast								Interpretation clarity								Sum
	7 radiologists scores							Mean score	7 radiologists scores							Mean score	
O	2	2	3	1	2	1	2	1.86	3	2	3	2	1	1	3	2.14	
O_1.0	3	2	2	3	3	1	3	2.43	3	2	2	3	3	1	3	2.43	
O_0.6	3	2	2	2	2	2	3	2.29	2	2	2	2	2	2	2	2.00	
O_0.1	2	2	2	2	3	3	2	2.29	2	2	2	1	3	3	1	2.00	
O_0.04	1	1	1	2	3	2	2	1.71	1	1	1	1	3	1	1	1.29	
Image	Shape								Margin								
	7 radiologists scores							Mean score	7 radiologists scores							Mean score	
O	3	2	3	2	2	1	3	2.29	3	3	3	2	3	2	3	2.71	9.00
O_1.0	3	3	2	3	3	2	3	2.71	2	3	2	3	3	2	3	2.57	10.14
O_0.6	2	2	2	2	2	2	3	2.14	1	2	2	2	3	2	3	2.14	8.57
O_0.1	2	2	2	2	3	3	3	2.43	2	3	2	2	3	3	3	2.57	9.29
O_0.04	2	3	1	1	3	2	1	1.86	2	3	2	1	3	2	3	2.29	7.14

4. CONCLUSIONS

Presented conception of diagnostic accuracy approximation should be treated as a hypothesis, a trial of simplification of unpractical ROC-based analyses in order to design useful tools for assessment of irreversible compressed medical images. Experimental diagnostic pattern for test mammograms seems to be reliable in the assumed limits of measure complexity and useful for compression applications because of the following reasons: subjective test was based on 2100 particular scores given by radiologists from different medical centres, test mammograms were selected solidly, test rules, a verification process, and 'gold standard' estimation were the results of experts' consultations and consensus, the analysis and general remarks were supported by the suggestions of radiologists collected during performed tests.

Convergence of image scores, comments and remarks given by radiologists in the subjective tests suggest that lossy wavelet compression (in an acceptable bit rates range) does not reduce diagnostic accuracy of original images. The comparison between original and reconstructed images did not demonstrate diagnostically important differences in radiologists' consenting opinion. The suggested acceptable bit rate value was even close to 0.1 bpp for certain tested mammograms. In many cases the rate differences up to 0.1 bpp due to lossy compression were smaller, in contrast to the differences among individual radiologist rates given for the same images, estimated over different periods of time. Therefore, the use of high degrees of irreversible compression of mammograms may be acceptable for diagnostic tasks.

More general objective of this study was to extract information (features) from mammograms and use them to improve lesion detection process. Defined diagnostic features may be applied in CAD system design. Radiologists participated in the experiments were trained to recognize the normal or abnormal appearance of mammograms taking into consideration these diagnostic features. Generally, the components of presented concepts seem to be important in more objective characteristics of diagnostic process. It could help radiologists to establish more objective criteria of diagnosis.

REFERENCES

[1] P.C. Cosman, R.M. Gray, R.A. Olshen, "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy", *Proc. of IEEE*, vol. 82, no. 6, 1994.

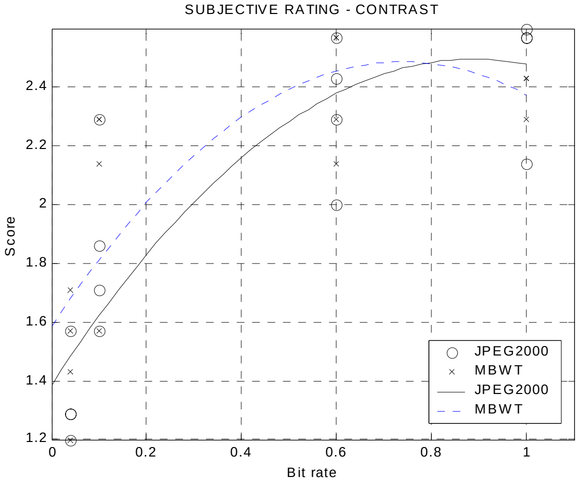
[2] B. Erickson, "Irreversible compression of medical images", *J. Digital Imaging*, vol. 15, no. 1, pp. 5-14, 2002.

[3] ISO/IEC 15444-1,2: JPEG2000 image coding system (2000) VM8.6.

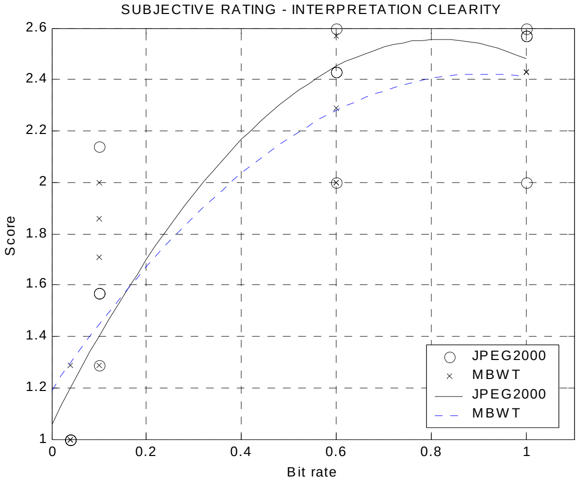
[4] A. Przelaskowski, "Details preserved wavelet-based compression with adaptive context-based quantisation", *Fundamenta Informaticae*, vol. 34, no. 4, pp. 369-388, 1998.

[5] A. Przelaskowski, "Numerical equivalent of accuracy as a measure of medical image quality", *presented at* XVII IMEKO World Congress*,* 2003.

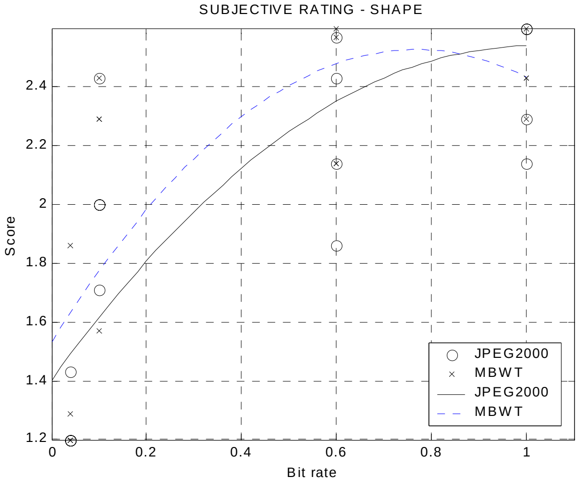
a)



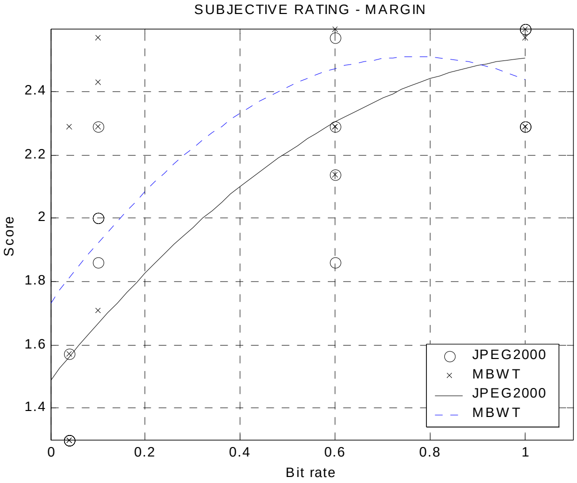
b)



c)



d)



e)

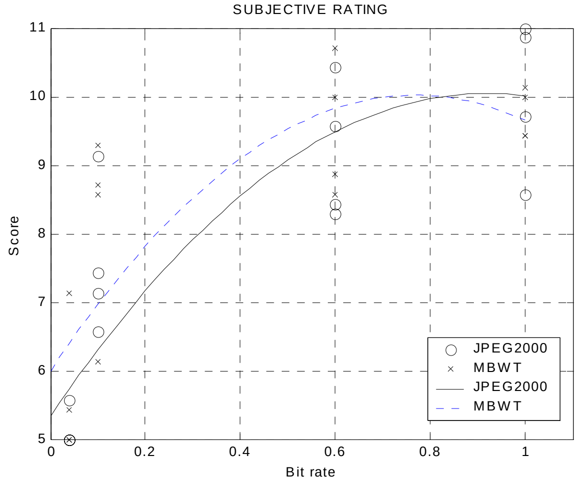


Fig. 2. Experimental results of compression efficiency evaluation. Diagnostic quality of mammograms compressed by two irreversible coders: JPEG2000 and MBWT was rated subjectively. The elements of given scores: contrast (a), interpretation clarity (b), shape (c), margin (d), and final *DQP* score (e) in a range of 0.04 to 1 bpp are presented and interpreted.

Authors: 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
Anna Kukuła, MD, Radiology Centre, Wolski Hospital, ul. Kasprzaka 17, 01-211 Warszawa, Poland, phone: +48 22 6327389