Effective wavelet-based compression method with adaptive quantization threshold and zerotree coding

Artur Przelaskowski, Marian Kazubek, Tomasz Jamrógiewicz

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

ABSTRACT

Efficient image compression technique especially for medical applications is presented. Dyadic wavelet decomposition by use of Antonini and Villasenor bank filters is followed by adaptive space-frequency quantization and zerotree-based entropy coding of wavelet coefficients. Threshold selection and uniform quantization is made on a base of spatial variance estimate built on the lowest frequency subband data set. Threshold value for each coefficient is evaluated as linear function of 9-order binary context. After quantization zerotree construction, pruning and arithmetic coding is applied for efficient lossless data coding. Presented compression method is less complex than the most effective EZW-based techniques but allows to achieve comparable compression efficiency. Specifically our method has similar to SPIHT efficiency in MR image compression, slightly better for CT image and significantly better in US image compression. Thus the compression efficiency of presented method is competitive with the best published algorithms in the literature across diverse classes of medical images.

Keywords: wavelet transform, image compression, medical image archiving, adaptive quantization

1. INTRODUCTION

Lossy image compression techniques allow significantly diminish the length of original image representation at the cost of certain original data changes. At range of lower bit rates these changes are mostly observed as distortion but sometimes improved image quality is visible. Compression of the concrete image with its all important features preserving and the noise and all redundancy of original representation removing is do required. The choice of proper compression method depends on many factors, especially on statistical image characteristics (global and local) and application. Medical applications seem to be challenged because of restricted demands on image quality (in the meaning of diagnostic accuracy) preserving. Perfect reconstruction of very small structures which are often very important for diagnosis even at low bit rates is possible by increasing adaptability of the algorithm. Fitting data processing method to changeable data behaviour within an image and taking into account a priori data knowledge allow to achieve sufficient compression efficiency. Recent achievements clearly show that nowadays wavelet-based techniques can realise these ideas in the best way.

Wavelet transform features are useful for better representation of the actual nonstationary signals and allow to use a priori and a posteriori data knowledge for diagnostically important image elements preserving. Wavelets are very efficient for image compression as entire transformation basis function set. This transformation gives similar level of data decorrelation in comparison to very popular discrete cosine transform and has additional very important features. It often provides a more natural basis set than the sinusoids of the Fourier analysis, enables widen set of solution to construct effective adaptive scalar or vector quantization in time-frequency domain and correlated entropy coding techniques, does not create blocking artefacts and is well suited for hardware implementation. Wavelet-based compression is naturally multiresolution and scalable in different applications so that a single decomposition provides reconstruction at a variety of sizes and resolutions (limited by compressed representation) and progressive coding and transmission in multiuser environments.

Wavelet decomposition can be implemented in terms of filters and realised as subband coding approach. The fundamental issue in construction of efficient subband coding techniques is to select, design or modify the analysis and synthesis filters.[1] Wavelets are good tool to create wide class of new filters which occur very effective in compression schemes. The choice of suitable wavelet family, with such criteria as regularity, linearity, symmetry, orthogonality or impulse and step response of corresponding filter bank, can significantly improve compression efficiency. For compactly supported wavelets corresponding filter length is proportional to the degree of smoothness and regularity of the wavelet. But

when the wavelets are orthogonal (the greatest data decorrelation) they also have non-linear phase in the associated FIR filters. The symmetry, compact support and linear phase of filters may be achieved by biorthogonal wavelet bases application. Then quadrature mirror and perfect reconstruction subband filters are used to compute the wavelet transform. Biorthogonal wavelet-based filters occurred very efficient in compression algorithms. A construction of wavelet transformation by fitting local defined basis transformation function (or finite length filters) into image data characteristics is possible but very difficult. Because of nonstationary of image data, miscellaneous image futures which could be important for good reconstruction, significant various image quality (signal to noise level, spatial resolution etc.) from different imaging systems it is very difficult to elaborate the construction method of the optimal-for-compression filters. Many issues relating to the choice of the most efficient filter bank for image compression remain still unresolved.[2] The demands of preserving the diagnostic accuracy in reconstructed medical images are exacting. Important high frequency coefficients which appear at the place of small structure edges in CT and MR images should be saved. Accurate global organ shapes reconstruction in US images and strong noise reduction in MN images is also required. It is rather difficult to imagine that one filter bank can do it in the best way. Rather choosing the best wavelet families for each modality is expected.

Our aim is to increase the image compression efficiency, especially for medical applications, by applying suitable wavelet transformation, adaptive quantization scheme and corresponding processed decomposition tree entropy coding. We want to achieve higher acceptable compression ratios for medical images by better preserving the diagnostic accuracy of images. Many bit allocation techniques applied in quantization scheme are based on data distribution assumptions, quantiser distortion function etc. All statistical assumptions built on global data characteristics do not cover exactly local data behaviour and important detail of original image, e.g., different texture small area may be lost. Thus we decided to build quantization scheme on the base of local data characteristics such a direct data context in two dimensions mentioned earlier. We do data variance estimation on the base of real data set as spatial estimate for corresponding coefficient positions in successive subbands. The details of quantization process and correlated coding technique as a part of effective simple wavelet-based compression method which allows to achieve high reconstructed image quality at low bit rates are presented.

2. THE COMPRESSION TECHNIQUE

Scheme of our algorithm is very simple: dyadic, 3 levels decomposition of original image (256×256 images were used) done by selected filters. For symmetrical filters symmetry boundary extension at the image borders was used and for asymmetrical filters - a periodic (or circular) boundary extension.

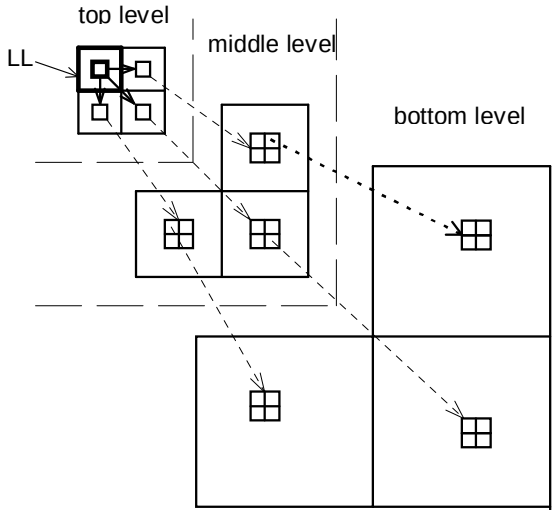


Figure 1. Dyadic wavelet image decomposition scheme. ▬ - horizontal relations, ┈┈ - parent - children relations. LL - the lowest frequency subband.

Our approach to filters is utilitarian one, making use of the literature to select the proper filters rather than to design them. We conducted an experiment using different kinds of wavelet transformation in presented algorithm. Long list of wavelet families and corresponding filters were tested: Daubechies, Adelson, Brislawn, Odegard, Villasenor, Spline, Antonini, Coiflet, Symmlet, Beylkin, Vaid etc.[3] Generally Antonini [4] filters occurred to be the most efficient. Villasenor, Odegard and Brislawn filters allow to achieve similar compression efficiency. Finally: Antonini 7/9 tap filters are used for MR and US image compression and Villasenor 18/10 tap filters for CT image compression.

2.1 Adaptive space-frequency quantization

Presented space-frequency quantization technique is realised as entire data pre-selection, threshold selection and scalar uniform quantization with step size conditioned by chosen compression ratio. For adaptive estimation of threshold and quantization step values two extra data structure are build. Entire data pre-selection allows to evaluate zero-quantized data set and predict the spatial context of each coefficient. Next simple quantization of the lowest frequency subband (LL) allows to estimate quantized coefficient variance prediction as a space function across sequential subbands. Next the value of quantization step is slightly modified by a model build on variance estimate. Additionally, a set of coefficients is reduced by threshold selection. The threshold value is increased in the areas with the dominant zero-valued coefficients and the level of growth depends on coefficient spatial position according variance estimation function.

Firstly zero-quantized data prediction is performed. The step size w is assumed to be constant for all coefficients at each decomposition level. For such quantization model the threshold value is equal to $w/2$. Each coefficient whose value is less than threshold is predicted to be zero-valued after quantization (insignificant). In opposite case coefficient is predicted to be not equal to zero (significant). It allows to create predictive zero-quantized coefficients P map for threshold evaluation in the next step. The process of P map creation is as follows:

$$\begin{array}{lll} if & c_i < w/2 & then \quad p_i = 0 \\ & else & \quad p_i = 1 \end{array}, \tag{1}$$

where $i = 1,2,...,m \cdot n$; $m,n -$ horizontal and vertical image size, c_i - wavelet coefficient value.

The coefficient variance estimation is made on the base of LL data for coefficients from next subbands in corresponding spatial positions. The quantization with mentioned step size w is performed in LL and the most often occurring coefficient value is estimated. This value is named MHC (mode of histogram coefficient). The areas of MHC appearance are strongly correlated with zero-valued data areas in the successive subbands. The absolute difference of the LL quantized data and MHC is used as variance estimate for next subband coefficients in corresponding spatial positions. We tested many different schemes but this model allows to achieve the best results in the final meaning of compression efficiency. The variance estimation is rather coarse but this simple adaptive model built on real data does not need additional information for reconstruction process and increases the compression efficiency. Let lc_i, i=1,2,...,lm, be a set of LL quantized coefficient values, lm - size of this set. Furthermore let mode of histogram coefficient MHC value be estimated as follows:

$$f(MHC) = \max_{lc_i \in Al} f(lc_i) \quad \text{and} \quad MHC \in Al, \tag{2}$$

where Al - alphabet of data source which describes the values of the coefficient set and $f(lc_i) = \frac{n_{lc_i}}{lm}$, n_{lc_i} - number of lc_i -valued coefficients. The normalised values of variance estimate ve_{si} for next subband coefficients in corresponding to i spatial positions (parent - children relations from the top to the bottom of zerotree - see fig. 1) are simply expressed by the following equation:

$$ve_{si} = \left| \frac{lc_i - MHC}{ve_{\max}} \right|. \tag{3}$$

These set of ve_{si} data is treated as top parent estimation and is applied to all corresponding child nodes in wavelet hierarchical decomposition tree.

9-th order context model is applied for coarser data reduction in ‘unimportant' areas (usually with low diagnostic importance). The unimportance means that in these areas the majority of the data are equal to zero and significant values are separated. If single significant values appear in these areas it most often suggests that these high frequency coefficients are caused by noise. Thus the coarser data reduction by higher threshold allows to increase signal to noise ratio by removing the noise. At the edges of diagnostically important structures significant values are grouped together and the threshold value is lower at this fields. *P* map is used for each coefficient context estimation. Noncausal prediction of the coefficient importance is made as linear function of the binary surrounding data excluding considered coefficient significance. The other polynomial, exponential or hyperbolic function were tested but linear function occurred the most efficient. The data context shown on fig. 2 is formed for each coefficient. This context is modified in the previous data points of processing stream by the results of the selection with the actual threshold values at these points instead of *w*/2 (causal modification). Values of the coefficient importance - *cim* are evaluated for each c_i coefficient from the following equation:

$$cim_i = coeff_1 \cdot (9 - \sum_{j=1}^{9} p_{i,j}), \text{ where } i = 1, 2, \ldots, m \cdot n. \tag{4}$$

Next the threshold value is evaluated for each c_i coefficient:

$$th_i = w / 2 \cdot (1 + cim_i \cdot w \cdot (1 - ve_{si})), \tag{5}$$

where $i = 1, 2, \ldots, m \cdot n$, *si* - corresponding to LL parent spatial location in lower decomposition levels.

The modified quantization step model uses the LL-based variance estimate to slightly increase the step size for less variance coefficients. Threshold data selection and uniform quantization is made as follows: each coefficient value is firstly compared to its threshold value and then quantized using *w* step for LL and modified step value mw_{si} for next subbands. Threshold selection and quantization for each c_i coefficient can be clearly described by the following equations:

$$\begin{array}{l} \textit{if} \quad c_i \in \mathrm{LL} \quad \textit{then} \quad c_i = c_i / w \\ \quad \textit{else} \\ \qquad \textit{if} \quad c_i < th_i \quad \textit{then} \quad c_i = 0 \\ \qquad\quad \textit{else} \qquad\quad c_i = c_i / mw_{si} \end{array}, \tag{6}$$

where

$$mw_{si} = w \cdot (1 + coeff_2 \cdot (1 - ve_{si})). \tag{7}$$

The $coeff_1$ and $coeff_2$ values are fitted to actual data characteristic by using a priori image knowledge and performing entire tests on groups of similar characteristic images.

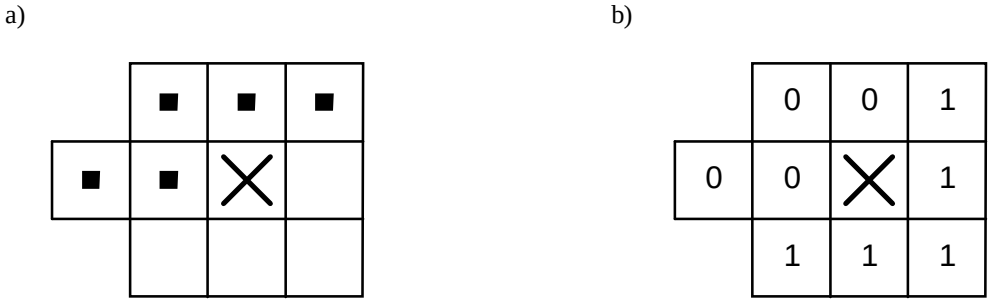


Figure 2. a) 9-order coefficient context for evaluating the coefficient importance value in procedure of adaptive threshold value estimation; ■ - points of coefficient context causal modification, b) example of binary *P* map context of single edge coefficient.

2.2 Zerotrees construction and coding

Sophisticated entropy coding methods which can significantly improve compression efficiency should retain progressive way of data reconstruction. Progressive reconstruction is simple and natural after wavelet-based decomposition. Thus the wavelet coefficient values are coded subband-sequentially and spectral selection is made typically for wavelet methods. The same scale subbands are coded as follows: firstly the lowest frequency subband, then right side coefficient block, down-left and down-right block at the end. After that next larger scale data blocks are coded in the same order. To reduce a redundancy of such data representation zerotree structure is built. Zerotree describes well the correlation between data values in horizontal and vertical directions, especially between large areas with zero-valued data. These correlated fragments of zerotree are removed and final data streams for entropy coding are significantly diminish. Also zerotree structure allows to create different characteristics data streams to increase the coding efficiency. We used simple arithmetic coders for these data streams coding instead of applied in many techniques bit map (from MSB to LSB) coding with necessity of applying the efficient context model construction. Because of refusing the successive approximation we lost full progression. But the simplicity of the algorithm and sometimes even higher coding efficiency was achieved. Two slightly different arithmetic coders for producing ending data stream were used.

2.2.1 Construction and pruning of zerotree

The dyadic hierarchical image data decomposition is presented on fig. 1. Decomposition tree structure reflects this hierarchical data processing and strictly corresponds to created in transformation process data streams. The four lowest frequency subbands which belong to the coarsest scale level are located at the top of the tree. These data have not got parent values but they are the parents for the coefficients in lower tree level of greater scale in corresponding spatial positions. These correspondence is shown on the fig. 1 as parent-children relations. Each parent coefficient has got four direct children and each child is under one direct parent. Additionally, horizontal relations at top tree level are introduced to describe the data correlation in better way.

The decomposition tree becomes zerotree when node values of quantized coefficients are signed by symbols of binary alphabet. Each tree node is checked to be significant (not equal to zero) or insignificant (equal to zero) - binary tree is built. For LL nodes way of significance estimation is slightly different. The MHC value is used again because of the LL areas of MHC appearance strong correlation with zero-valued data areas in the next subbands. Node is signed to be significant if its value is not equal to MHC value or insignificant if its value is equal to MHC. The value of MHC must be sent to a decoder for correct tree reconstruction.

Next step of algorithm is a pruning of this tree. Only the branches to insignificant nodes can be pruned and the procedure is slightly other at different levels of the zerotree. Procedure of zerotree pruning starts at the bottom of wavelet zerotree. Sequential values of four children data and their parent from higher level are tested. If the parent and the children are insignificant - the tree branch with child nodes is removed and the parent is signed as pruned branch node (PBN). Because of this the tree alphabet is widened to three symbols. At the middle levels the pruning of the tree is performed if the parent value is insignificant and all children are recognised as PBN. From conducted research we found out that adding extra symbols to the tree alphabet is not efficient for decreasing the code bit rate. The zerotree pruning at top level is different. The checking node values is made in horizontal tree directions by exploiting the spatial correlation of the quantized coefficients in the subbands of the coarsest scale - see fig. 1. Sequentially the four coefficients from the same spatial positions and different subbands are compared with one another. The tree is pruned if the LL node is insignificant and three corresponding coefficients are PBN. Thus three branches with nodes are removed and LL node is signed as PBN. It means that all its children across zerotree are insignificant. The spatial horizontal correlation between the data at other tree levels is not strong enough to increase the coding efficiency by its utilisation.

2.2.2 Making three data streams and coding

Pruned zerotree structure is handy to create data streams for ending efficient entropy coding. Instead of PBN zero or MHC values (nodes of LL) additional code value is inserted into data set of coded values. Also bit maps of PBN spatial distribution at different tree levels can be applied. We used optionally only PBN bit map of LL data to slightly increase the coding efficiency. The zerotree coding is performed sequentially from the top to the bottom to support progressive reconstruction. Because of various quantized data characteristics and wider alphabet of data source model after zerotree pruning three separated different data streams and optionally fourth bit map stream are produced for efficient data coding. It is well known from information theory that if we deal with a data set with significant variability of data statistics and

different statistics (alphabet and estimate of conditional probabilities) data may be grouped together it is better to separate these data and encode each group independently to increase the coding efficiency. Especially is true when context-based arithmetic coder is used. The data separation is made on the base of zerotree and than the following data are coded independently:

- the LL data set which has usually smaller number of insignificant (MHC-valued) coefficients, less PBN and less spatial data correlation than next subband data (word- or charwise arithmetic coder is less efficient then bitwise coder); optionally this data stream is divided on PBN distribution bit map and word or char data set without PBNs,
- the rest of top level (three next subbands) and middle level subband data set with a considerable number of zero-valued (insignificant) coefficients and PBN code values; level of data correlation is greater, thus word- or charwise arithmetic coder is efficient enough,
- the lowest level data set with usually great number of insignificant coefficients and without PBN code value; data correlation is very high.

Urban Koistinen arithmetic coder (DDJ Compression Contest public domain code accessible by internet) with simple bitwise algorithm is used for first data stream coding. For the second and third data stream coding 1-st order arithmetic coder built on the base of code presented in Nelson book [5] is applied. Urban coder occurred up to 10% more efficient than Nelson coder for first data stream coding. Combining a rest of top level data and the similar statistics middle level data allows to increase the coding efficiency approximately up to 3%.

The procedure of the zerotree construction, pruning and coding is presented on fig. 3.

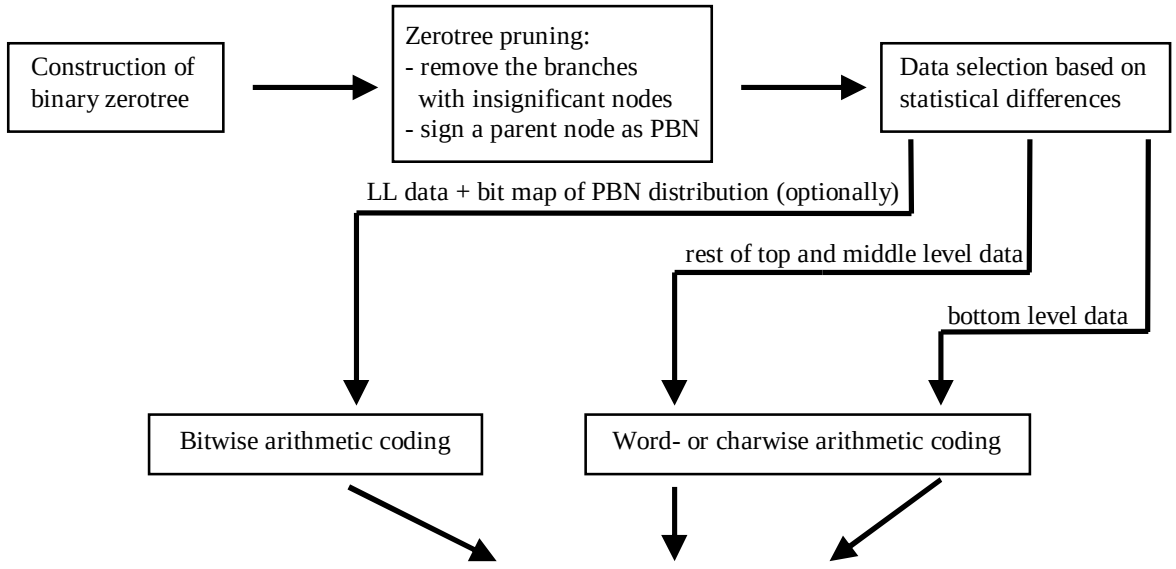


Figure 3. Quantized wavelet coefficients coding scheme with using zerotree structure. PBN - pruned branch node.

3. TESTS, RESULTS AND DISCUSSION

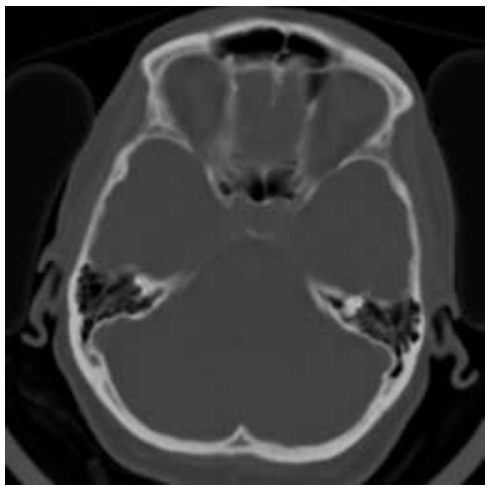
In our tests many different medical modality images were used. For chosen results presentation we applied three 256×256×8-bit images from various medical imaging systems: CT (computed tomography), MR (magnetic resonance) and US(ultrasound) images. These images are shown on fig. 4. Mean square error - MSE and peak signal to noise ratio - PSNR were assumed to be reconstructed image quality evaluation criteria. Subjective quality appreciation was conducted in very simple way - only by psychovisual impression of the non-professional observer.

Application of adaptive quantization scheme based on modified threshold value and quantization step size is more efficient than simple uniform scalar quantization up to 10% in a sense of better compression of all algorithm. Generally applying zerotree structure and its processing improved coding efficiency up to 10% in comparison to direct arithmetic coding of quantized data set.

The comparison of the compression efficiency of three methods: DCT-based algorithm,[6,7] SPIHT [8] and presented compression technique, called MBWT (modified basic wavelet-based technique) were performed for efficiency evaluation of MBWT. The results of MSE and PSNR-based evaluation are presented in table 1. Two wavelet-based compression techniques are clearly more efficient than DCT-based compression in terms of MSE/PSNR and also in our subjective evaluation for all cases. MBWT overcomes SPIHT method for US images and slightly for CT test image at lower bit rate range.

The concept of adaptive threshold and modified quantization step size is effective for strong reduction of noise but it occurs sometimes too coarse at lower bit rate range and very small details of the image structures are put out of shape. US images contain significant noise level and diagnostically important small structures do not appear (image resolution is poor). Thus these images can be efficiently compressed by MBWT with image quality preserved. It is clearly shown on fig. 5. An improvement of compression efficiency in relatio to SPIHT is almost constant at wide range of bit rates (0.3 - 0.6 dB of PSNR).

a) 

b) 

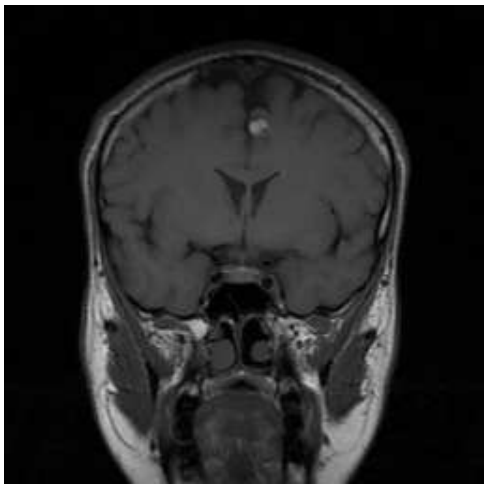
c) 

Figure 4. Examples of images used in the tests of compression efficiency evaluation. The results presented in table 1 and on fig. 5 were achieved for those images. The images are as follows: a) echocardiography image, b) CT head image, c) MR head image.

Table 1. Comparison of the three techniques compression efficiency: DCT-based, SPIHT and MBWT. The bit rates are chosen in diagnostically interesting range (near the borders of acceptance).

Modality - bit rate	DCT-based		SPIHT		MBWT	
	MSE	PSNR[dB]	MSE	PSNR[dB]	MSE	PSNR[db]
MRI - 0.70 bpp	8.93	38.62	4.65	41.45	4.75	41.36
MRI - 0.50 bpp	13.8	36.72	8.00	39.10	7.96	39.12
CT - 0.50 bpp	6.41	40.06	3.17	43.12	3.18	43.11
CT - 0.30 bpp	18.5	35.46	8.30	38.94	8.06	39.07
US - 0.40 bpp	54.5	30.08	31.3	33.18	28.3	33.61
US - 0.25 bpp	91.5	28.61	51.5	31.01	46.8	31.43

The level of noise in CT and MR images is lower and small structures are often important in image analysis. That is the reason why the benefits of MBWT in this case are smaller. Generally compression efficiency of MBWT is comparable to SPIHT for these images. Presented method lost its effectiveness for higher bit rates (see PSNR of 0.7 bpp MR representation) but for lower bit rates both MR and CT images are compressed significantly better. Maybe the reason is that the coefficients are reduced relatively stronger because of its importance reduction in MBWT threshold selection at lower bits rate range.

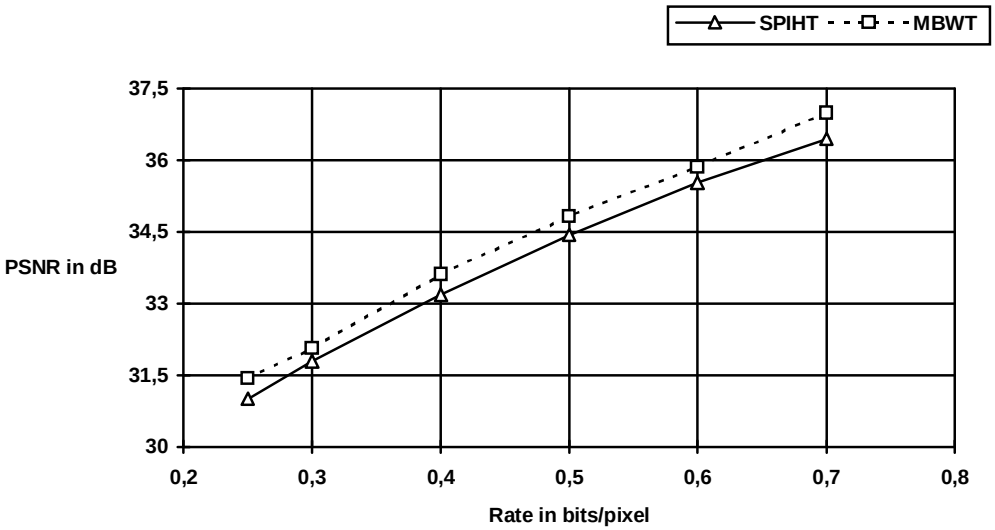


Figure 5. Comparison of SPIHT and presented in this paper technique (MBWT) compression efficiency at range of low bit rates. US test image was compressed.

4. CONCLUSIONS

Adaptive space-frequency quantization scheme and zerotree-based entropy coding are not time-consuming and allow to achieve significant compression efficiency. Generally our algorithm is simpler than EZW-based algorithms [9] and other algorithms with extended subband classification or space -frequency quantization models [10] but compression efficiency of presented method is competitive with the best published algorithms in the literature across diverse classes of medical images. The MBWT-based compression gives slightly better results than SPIHT for high quality images: CT and MR and significantly better efficiency for US images. Presented compression technique occurred very useful and promising for medical applications. Appropriate reconstructed image quality evaluation is desirable to delimit the acceptable lossy compression ratios for each medical modality. We intend to improve the efficiency of this method by: the design a construction method of adaptive filter banks and correlated more sufficient quantization scheme. It seems to be possible by

applying proper a priori model of image features which determine diagnostic accuracy. Also more efficient context-based arithmetic coders should be applied and more sophisticated zerotree structures should be tested.

REFERENCES

1. Hui, C. W. Kok, T. Q. Nguyen, „Image Compression Using Shift-Invariant Dydiadic Wavelet Transform", subbmited to IEEE Trans. Image Proc., April 3nd, 1996.
2. J. D. Villasenor, B. Belzer and J. Liao, „Wavelet Filter Evaluation for Image Compression", IEEE Trans. Image Proc., August 1995.
3. A. Przelaskowski, M.Kazubek, T. Jamrógiewicz, „Optimalization of the Wavelet-Based Algorithm for Increasing the Medical Image Compression Efficiency", submitted and accepted to *TFTS'97 2nd IEEE UK Symposium on Applications of Time-Frequency and Time-Scale Methods*, Coventry, UK 27-29 August 1997.
4. M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, „Image coding using wavelet transform", IEEE Trans. Image Proc., vol. IP-1, pp.205-220, April 1992.
5. M. Nelson, *The Data Compression Book*, chapter 6, M&T Books, 1991.
6. M. Kazubek, A. Przelaskowski and T. Jamrógiewicz, „Using A Priori Information for Improving the Compression of Medical Images", *Analysis of Biomedical Signals and Images*, vol. 13, pp. 32-34, 1996.
7. A. Przelaskowski, M. Kazubek and T. Jamrógiewicz, „Application of Medical Image Data Characteristics for Constructing DCT-based Compression Algorithm", *Medical & Biological Engineering & Computing*, vol. 34, Supplement I, part I, pp.243-244, 1996.
8. A. Said and W. A. Pearlman, „A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees", submitted to IEEE Trans. Circ. & Syst. Video Tech., 1996.
9. J. M. Shapiro, „Embedded Image Coding Using Zerotrees of Wavelet Coefficients", IEEE Trans. Signal Proces., vol. 41, no.12, pp. 3445-3462, December 1993.
10. Z. Xiong, K. Ramchandran and M. T. Orchard, „Space-Frequency Quantization for Wavelet Image Coding", IEEE Trans. Image Proc., to appear in 1997.