L. Yaroslavsky. Fundamentals of Digital Image Processing. Course 0555.3230

Lec. 6. Principles of image coding

The term “image coding” or “image compression” refers to processing image digital data aimed at achieving the lowest possible volume of the data, measured in number of bits per pixel, provided appropriate quality of image reconstruction from the encoded data. Image reconstruction quality is assessed by the image end user. Three classes of image end users should be distinguished: collective users, such as viewers in image broadcasting in television and internet, expert user such as image interpreters in specific image related applications, and automatons.

Image coding methods can be classified into two large categories: “loss-less” compression methods and “lossy” compression methods. Loss-less compression methods allow exact reconstruction of initial digital signals from the compressed data. In lossy compression, initial digital images are not exactly reconstructed from the compressed data, although reconstructed images preserve the quality required for applications.

Classification diagram of image coding methods is presented in Fig. 6.1. Coding methods are classified in into two groups: methods in which data decorrelation and quantization are carried out separately one after another and methods in which one can not separate decorrelation from quantization. The latter attempt to implement, though in a reduced form, the idea of encoding of signals as a whole to achieve the theoretical minimum of bits.

In decorrelation-then-quantization methods, data decorrelation is carried out with linear transforms. Two large families of decorrelating linear transforms are in use: “predictive” and orthogonal transforms. In predictive transforms, for each pixel, its “predicted” value is generated in a certain way from other pixels and a prediction error is computed as a difference between the pixel value and its “predicted” value. The prediction error is then sent to a scalar quantizer. The simplest method that implements decorrelation by prediction is known as ***DPCM*** (from “Differential Pulse Code Modulation). Block diagram of DPCM coding and decoding are shown in Fig. 6.2. DPCM assumes that images are scanned row-wise/column-wise and, for each current pixel $a_{k,l}$, prediction error is computed as (Fig. 6.3)

$$\varepsilon_{k,l} = a_{k,l} - \bar{a}_{k,l} = a_{k,l} - c_{-1,0}a_{k-1,l} - c_{0,-1}a_{k,l-1} - c_{-1,-1}a_{k-1,l-1} + c_{-1,1}a_{k-1,l+1} . \qquad (6.1)$$

where $\bar{a}_{k,l}$ is a prediction value. Prediction coefficients $\{c\}$ are usually chosen so as to minimize prediction error standard deviation as computed for an ensemble of pixels subjected to coding. For instance, in a natural assumption that images are statistically isotropic in all directions,

$$c_{-1,0} = c_{0,-1} = c_1;\ c_{-1,-1} = c_{-1,1} = c_2 , \qquad (6.2)$$

and c_1 and c_2 are found from the equation:

$$\begin{cases} c_1(1+\rho_d) + c_2\dfrac{3\rho_{hv}+\rho_{2hv}}{2} = \rho_{hv} \\ c_1\dfrac{3\rho_{hv}+\rho_{2hv}}{2} + c_2(1+\rho_{2h}) = \rho_d \end{cases} , \qquad (6.3)$$

where

$$\rho_{hv} = \frac{AV(a_{k-1,l}a_{k,l} + a_{k,l-1}a_{k,l})}{2AV((a_{k,l})^2)}; \quad \rho_d = \frac{AV(a_{k-1,l-1}a_{k,l})}{AV((a_{k,l})^2)}; \qquad (6.4)$$

$$\rho_{2h} = \frac{AV(a_{k-1,l-1}a_{k-1,l+1})}{AV((a_{k,l})^2)}; \quad \rho_{2hv} = \frac{AV(a_{k,l-1}a_{k-1,l+1})}{AV((a_{k,l})^2)}; \qquad (6.5)$$

are correlation coefficients of pixels with their immediate neighbors in vertical or horizontal direction, and of pixels with their immediate neighbor in diagonal direction. Fig. 6-4 illustrates how predictive decorrelation reduces dynamic range and entropy of signal-to-be-quantized.

DPCM with no quantization is used for loss-less image coding. With quantization, substantially larger compression is possible although images are reconstructed with losses. Quantization artifacts can be reduced if prediction error is computed from already quantized data as it is done in ***DPCM with feedback***

(Fig. 6-5). From the classification point of view, DPCM with feedback can be regarded as a combined discretization/quantization method.

In DPCM, the prediction is practically made over only immediate vicinity of pixels. More efficient decorrelation is achieved when larger spatial neighborhood of pixels are involved in the prediction as it is implemented in multi-resolution image expansion. This method, depending on the implementation, is known under different names: ***pyramid coding***, ***sub-band decomposition coding***, ***wavelet coding***. In this method, image transform coefficients obtained at each resolution (Fig. 6-6) are optimally quantized and then statistically encoded. Figs. 6-6 – 6-8 illustrate principles of such image multi-resolution decomposition.

Decorrelation with orthogonal transforms is an alternative to predictive decorrelation. Redundancy of the image signals exhibits itself in transform domain mostly in compaction of signal energy in small number of transform coefficients and in contraction of the dynamic range of higher order coefficients. In practice, transforms such as DFT, DCT, Walsh transforms that can be computed with fast transform algorithms are used.

Two versions of transform coding are known: frame-wise and block-wise ones. In frame-wise coding, image frame as a whole is transformed, transform coefficients with very low energy are truncated and the others are quantized. In block-wise coding, image is split into blocks and individual blocks are transformed separately. For every block, low energy block transform coefficients are truncated and the others are quantized. Truncation and quantization of transform coefficients is either predefined by a special quantization table (***zonal quantization***) or adaptively changed for every block.

In terms of energy compaction, frame-wise transforms are, in principle, more efficient then block-wise ones. However, block-wise transform requires much less computations and can even be implemented in an inexpensive hardware. Moreover, it is well suited to spatial image inhomogeneity and allows using adaptive quantization in which way truncating and quantizing coefficients is optimized for each individual block.

Among the block transforms, DCT had proved to be the best one and it is put in the base of image compression standards H.261, H.262, H.263, H.320, JPEG and MPEG. The principle of image DCT block coding is illustrated in Fig. 6-9.

In the image block transform coding, image is split up into blocks. DCT of each block is then computed and spectral coefficients are quantized individually according to a quantization table of zonal quantization that specifies the number of quantization levels allocated to each coefficients as a function of the coefficient indices. The table is built according to typical energy distribution of the transform coefficients: the higher is variance of the coefficient, the larger number of quantization levels is allocated for quantization of this coefficient. Obtained 2-D arrays of quantized numbers for each block are then converted, for transmitting and storing, into 1-D sequences by zigzag scanning of the arrays.

Quantization of DCT coefficients causes certain distortions in reconstructed images. A typical artifact caused by quantization in image block transform coding is the so-called blocking effect: visible discontinuities at the borders of the blocks. These discontinuities can be substantially mitigated if blocks are overlapping. A most known implementation of this idea is lapped transform coding .

While in coding still images predictive or orthogonal transform decorrelation are mostly used, hybrid combination of predictive and transform coding has also proved its usefulness. For instance, JPEG standard for image coding assumes using DPCM for coding dc components of image blocks and in video coding intra-frame redundancy is removed with block DCT coding whereas for removing inter-frame redundancy the ideas of predictive decorrelations are used in a form of object motion compensation.

Combined decorrelation/quantization methods coding are exemplified by DPCM with feedback, adaptive discretization and vector quantization. Adaptive discretization assumes taking samples of the signal only if difference in the signal values from the previous sample exceeds a quantization interval. For images, such coding results in generating an image "contour map". Vector quantization is an exact implementation of the general coding signals as a whole. However, owing to computer memory and computational capacity limitations it is applied to groups of signal/image samples small enough to match these limitations.

Statistical* (*entropy) coding in itself is ***loss-less coding*** because it does not assume introducing any irreversible changes (such as quantization) into the data and enables perfect restoration of initial signal representation coefficients.

Two mutually complement methods of binary statistical coding are known and overwhelmingly used: ***variable length coding*** (***VL-coding***) and ***coding of rare symbols***. In data coding jargon, objects of encoding are called symbols. VL-coding is aimed at generating, for each symbol to be coded, a binary code word of the length as close to logarithm of the inverse to its probability. Originally suggested by C. Shannon and Fano, it was later improved by Huffman, and the Huffman's coding procedure had become the preferred implementation of VL-coding.

VL-Huffman coding is an iterative process, in which, at each iteration, to codes of two symbols with the least probabilities bits 0 and 1, correspondingly, are added and then the symbols are united in a new auxiliary symbol whose probability is sum of the united symbols. The iteration process is repeated with respect to the modified at the previous iteration set of symbols until all symbols are encoded. If probabilities of symbols are integer power of ½, such a procedure generates binary codes with number of bits per symbol exactly equal to binary logarithm of the inverse to its probability, and therefore average number of bits per symbol is equal to the entropy of the symbol ensemble. An example of Huffman-coding is illustrated in the table for 8 symbols (A to H).

As one can see, when one of the symbols of the ensemble has probability that is much higher than ½ while common probability of others is much less than ½, Huffman coding is inefficient because it does not allow to allocate to symbols less than one bit. In such cases, coding of rare symbols is applied.

Table An example of Huffman coding of 8 symbols

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Iteration	***P***(A)= 0.49	***P***(B)= 0.27	***P***(C)= 0.12	***P***(D)= 0.065	***P***(E)= 0.031	***P***(F)= 0.014	***P***(G)= 0.006	***P***(H)= 0.004
1-st	-	-	-	-	-	-	0	1
							P(GH)=0.01	
2-nd	-	-	-	-	-	0	1	
						P(FGH)=0.024		
3-d	-	-	-	-	0	1		
					P(EFGH)=0.055			
4-th	-	-	-	0	1			
				P(DEFGH)=0.12				
5-th	-	-	0	1				
			P(CDEFGH)=0.24					
6-th	-	0	1					
		P(BCDEFGH)=0.51						
7-th	0	1						
Binary code	**0**	**10**	**110**	**1110**	**11110**	**111110**	**1111110**	**1111111**
Entropy ***H*****=1.9554**								
Average number of bits per symbol: **1.959**								

There are two most popular varieties of methods for coding rare symbols: ***run length coding*** and ***coding co-ordinates of rare symbols***. In run length coding, it is assumed that symbols to be coded form a sequence in which runs of the most frequent symbol occur. The encoder computes length of the runs and generates a new sequence of symbols in which the runs of the most frequent initial symbol are replaced with new auxiliary symbols that designate the run lengths. This new sequence can then be subjected to VL- coding if necessary. Run length coding is an essentially 1-D procedure. For coding 2-D data, it is applied after 2-D data are converted into 1-D data by means of a zigzag scanning. Principle of zigzag scanning is illustrated below in Fig. 6-9.

Coding coordinates of rare symbols is an alternative to the run length coding. In this method, positions of rare symbols (others than the most frequent one) are found. Then a new sequence of symbols is generated from the initial one in which all occurrences of the most frequent symbol are removed and auxiliary symbols are added to each occurrence of rare symbols that designate their position in the initial sequence. Coding coordinates of rare symbols is less sensitive to errors in transmitting the binary code than run length coding. For the former, transmission errors cause only localized errors in the decoded symbol sequence while in the latter they result in shifts of entire decoded symbol sequence after the erroneous one. This property is of especial importance in image coding for transmission where transmission channel errors, in case of run length coding, may cause substantial deformation of object boundaries.

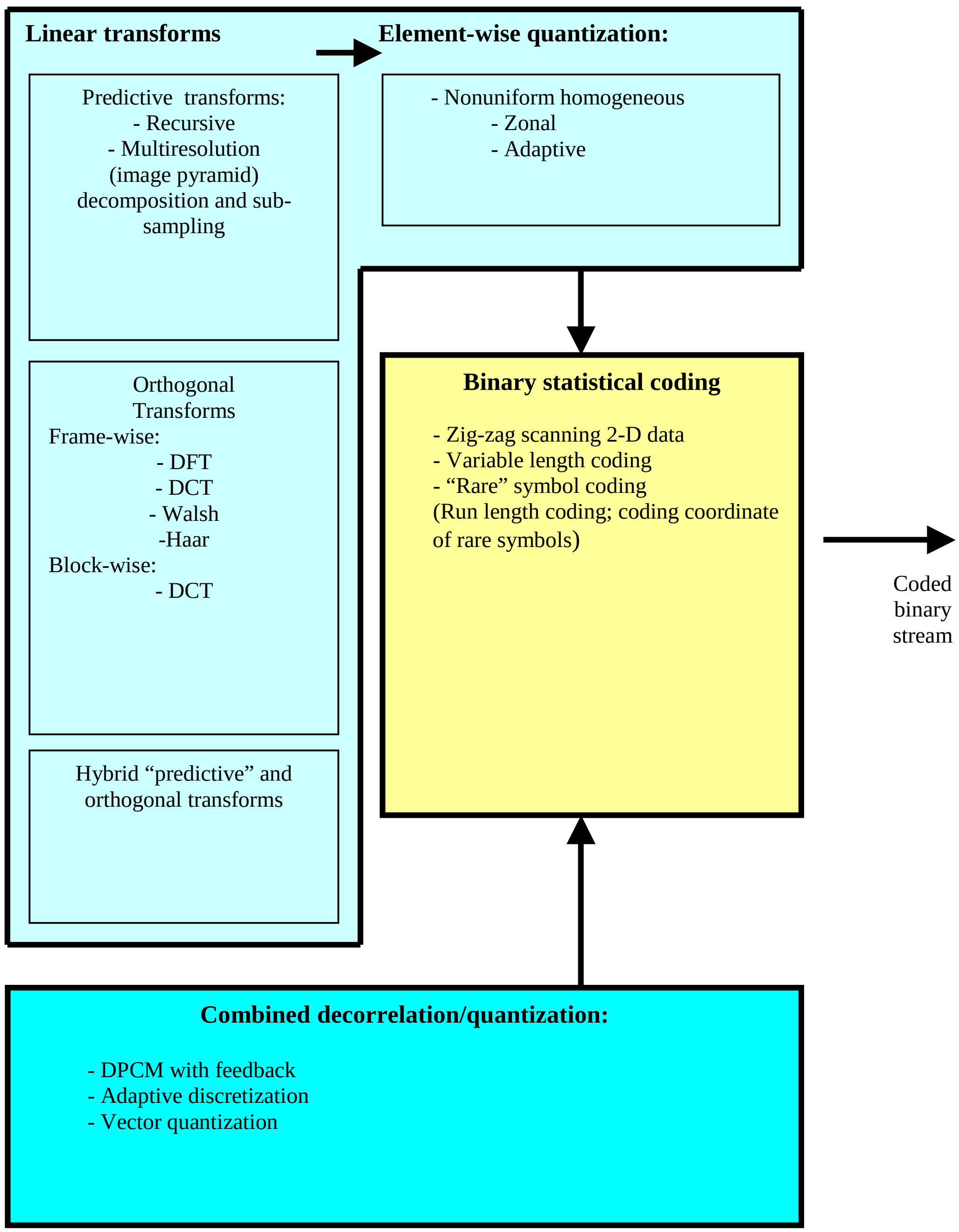
Exercices and questions for selftesting

1. What is the purpose of image coding?
2. How can one characterize end users in image coding?
3. What are “loss-less” and “lossy” coding?
4. Describe the principle of DPCM coding.
5. For one-step prediction:

$$\boldsymbol{\varepsilon_k} = \boldsymbol{a_k} - \boldsymbol{h}_1 \boldsymbol{a_{k-1}},$$

derive a formula for optimal prediction weight coefficients.

6. Observe, in a Matlab experiment, reduction of the prediction error dynamic range
7. Test, in a Matlab experiment with DPCM image coding, possible image compression and the trade-off between compression rate and reconstruction image quality. Discuss the selection of the dynamic range and using P-th low quantization of the prediction error
8. Explain principles of image multi-resolution decomposition and coding
9. Describe principles of block tranform coding.
10. What is zig-zag encoding?
11. Test, in a Matlab experiment with DCT block image coding, possible image compression and the trade-off between compression rate and reconstruction image quality. Explain quantization table. Observe blocking effects for low data rate.



igure 6.1. Classification of digital data compression methods

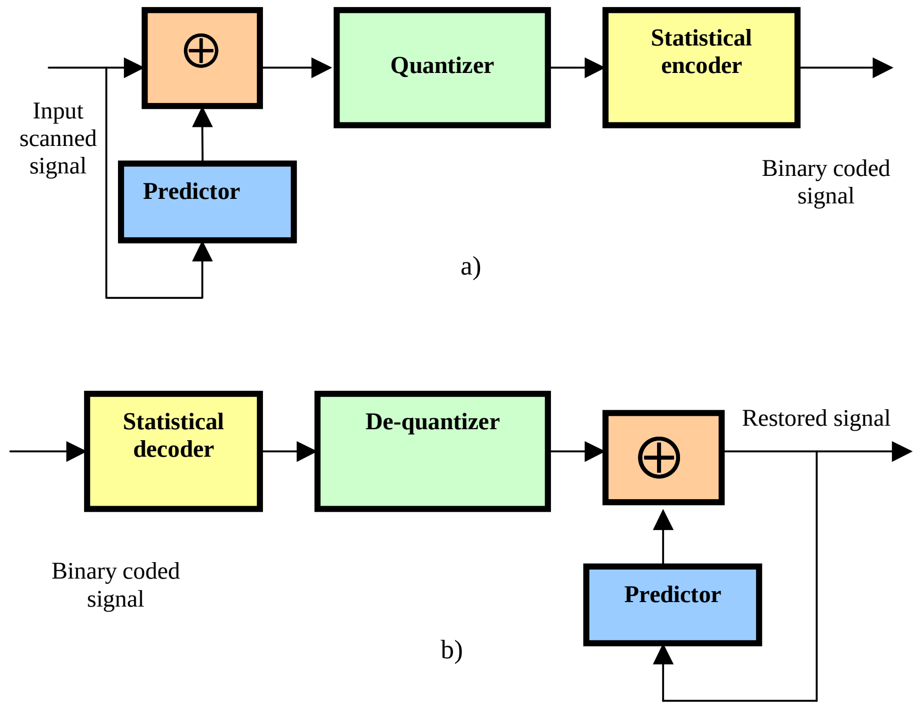


Figure 6-2. Block digram of DPCM coding (a) and decoding (b)

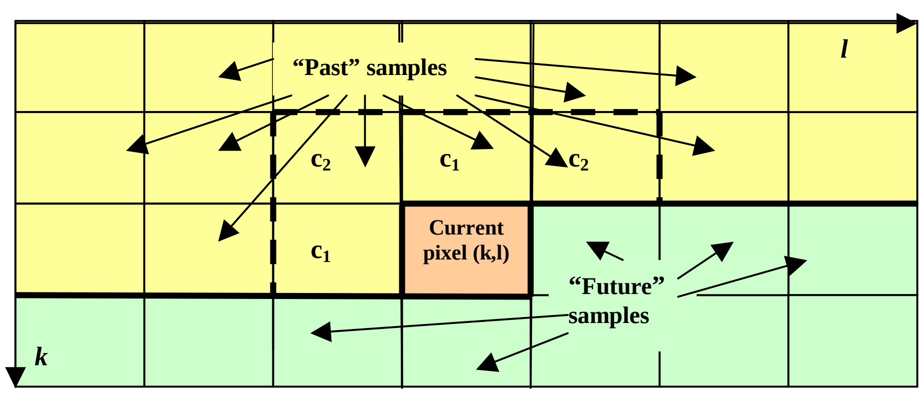


Figure 6-3. 2-D prediction for row-column image scanning method

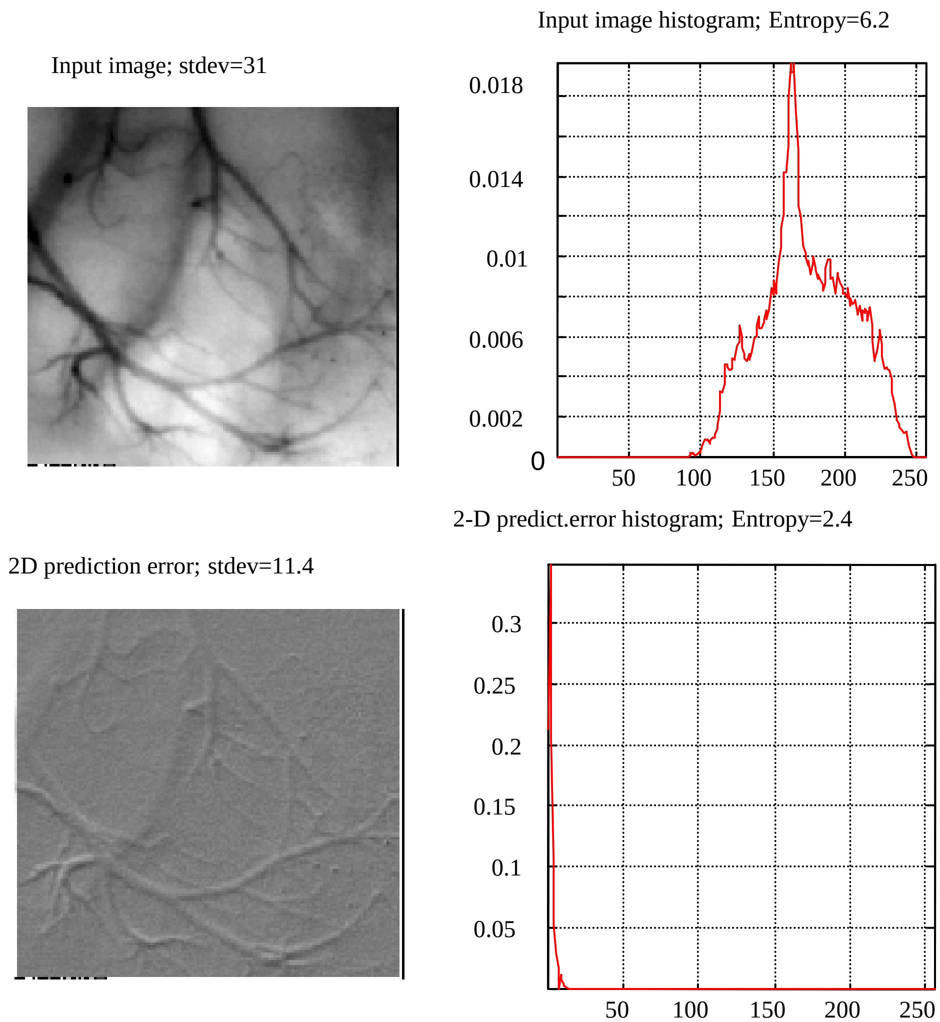


Figure 6-4. Dynamic range, standard deviation and entropy reduction by predictive decorrelation in DPCM

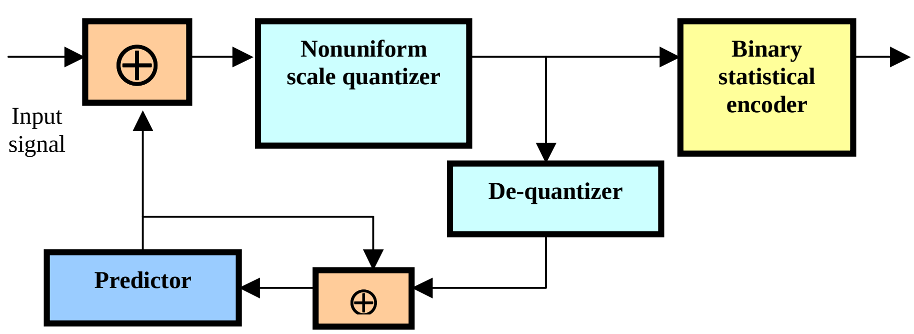
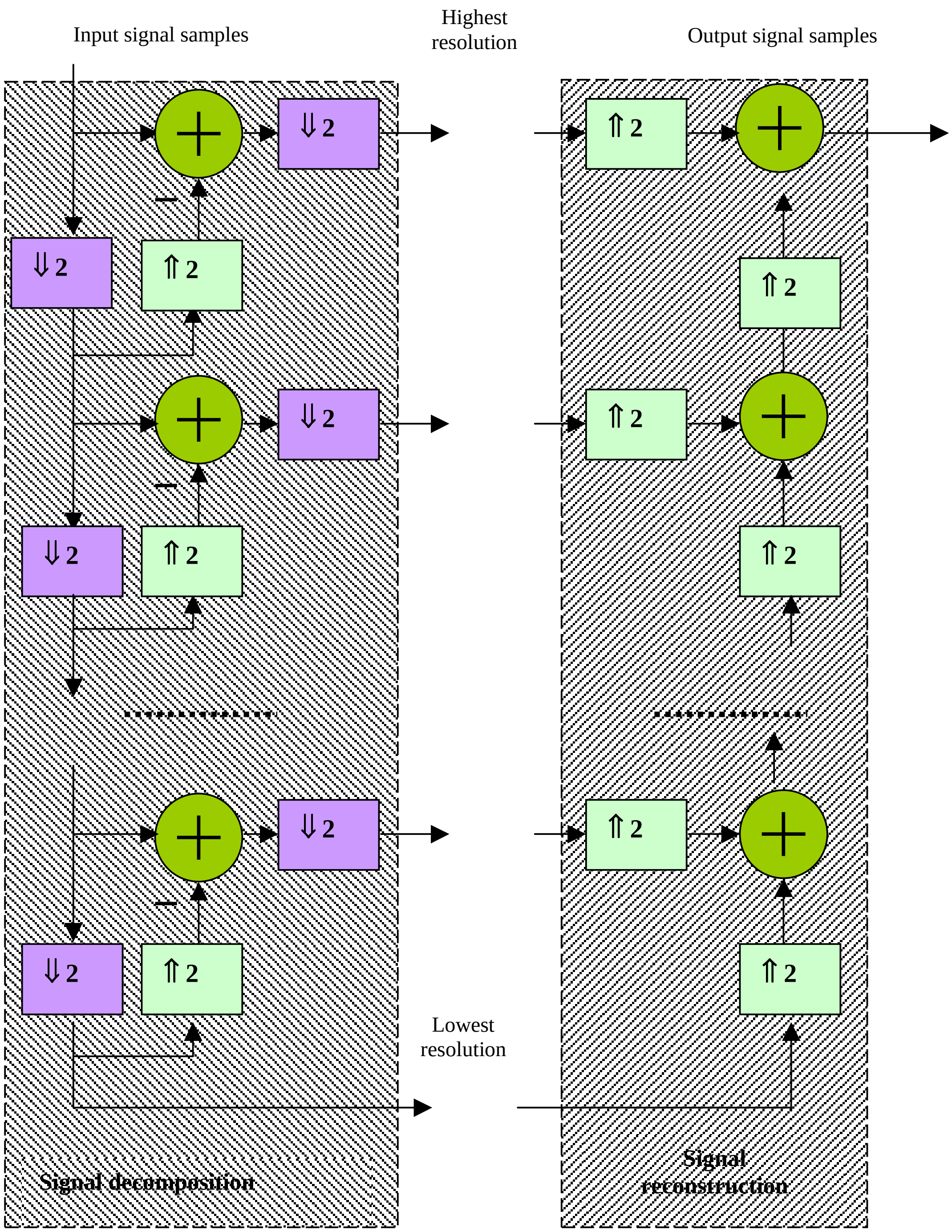


Figure 6-5. Block diagram of DPCM with feedback.



⇓2	Unit that performs signal low pass filtering and 2-fold decimation of signal samples
⇑2	Unit that performs signal 2-fold zooming and interpolation

Fig. 6.6. The principle of signal coding by sub-band decomposition

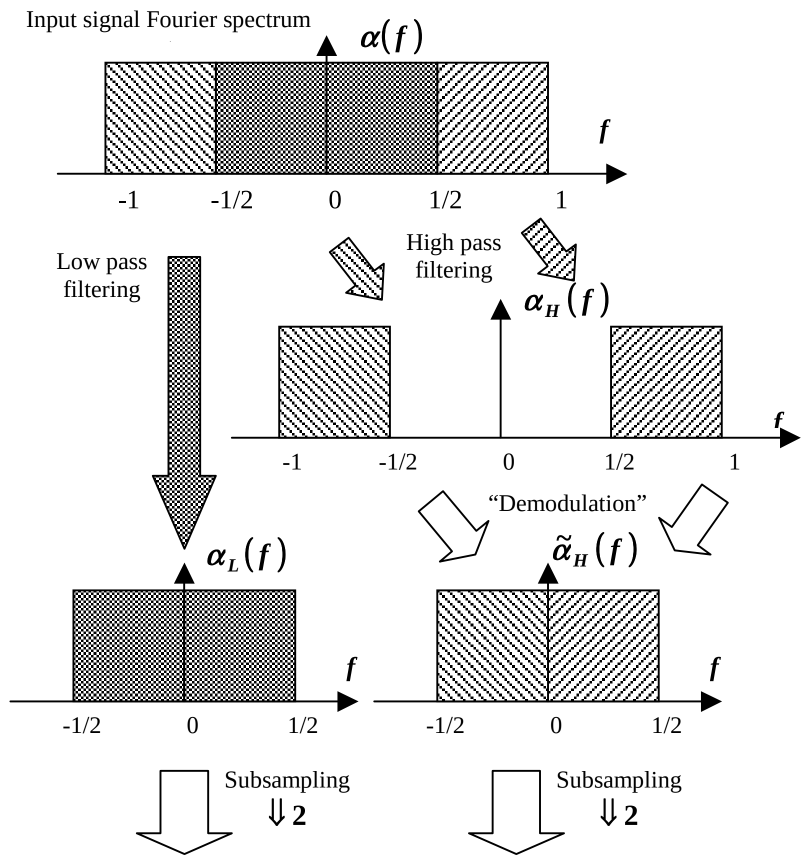


Fig. 6-7. Multi-resolution sampling explained for 10-D signal in Rourier domain

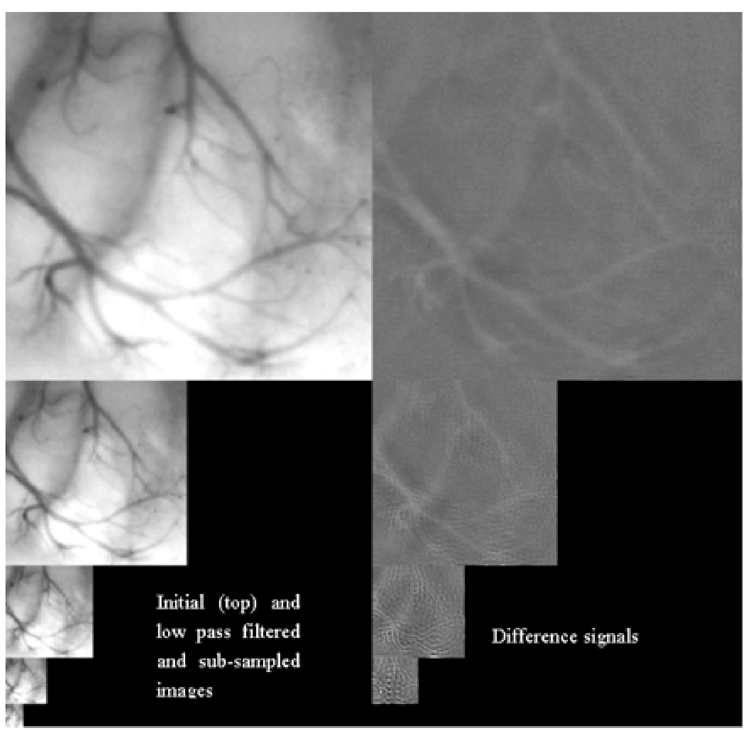


Figure 6-8. Image pyramid

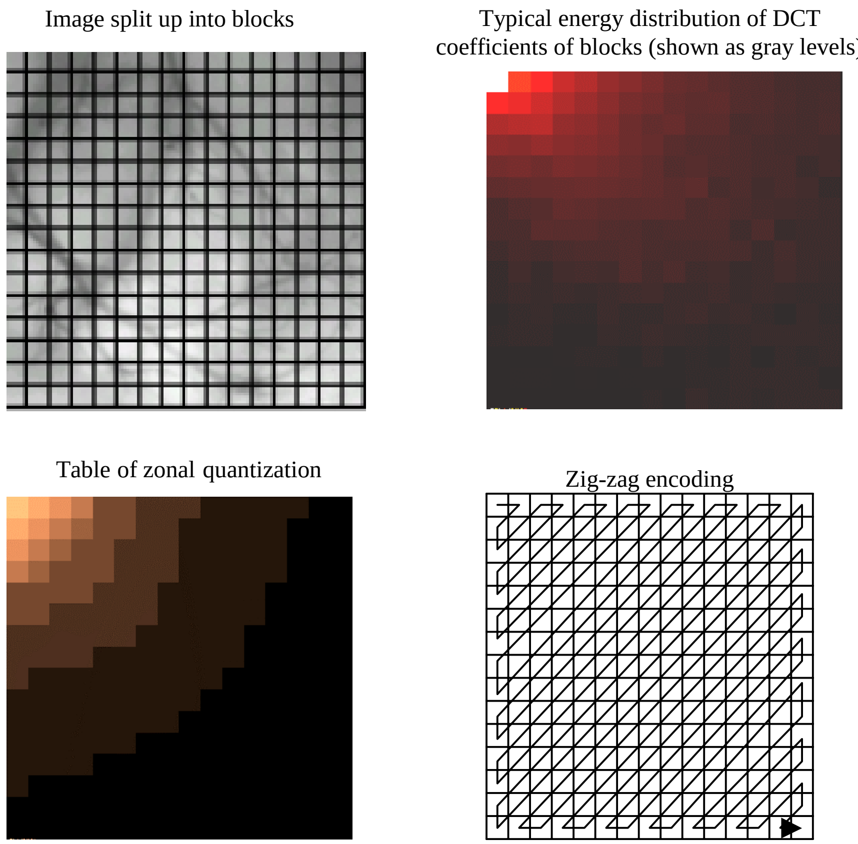


Figure 6-9 . Principle of DCT block transform coding