ROI-BASED IMAGE CODING USING MULTIRESOLUTION NEURAL NETWORKS

Vassilios Alexopoulos and Stefanos Kollias
Division of Computer Science
Department of Electrical and Computer Engineering
National Technical University of Athens
Heroon Polytechniou 9, 15773 Zographou, Greece
Tel: +30 1 772 2491; fax: +30 1 772 2459
e-mail: valex@image.ece.ntua.gr, stefanos@cs.ntua.gr

ABSTRACT

In this paper is presented a ROI-based multiresolution coding scheme, whose main importance is that it achieves both high compression ratios and good reconstruction of the images. It uses optimal, in the mean square error sense, analysis and synthesis filters in the most significant areas (Regions of Interest) while conventional ones are used in the rest of the image. A linear autoassociative neural network architecture is proposed to compute the filters for optimal reconstruction of the images based on low resolution approximations of these. The characteristics of optimal filters are examined in 'head and shoulder' videoconferencing images.

1 INTRODUCTION

The subband approach to image and video coding has become very popular recently. An effective subband coder should remove the redundancy and incorporates statistical as well as perceptual criteria into the coding procedure. Subband coding consists of splitting an image into bands and using different techniques to code these. After transmission through one or more channels and by band interpolation the original image is properly reconstructed.

The above mentioned coding strategy, however, compute the perfect reconstruction filterbanks according to metrics and statistics extraxted globally over the whole image. In this way any a-priory knowledge about the image content and spatial variations in the image is disregarded. In this paper we propose a scheme that determines *Regions of Interest* (ROIs) in the images prior to coding which achieves variable spatial reconstruction throughout the image, by differentiating the filterbanks used for coding the ROI and non-ROI image areas.

Section 2 describes the use of ROI for further reduction of the size of the input data and section 3 defines multiresolution image analysis techniques. An efficient technique for obtaining low resolution images in an optimal way is presented in section 4. Linear autoassociative neural networks are used for optimal computation of filterbanks when performing optimal multiresolution analysis of ROIs. Such networks have been recently presented in the framework of principal component analysis [1]. Section 5 presents simulation studies which illustrate the performance of the above procedure.

2 ROI-BASED IMAGE CODING

Applications related to image data storage or transmission through low bandwidth telecommunication channels, involve a trade-off between image quality and compession ratio or coding complexity. Most coding techniques reach decisions according to metrics and statistics extracted from the whole image in order to globally minimise the reconstruction error. In contrast with the usual broadcast transmission (TV), where all the regions of image frames to be processed are considered of equal importance, definition of (ROI) in the image can be proved very useful for specific applications. Such significant applications ranging from specific image transmission applications (teleconference and telemedicine systems) to multimedia database systems.

An efficient ROI-based compression scheme has the advantage of the differentiation on coding of specific regions in the image, considering their importance on the human visual system. ROIs generally correspond to areas or foreground objects which are of extreme interest for recognition or coding/reconstruction purposes and should be coded with maximum precision; a tolerably lossy reconstruction can be used in the rest image areas which are of minimal contribution to the perceived information.

A region of interest, or mask, is a binary image the same size as the image we want to filter. Selection of ROIs can be performed, either through user interaction (by defining a polygon that encloses the desired area), or automatically when considering specific applications. A neural network is presented below for performing the ROI selection by an hierarchical block-oriented two level architecture [2].

In particular, for selecting the ROIs the images are first separated in blocks of 8×8 pixels and these blocks are then DCT-transformed. The first level of the architecture automatically selects all edges appearing in the examined images and classifies the corresponding blocks

to the ROI category. This edge selection is due to the fact that, in most applications, the edges existing in the image belong to regions that are of major importance for recognition or classification purposes. This level consists of a feedforward network which performs the frequency dependent edge detection task, classifying each image block to a 'shade' or 'edge' category. 'Shade' blocks correspond to homogeneous areas, while 'edge' blocks generally include significant high frequency content. To accomplish this task, the network accepts at its input the computed DCT coefficients of each image block.

Supervised learning has been adopted for training the network to perform edge detection. According to it, a predefined training set of characteristic images is selected to which conventional spatial edge detection operators, such as Sobel or gradient ones, or more advanced morphological operators are applied. Following this selection the images are divided in blocks which are DCT transformed and labelled as 'shade' or 'edge' ones; then are used to train the network. After training, the network is able to classify each block of images, that are similar to the ones used for training, to an 'edge' or 'shade' category. If, however, the block is found to belong to an homogeneous region, no decision is taken, but the block is subsequently fed as input to the second level of the proposed architecture, which consists of another network that finally classifies it to a ROI or not.

The second network also uses the computed DCT coefficients of the block as input features. The number, however, of these features is generally smaller than the corresponding number of the first network input units. In particular, the input features that are chosen to feed the second network are the DC coefficient and a small number of AC coefficients following the well-known zig zag DCT scanning of each image component. In cases where edges do not play an important role, e.g., in texture-like images, it is possible to overpass the first level of the hierarchical architecture, focusing on the results of the second level. It can be verified [2] that during MC-DCT (Motion Compensated-DCT) coding with ROIs the PSNR (Peak Signal to Noise Ratio) is improved.

Whenever a block is classified as one of high importance, i.e. belongs to a ROI, it can be further treated, with maximum accuracy, as the main information/feature to be used for classification or recognition. In the other case, the block may be even disregarded in the following recognition procedure. In case of multiresolution image decomposition, a ROI-based scheme can be implemented, by differentiating the filterbanks used for coding the ROI and non-ROI image areas. In particular optimal filters are used in ROI areas while conventional ones are used in the rest of the image.

3 MULTIRESOLUTION IMAGE ANALYSIS

The basic concept of signal decomposition is to divide the signal spectrum into its subspectra and, then, to treat those subspectra individually. The decomposition of the signal spectrum into subbands provides the monitoring of signal energy components and the processing of the subbands independently. Representation of signals at many resolution levels has gained much popularity especially with the introduction of the discrete wavelet transform, implemented by filter banks using quadrature mirror filters (QMFs) [3]. In image processing the above are equivalent to subband processing. The basic idea of subband coding is to split the Fourier spectrum of an image into nonoverlapping bands, and then inverse transform each subband to obtain a set of bandpass images which can perfectly reconstruct the original image. Multiresolution decompositions result in approximation images which are low resolution replicas of the images and in a set of detail images which contain more detailed information as resolution is gradually increasing.

Let x_0 denote an $N \times N$ image representation. Using appropriate FIR perfect reconstruction filters $h_L(n)$ and $h_H(n)$, where $h_L(n)$ generally is a low-pass and $h_H(n)$ a high-pass filter, we can split the image into four lower resolution images of about $\frac{N}{2} \times \frac{N}{2}$ size. It is possible to use non-separable analysis (and synthesis) filters to perform the multiresolution decomposition. Applying, for example, the low pass filter $h_{LL}(m,n)$ we get the approximation image at the lower resolution level $j = -1$, denoted as x_{-1}^{LL}, where

$$x_{-1}^{LL}(m,n) = \sum_{k=1}^{N}\sum_{l=1}^{N} h_{LL}(2m-k, 2n-l)x_0(k,l) \quad (1)$$

By applying all other possible combinations of the above FIR filters, we get three lower resolution detail images, denoted as x_{-1}^{LH}, x_{-1}^{HL}, x_{-1}^{HH}. Perfect reconstruction of the original image $x_0(k,l)$ can be achieved through synthesis of all four subband components. By using only the approximation image $x_{-1}^{LL}(m,n)$ and synthesis filter $f_{LL}(m,n)$ it is possible to obtain an approximate reconstruction $x_0(k,l)$ of the original image as follows

$$x_0(m,n) = \sum_{k=1}^{\frac{N}{2}}\sum_{l=1}^{\frac{N}{2}} f_{LL}(m-2k, n-2l)x_{-1}^{LL}(k,l) \quad (2)$$

Optimal design of the analysis and synthesis h and f and filters in specific applications is examined next.

4 OPTIMAL FILTER DESIGN USING NEURAL NETWORKS

The design of perfect reconstruction filter banks is based on the assumption that all the subband signals are available to the interpolation bank with infinite precision. This is not, however, true, when only a part of subband components, and particularly only one of them, is

used for reconstruction; in this case which is frequently met when only the approximation, e.g., image is used for compression purposes, perfect reconstruction filters lose their optimality. Design techniques for analysis and synthesis filters that perform optimal reconstruction of an original image from a low-resolution representation of it have been recently proposed in [4]. Based on the minimization of the mean squared error between the original signal and the low-resolution representation of it, the filters are optimally adjusted to the statistics of the input images, so that most of the signal's energy is concentrated in the low resolution subband component.

Another case in which perfect reconstruction filters lose their optimality is the presence of additive, or quantization noise. In [5] appropriate formulae for the optimal reconstruction filters are desired using the mean-squared error as minimized criterion.

Let us concentrate next on the problem of generating four subband components from each image only one of which is retained, as the low resolution representation. Let the M-dimensional vector $\mathbf{x}(m,n)$ denote the vectorized $P \times P$ blocks of the input image $x_0(m,n)$, with $M = P^2$, the Q-dimensional vector $\mathbf{y}(m,n)$ denote the corresponding $L \times L$ blocks of the low-resolution representation $x_{-1}(m,n)$ also in vectorized form with $Q = L^2$ and finally the M-dimensional vector $\hat{\mathbf{x}}(m,n)$ represent the reconstructed vectorized image blocks.

The above vector notations are adopted, so that it be possible to denote the whole convolutional analysis and synthesis operations as multiplications of the above defined vectors by appropriate matrices, say H and F respectively. In particular Eqs.(1) and (2) can be written as

$$\mathbf{y}(m,n) = H\mathbf{x}(m,n) \tag{3}$$

$$\hat{\mathbf{x}}(m,n) = F\mathbf{y}(m,n) \tag{4}$$

Straightforward but tedious calculating provides analytical expressions of the $(Q \times M)$ and $(M \times Q)$ H and F matrices in terms of the, say $(J \times J)$, optimal filters h and f respectively. If for example, $M = 8$ and $J = 4$, then matrix H has the following structure

$$H = \begin{bmatrix} H_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ H_1 & H_2 & H_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & H_0 & H_1 & H_2 & H_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & H_0 & H_1 & H_2 & H_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & H_0 & H_1 & H_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & H_0 \end{bmatrix} \tag{5}$$

where H_0 is a submatrix of the form

$$H_0 = \begin{bmatrix} h_{30} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ h_{10} & h_{20} & h_{30} & 0 & 0 & 0 & 0 & 0 \\ 0 & h_{00} & h_{10} & h_{20} & h_{30} & 0 & 0 & 0 \\ 0 & 0 & 0 & h_{00} & h_{10} & h_{20} & h_{30} & 0 \\ 0 & 0 & 0 & 0 & 0 & h_{00} & h_{10} & h_{20} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & h_{00} \end{bmatrix} \tag{6}$$

where $h_{00}, h_{10}, h_{20}, h_{30}$ form the first column of filter $h(k,l)$. Matrices H_1, H_2 and H_3 are formed in exactly the same way using the corresponding columns of filter h, while extension to other values of M and J can easily be obtained. Matrix F is similarly formed in terms of corresponding matrices F_0, F_1, F_2 and F_3.

Based on Eqs. (3) and (4), we propose next to use a feedforward neural network to compute the optimal $J \times J$ analysis and synthesis filters, h and f respectively, through minimization of the mean squared difference between the original and reconstructed images. The network contains one hidden layer and linear hidden and output units. In particular the network accepts at its input the M-dimensional input image vector $\mathbf{x}$, uses Q hidden units and is trained to produce a reconstructed vector, at its M output units, that is equal to the input vector. As a consequence, the network operates in autoassociative form and during training is provided with the same input and desired output image blocks, in which the particular image, or a sequence of images, has been separated into; a backpropagation variant (see, for example, [6]) with a linear activation function can be the training algorithm. It is desired that the interconnection weights between the hidden units and the network inputs form a matrix W_{IH} equal to matrix H defined above in terms of the optimal filter h, while the interconnection weights between the output and hidden units form a matrix W_{HO} equal to the corresponding matrix F, so that the network implements the operations described above. In the following, we impose appropriate constrains in the proposed network architecture, so that it is able to solve the filterbank design problem.

Based on the fact [4] that the optimal synthesis filter is related to the analysis one through

$$\mathbf{F}(\omega_1, \omega_2) = \mathbf{H}(\omega_1, \omega_2)^{T*} \tag{7}$$

in the frequency domain, or equivalently in the spatial domain

$$f(m,n) = h(-m,-n) \tag{8}$$

the following constraint on the network structure is easily verified

$$W_{HO} = W_{IH}^T \tag{9}$$

Moreover, in order to force matrices W_{IH} and W_{HO} obtain the required forms (as, for example, the ones given in Eqs. (5) and (6) for the analysis matrix filter H), the weights corresponding to zero entries in the matrices are fixed to zero during training. Furthermore, when a specific weight of matrix W_{IH} (similarly for W_{HO}) is updated, its value is copied to all other weights that correspond to the same sample value of the optimal analysis filter $h(m,n)$, as determined by Eqs.(5) and (6); this procedure is the same as the one used for training time-delay networks, where the need for copying the updated weight values to groups of weights with identical values also arises.

5 SIMULATION RESULTS

The ability of neural networks to compute optimal filterbanks on specific regions of interest is exploited in this paper using various videoconference scenes ('Miss America' and 'Claire'). Optimal filters are computed first on the 'head and shoulder' part of the images and then on specific ROIs of these images, namely the area of speaker eyes and the area of mouth. The results explore the spectral characteristics of these filters comparing them to each other. A classification map shown in Figure 1 and figure 2 shows an original frame from the 'Claire' sequence, the reconstructed image with conventional filter banks and the corresponding reconstructed image with optimal filters.

We computed optimal resolution reduction filters using ROI blocks from the difference images, such as the ones shown in Figure 3. The frame difference images contain high frequency information for which conventional filters are not optimal. Such images are useful for recognition purposes as well.

Figure 1: Classification map

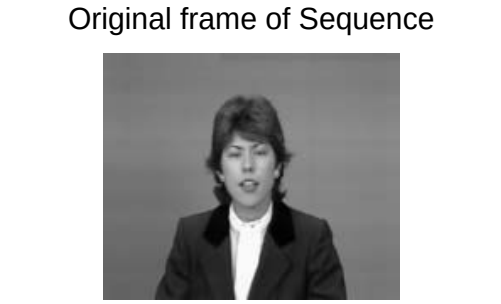

Figure 2:

Figure 3: Difference Image

6 CONCLUSIONS

Linear autoassociative networks with constraints has been used for optimally selecting filterbanks and implementing multiresolution image analysis. A feedforward neural network architecture has been also used for adaptively selecting regions of interest. The implementation of optimal filtering only in ROIs results in effective coding with reduction of the amount of information. Combination of these techniques for classification and recognition purposes is possible as well.

References

[1] V. Alexopoulos and S. Kollias, 'Optimal Multiresolution Analysis Using Principal Component Neural Networks', *Proc. of ECCTD '95*, pp. 323-326, 1995.

[2] D. Kalogeras and S. Kollias, 'Low Bit Rate Coding of Image Sequences Using Neural Networks', *ICNN '95*, Perth, Australia, 1995.

[3] S. Mallat, 'A Theory for Multiresolution Signal Decomposition: The Wavelets Representation', *IEEE Trans. on PAMI*, vol. 11, pp. 674-693, 1989.

[4] A. Tirakis, A. Delopoulos and S. Kollias, '2-D Filter Bank For Optimal Reconstruction Using Limited Subband Information', *IEEE Trans. on Image Proc.*, August 1995.

[5] A. Delopoulos and S. Kollias, 'Optimal Filterbanks for Signal Reconstruction from Noisy Subband Components', *IEEE Trans. on Signal Proc.*, Feb. 1996.

[6] S. Kollias and D. Anastasiou, 'An Adaptive Least Squares Algorithm for Efficient Training of Artificial Neural Networks', *IEEE Trans on CAS*, vol. 36, pp. 1092-1101, 1989.