Available online at www.sciencedirect.com
SCIENCE DIRECT
Engineering Applications of Artificial Intelligence 18 (2005) 383–392
Engineering Applications of
ARTIFICIAL INTELLIGENCE
www.elsevier.com/locate/engappai

Medical image compression using topology-preserving neural networks

Anke Meyer-Bäse[a], Karsten Jancke[a], Axel Wismüller[b], Simon Foo[a,*], Thomas Martinetz[c]

[a]Department of Electrical Engineering, Florida State University, Tallahassee, FL 32310-6046, USA
[b]Department of Diagnostic Radiology, University of Munich, 80336 Munich, Germany
[c]Institute for Neuro- and Bioinformatics, University of Luebeck, Luebeck 23569, Germany

Received 24 October 2003; received in revised form 12 October 2004; accepted 16 October 2004
Available online 24 November 2004

Abstract

A novel method based on topology-preserving neural networks is used to implement vector quantization for medical image compression. The described method is an innovative image compression procedure, which differentiates itself from known systems in several ways. It can be applied to larger image blocks and represents better probability distribution estimation methods. A transformation-based operation is applied as part of the encoder on the block-decomposed image. The quantization process is performed by a "neural-gas" network which applied to vector quantization converges quickly to low distortion errors and reaches a distortion error lower than that resulting from Kohonen's feature map or the LBG algorithm. To study the efficiency of our algorithm, we blended mathematical phantom features into clinically proved cancer free mammograms. The influence of the neural compression method on the phantom features and the mammo-graphic image is not visually perceptible up to a high compression rate.

Keywords: Medical image compression; Topology preserving neural networks

1. Introduction

The importance of coding and compression in transmission and storage is well established. The goal of image compression is to reduce the data amount and to obtain a low bit rate digital representation without affecting the perceptual image quality.

A compression system typically consists of a signal decomposition such as Fourier or wavelet, a quantization operation on the coefficients, and finally lossless or entropy coding such as Huffman or arithmetic coding. Decompression reverses the above process; although if quantization is used, the system will be lossy in the sense that the image will not be perfectly reconstructible from the digital representation. Quantization is only approximately reversible.

Transformed-based techniques have been proposed for the efficient reduction of the high redundancy usually encountered in real life images (Netravali and Haskell, 1988; Tzovaras and Strintzis, 1998; Strintzis, 1998). Unsupervised neural networks can perform nonlinear principal component analysis as a transform-based method in image compression (Oja et al., 1991; Tzovaras and Strintzis, 1998). They outperform linear principal component analysis, and are relatively easy to implement.

Another common method to compress images is to code them through vector quantization (VQ) techniques (Gray, 1984). Self-organized Kohonen maps have been used to achieve the VQ process of image compression (Amerijckx et al., 1998). They represent an efficient compression scheme based on the fact that consecutive

*Corresponding author. Tel.: +1 850 410 6474; fax: +1 850 410 6479.
E-mail address: foo@eng.fsu.edu (S. Foo).

0952-1976/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2004.10.004

blocks in an image are often similar, and thus coded by similar codewords with a VQ algorithm.

Early studies of lossy compressed medical images performed compression using variations on the standard discrete cosine transform (DCT) coding algorithm combined with scalar quantization and lossless coding.

More recent studies of efficient lossy image compression algorithms have used subband or wavelet decompositions combined with scalar or vector quantization (Woods, 1991; Said and Pearlman, 1996; Pal et al., 2000). These algorithms provide several potential advantages over traditional Fourier-type decompositions, including better concentration of energy, better decorrelation for a wider class of signals. The wavelet-based approach, however, has a filter whose length varies as a function of resolution. As a result, the errors they introduce are not well-localized and can appear as ringing distortion. In Betts et al. (1998) and Perlmutter et al. (1997) was used a compression algorithm (Shapiro's embedded zerotree algorithm (Shapiro, 1993)) of the subband/pyramid/wavelet coding class. New forms of multiresolution vector quantization algorithms were investigated in Cosman et al. (1996) and Li et al. (1996). The vector generalizations of the embedded zero-tree wavelet technique used in Cosman et al. (1996) incurred additional complexity and made this approach inferior to the scalar wavelet coding selected. The non-wavelet multiresolution technique used in Li et al. (1996) provided significant improvements at modest complexity cost for low resolution images, e.g., for medical images reduced in size for progressive viewing during the rendering of the full image.

The goal of this paper is to present and evaluate a new lossy neural-based compression method for medical images. The proposed cascaded technique combines a transformation-based neural network with a VQ-neural network, the "neural gas" network.

The described method is an innovative medical image compression procedure, which differentiates itself from known systems in several ways. Existing medical image compression methods are mostly vector generalizations of the embedded zero-tree wavelet technique (Shapiro, 1993; Said and Pearlman, 1996). They perform poorly in comparison to sophisticated pattern recognition techniques and involve only operations on small 2×2 pixel blocks. The proposed neural-based approach is extended to larger blocks and represents a better probability distribution estimation method.

2. The global compression scheme

Transform-based algorithms for lossy compression transform an image to concentrate signal energy into a few coefficients, quantize the coefficients, and then entropy code them after assigning special symbols for runs of zeros.

The linear principal component analysis realized by the Karhunen–Loeve transform (KLT) is the optimal linear transform for energy compaction into a few coefficients. A linear principal component analysis (PCA) can be easily implemented by a class of autoassociative neural networks (Haykin, 1994; Oja, 1982).

The proposed method combines a PCA-type neural network, based on Hebbian learning, with a proposed data compressor ("neural-gas" network) performing vector quantization. Applied to vector quantization, the "neural-gas" network converges quickly to low distortion errors and reaches a distortion error lower than that resulting from Kohonen's feature map or the LBG (Linde, Buzo, Gray) algorithm (Linde et al., 1980).

The PCA transform is applied to image blocks which are treated like data vectors and performs a projection of the data vector on a lower dimensional space.

A single layer neural network with L input neurons and M linear output neurons where $M \ll L$ is performing the PCA analysis based on three different algorithm. The output of the network $\mathbf{y}$ is defined by $\mathbf{y} = \mathbf{W}^{\mathbf{T}}\mathbf{x}$. $\mathbf{x}$ is the input vector and $\mathbf{W}$ is the weight matrix of the network.

Depending on the type of the activation function (linear or nonlinear) we implemented three different PCA algorithms known as Generalized Hebbian Algorithm (GHA) (Sanger, 1989), Oja's symmetric algorithm (OJA) (Oja, 1989) and nonlinear PCA (NLPCA) (Karhunen and Jourtensalo, 1994).

Oja's symmetric algorithm will provide the convergence of the weight matrix to a subspace spanned by the most significant eigenvectors of the input data covariance matrix. The outputs of the network will then be decorrelated and their variance maximized, so the information content of the L-dimensional vectors will be maximally transferred by the M-dimensional output vectors in the mean square sense.

The weight adaptation equations for the feedforward weight matrix $\mathbf{W}$ for OJA's and NLPCA algorithm are given below

$$\text{OJA}: \mathbf{W}(n+1) = \mathbf{W}(n) + \eta[\mathbf{I} - \mathbf{W}(n)\mathbf{W}^{\mathrm{T}}(n)]\mathbf{x}(n)\mathbf{x}^{\mathrm{T}}(n)\mathbf{W}(n), \tag{1}$$

$$\text{NLPCA}: \mathbf{W}(n+1) = \mathbf{W}(n) + \eta[\mathbf{x}(n)g(\mathbf{x}^{\mathrm{T}}(n)\mathbf{W}(n)) \tag{2}$$

$$-\mathbf{W}(n)g(\mathbf{W}^{\mathrm{T}}(n)\mathbf{x}(n))g(\mathbf{x}^{\mathrm{T}}(n)\mathbf{W}(n))], \tag{3}$$

where μ is the learning rate, $\mathbf{x}$ is the input vector and g is usually a monotone odd nonlinear function. Here the function $g(t)$ is applied separately to each component of its argument vector. The advantage of using