Objective assessment of MPEG-video quality: a neural-network approach

Paolo Gastaldo*, Stefano Rovetta#, and Rodolfo Zunino*

*DIBE - Dept. Biophysical and Electronic Engineering - University of Genoa
Via all'Opera Pia 11a - 16145 Genova - Italy
e-mail: {gastaldo, zunino}@dibe.unige.it

INFM, DISI - Dept. of Computer and Information Sciences - University of Genoa
Via Dodecaneso 35, 16146 Genova – Italy
e-mail: ste@disi.unige.it

Abstract

The increasing use of compression standards in broadcasting digital TV has raised the need for established criteria to measure perceived quality. This paper presents a methodology using Circular Back-Propagation (CBP) neural networks for the objective quality assessment of MPEG video streams. Objective features are continuously extracted from compressed video streams on a frame-by-frame basis; they feed the CBP network estimating the corresponding perceived quality. The resulting adaptive modeling of subjective perception supports a real-time system for monitoring displayed video quality.

1 Introduction

The recent increasing success of digital TV has stimulated the research for objective, automated methods to assess the user-end perception of broadcasting. Quality may bias a customer's choices of advanced pay on-demand services. In addition, the number of coders on the market will increase in the next years, hence both manufacturers and broadcasters will invariably face the problem of comparing the user-level quality of video.

The underlying technical problem is to estimate the effects of the visual artifacts brought about by digital encoding. Subjective assessment methods [1] attempt to evaluate the perceived video quality by asking human assessors to score the quality of a series of test scenes. These methods can yield accurate results. However, subjective testing is complex and does not allow real-time monitoring. Objective quality assessment aims to emulate human response to perceived quality by processing numerical quantities worked out from video streams. As a result, this technique no longer requires inputs from human operators.

A variety of methods for objective quality assessment of digital TV have been proposed in the literature [2-3]. Most approaches are based on decompressed video [4-7]: objective parameters are worked out by comparing pictures at the receiver end with original scenes. A method that does not involve the original video is described in [8].

From a scientific perspective, most of the above papers approached the problem of human perception of quality as a modeling one. As compared with those works, the present approach aims to produce a method to mimic such perception. This paper presents a method using the "Circular Back-Propagation" neural network [9] for automatic evaluation of subjective assessment. The network operates on compressed data only; this removes the need for any information about either the original video or the decoding process. From an engineering standpoint, the adaptive neural framework decouples the evaluation task from the specific video source and from decoder issues as well.

2 Feed-Forward CBP Networks

Feed-forward neural networks provide a straightforward paradigm to map feature vectors (describing video frames) into the corresponding quality assessments. Such a problem setting treats the quality scorings used for training as an ordered, discrete set of labels, whereas any intermediate values in the associate network output are allowed. In this sense, efficiency requirements as well as generalization issues ultimately lead to the problem of properly sizing the number of neurons in the NN.

MultiLayer Perceptrons (MLPs) can efficiently tackle problems in which the target-mapping function can be supported by few units with global scope; in MLPs, those

0-7803-7044-9/01/$10.00 ©2001 IEEE

elements are encoded by the sigmoid functions within hidden units. Conversely, if the target mapping can be best expressed as a superposition of locally-tuned components, radial-basis function (RBF) networks will typically perform much more efficiently. As a result, the unknown characteristics of the problem-related target mapping further complicate the problem of selecting the nature and the number of hidden units.

A solution to this specific problem has been proposed in [9]. The "Circular Back Propagation" (CBP) network extends the multilayer perceptron by including one additional input with its associated weight. Such an input just sums the squared values of all the other network inputs. As proved by CBP theory, the additional unit allows the overall network to adopt the standard, sigmoidal behaviour, or to drift smoothly to a bell-shaped radial function, which approximates - but is not - a Gaussian. At the same time, the limited increase in the network parameters does not affect its expected generalization performance, as it has been proved that the Vapnik-Chervonenkis dimension (VC-dim) [10] of the augmented, circular perceptron increases by one unit [9].

The CBP model adopted for this research can be formally described as follows. An MLP architecture combines two functional layers including n_h and n_o units, respectively. The conventional sigmoidal function is denoted by $\sigma(x) = (1+e^{-x})^{-1}$. The input layer connects the n_i input values to each unit of the hidden layer. The j-th "hidden" neuron performs the following transformations on the input values (j=1,..., n_h):

$$r_j = w_{j,0} + \sum_{i=1}^{n_i} w_{j,i} x_i + w_{j,n_i+1} \sum_{i=1}^{n_i} x_i^2 \qquad (1)$$

$$a_j = \sigma(r_j) \qquad (2)$$

The input features x_i (i=1 ... n_i) combine with the associated weights w_{ji} (i=0 ... n_i+1) and feed the j-th hidden unit. The terms r_j and a_j denote the neuron *stimulus* and *activation*, respectively. The last term in expression (1) actually augments the conventional MLP up to the CBP model.

The *output* layer provides the actual network responses, y_k, by the following transformations (k = 1,..., n_o):

$$r_k = w_{k,0} + \sum_{j=1}^{n_h} w_{k,i} a_j \qquad (3)$$

$$y_k = \sigma(r_k) \qquad (4)$$

Theory proves [9] that this model is the most efficient polynomial extension of MLPs with linear stimulus, and formally encompasses the RBF network model as well. The strict relationship of CBP to Vector-Quantization networks has been analyzed in [11], showing that the model ensures a notable representation effectiveness with a very small increase in the number of parameters.

The crucial feature that makes the CBP model suitable for the video quality-assessment task is its ability to switch autonomously between the different representation paradigms (MLP or RBF), as conventional back-propagation algorithms [12] can be adopted for weight adjustment. The resulting weight configuration ultimately sets the most suitable representation setting for the mapping problem, and is only driven by training data, independently of any a-priori assumption on the observed domain.

3 Neural-Network Based Assessment of Video Quality

The present work applies CBP networks to the automated quality evaluation of MPEG-2 [13] video streams. The resulting objective assessment system aims to estimate perceived quality by processing data extracted from video streams only. Hence, the system follows a "no-reference" paradigm.

Figure 1 shows a schematic representation of the system. Objective features are worked out directly from MPEG-2 bitstreams (i.e., without any decoding), and feed the neural network to obtain quality ratings. The system operates on a frame-by-frame basis and yields a continuous output; as such, it provides a real-time monitoring tool for displayed video quality. Thus, the neural network is entrusted to mimic the subjective, Single-Stimulus Continuous Quality Evaluation (SSCQE) method [14], recording continuous assessments of picture quality provided by human observers. The crucial advantage of the approach lies in generating quality ratings without decoding the video stream. Indeed, the objective metric supported by the neural system relies entirely on a representation format – the compressed bitstream – that bypasses the need for human assessors' rating process altogether. This greatly improves the method's effectiveness especially in terms of real-time performance, as one can get an estimate of perceived quality at transmission time.

For the reader's convenience, we recall that MPEG-2 attains still-image quality by standard DCT compression; motion information is treated by dividing each frame (picture) into several macroblocks (holding 16x16 pixels each), and by encoding the apparent movement of macroblocks within time-consecutive frames.

3.1 Features for Objective Quality Assessment

The objective metric set plays a crucial role for the effectiveness of the overall methodology. Since the present approach does not imply any a-priori assumption on the significance of the encoding parameters, a quite large set of

features is extracted from video streams (Appendix A lists the objective features worked out from the MPEG-2 compressed stream). The purpose is to collect as much information as possible.

In principle, one expects that a considerable number of all the above features will be discarded, either because they do not carry significant information or because they are mutually correlated. Thus, an a-posteriori statistical analysis drives the feature-selection criterion. The feature-selection algorithm is outlined in Fig. 2.

As a result of the above procedure, the set Z includes the features that, due to their asymmetrical distribution, are unlikely to stem from a Gaussian distribution. The purpose is to single out the statistically significant objective descriptors, under the (practically reasonable) assumption that non-informative quantities most often exhibit a Gaussian distribution.

3.2 Feature Run-Time Sampling

The mechanism generating the input features x_i must take into account known mechanisms specific for human perception (Fig. 3). Hence, feature values are not extracted from each sequence frame. In more detail, the following quantities are used to parameterize these mechanisms:

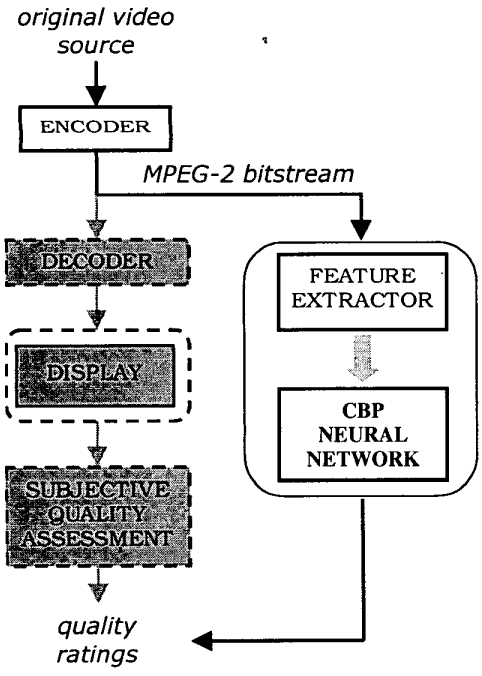

Figure 1: Objective assessment using CBP neural network.

– Δ ("assessor's response time" [15]): it refers to the

Feature-selection algorithm

Input:

Ψ : a library $\{\Psi_1, \ldots, \Psi_L\}$ of L test streams, composed of P frames each;

F_k: the set of objective features (k=1 … N_f);

$f_{k\Psi_i}^{(j)}$: the value measured by F_k for the j^{th} frame of the i–th stream Ψ_i.

0. (extract features from the library of test streams Ψ)
   For k=1 … N_f:

$$\Phi_k = \left\{ f_{k\Psi_1}^{(1)}, \ldots, f_{k\Psi_1}^{(P)}, \ldots, f_{k\Psi_L}^{(1)}, \ldots, f_{k\Psi_L}^{(P)} \right\}$$

1. (rescaling Φ_k)
   For k=1 … N_f:

   Compute the .05 and the .95 percentiles, $x_{0.05}^{(k)}$, $x_{0.95}^{(k)}$, respectively, for the values in Φ_k

   Build up a set $\underline{\Phi}_k$ by re-scaling each element of Φ_k into the range [-1,1]:

$$\underline{\Phi}_k = \left\{ \underline{f}_{kz} \right\} \qquad z = 1 \ldots P, \ldots, P(L-1)+1 \ldots P \cdot L;$$

   where:

$$\underline{f}_{kz} = \frac{\left(f_{k\Psi_q}^{z-Pq} - x_{0.05}^{(k)} \right)}{\left(x_{0.95}^{(k)} - x_{0.05}^{(k)} \right)}$$

2. (compute descriptive statistics)
   Create two sets (k=1 … N_f):

$$\Sigma = \{skew_k\} \text{ where } skew_k = \text{skewness}(\underline{\Phi}_k);$$

$$\mathrm{K} = \{kurt_k\} \text{ where } kurt_k = \text{kurtosis}(\underline{\Phi}_k);$$

3. (threshold settings)
   Compute $skew_{thr}$ and $kurt_{thr}$, as:

   $skew_{thr}$: 0.5 percentile of Σ;

   $kurt_{thr}$: 0.5 percentile of K.

4. (feature selection)
   For k=1 … N_f:

   Compile the feature set, Z, holding the objective features that satisfy:

$$F_k \in \mathrm{Z} \Leftrightarrow (skew_k > skew_{thr}) \text{ AND } (kurt_k > kurt_{thr})$$

Output:
Set Z

Figure 2: Feature-selection algorithm.

delay between the subjective judgment and the last frame that has influenced it;
- N ("recency effect" [3, 16]): frames that contribute to generating a single score;
- W ("masking phenomenon" [17]): time-consecutive frames tend to interfere with one another; thus, W groups of consecutive frames yield a single feature vector $\vec{x}$.

The input vector $\vec{x}$ includes n_i features $\hat{f}_{k\Psi_L}^{(j,W)}$ ($F_k \in Z$) defined as follows:

$$\hat{f}_{k\Psi_L}^{(j,W)} = \rho(f_{k\Psi_L}^{(j)} \cdots f_{k\Psi_L}^{(j+W)}) \qquad (5)$$

where $\rho(f_1 \ldots f_k)$ is a family of operators, with ρ_h, ρ_s, and ρ_m respectively the highest, the smallest and the mean values over the interval.

3.3 The Neural Network Approach

Several features characterizing video streams jointly affect subjective judgments; possibly non-linear relationships and partly unknown mechanisms may complicate the process modelling. These effects actually seem to have sometimes been underevaluated in the literature, and the major advantage of a neural-network approach lies in the ability to deal with multidimensional data representing complex relationships. By decoupling the feature-selection task from the design of an explicit mathematical model, one obtains the crucial advantage of avoiding a-priori assumptions on the significance of objective measures.

In the present approach, CBP networks map feature vectors into quality ratings. The mapping function is learned from examples by means of an iterative training algorithm, and a single output neuron in the NN yields the quality assessment for a given input vector.

The network configuration (i.e., the number of hidden units) has been designed by using a specific initialization technique that exploits the equivalence of the CBP model to Vector-Quantization paradigms [11]. In particular, a VQ preliminary phase using the "Plastic Neural Gas" algorithm [18] made it possible to assess the proper number of prototype vectors to represent the available sample distribution. In the subsequent network set-up phase, the number and the space positions of those prototypes were mapped directly into the specific CBP network configuration according to the formalism described in [11]. Thus the initial setting of the network weights proved most effective in accelerating the convergence of the overall training process, as compared with a conventional random setting.

The CBP network training uses an accelerated variant [19] of the classical back-propagation algorithm. The possibility of using conventional techniques to train an advanced network structure is the major advantage of the CBP model. The network cost function is expressed as:

$$e(\vec{w}) = \frac{1}{n_o n_p} \sum_{m=1}^{n_p} \sum_{k=1}^{n_o} \left(t_k^{(m)} - y_k^{(m)} \right) \qquad (6)$$

where n_p is the number of training patterns and t_k is the actual quality assessment derived experimentally from the human scoring panel.

4 Experimental Results

The CBP model for objective quality assessment was tested experimentally by using a library of MPEG-2 videos provided by the Research Center of the Italian Radio and Television Corporation (RAI). The testbed included twelve frame-coded sequences, each 70-sec long; the picture size

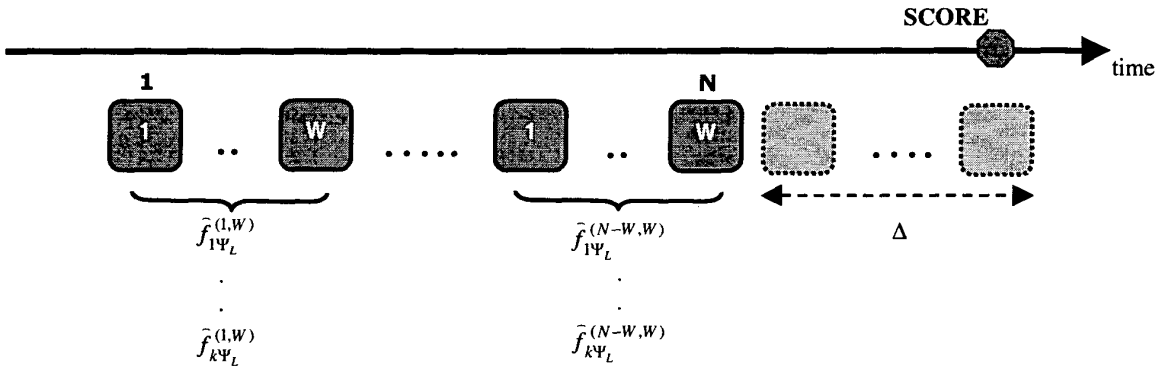

Figure 3: Feature run-time sampling process according to perceptual mechanism.

was 720x576 pixels. The sequence contents varied from fiction to sport and were encoded at different bit rates in the range [4, 8] Mbits/sec.

The assessments for each sequence were collected from non-expert viewers; the subjective tests were performed with an SSCQE technique at a sampling rate of two scores per second. The quality ratings were represented by a continuous scale ranging in [-1, 1].

4.1 Experimental Setup

The neural-network training process involved the set of features Z that the statistical analysis selected from the global feature set F_k listed in Appendix A. The dimensionality of the input data space was further reduced with the feature-selection technique described in [20]. The eventual 4-dimensional feature space covered the quantities: $\rho_s(Nn\_bits)$, $\rho_h(Xq\_scale(1))$, $\rho_h(Xmv(1))$, and $\rho_h(Smv\_dev\_std)$.

Training patterns were generated by the run-time sampling process presented in Section 3, with N=24, W=6 and Δ=17. The plastic VQ algorithm processed the training samples to design the neural network configuration; the resulting value n_h=15 set the number of hidden units in the feedforward structure.

4.2 Results

Figure 4 and Fig. 5 shows test results obtained for the selected feature set. Figure 4 compares the quality ratings by human assessors with the corresponding outputs of the neural network; for display clarity, the actual ratings are sorted in increasing order. The picture shows an asymmetric distribution of subjective scores; lower scores appear subsampled, thus they are subject to greater errors due to the lower statistical confidence. Nevertheless, the CPB neural network attained an average error μ_{err}=0.001 on the test set. Figure 5 plots the error distribution together with the related best-fitting Gaussian approximation ($\overline{\mu}$=0, $\overline{\sigma}^2$=0.066).

5 Conclusions

The present paper has presented an automated objective quality assessment method using CBP networks. The neural-network model is specifically tuned to learn the perceptual phenomenon from examples, and exploits a known effective augmentation of standard BP networks. A crucial advantage of the proposed methodology is the system ability to handle compressed video streams. Avoiding the need for decompressed pictures enhances the method's effectiveness in real-time production applications.

Experimental evidence confirmed the validity of the approach, as the system always provided a satisfactory, continuous-time approximation for the actual scoring curves related to test videos.

Appendix A

An MPEG-2 bitstream has a hierarchical structure that allows one to get information at multiple levels: sequence,

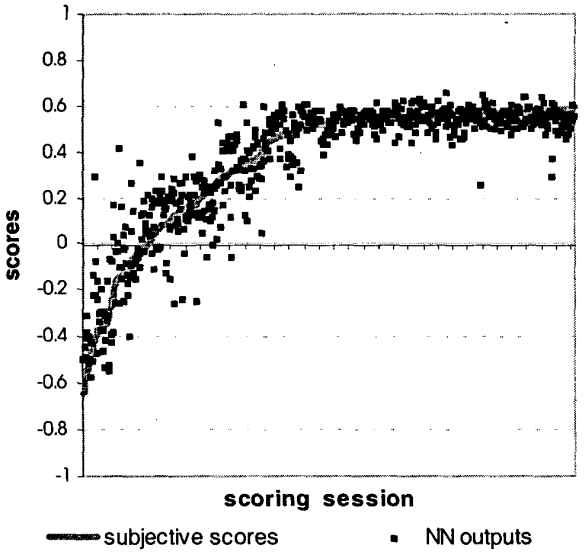

Figure 4: Neural Network outputs compared with human quality ratings.

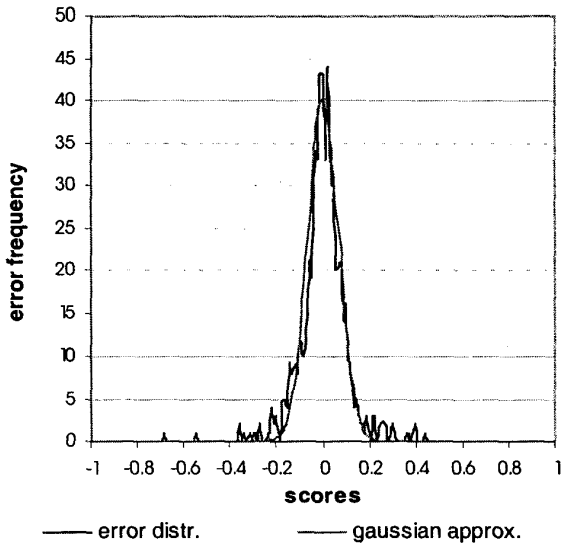

Figure 5: Error distribution.

group of pictures, picture, slice, macroblock and block. Objective features have been designed to characterize the stream at the picture level.

The following quantities are defined:

- $$energy = \frac{1}{256}\sum_{i=0}^{16}\sum_{j=0}^{16}(mb_{DCT}[i][j])^2 \quad (7)$$

  where $mb_{DCT}[i][j]$ are the DCT coefficients of a P or B macroblock. This quantity gives the energy of the correction to the predicted macroblock.

- $$q\_mv = \frac{q\_scale}{1+\langle|m\_v|\rangle} \quad (8)$$

  where *q_scale* is the quantiser-scale factor in a macroblock, and $\langle|m\_v|\rangle$ is the mean amplitude value of motion vectors in the same macroblock.

- $$e\_mv = energy \cdot \langle|m\_v|\rangle \quad (9)$$

  where *e_mv* is defined as the weighted energy of a macroblock.

The set F_k of objective features includes *Nn_ bits* (number of bits per picture) and the following three class of measurements:

1 - Percentages of macroblocks (MB) or blocks in a picture: *Pmb_no_pred* (MB with no motion vectors); *Pmb_fwd* (MB with forward motion vector only); *Pmb_back* (MB with backward motion vector only); *Pmb_bidir* (MB with both forward and backward motion vectors); *Pmb_I* (intra MB); *Pmb_sk* (skipped MB); *Pb_sk_lum* (skipped luminance blocks); *Pb_sk_chr* (skipped luminance blocks).
2 – Statistical figures of parameters extracted from a picture: *Smv_av* (mean of motion vectors); *Sq_scale_av* (mean of *q_scale* factors); *Senergy_av* (mean of energy); *Smv_dev_std* (standard deviation of motion vectors); *Sq_scale_dev_std* (standard deviation of *q_scale* factors); *Senergy_dev_std* (standard deviation of *energy*).
3 - Percentiles of parameters extracted from a picture, where α fixes the percentile: *Xmv(α)* (motion vector abs. value); *Xq_scale(α)* (*q_scale* factors); *Xenergy(α)* (*energy*); *Xq_mv(α)* (*q_mv*); *Xe_mv(α)* (*e_mv*).

References

[1] ITU-R BT.500, Methodology for the subjective assessment of the quality of television pictures.
[2] S. Olsson, M. Stroppiana and J. Baïna, "Objective methods for assessment of video quality: state of the art", *IEEE Trans. on Broadcasting*, vol. 43, no. 4, pp. 487-95, Dec. 1997.
[3] W. Y. Zou and P. J. Corriveau, "Methods for evaluation of digital television picture quality", *IEEE Broadcast Tech. Soc.* – 4th meeting G-2.1.6, Audio Video Techniques Committee G-2.1, Compression and Processing Subcommittee -- Boulder, Colorado, Doc.~G-2.1.6/28, May 1997.
[4] A. Pessoa, A. Falcão, R. Nishihara, A. Silva, and A. Lotufo, "Video quality assessment using objective parameters based on image segmentation", *SMPTE Journal*, pp. 865-872, Dec. 1999.
[5] T. Hamada, S. Miyaji, and S. Matsumoto, "Picture quality assessment system by three-layered bottom-up noise weighting considering human visual perception", *SMPTE Journal*, pp. 20-6, Jan. 1999.
[6] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video systems", *SPIE – Int. Symposium on Voice, Video and Data Communications*, Sept. 1999.
[7] K. T. Tan and M. Ghanbari, "A multi-metric objective picture quality measurement model for MPEG video", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 10, no. 7, pp. 1208-13, Oct. 2000.
[8] T. Vlachos, "Detection of blocking artifacts in compressed video", *Electronics Letters*, vol. 36, no. 13, pp. 1106-8, June 2000.
[9] S. Ridella, S. Rovetta and R. Zunino, "Circular back-propagation networks for classification", *IEEE Trans. on Neural Networks*, vol. 8, no. 1, pp. 84-97, Jan. 1997.
[10] V. N. Vapnik, *Statistical learning theory*, Wiley & Sons, New York, 1998.
[11] S. Ridella, S. Rovetta and R. Zunino, "Circular Backpropagation networks embed Vector Quantization" *IEEE Trans. on Neural Networks*, vol. 10, no. 4, pp. 972-975, July 1999.
[12] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink and D. L. Alkon, "Accelerating the convergence of the back propagation method", *Biol. Cybern.*, vol. 59, pp. 257-263, 1988.
[13] ISO/IEC 13818-2: Information technology: Generic coding of moving pictures and associated audio video information: Video, 1994.
[14] ITU-R 11E/9, Introduction of a new method fo single stimulus continuous quality evaluation (SSCQE), Draft revision of Rec. ITU-R BT.500-7, ITU-R SG 11/E Document 11/21, June 1996.
[15] R. Aldridge and D. Pearson, "A calibration method for continuous video quality (SSCQE) measurements" *Signal Processing: Image Communication*,16, pp.321-32, 2000.
[16] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands and D. Pearson, "Recency effect in the subjective assessment of digitally-coded television pictures" *Proc. MOSAIC Workshop Advanced Methods for the Evaluation of Television Picture Quality*, Institute for Perception Research, Sept. 1995.
[17] H. R. Schiffman, *Sensation and perception - An integrated approach*, Fourth edition, John Wiley & Sons, inc., 1996.
Trans. on Neural Networks, vol.12, no.2, March 2001, in press.
[18] S. Rovetta and R. Zunino, "Efficient training of Neural Gas vector quantizers with analog circuit implementation", *IEEE Trans. on Circuits and Systems II*, vol.46, No.6, pp.688-698, June 1999.
[19] T. P. Vogl, J. K. Mangis, A. K. Rigler, W. T. Zink and D. L. Alkon, "Accelerating the convergence of the back propagation method", *Biol. Cybern.*, vol. 59, pp. 257-263, 1988.
[20] G. P. Drago, S. Ridella, "Pruning with interval arithmetic perceptron", Neurocomputing, vol. 18, pp. 229-46, 1998.