Journal of Electronic Imaging 11(3), 365–374 (July 2002).

Objective assessment of MPEG-2 video quality

Paolo Gastaldo
Rodolfo Zunino
University of Genoa
Department of Biophysical and Electronic Engineering
Via all'Opera Pia 11a
16145 Genova, Italy
E-mail: zunino@dibe.unige.it

Stefano Rovetta
University of Genoa
INFM, Department of Computer and Information Sciences
Via Dodecaneso 35
16146 Genova, Italy

Abstract. *The increasing use of video compression standards in broadcasting television systems has required, in recent years, the development of video quality measurements that take into account artifacts specifically caused by digital compression techniques. In this paper we present a methodology for the objective quality assessment of MPEG video streams by using circular back-propagation feedforward neural networks. Mapping neural networks can render nonlinear relationships between objective features and subjective judgments, thus avoiding any simplifying assumption on the complexity of the model. The neural network processes an instantaneous set of input values, and yields an associated estimate of perceived quality. Therefore, the neural-network approach turns objective quality assessment into adaptive modeling of subjective perception. The objective features used for the estimate are chosen according to the assessed relevance to perceived quality and are continuously extracted in real time from compressed video streams. The overall system mimics perception but does not require any analytical model of the underlying physical phenomenon. The capability to process compressed video streams represents an important advantage over existing approaches, like avoiding the stream-decoding process greatly enhances real-time performance. Experimental results confirm that the system provides satisfactory, continuous-time approximations for actual scoring curves concerning real test videos.* © 2002 SPIE and IS&T.
[DOI: 10.1117/1.1479703]

1 Introduction

The shift from analog to digital techniques has allowed TV broadcasters to offer new advanced services. Nevertheless, the technical quality of the video displayed may still compromise the success of digital TV production. The crucial issue is that digital encoding brings about specific visual artifacts; hence, traditional techniques for evaluating analog signals often prove ineffective in measuring the perceived quality of a digitally compressed video.

There exist several techniques for assessing the quality perceived by viewers. Subjective methods[1] simply ask human assessors to score the quality of a series of test scenes. Up to now subjective tests have been the basic tools with which to characterize video quality, despite the complexity, cost, and varying results of such tests.

From a different perspective, objective quality assessment aims to emulate human response to perceived quality by processing numerical quantities that describe video streams. As a result, this technique no longer requires inputs from human operators. Thus, objective assessment leads to deterministic models and makes real-time monitoring of perceived quality feasible. The need for objective measures in the area of digital TV has a commercial rationale, too: the number of coders on the market will increase in the next years, hence both manufacturers and broadcasters will necessarily face the problem of comparing video quality at the user's level. Several objective methods have been proposed in the literature.[2–11] Most approaches are based on decompressed video: objective parameters are derived by comparing pictures with original scenes at the receiver end. The comparison is made either in feature space or in the picture domain, and typically applies differencing methods.[12] Other recent approaches measure blocking artifacts without using reference images;[13–15] in addition, commercial tools have already been issued that can measure the quality of MPEG-2 video streams without referring to original sequences.[16]

From a scientific perspective, most of the above approaches aim at modeling perceived quality and imply some *a priori* assumptions of the underlying mathematical model. These simplifying hypotheses may somehow affect the general validity of results; in this respect, one should also consider that no valid model of human perception seems to have been developed yet, due to the highly nonlinear nature of the phenomena involved.

In this paper we present a method that uses neural networks[17] for automated evaluation of subjective assess-

Paper 01057 received Sep. 4, 2001; revised manuscript received Feb. 4, 2002; accepted for publication Feb. 4, 2002.
1017-9909/2002/$15.00 © 2002 SPIE and IS&T.

ment. Previous neural-based approaches to MPEG quality evaluation mainly addressed video coding control, and exploited neural networks to implement quality/rate strategies.[18,19] Those works used conventional, specifically tuned neural models (either multilayer perceptrons or radial-basis function networks), which entered the control process at the input (encoder) end. The research presented in this paper, instead, focuses on real-time monitoring of perceived quality at the decoder end, and exploits an advanced, flexible neural model. The network operates on compressed data only: it processes numerical features extracted continuously from the video stream, and generates the associated quality rating. This mode of operation 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 both the specific video source and decoder issues.

The present approach bypasses the objective of a deeper insight into the mechanism of quality perception; indeed, it aims to mimic such a perception. This goal is attained by using circular back-propagation (CBP) neural networks.[17] These networks support a general paradigm to deal with complex mathematical models, and remove the need for any *a priori* assumption aimed at simplifying an analytical model.

The paper is organized as follows. In Sec. 2 we describe the architecture of the neural-network-based system for video-quality evaluation, and the criteria for feature selection and feature run-time sampling. In Sec. 3 we outline the neural model adopted and the advantages of using this network in the multimedia application considered. In Sec. 4 are the experimental results, demonstrating the method's validity under different conditions and for different input sources. Some concluding remarks are made in Sec. 5.

2 Objective Assessment of Video Quality

The proposed approach aims at an automated quality evaluation of MPEG-2 bit streams.[20] The method can be regarded as being ‘‘objective’’ since it operates on numerical quantities (features) that are worked out directly from MPEG-2 bit streams and feed a neural network to obtain quality ratings. Figure 1 shows a schematic representation of the overall system.

The model 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. Therefore, the objective system lies within the single-stimulus continuous quality evaluation (SSCQE) paradigm,[21] requiring that assessments of picture quality be continuously recorded (in standard cases, by human observers). The technical framework for the evaluation schema adopted is a single-ended, ‘‘No-Reference’’ paradigm. The system does not require uncompressed original videos, unlike ‘‘Full Reference’’ or ‘‘Reduced Reference’’ approaches, which also involve considering the source of the video in the evaluation process.

In the design of objective-assessment systems, one should take into account that (1) several features that characterize video streams jointly affect subjective judgments, and (2) nonlinear relationships and unknown mechanisms may complicate the modeling process. The CBP network provides a paradigm by which to deal with multidimensional data characterized by complex relationships. The effectiveness of the neural-network approach lies in its capability to decouple the problem of feature selection from the design of an explicit mathematical model. The neural network directly yields the quality assessments associated with input vectors of extracted features; the function that maps feature vectors into quality ratings is learned from examples by use of an iterative training algorithm. Therefore, the design of the objective metric set is not involved in the setup of the mapping function.

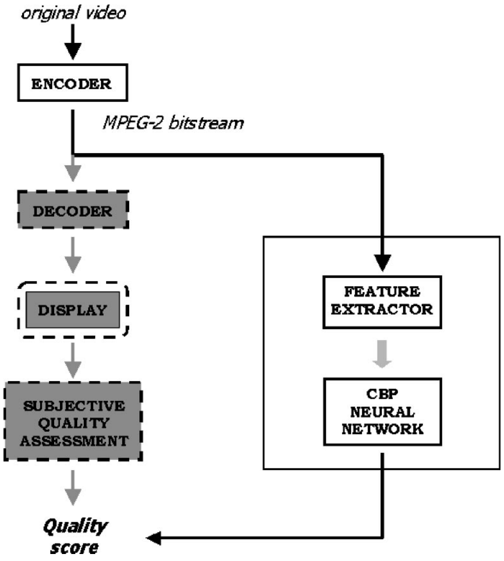


Fig. 1 Single-ended system for automated quality assessment. The neural network yields a continuous-time evaluation of perceived quality.

The implicit neural metrics rely entirely on a representation support—the compressed bit stream—that bypasses the need for human assessors' ratings altogether. This greatly improves the method's real-time performance, because the broadcaster can monitor perceived quality at transmission time. Handling compressed video privileges the relevance of blocking effects; this actually supports the basic model's performance, since blockiness represents the most significant visual impairment. Nevertheless, the bit stream carries complete information about the coded video (including detail-related quantities such as quantization matrices), hence the neural quality-evaluation system can reproduce perceived quality assessments quite accurately, especially since it can manage all the information available to the end user.

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

2.1 *Feature Selection for Objective Quality Assessment*

The set of processed features plays a crucial role for the effectiveness of the overall methodology. A single-ended paradigm requires that quite a large set of parameters be extracted *a priori* from video streams. Examples of such quantities are the number of bits per picture and the mean value over a picture of motion vector absolute values. The Appendix lists the complete feature set worked out from MPEG-2 compressed streams.

A subsequent statistical analysis, therefore, must sort out truly significant features. As expected, a considerable portion of all the features could be discarded because they either do not carry important information or are mutually correlated. Anyway, the present approach does not imply any *a priori* assumption of the significance of the encoding parameters, and only an *a posteriori* statistical analysis drives the feature-selection criterion. To this end, the following quantities are defined:

$\mathbf{\Psi}$ is a library $\{\psi_1,\ldots,\psi_L\}$ of L test streams, composed of P frames each;

$f_k^{(j)}(\psi_i)$ is the value assumed by the kth feature for the jth frame in the ith stream, ψ_i.

The feature-selection algorithm can be outlined as follows:

0. (input): a set of measured values, F_k, for each objective feature:

$$F_k=\{f_k^{(j)}(\psi_i), i=1,...,L,\ \ j=1,...,P\},\ \ k=1,...,N_f. \tag{1}$$

1. (Rescaling): For $k=1,...,N_f$:

1.a compute the 0.05 and 0.95 percentiles, $x_{0.05}^{(k)}$ and $x_{0.95}^{(k)}$, respectively, for the values in F_k;

1.b build up a normalized set $\underline{F}_k$ by rescaling each element of F_k into the range $[-1,\ 1]$:

$$\underline{F}_k=\{\underline{f}_{ijk}; i=1,...,L;\ \ j=1,...,P\}, \tag{2}$$

where

$$\underline{f}_{ijk}\overset{\text{def}}{=}2\,\frac{[f_k^{(j)}(\psi_i)-x_{0.05}^{(k)}]}{(x_{0.95}^{(k)}-x_{0.05}^{(k)})}-1. \tag{3}$$

2. (Descriptive statistics): Compute the two sets and the associated threshold values:

$S=\{\text{skew}_k; k=1,...,N_f\}$ where $\text{skew}_k=\text{skewness}(\underline{F}_k)$; $\text{skew}_{\text{thr}}=0.5$ percentile of S;

$K=\{\text{kurt}_k; k=1,...,N_f\}$ where $\text{kurt}_k=\text{kurtosis}(\underline{F}_k)$; $\text{kurt}_{\text{thr}}=0.5$ percentile of K.

3. (Feature selection): Compile the feature set, Z, keeping the objective features that satisfy

$$f_k\in Z \Leftrightarrow (\text{skew}_k>\text{skew}_{\text{thr}})\ \text{ and }\ (\text{kurt}_k>\text{kurt}_{\text{thr}}); \qquad k=1,...,N_f. \tag{4}$$

As a result, set Z includes the features that, due to their asymmetrical distribution, are unlikely to stem from a Gaussian distribution; this selection criterion can be justified as follows.

The main goal of the above procedure is to drive selection of the neural-network input vector. In principle, one might feed the neural network with the whole set of objective features; in fact, such a large number of inputs (1) would increase the complexity of the neural network and (2) might cause poor generalization due to overfitting problems. Thus, an empirical criterion that supports the feature-selection process is needed.

The present procedure uses skewness and kurtosis as paradigms to characterize the statistical activity of the features. The underlying hypothesis is that quantities with a non-normal distribution are most likely to be informative; of course, one must be aware that, in principle, normally distributed features can provide useful information as well. Therefore, ultimate validation of the antinormal selection will only stem from testing the empirical performance of the quality-evaluation system on the tentative feature set.

In this respect, the algorithm described has been preferred to alternative approaches [e.g., principal component analysis (PCA)[22]], mainly for the high data dimensionality involved in such methods. Numerical precision issues in working out eigenvectors, in particular, the presence of possible outliers, sometimes may affect the performance of PCA in high-dimensional domains. By contrast, the exploratory projection pursuit[23] is a method that uses the same paradigm as the proposed algorithm.

2.2 *Feature Run-Time Sampling*

The objective-assessment system generates continuous-time quality ratings. In principle, one can feed the CBP network with the feature values continuously extracted from each sequence frame. In fact, some specific mechanisms of human perception should be taken into account: (1) the assessor's reaction times are subject to delays,[24–26] (2) the most recent segment of a sequence has a greater effect on the instantaneous quality rating,[24,27] and (3) time-consecutive frames tend to interfere with one another.[28] In the literature, such phenomena are known as ''the assessor's response time,'' ''time-weighted averaging,'' and ''the masking phenomenon,'' respectively.

All of these aspects have been parametrized in the feature-extraction process (Fig. 2). The parameter Δ refers to the delay between the subjective judgment and the last frame that has influenced it. To compensate for time-weighted averaging, a set of N frames generates a single score. Within this set, groups of W consecutive frames make up a single feature vector, thus accounting for the masking phenomenon. In order to preserve information about the perceived quality over each interval of W pictures, the eventual input vector, $\mathbf{x}$, to the neural network includes, for each feature selected by the above-described analysis, one of the three values worked out as follows:

$$\begin{aligned}\hat{f}_{kL}^{(j)}(\psi_i)&=\max\{f_k^{(j)}(\psi_i),\ldots,f_k^{(j+W-1)}(\psi_i)\},\\ \hat{f}_{kS}^{(j)}(\psi_i)&=\min\{f_k^{(j)}(\psi_i),\ldots,f_k^{(j+W-1)}(\psi_i)\},\\ \hat{f}_{kA}^{(j)}(\psi_i)&=\text{avg}\{f_k^{(j)}(\psi_i),\ldots,f_k^{(j+W-1)}(\psi_i)\}.\end{aligned} \tag{5}$$

For instance, with the feature ''number of bits per picture,'' interest is in the smallest value over a set of W frames, because it is expected that the smaller the number of bits, the larger the degradation of the picture.

3 Neural Networks for Quality Estimation

Feedforward neural networks (NNs) map the feature vectors that describe video frames into the associated quality assessments. This problem formulation treats the quality

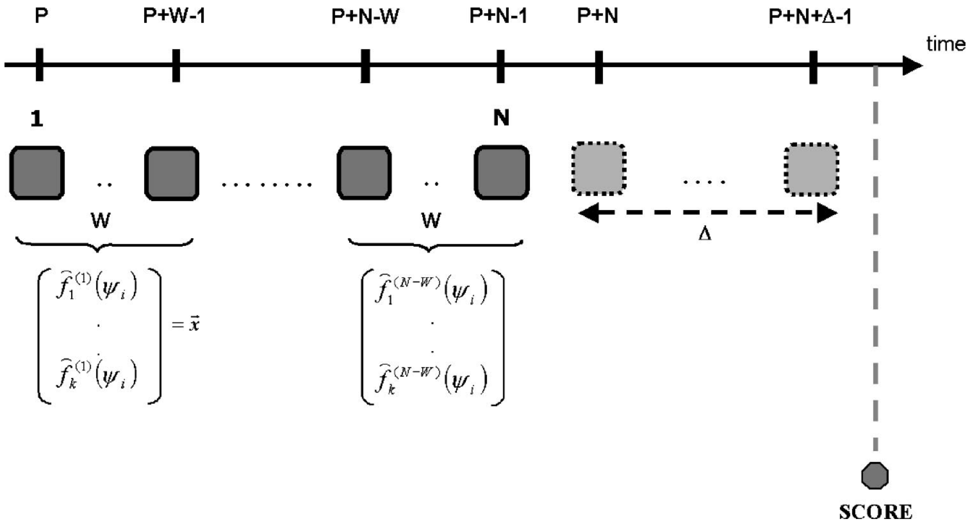

P
P+W-1
P+N-W
P+N-1
P+N
P+N+Δ-1
time
1
N
W
W
Δ
$\begin{pmatrix} \hat{f}_1^{(1)}(\psi_i) \\ \cdot \\ \hat{f}_k^{(1)}(\psi_i) \end{pmatrix} = \vec{x}$
$\begin{pmatrix} \hat{f}_1^{(N-W)}(\psi_i) \\ \cdot \\ \hat{f}_k^{(N-W)}(\psi_i) \end{pmatrix}$
SCORE

Fig. 2 Feature run-time sampling process according to perceptual mechanisms.

scores used for training as a discrete set of scalar labels, and network outputs are reported as scalar quantities. In this sense, efficiency requirements (i.e., the storage size of the parameters) and generalization issues (i.e., NN performance over data not used for training) ultimately result in the problem of properly sizing the number of neurons in the NN.

3.1 *CBP Architecture*

Neural-network research has shown that multilayer perceptrons (MLPs)[29] can efficiently tackle problems in which the target-mapping function can be supported by few parameters with a global scope. Instead, if the target-mapping process can be best expressed as a superposition of locally tuned contributions, radial-basis-function (RBF) networks[29] typically perform better. This implies that the unknown characteristics of the specific mapping problem further complicate the choice of the nature and size of the NN. The basic advantage of the circular back-propagation model is that it has been proved[17] to encompass both MLP and RBF paradigms; the choice of the more appropriate representation is implicit because it is performed during the training process and depends on the empirical problem at hand.

A CBP network includes a two-layer architecture (Fig. 3). The input layer connects the n_i input values (features) to each neuron of the "hidden layer." The uth "hidden" neuron first computes a linear combination of input values, which are weighted by coefficients $\{w_{u,k};u=1,...,n_h;k=1,...,n_i\}$:

$$r_u = w_{u,0} + \sum_{k=1}^{n_i} w_{u,k}x_k + w_{u,n_i+1}\sum_{k=1}^{n_i} x_k{}^2; \quad u=1,...,n_h. \tag{6a}$$

Then each neuron performs a nonlinear, sigmoidal transformation of the result:

$$a_u = \sigma(r_u); \quad u=1,...,n_h, \tag{6b}$$

where $\sigma(x)=(1+e^{-x})^{-1}$. The terms r_u and a_u are usually called the neuron *stimulus* and *activation*, respectively. The *output* layer provides the actual network responses, y_v, by a similar transformation:

$$r_v = w_{v,0} + \sum_{u=1}^{n_h} w_{v,u}a_u; \quad v=1,...,n_o. \tag{7a}$$

$$y_v = \sigma(r_v); \quad v=1,...,n_o. \tag{7b}$$

A quadratic cost function measures the distortion between the actual network output(s) and the expected reference output(s) on a sample of training patterns. The cost is expressed as

$$E = \frac{1}{n_o n_p} \sum_{l=1}^{n_p} \sum_{v=1}^{n_o} (t_v^{(l)} - y_v^{(l)})^2, \tag{8}$$

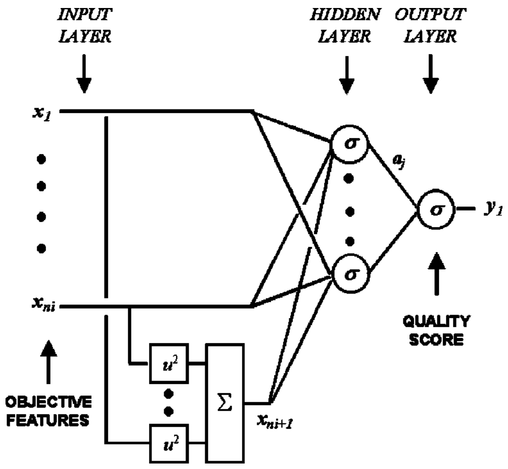


Fig. 3 Schematic representation of a CBP architecture. The CBP model includes one additional input to the standard MLP.

where n_p is the number of training patterns, and t_v are the desired training outputs. In the present application, $n_o = 1$ and the expected output is given by the quality assessment (score) measured experimentally from a human panel. An alternative to Eq. (8) is the *threshold cost* function $E_T(\gamma)$:

$$E_T(\gamma) = \frac{1}{n_o n_P} \sum_{l=1}^{n_P} \sum_{v=1}^{n_o} g(|t_v^{(l)} - y_v^{(l)}|);$$

$$g(x) = \left\{ \begin{matrix} 0 \Leftrightarrow x \leqslant \gamma, \\ 1 \Leftrightarrow x > \gamma \end{matrix} \right\}, \tag{9}$$

where the distortion cost is expressed as the percentage of outputs y_v that differ from the expected score t_v in more than a fixed threshold γ.

Training algorithms usually aim to minimize Eq. (8) mainly because, for that cost formulation, one can derive a gradient expression and use conventional gradient-descent techniques. The back-propagation algorithm[17] is by far the most widely used and most effective method for weight optimization in feedforward neural networks, and is adopted for CBP training as well.

From a structural perspective, the quadratic term in expression (6a) sets the difference between the CBP model and a conventional MLP. Such augmentation is attained by simply including one additional input (Fig. 3), which just sums the squared values of all the other network inputs. The additional unit allows the overall network to exhibit standard, sigmoidal behavior, or to drift smoothly to a bell-shaped, Gaussian-like radial function; this makes the CBP model able to choose autonomously from MLP and RBF representation paradigms. At the same time, the limited increase in the network parameters does not affect the expected generalization performance of the model.[17] The weight configuration resulting from the network-training process ultimately fixes the most suitable representation setting for the mapping problem.

The effectiveness of a neural network-based approach may not be intuitively obvious, especially when considering its degree of correspondence with human visual perception. The connection between the CBP architecture and visual perception mainly lies in the capability of the empirically trained network to catch some of the nonlinearities inherent in human perception. The resulting model is implicitly buried in the network parameters, hence most likely it proves difficult to interpret. As a natural feature of any empirical model, system effectiveness will strongly depend on the adequacy and completeness of training data.

3.2 *Neural Network Setup*

The network configuration (i.e., the number of hidden units) has been designed by use of a specific initialization technique that exploits the equivalence of the CBP model to vector-quantization (VQ) paradigms.[30] In particular, a VQ preliminary phase using the plastic neural gas algorithm[31] assessed the proper number of reference vectors to represent the available sample distribution. The subsequent configuration phase directly plugged the number and space positions of the VQ vectors in the CBP network.[30] That weight initialization proved most effective in accelerating the convergence of the overall training process, compared with conventional random initialization methods.

The CBP network training applies an accelerated variant[32] of the back-propagation algorithm. The possibility of using conventional techniques to train an advanced network structure is the major advantage of the CBP model.

4 Experimental Results

The effectiveness of the neural approach to objective quality assessment was verified experimentally by a library of MPEG-2 videos provided by the Research Center of the Italian Radio and Television Corporation (RAI). The testbed included 12 frame-coded MP@ML sequences, each 70 s long; the picture size was 720×576 pixels. The sequence contents varied from fiction to sport, and were encoded at different bit rates in the range of 4–8 Mbits/s.

Assessments of all the sequences were collected by nonexpert viewers; the subjective tests were carried out with a SSCQE technique at a sampling rate of two scores per second. Quality ratings were represented by a continuous scale ranging in $[-1, 1]$.

4.1 *Experimental Setup*

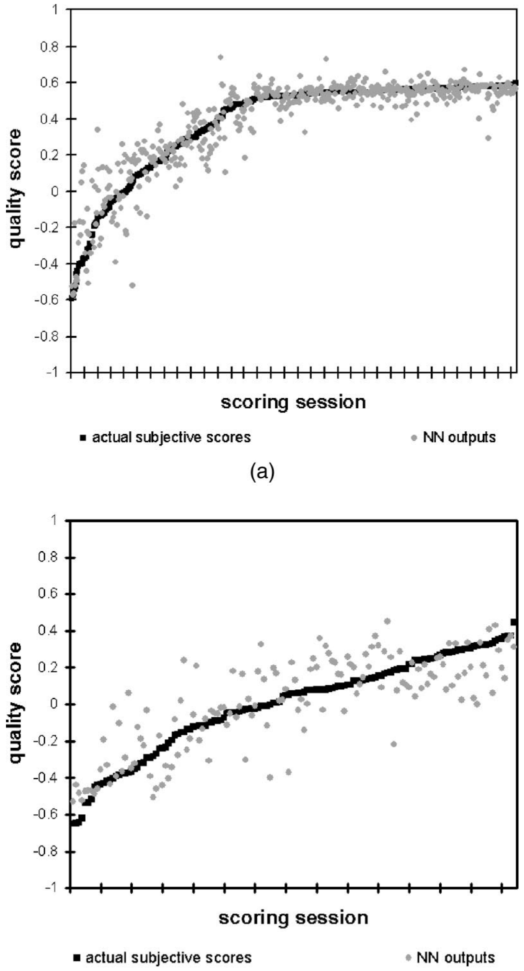
The neural-network training process involved the Z set of features that the statistical analysis had selected from the global feature set listed in the Appendix. In order to enhance the CBP network's generalization performance, the dimensionality of the input data space was further reduced with a feature-selection technique.[33] The eventual four-dimensional feature space covered the quantities *N bits*, *Xq_scale(1)*, *Xmv(1)*, and *Smv_dev_std*. The objective metric handles information about intracoding (quantization factors, number of bits per picture) and intercoding (motion vectors) properties of the video stream; therefore, as anticipated in Sec. 2, the quality-evaluation system can manage global characteristics of the video.

The data set included 1320 patterns generated by the run-time sampling process presented in Sec. 2, with $N = 24$, $W = 6$ and $\Delta = 17$. The numerical values of these parameters were determined by using standard values proposed in the literature.[24–28] The training and test sets were obtained by dividing the data set into two subsets of 820 and 500 patterns, respectively.

In order to avoid overfitting problems, the number of nodes in the hidden layer was chosen by using the plastic VQ algorithm, which processed the training samples to design the neural network configuration. The resulting value $n_h = 14$ set the number of hidden units in the feedforward structure.

4.2 *Results*

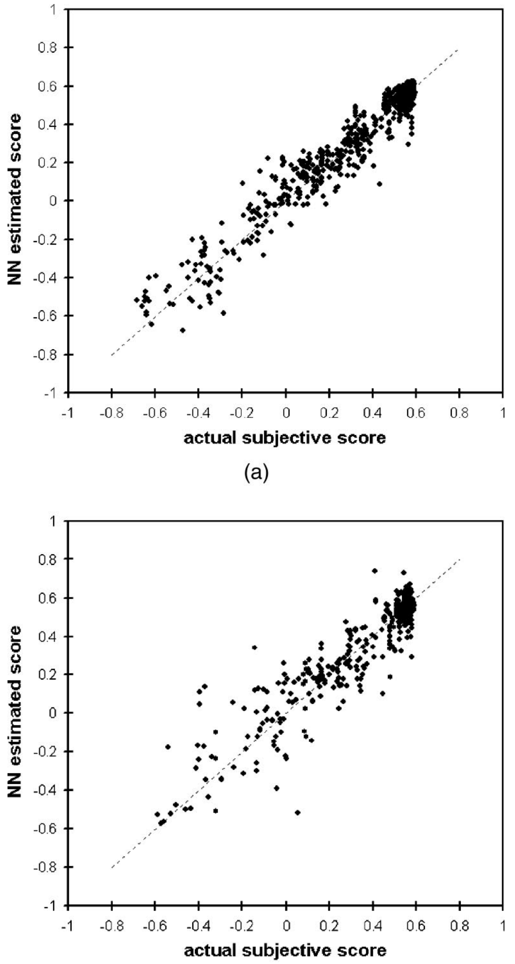
Figure 4 shows the test results obtained for the selected feature set. Figures 4(a) and 4(b) compare the quality ratings by human assessors with the corresponding outputs of the neural network. For display clarity, the human ratings were sorted in increasing order, each point on the x axis representing a single evaluation event. Figure 4(a) shows the plot of the numerical results obtained for the whole library of test videos, that is, the 12 MPEG-2 video streams. It also presents an asymmetric distribution of subjective scores, 44% of the original scores exceed 0.5. Since



(a)

(b)

Fig. 4 Neural-network scoring performance: (a) results obtained on the whole test library; (b) results obtained on sport videos.



(a)

(b)

Fig. 5 Correlation between the actual subjective score and the estimated objective score: (a) training results; (b) test results.

the lower scores appear subsampled, they are subject to larger errors due to the lower statistical confidence. Nevertheless, the CPB neural network attained an average error $\hat{\mu}_{\text{err}} = -0.001$ over the test set. The average error over the absolute values of the prediction errors was $\hat{\mu}_{|\text{err}|} = 0.06$.

Figure 4(b) presents results obtained for a subset of the testbed. The subset includes videos with sport contents only. Figure 4(b) shows that human quality ratings exhibit a higher variance for this kind of video. In addition, the sequences with sport contents are a small subset of the test library, hence the neural network suffered from larger errors due to the lower statistical confidence. In this case, the neural-network system achieved $\hat{\mu}_{\text{err}} = -0.01$ and $\hat{\mu}_{|\text{err}|} = 0.12$.

In order to show the generalization ability of the model, Fig. 5 shows a comparison of training and test results obtained with the complete testbed. Figure 5(a) gives a scatter plot of the training results, with the actual subjective score as the *x* axis and the estimated objective score as the *y* axis. Pearson's correlation coefficient for the training results takes on a value of 0.97; the Spearman rank order correlation, a nonparametric and distribution free test, gives a value of 0.85 as the correlation coefficient. Figure 5(b) shows the scatter plot for the test results, which exhibit a slightly lower correlation; Pearson's correlation coefficient takes a value of 0.93 and the Spearman rank order correlation coefficient takes a value of 0.8.

The significance of these results is supported by the comparison with the experimental results obtained by picture appraisal rating (PAR),[16] a single-ended quality measure for MPEG used in a commercial product for video quality control. PAR achieves a Pearson correlation coefficient of 0.93 between the estimated outputs and the peak signal-to-noise ratio (PSNR), which is used as a reference measure of quality and is worked out on the difference between original and decoded frames. Compared with PAR, the proposed neural-based approach obtains on test results the same correlation coefficient between estimated outputs and reference quality measures. Furthermore, the present work uses as reference quality ratings the subjective scores rather than an objective measure such as the PSNR. In this sense, it can be asserted that the neural network yields a more reliable estimate of video quality as perceived by human assessors.

Figure 6(a) shows a plot of the error distribution obtained for the whole library of test videos. The graph presents the distribution together with the related best-fitting

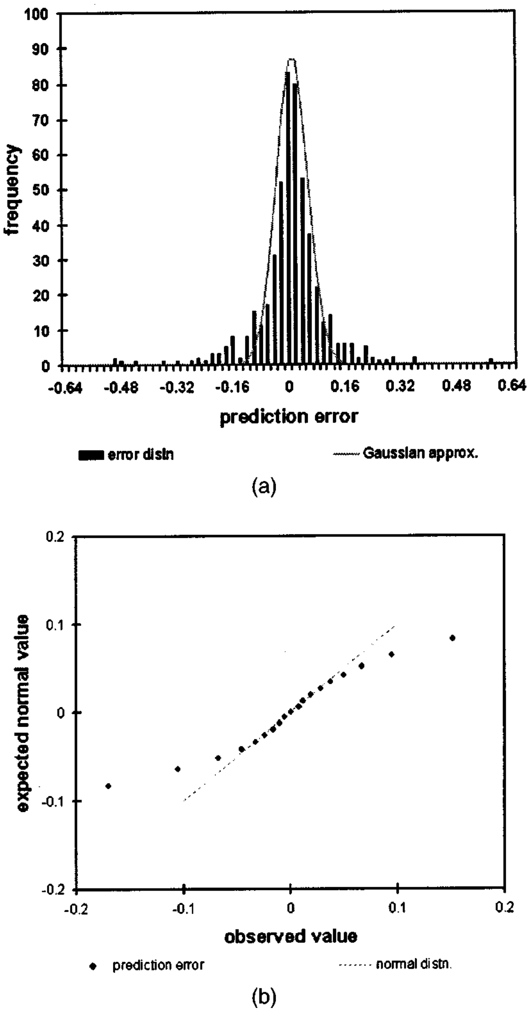


Fig. 6 Error distribution obtained for the whole library of test streams: (a) actual error distribution together with $N(0,0.05)$; (b) comparison of quantiles of the normal distribution $N(0,0.05)$ vs the corresponding sample quantiles of the prediction errors.

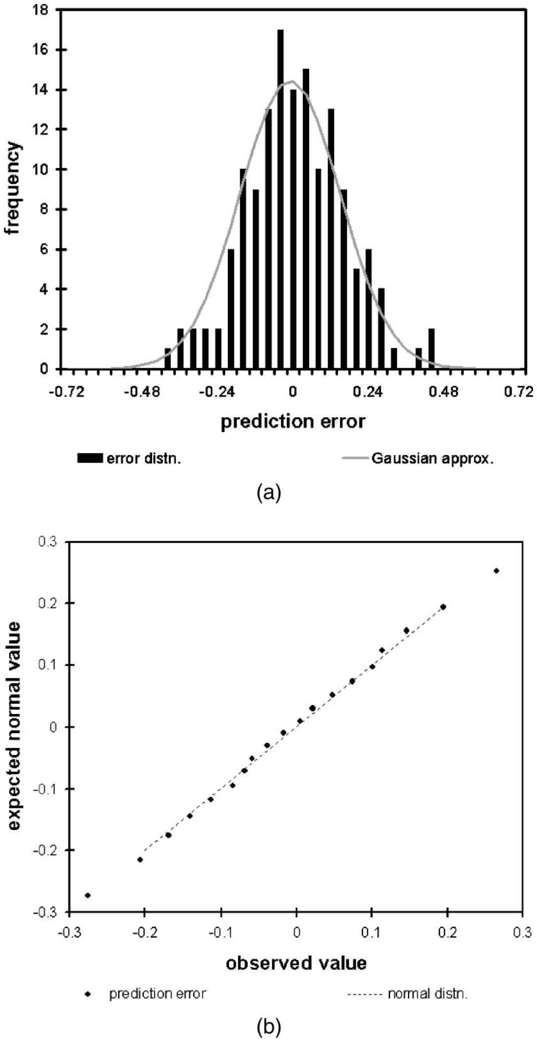


Fig. 7 Error distribution obtained for video streams with sport contents: (a) actual error distribution together with $N(0,0.13)$; (b) quantiles of the normal distribution $N(0,0.13)$ vs the corresponding sample quantiles of the prediction errors.

Gaussian approximation $N(0,0.05)$. The correctness of the Gaussian assumption was verified by carrying out a Kolmogorov–Smirnov (KS) normality test, which satisfied the null hypothesis to a high degree of confidence ($p > 0.95$). The $Q-Q$ plot shown in Fig. 6(b) confirms that the error distribution follows a normal distribution. The graph demonstrates that almost all the actual observed values of the prediction errors lie on the dashed line that represents the Gaussian distribution $N(0,0.05)$.

Figure 7(a) shows the error distribution obtained for test videos with sport contents. The actual error distribution is plotted together with the associate best-approximating Gaussian distribution $N(0,0.16)$. Figure 7(b) presents the corresponding $Q-Q$ plot, which strengthens the hypothesis about a normal distribution, as most of the observed values lie on the dashed line indicating Gaussian distribution $N(0,0.16)$.

The overall numerical results are summarized in Table 1, which also gives the costs e and $e_T(\gamma)$ derived from the neural-network test.

The analysis of the confidence interval (CI) for μ_{err} confirms the method's effectiveness. For large sample sizes n, the $1-\alpha$ CI for a distribution with unknown mean μ and unknown variance σ^2 can be approximated by

$$\hat{\mu} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}, \tag{10}$$

Table 1 Test results.

	Complete set	Sport content only
$\hat{\mu}_{\lvert\mathrm{err}\rvert}$	0.06	0.12
$\hat{\mu}_{\mathrm{err}}$	−0.001	−0.01
$\hat{\sigma}^2{}_{\mathrm{err}}$	0.01	0.02
e	0.01	0.0251
$e_T(0.15)$	0.1022	0.3011

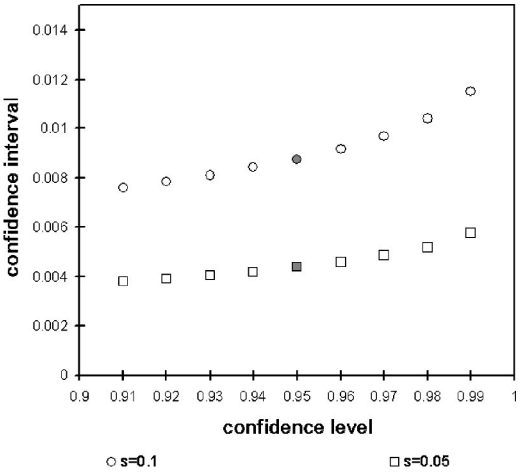


Fig. 8 Plot of the confidence interval for sample mean $\hat{\mu}_{\text{err}}$ as a function of the confidence level 1-α.

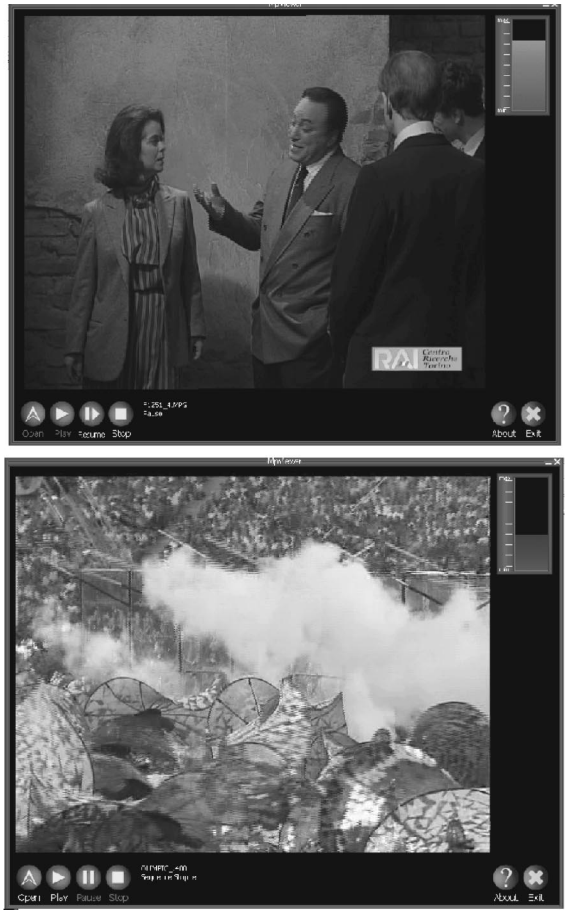


Fig. 9 Video quality analyzer.

where $\hat{\mu}$ is the sample mean, $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution $N(0,1)$, and s is the sample standard deviation. The graph in Fig. 8 plots $z_{\alpha/2}s/\sqrt{n}(=|\hat{\mu}_{\text{err}}-\mu_{\text{err}}|)$ as a function of the confidence level $1-\alpha$ for the test result obtained on the complete set $(n=500)$. Figure 8 compares the results obtained by assuming s to be

- the sample standard deviation measured on the whole sample set $(s\cong 0.1)$;
- the estimate of the sample standard deviation by using the Gaussian approximation of the error distribution $(s\cong 0.05)$.

The graph shows that the 0.95 CI is $\hat{\mu}_{\text{err}}\pm 0.0087$ for $s=0.1$ and $\hat{\mu}_{\text{err}}\pm 0.0043$ for $s=0.05$.

5 Conclusions

In this work we have presented an automated method for objective quality assessment by use of neural networks (Fig. 9). The evaluation system handles MPEG-2 video streams. Numerical observations are computed for each frame of the processed MPEG sequence and enter the neural network consisting of a circular back-propagation architecture. The statistical model supported by the trained neural network yields an output scalar value, which provides a numerical representation of perceived quality.

The major result of the proposed method is the possibility of reproducing human perception consistently by using quantitative, data-driven models. The neural-network model is specifically tuned to learn the perceptual phenomenon from examples, and exploits a known effective augmentation of standard back-propagation (BP) networks.

A crucial advantage of the methodology described is the system's capability to handle compressed video streams. Avoiding the need for decompressed pictures enhances the method's effectiveness in real-time production applications.

The experimental setup involved both a training phase with observations collected from evaluation panels and generalization testing using sequences and the associated quality assessments not included in the training sets. Experimental evidence confirmed the validity of the approach, because the system always provided satisfactory, continuous-time approximations for the actual scoring curves related to test videos.

Acknowledgments

The authors wish to thank P. Badino and P. Tedesco for their assistance in developing the method and performing the experiments described in this paper.

Appendix: Objective Features

An MPEG-2 bit stream has a hierarchical structure that allows one to get information at multiple levels, i.e., sequence, group of pictures, picture, slice, macroblock and block. In the present work, objective features have been chosen to characterize the stream at the picture level.

The following quantities are defined:

- $$\text{energy}=\frac{1}{256}\sum_{i=0}^{16}\sum_{j=0}^{16}(mb_{\text{DCT}}[i][j])^2, \tag{A1}$$

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

Table 2 Features worked out from MPEG stream.

Feature name	Feature description
	Percentage (macroblocks)
Pmb_no_pred	n_{mb}=macroblocks with no motion vectors
Pmb_fwd	n_{mb}=macroblocks with forward motion vector only
Pmb_back	n_{mb}=macroblocks with backward motion vector only
Pmb_bidir	n_{mb}=bidirectional macroblocks
Pmb_I	n_{mb}=intramacroblocks
Pmb_skipped	n_{mb}=skipped macroblocks
	Percentage (blocks)
Pb_sk_luma	n_b=skipped luminance blocks
Pb_sk_chroma	n_b=skipped chrominance blocks
	Statistical figures
Smv_mean	mean-$p_i=\|$motion vector$\|$
Sq_scale_mean	mean-$p_i=q\_scale$
Senergy_mean	mean-$p_i=energy$
Smv_dev_std	standard deviation-$p_i=\|$motion vector$\|$
Sq_scale_dev_std	standard deviation-$p_i=q\_scale$
Senergy_dev_std	standard deviation-$p_i=energy$
Smv_var	variance-$p_i=\|$motion vector$\|$
Sq_scale_var	variance-$p_i=q\_scale$
Senergy_var	variance-$p_i=energy$
	Percentile
Xmv(α)	p_i=mean of $\|$motion vector$\|$
Xq_scale(α)	$p_i=q\_scale$
Xenergy(α)	$p_i=energy$
Xq_mv(α)	$p_i=q\_mv$
Xe_mv(α)	$p_i=e\_mv$

- $$q\_mv=\frac{q\_\text{scale}}{1+\langle|m\_v|\rangle}, \tag{A2}$$

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

- $$e\_mv=\text{energy}\cdot\langle|m\_v|\rangle, \tag{A3}$$

where e_mv is the weighted energy of a macroblock.

Table 2 lists the objective features worked out from the coded bit stream. The following four classes of measures can be identified.

- Percentage of macroblocks: Features are defined as

$$f_k=\frac{n_{mb}}{n_t^{(mb)}}, \tag{A4}$$

where n_{mb} is the number of macroblocks of the type specified in the second column of Table 2, and $n_T^{(mb)}$ is the total number of macroblocks in the picture.

- Percentage of blocks: Features are defined as

$$f_k=\frac{n_b}{n_T^{(b)}} \tag{A5}$$

where n_b is the number of blocks of the type specified in Table 2, and $n_T^{(b)}$ is the total number of blocks in the picture.

- Statistic features are defined as

$$f_k=\begin{cases}\text{mean}(\mathbf{p})\\ \text{std deviation}(\mathbf{p})\\ \text{variance}(\mathbf{p})\end{cases} \tag{A6}$$

where $\mathbf{p}$ is a vector of values p_i computed on each macroblock of the picture; p_i is given in Table 2.

- Percentiles: Features are defined as

$$f_k=x_\alpha(\mathbf{p}) \tag{A7}$$

where x_α is the α percentile of $\mathbf{p}$.

The last feature included in the objective set is *N*bits, i.e., the number of bits per picture.

References

1. International Telecommunication Union, "Methodology for the subjective assessment of the quality of television pictures," Recommendation No. BT.500-10, Geneva, Switzerland (2000).
2. A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "An objective video quality assessment system based on human perception," in *Human Vision, Visual Processing, and Digital Display IV, Proc. SPIE* **1913**, 15–26 (1993).
3. M. Ardito and M. Visca, "Correlation between objective and subjective measurements for video compressed systems," *SMPTE J.* **105**(12), 768–773 (1996).
4. S. Olsson, M. Stroppiana, and J. Baïna, "Objective methods for assessment of video quality: State of the art," *IEEE Trans. Broadcasting* **43**(4), 487–495 (1997).
5. W. Y. Zou and P. J. Corriveau, "Methods for evaluation of digital television picture quality," IEEE Broadcast Technical Society, 4th Meeting, Audio Video Techniques Committee, Compression and Processing Subcommittee, Doc. No. G-2.1.6/28 (1997).
6. G. W. Cermak, S. Wolf, E. P. Tweedy, M. H. Pinson, and A. Webster, "Validating objective measures of MPEG video quality," *SMPTE J.* **107**(4), 226–235 (1998).
7. A. Pessoa, A. Falcão, R. Nishihara, A. Silva, and A. Lotufo, "Video quality assessment using objective parameters based on image segmentation," *SMPTE J.* **108**(12), 865–872 (1999).
8. T. Hamada, S. Miyaji, and S. Matsumoto, "Picture quality assessment system by three-layered bottom-up noise weighting considering human visual perception," *SMPTE J.* **108**(1), 20–26 (1999).
9. S. Wolf and M. H. Pinson, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video systems," in *SPIE International Symposium on Voice, Video and Data Communications, Proc. SPIE* **3845**, 266–277 (1999).
10. K. T. Tan and M. Ghanbari, "A multi-metric objective picture quality measurement model for MPEG video," *IEEE Trans. Circuits Syst. Video Technol.* **10**(7), 1208–1213 (2000).
11. A. B. Watson, J. Hu, and J. F. McGowan III, "Digital video quality metric based on human vision," *J. Electron. Imaging* **10**(1), 20–29 (2001).
12. D. K. Fibush, "Practical application of objective picture quality measurements," *SMPTE J.* **108**(1), 10–19 (1999).
13. T. Vlachos, "Detection of blocking artifacts in compressed video," *Electron. Lett.* **36**(13), 1106–1108 (2000).
14. H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Process. Lett.* **4**(11), 317–320 (1997).
15. Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," *Proc. IEEE* **3**, 981–984 (2000).

16. M. Knee, ''The picture appraisal rating (PAR)—A single-ended picture quality measure for MPEG-2,''in *Proc. Int. Broadcasting Convention, Amsterdam, The Netherlands* (2000).
17. S. Ridella, S. Rovetta, and R. Zunino, ''Circular back-propagation networks for classification,'' *IEEE Trans. Neural Netw.* **8**(1), 84–97 (1997).
18. L. Luo, Y. Lu, C. Zou, and Z. He, ''Image sequence macroblocks classification using neural networks,'' *Signal Process.* **69**(2), 191–198 (1998).
19. F.-H. Lin and R. M. Mersereau, ''Rate-quality tradeoff MPEG video encoder,'' *Signal Process. Image Commun.* **14**(4), 297–309 (1999).
20. ISO/IEC 13818-2, ''Information technology: Generic coding of moving pictures and associated audio video information,'' Video (1994).
21. International Telecommunication Union, ''Introduction of a new method for single stimulus continuous quality evaluation (SSCQE),'' Draft revision of Rec. No. ITU-R BT.500-7, ITU-R SG 11/E document 11/21 (1996).
22. I. T. Joliffe, *Principal Component Analysis*, Springer, New York (1986).
23. J. H. Friedman and J. W. Tukey, ''A projection pursuit algorithm for exploratory data analysis,'' *IEEE Trans. Comput.* **C-23**(9), 881–890 (1974).
24. H. de Ridder and R. Hamberg, ''Continuous assessment of image quality,'' *SMPTE J.* **106**(2), 123–128 (1997).
25. N. Narita, Y. Sugiura, and I. Yuyama, ''Time sensitive evaluation of the quality of digital coded sequences,'' *SMPTE J.* **108**(1), 32–38 (1999).
26. R. Aldridge and D. Pearson, ''A calibration method for continuous video quality (SSCQE) measurements,'' *Signal Process. Image Commun.* **16**(3), 321–332 (2000).
27. H. de Ridder and R. Hamberg, ''Continuous assessment of perceptual image quality,'' *J. Opt. Soc. Am. A* **12**(12), 2573–2577 (1995).
28. H. R. Schiffman, *Sensation and Perception: An Integrated Approach*, 4th ed., Wiley, New York (1996).
29. J. Hertz, A. Krogh, and R. Palmer, *Introduction to Neural Computation*, Prentice–Hall, Englewood Cliffs, NJ (1992).
30. S. Ridella, S. Rovetta, and R. Zunino, ''Circular backpropagation networks embed vector quantization,'' *IEEE Trans. Neural Netw.* **10**(4), 972–975 (1999).
31. S. Ridella, S. Rovetta, and R. Zunino, ''K-Winner machines for pattern classification,'' *IEEE Trans. Neural Netw.* **12**(2), 371–385 (2001).
32. 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.* **59**, 257–263 (1988).
33. G. P. Drago and S. Ridella, ''Pruning with interval arithmetic perceptron,'' *Neurocomputing* **18**, 229–246 (1998).

Paolo Gastaldo received his ''Laurea'' degree in electronic engineering in 1998 from Genoa University. He is currently with the Electronic Systems and Networking Group of the Department of Biophysical and Electronic Engineering, University of Genoa, and is working toward his PhD in space science and engineering. His research interests include neural network implementation and applications, advanced multimedia signal processing, and digital television.

Rodolfo Zunino obtained his ''Laurea'' degree in electronic engineering from Genoa University in 1985. From 1986 to 1995 he was a research consultant with the Department of Biophysical and Electronic Engineering at Genoa University. He is currently at the same Department as an associate professor of industrial electronics. His main scientific interests include electronic systems for neural networks, efficient models for data representation and learning, advanced techniques for multimedia data processing, and distributed-control methodologies.

Stefano Rovetta received his ''Laurea'' degree in electronic engineering in 1993 and his PhD degree in models, methods and tools for electronic and electromagnetic systems in 1997 from the University of Genoa. He held the position of postdoctoral researcher for the Electronic Systems and Networking Group of the Department of Biophysical and Electronic Engineering, University of Genoa. He has been an invited professor of operating systems at the University of Siena. He is currently assistant professor at the Department of Computer and Information Sciences at Genoa University. His research interests include electronic circuits and systems, and neural network theory, implementation and applications.