OBJECTIVE VIDEO QUALITY ASSESSMENT: 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, rovetta, zunino}@dibe.unige.it

ABSTRACT

The paper presents a tool that uses feed-forward neural-network technologies for the objective quality assessment of MPEG-2 video. The system operates on a frame-by-frame basis and yields continuous output, providing a real-time monitoring tool for displayed video quality. Features used for the estimation have been designed according to their assessed relevance to perceived quality.

Keywords: objective quality assessment, neural network, MPEG-2 video.

1 INTRODUCTION

The increasing use of video compression standards in broadcasting TV systems raised, in recent years, the need for techniques to measure video quality. Novel methods must take into account the specific artifacts introduced by digital compression techniques.

Objective quality assessment aims to emulate human perceived quality by extracting numerical quantities from video streams. As opposed to subjective assessment methods [1], no inputs from human evaluators are required. As a result, objective assessment leads to deterministic models of perceived quality; such method should anyway keep human scores as a reference to ensure consistency with subjective results.

A variety of methods for objective quality assessment of digital TV has been proposed in the literature. Most are based on decompressed video [2-4]**,** and tackle human perception of quality as a modeling problem. A method that does not involve the original video is described in [5], where data are extracted from decoded frames.

This paper presents a method using a neural network (NN) for automatic evaluation of MPEG-2 [6] video quality, that operates on compressed data only. This removes the need for any information about either the original video or the decoding process. In other words, a single-ended paradigm is adopted. From an engineering perspective, the adaptive neural framework decouples the evaluation task from both the specific video source and any decoder features.

2 FEED-FORWARD NEURAL NETWORK (CBP)

Feed-forward neural networks provide a straightforward paradigm to map feature vectors (describing video frames) into the corresponding quality assessments. This problem setting uses quality scores by human evaluators as a training set; these represent an ordered, discrete set of labels. Conversely, any intermediate values are allowed in network outputs. Efficiency requirements, as well as generalization issues, ultimately lead to the problem of properly sizing the number of neurons in the NN.

In this research, the score-mapping task is supported by an augmented version of standard Multilayer Perceptrons (MLPs), namely, the "Circular BackPropagation" (CBP) [7] model. The CBP model extends the MLP by including one additional input and its associated weight. That input simply sums the squared values of all other network inputs. Theory proves [7] that this model is the most efficient polynomial extension of MLPs with linear stimulus, and formally encompasses the RBF network model, as well.

The CBP model can be formally described as follows. A MLP architecture combines two functional layers including n_h and n_o units, respectively. The conventional sigmoidal function will be denoted as $\sigma(x) = (1+e^{-x})^{-1}$.

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 input values ($j = 1,\ldots, n_h$):

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

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

Input features x_i (i=1 … n_i) feed the *j*-th hidden unit by the associated weights w_{ji} (i=0 … n_i+1). 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 to the CBP model.

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

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

$$y_k = \sigma(r_k) \; ; \quad k = 1,\ldots, n_o; \qquad (4)$$

The strict relationship of CBP with Vector-Quantization networks has been analyzed in [8], showing that the model ensures a notable representation effectiveness with a very limited 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). The conventional back-propagation algorithm can still be adopted for weight adjustment independently of the underlying representation model, which is actually driven by the training process [9]. 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 approach is specifically aimed at using CBP networks for the evaluation of MPEG-2 streams. Objective features are worked out from MPEG-2 bitstreams directly (i.e., without any decoding), and feed the NN to obtain quality ratings (Fig. 1). The system operates on a frame-by-frame basis and yields continuous output. As such, it provides a real-time monitoring tool for displayed video quality. Thus the neural network eventually mimics the subjective, Single-Stimulus Continuous Quality Evaluation paradigm (SSCQE) [10], recording continuous assessments of picture quality from human observers.

The crucial advantage of the neural-network approach is that no decoding phase is required for attaining scoring outputs. The NN is trained to associate bitstream-related quantities with quality scorings, hence the underlying, implicit modelling process does not require decoded pictures altogether. This greatly improves the method's effectiveness especially in terms of real-time performance, since one can get an estimate of perceived quality at transmission time.

For the reader's convenience, we remind 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 set of processed features play a crucial role in the effectiveness of the overall methodology. To avoid a-priori assumptions on the significance of encoding parameters, a quite large set of features (Appendix A) have been extracted from video streams. The purpose is to collect as much information as possible. Truly significant features are then selected by an a-posteriori statistical analysis. The following quantities are defined:

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

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

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

The feature-selection algorithm can be outlined as follows:

0. ***(Input)*** **-** $\Phi_k = \left\{ f_{k\Psi_1}^{(1)}, \ldots, f_{k\Psi_1}^{(P)}, \ldots, f_{k\Psi_L}^{(1)}, \ldots, f_{k\Psi_L}^{(P)} \right\}$, sets of features associated with each k-th sequence frame.

1. ***(Rescaling)*** **-** For k=1 … N_f:

- compute the 0.05 and the 0.95 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:

$$\underline{f}_{kz} = 2 \frac{(f_{k\Psi q}^{z-Pq} - x_{0.05}^{k})}{(x_{0.95}^{k} - x_{0.05}^{k})} - 1 \qquad q = \lfloor (P+z-1)/P \rfloor \quad (5)$$

2. ***(Descriptive statistics)*** **-** Create two sets (k=1 … N_f):

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

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

3. ***(Threshold setting)*** **-** Compute the thresholds values, sk_{thr} (0.5 percentile of Σ) and ku_{th} (0.5 percentile of K).

4. ***(Feature selection)*** - Compile the feature set, Z, holding the objective features that satisfy:

$$F_k \in \mathrm{Z} \Leftrightarrow (skew_k > sk_{thr}) \; AND \; (kurt_k > ku_{thr}) = TRUE$$

k=1 … N_f; (6)

The set Z includes those features that are unlikely to stem from a Gaussian distribution, mainly because of their asymmetrical distribution.

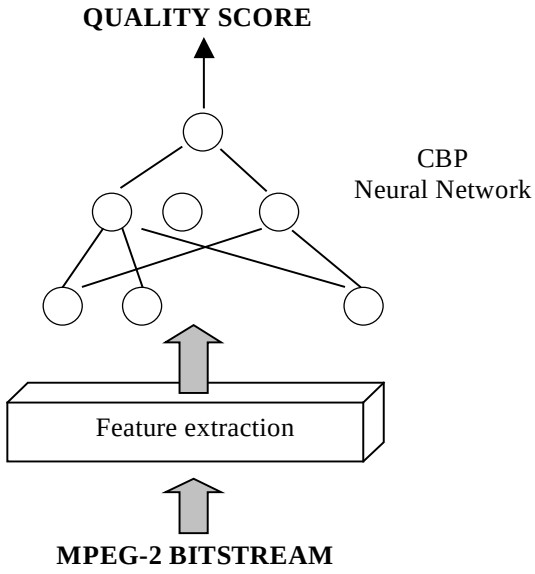

Figure 1 NN system for quality assessment.

3.2 The Neural Network Approach

Several features characterizing video streams jointly contribute to form subjective judgements; in addition, possibly non-linear laws ruling unknown mechanisms may play some role. These effects sometimes seem to have been underevaluated. The main advantage of the neural network approach lies in the ability to take these peculiarities into account.

In the present approach, CBP networks must map the feature vectors described in the previous Section into quality ratings. The mapping function is learned from examples by means of an iterative algorithm. Thus the task of features selection is decoupled from the problem of designing an explicit mathematical model. The main advantage lies in discarding a-priori assumption on the significance of objective measurements. Objective features are selected by statistical criteria that do not involve the neural network training phase.

The unique output neuron in the NN yields the quality assessment for a given input vector. Neural network training uses the back-propagation algorithm. The possibility to adopt conventional techniques to train an advanced network structure is a major advantage of the CBP model. The network cost function is expressed as:

$$e(\vec{w}) = \frac{1}{n_o n_p} \sum_{m=1}^{n_p} \sum_{k=1}^{n_o} \left(t_k^{(m)} - y_k^{(m)} \right) \qquad (7)$$

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

4 EXPERIMENTAL RESULTS

The method performance for practical applications have been evaluated by using a library of test videos provided by National RAI Research Center. The library included twelve MPEG-2 frame-coded sequences, with frame size 720x576 pixel, and bitrate from 4 to 8 Mbits/sec. Subjective assessment of the quality was developed with the SSCQE method, with a score sample interval of half a second.

4.1 Experiment Setup

The proposed objective assessment system has been designed to generate a continuous time quality rating. To this purpose input features x_i have been worked out taking into account the influence of perceptual mechanisms. Figure 2 shows how assessors response time (*M=17*), averaged picture quality ratings (*N=24*) and visual masking phenomenon (*W=6*) have been modelled. Each group of *W* frames gives a pattern of n_i features $\hat{f}_{k\Phi_L}^{(j,W)}$ ($F_k \in Z$) defined as follows:

$$\hat{f}_{k\Psi_L}^{(j,W)} = \rho(f_{k\Psi_L}^{(j)} \cdots f_{k\Psi_L}^{(j+W)}) \qquad (8)$$

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

4.2 Results

The experiment has involved a training and a data set consisting of 665 patterns each. Best results have been obtained with n_h=15 and an input vector including 4 features: $\vec{x} = \{\rho_s(Nn\_bits), \rho_h(Xq\_scale(1)), \rho_h(Xmv(1)), \rho_h(Smv\_dev\_std)\}$.

Figure 3 shows the plot of sorted assessors quality ratings with the correspondent outputs yielded by the neural network system. The asymmetric distribution of subjective scores, the 44% of quality ratings are over 0.5, clearly gives rise to higher errors where the statistic confidence is lower. Nevertheless the NN achieved on the test set an average error of –0.00324, with a standard deviation of 0.10231.

The plot given in fig. 4 confirms system effectiveness. The curve shows that an accuracy ε=0.0175 has been obtained with confidence p=0.95(=1-δ).

5 CONCLUSIONS

This paper has described an objective quality system based on feed-forward neural networks. The most relevant outcome of the presented research is the possibility of reproducing perception consistently by means of

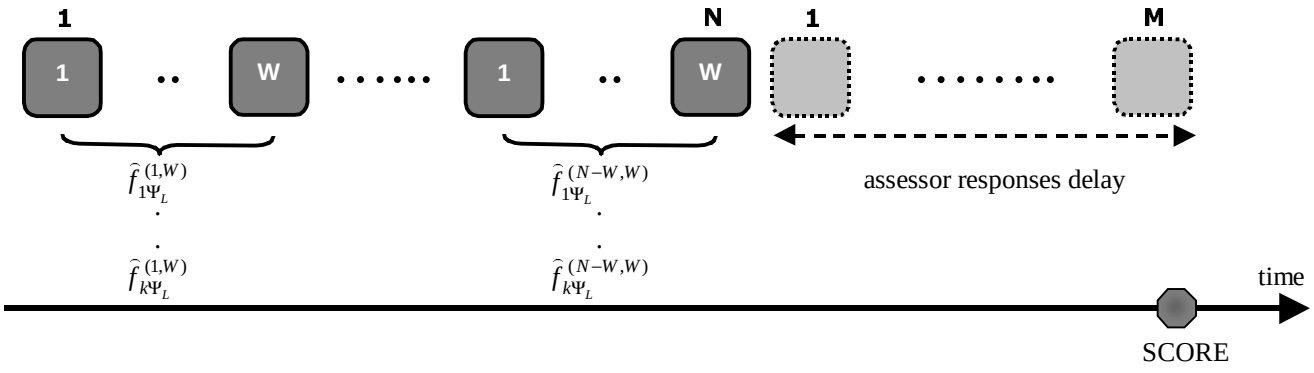

Figure 2 Feature extraction according to perceptual mechanisms.

quantitative, data-driven models. Processing compressed video streams also enhances the method's effectiveness in real-time production applications.

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

APPENDIX A

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).

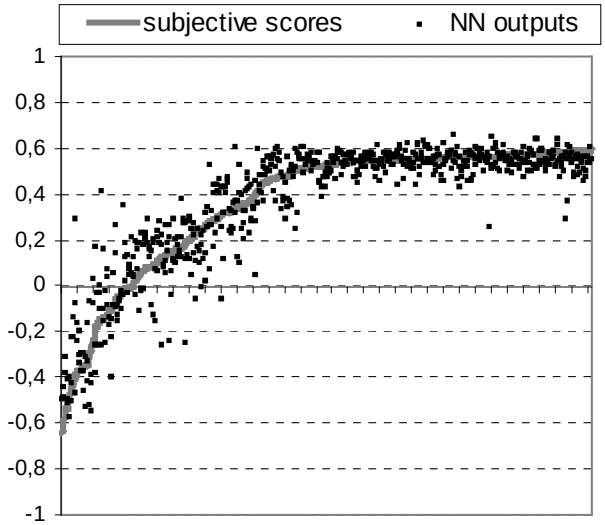

Figure 3 Quality ratings and corresponding NN outputs.

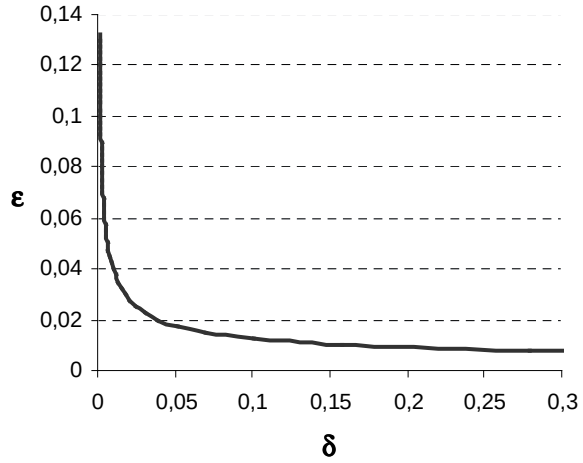

Figure 4 Accuracy ε versus confidence p=1-δ.

2 - Statistic on parameters extracted from a picture: *Smv_av* (mean of motion vectors); *Sq_scale_av* (mean of *q_scale* factors); *Smv_dev_std* (standard deviation of motion vectors); *Sq_scale_dev_std* (standard deviation of *q_scale* factors); *Smv_var* (variance of motion vectors); *Sq_scale_var* (variance of *q_scale* factors).

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).

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] T. Hamada, S. Miyaji, and S. Matsumoto, "Picture quality assessment system by three-layered bottom noise weighting considering human visual perception", SMPTE Journal, vol. 108, no. 1, pp. 20-6, Jan. 1999.

[4] 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, n. 7, pp. 1208-13, Oct. 2000.

[5] T. Vlachos, "Detection of blocking artifacts in compressed video", Electronics Letters, vol. 36, no. 13, pp. 1106-8, June 2000.

[6] ISO/IEC 13818-2: Information technology: Generic coding of moving pictures and associated audio video information: Video, 1994.

[7] 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.

[8] 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.

[9] S. Ridella, S. Rovetta and R. Zunino, "Adaptive internal representation in circular backpropagation networks" Neural Computing and Applications, No. 3, pp. 222-233, 1995.

[10] ITU-R 11E/9, Introduction of a new method for 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.