AN ARTIFICIAL NEURAL NETWORK FOR QUALITY ASSESSMENT IN WIRELESS IMAGING BASED ON EXTRACTION OF STRUCTURAL INFORMATION

Ulrich Engelke and Hans-Jürgen Zepernick

Blekinge Institute of Technology
PO Box 520, SE-372 25 Ronneby, Sweden
E-mail: {ulrich.engelke, hans-jurgen.zepernick}@bth.se

ABSTRACT

In digital transmission, images may undergo quality degradation due to lossy compression and error-prone channels. Efficient measurement tools are needed to quantify induced distortions and to predict their impact on perceived quality. In this paper, an artificial neural network (ANN) is proposed for perceptual image quality assessment. The quality prediction is based on structural image features such as blocking, blur, image activity, and intensity masking. Training and testing of the ANN is performed with reference to subjective experiments and the obtained mean opinion scores (MOS). It is shown that the proposed ANN is capable of predicting MOS over a wide range of image distortions. This applies to both cases, when reference information about the structure of the original image is available to the ANN but also in absence of this knowledge. The considered ANN would therefore be well suited for combination with link adaption techniques.

Index Terms— Artificial neural network, image quality assessment, feature extraction, communication systems.

1. INTRODUCTION

The deployment of third-generation mobile networks has led to a higher adoption of digital multimedia applications such as audio, image, and video. However, the data suffers from impairments through both lossy source encoding and transmission over error-prone channels, eventually resulting in a degradation of quality. Combating these losses requires them to be measured accurately. Traditionally, this has been done with measures like the bit error rate (BER). It has been shown that this type of measures does not necessarily correlate well with the quality as perceived by humans. Therefore, user-oriented objective quality evaluation, taking into account human sensitivity to certain distortions, has received increased attention.

Two approaches have been generally followed in the design of objective image quality metrics which in [1] are referred to as the psychophysical approach and the engineering approach. Metrics following the former approach are mainly based on incorporation of various aspects of the human visual system (HVS). Metrics based on the latter approach utilize image analysis and feature extraction algorithms to perform the quality prediction. These metrics can then be related to human perception by performing subjective experiments.

The most widely used image quality measure is the peak signal-to-noise ratio (PSNR) because of its simplicity and ability to measure distortions over a wide range. However, PSNR is unable to accurately quantify structural distortions and does not account for non-linearities and saturation effects in human vision. Hence, its prediction performance often does not agree with the quality as perceived by human observers. Also, PSNR as a full-reference (FR) metric requires the original image being available for quality prediction. This is generally not the case in a communication system where the receiver does not have access to the original image. In such cases no-reference (NR) or reduced-reference (RR) metrics are preferably used. The former utilizes solely the distorted image for quality evaluation whereas the latter uses additionally a set of extracted features from the reference image (see Fig. 1).

In this paper, image quality assessment is based on feature extraction algorithms accounting for blocking, blur, image activity and intensity masking. This approach is supported by the fact that the HVS is highly adapted to the extraction of structural information [2]. The goal is then to use the feature measures along with mean opinion scores (MOS) obtained in subjective experiments to train and test an artificial neural network (ANN) for image quality prediction. Link adaptation techniques in wireless multimedia systems may benefit from such an ANN.

The paper is organized as follows. In Section 2, the image distortion process, subjective experiments, and feature extraction are described. Section 3 discusses the ANN design, training, and testing. In Section 4, an evaluation of the ANN performance is presented. Section 5 concludes the paper.

2. SUBJECTIVE & OBJECTIVE IMAGE ANALYSIS

2.1. Image Distortion Process

A set $\mathcal{I}_{ref}$ of $L = 7$ reference monochrome images in Joint Photographic Experts Group (JPEG) format was chosen to ac-

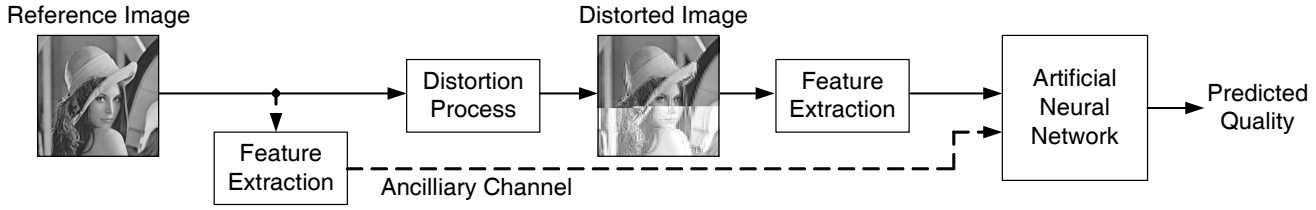

Fig. 1. Network scenario with ANN as no-reference (solid line) or reduced-reference (dashed line) image quality predictor.

count for different textures and complexity. A simple simulation model of a wireless system was used in order to generate a wide range of image distortions. The model comprised of a Rayleigh fading channel with additive white Gaussian noise, a (31,21) Bose-Chaudhuri-Hocquenghem code for error protection and binary phase shift keying as modulation technique.

2.2. Subjective Experiments

The impact of different image distortions on human perception is based on data from two subjective experiments. These were conducted according to ITU-R Rec. BT.500-11 [3] with each experiment involving 30 non-expert observers. The first experiment took place at the Western Australian Telecommunication Research Institute in Perth, Australia. The test persons were shown the distorted images from a set $\mathcal{I}_1$ of size $J = 40$ along with their references. The 30 votes for each distorted image were accumulated to built the MOS vector $\mathbf{s}_1 = [s_j^{(1)}]_{1\times J}$ with $s_j^{(1)} \in [0, 100]$ denoting the MOS value of the j^{th} image in $\mathcal{I}_1$. The second experiment was conducted at the Blekinge Institute of Technology in Ronneby, Sweden. Accordingly, 30 test persons were presented the images from a different set $\mathcal{I}_2$ of size $J = 40$ resulting in a MOS vector $\mathbf{s}_2 = [s_j^{(2)}]_{1\times J}$ with the MOS value of the j^{th} image in $\mathcal{I}_2$ given by $s_j^{(2)} \in [0, 100]$. The test procedure and results of both experiments are extensively reported in [4].

2.3. Feature Extraction

To obtain information about structural degradation in the images that can subsequently be mapped to perceptual image quality, we extracted the following five features for each of the images in the three sets $\mathcal{I}_{ref}$, $\mathcal{I}_1$ and $\mathcal{I}_2$:

- $\tilde{f}_1 \triangleq$ Blocking (Wang et al. [5])
- $\tilde{f}_2 \triangleq$ Blur (Marzilliano et al. [6])
- $\tilde{f}_3 \triangleq$ Edge-based image activity (Saha et al. [7])
- $\tilde{f}_4 \triangleq$ Gradient-based image activity (Saha et al. [7])
- $\tilde{f}_5 \triangleq$ Intensity masking

Accordingly, three matrices containing these feature measures may be defined as

$$\widetilde{\mathbf{F}}_{ref} = [\tilde{f}_{i,l}^{(ref)}]_{I\times L},\ \widetilde{\mathbf{F}}_1 = [\tilde{f}_{i,j}^{(1)}]_{I\times J},\ \widetilde{\mathbf{F}}_2 = [\tilde{f}_{i,j}^{(2)}]_{I\times J} \quad (1)$$

where $\tilde{f}_{i,l}^{(ref)}$, $\tilde{f}_{i,j}^{(1)}$, and $\tilde{f}_{i,j}^{(2)}$, respectively, denote the i^{th} feature measure of the l^{th} and j^{th} image in $\mathcal{I}_{ref}$, $\mathcal{I}_1$, and $\mathcal{I}_2$. Also, the dimensions of these matrices relate to the number of features, $I = 5$, the number of reference images, $L = 7$, and the number of test images, $J = 40$. Given the matrices of (1), a partitioned matrix containing the features of the total of $K = L + 2J = 87$ images may be introduced as

$$\widetilde{\mathbf{F}}_{tot} = [\tilde{f}_{i,k}^{(tot)}]_{I\times K} = [\widetilde{\mathbf{F}}_{ref}|\widetilde{\mathbf{F}}_1|\widetilde{\mathbf{F}}_2] \quad (2)$$

In order to obtain a defined and finite feature space, the feature measures were normalized into an interval using an extreme value normalization [8]

$$f_{i,k}^{(tot)} = \frac{\tilde{f}_{i,k}^{(tot)} - \min\limits_{k=1,\cdots,K}\{\tilde{f}_{i,k}^{(tot)}\}}{\delta_i}, \quad i = 1, \cdots, I \quad (3)$$

where the denominator is computed as

$$\delta_i = \max_{k=1,\cdots,K}\{\tilde{f}_{i,k}^{(tot)}\} - \min_{k=1,\cdots,K}\{\tilde{f}_{i,k}^{(tot)}\} \quad (4)$$

and as a consequence, we have $\forall i, k : 0 \leq f_{i,k}^{(tot)} \leq 1$.

In the case of an RR scenario, the absolute difference between the normalized features of the distorted and reference image may be used to quantify changes in image quality as

$$\Delta f_{i,j}^{(1)} = |f_{i,j}^{(1)} - f_{i,l}^{(ref)}| \ \text{ and } \ \Delta f_{i,j}^{(2)} = |f_{i,j}^{(2)} - f_{i,l}^{(ref)}| \quad (5)$$

to build the elements of the following delta-feature matrices

$$\Delta\mathbf{F}_1 = [\Delta f_{i,j}^{(1)}]_{I\times J} \quad \text{and} \quad \Delta\mathbf{F}_2 = [\Delta f_{i,j}^{(2)}]_{I\times J} \quad (6)$$

3. THE NEURAL NETWORK APPROACH

In view of the results obtained from the subjective experiments and the related structural image information as reported above, the overall aim is to design an ANN that can assess and quantify image quality in terms of predicted MOS. Accordingly, the favorable ANN needs to be trained to find associations between input signals (image features) and a corresponding desired response (predicted MOS). Clearly, the trained neural network should not only be able to map known inputs to known outputs but should also be able to associate unknown inputs to meaningful outputs. In the sequel, we will present the considered feed-forward network architecture and describe its training and testing.

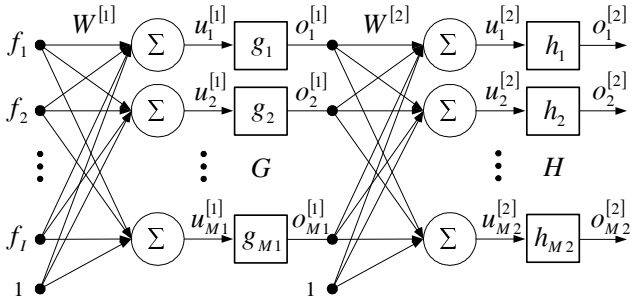

Fig. 2. Fully-connected two-layer neural network structure.

3.1. Feed-forward Network Architecture

In general, a feed-forward ANN consists of multiple layers, in particular an input layer, an output layer, and one or several hidden layers. Each of the layers contains various amounts of neurons. These are processing units composed of a summation part and a transfer function. In a fully-connected network all neurons in a hidden layer have a weighted interconnection to the neurons in the previous and successive layer.

A fully-connected two-layer network architecture with M_1 and M_2 neurons in the first and second layer, respectively, is illustrated in Fig. 2. Here, f_i denotes the i^{th} feature at the network input. The interconnection weights, including biases, to the neurons are stored in the matrices

$$\mathbf{W}^{[1]} = [w^{[1]}_{m_1,i}]_{M_1\times(I+1)},\ \mathbf{W}^{[2]} = [w^{[2]}_{m_2,m_1}]_{M_2\times(M_1+1)} \quad (7)$$

The activation functions in the first and second layer are given as G and H, respectively. The inputs to the activation functions are denoted as $u^{[n]}$ and the outputs as $o^{[n]}$. In general, the superscripts $(\cdot)^{[n]}$ denote the n^{th} layer in the network.

The choice of a suitable architecture (number of layers, neurons per layer, activation functions) is crucial to the performance of an ANN for an intended application. Neural networks of too high complexity tend to easily overfit which means that they function well on the training set but show weak performance on unknown input data. On the other hand, networks of too low complexity might result in large errors for both training and generalization. However, it is well known that any continuous function can be approximated sufficiently well by a two-layer network architecture given a non-linear, differentiable transfer function and sufficient neurons in the first layer and a linear transfer function in the second layer [9]. In view of this finding, we designed a fully-connected two-layer feed-forward network containing one hidden and one output layer. The differentiable bipolar sigmoid function was chosen as activation function g for all neurons in the hidden layer. A linear activation function h was used for the single output neuron. There is no strict design rule regarding the number of neurons in the hidden layer but in our case a choice of 8 neurons has provided best performance.

3.2. Network Training and Testing

Let us refer to the columns of the matrices $\mathbf{F}_1$ and $\mathbf{F}_2$ containing the features of the distorted images of sets $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively, as feature vectors $\mathbf{f}$. Similarly, let us refer to the columns of the delta-feature matrices $\Delta\mathbf{F}_1$ and $\Delta\mathbf{F}_2$ as delta-feature vectors $\Delta\mathbf{f}$. Accordingly, we have 80 feature vectors $\mathbf{f}$ and 80 delta-feature vectors $\Delta\mathbf{f}$ available as network inputs. It should be noted that the feature vectors are used for NR image assessment while the delta-feature vectors support RR image assessment. The related MOS values representing the desired network responses are contained in the MOS vectors $\mathbf{s}_1$ and $\mathbf{s}_2$ deduced from the subjective experiments.

To train the network and also test its ability to generalize unknown inputs we need to split the available feature vectors into two subsets, a training and a test set. The size of the training subset has been chosen as $P = 60$ with 30 feature vectors randomly selected from each of $\mathbf{F}_1$ and $\mathbf{F}_2$. Similarly, this has been done with the delta-features $\Delta\mathbf{F}_1$ and $\Delta\mathbf{F}_2$. The selection was constrained such that the training set contains the minima and maxima of each of the 5 features and delta-features. Therewith, the networks generalization to new input data is eased to an interpolation problem rather than extrapolation to unknown data which might exceed the training data. The training sequences for the NR and RR image quality assessment along with the related MOS are given by

$$\mathbf{F}_{tr} = [f^{(tr)}_{i,p}]_{I\times P},\ \Delta\mathbf{F}_{tr} = [\Delta f^{(tr)}_{i,p}]_{I\times P},\ \mathbf{s}_{tr} = [s^{(tr)}_{p}]_{1\times P} \quad (8)$$

The remaining $Q = 20$ feature, delta-feature, and MOS vectors, respectively, were used to obtain the test sequences:

$$\mathbf{F}_{ts} = [f^{(ts)}_{i,q}]_{I\times Q},\ \Delta\mathbf{F}_{ts} = [\Delta f^{(ts)}_{i,q}]_{I\times Q},\ \mathbf{s}_{ts} = [s^{(ts)}_{q}]_{1\times Q} \quad (9)$$

Due to the relatively small set of training sequences, the networks capability to generalize unknown data is restricted. Therefore, special methods have to be used to improve the generalization of the network. The most widely used techniques are early stopping and Bayesian regularization. The former method requires the data to be divided into three subsets, a training, validation, and test set. On the other hand, Bayesian regularization only needs a training and a test set and is therefore preferably used on smaller data sets. We used the Levenberg-Marquardt algorithm together with Bayesian regularization to train our network. To get the best performance with Bayesian regularization during training we scaled both network inputs and targets to fall in the range $[-1, 1]$. In a post-processing step the MOS have been reverted to fall into their original interval $[0, 100]$. In supervised training the output of the second layer $o^{[2]}$ is compared to the MOS, the desired response s, to establish the error $e = s - o^{[2]}$ which is used to update the network weights $\mathbf{W}^{[1]}$ and $\mathbf{W}^{[2]}$. The trained network is then applied with fixed weights and biased input $\mathbf{f}_b = [\mathbf{f}^T|1]^T$ providing the predicted MOS, p, which is calculated as

$$p = o^{[2]} = H\Big[\mathbf{W}^{[2]} \cdot G\big[\,\mathbf{W}^{[1]}\mathbf{f}_b\,\big]\Big] \quad (10)$$

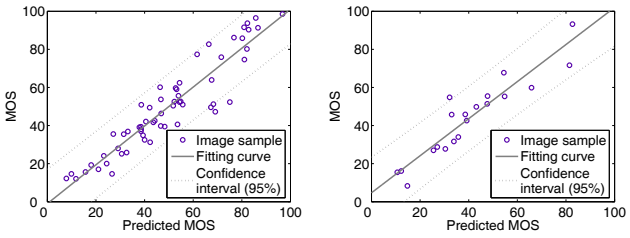

Fig. 3. Linear curve fitting for NR approach: (a) network training with 60 images, (b) network testing with 20 images.

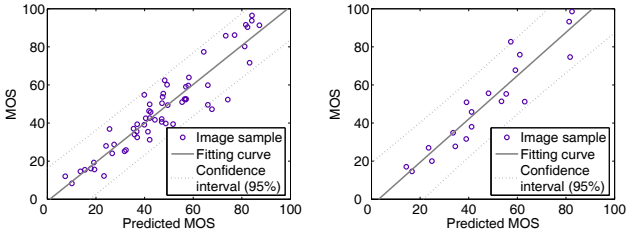

Fig. 4. Linear curve fitting for RR approach: (a) network training with 60 images, (b) network testing with 20 images.

4. NETWORK PERFORMANCE EVALUATION

A linear regression between the predicted MOS, p, at the network output and the MOS, s, obtained from the subjective experiments has been performed. The relationship between them can be expressed as $s = \alpha p + \beta$, where α is called the slope and β the y-intercept of the regression function. For a perfect fit where predicted MOS, p, equals MOS, s, the slope α would be 1 and the y-intercept β would be 0. The results of the curve fitting for the training and test sets are shown in Fig. 3 and Fig. 4, respectively, for NR and RR assessment.

To quantify the accuracy by which the designed ANN predicts MOS has been determined using the Pearson linear correlation coefficient r. The results are summarized in Table 1. The correlation coefficients r_{tr} and r_{ts} and the fitting curve parameters α and β demonstrate a very good prediction performance of the network for both the training (tr) and testing (ts) processes. This shows the networks strong ability to generalize to unknown inputs. It is also noted that the proposed ANN outperforms the previously reported results in [4] which relate to RR image quality evaluation based on weighted feature difference values with a correlation value of 0.869.

It can also be observed from the table that the correlation values for NR and RR assessment are very similar. This suggests that information about the changes in the structural information is not necessarily needed to enhance prediction performance of the ANN as compared to the feature values of the distorted image. Thus, one may deploy the ANN as an NR image quality predictor to save the feature extraction on the reference image as well as transmission of these features over an ancillary channel.

Table 1. Prediction performance.

	r_{tr}	α_{tr}	β_{tr}	r_{ts}	α_{ts}	β_{ts}
NR	0.933	1.02	-1.011	0.931	0.973	4.686
RR	0.932	1.022	-1.018	0.932	1.137	-3.35

5. CONCLUSIONS

In this paper, we designed an ANN for perceptual image quality assessment considering both NR and RR metrics. The feature-based ANN design takes advantage of structural image information. The network was trained and tested using MOS obtained in subjective experiments. An analysis of the prediction performance of the ANN revealed strong ability of the network to associate structural features to perceived image quality in terms of predicted MOS. This applies to both, NR and RR quality assessment. As such, the ANN may be combined with link adaption techniques that can adapt to system dynamics as observed in wireless communications.

6. REFERENCES

[1] H. R. Wu and K. R. Rao (Ed.), *Digital Video Image Quality and Perceptual Coding*, CRC Press, 2006.

[2] Z. Wang et al, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. on Image Processing*, pp. 600–612, April 2004.

[3] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Rec. BT.500-11, 2002.

[4] T. M. Kusuma, *A Perceptual-based Objective Quality Metric for Wireless Imaging*, Ph.D. thesis, Curtin University of Technology, Perth, Australia, 2005.

[5] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. of IEEE ICIP*, Sept. 2002, pp. 477–480.

[6] P. Marziliano et al, "A no-reference perceptual blur metric," in *Proc. of IEEE ICIP*, Sept. 2002, pp. 57–60.

[7] S. Saha and R. Vemuri, "An analysis on the effect of image features on lossy coding performance," *IEEE Signal Processing Letters*, pp. 104–107, May 2000.

[8] J.-R. Ohm, *Multimedia Communication Technology*, Springer, 2004.

[9] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.