Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 20 (2005) 643–661
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Objective quality assessment of displayed images by using neural networks

Paolo Gastaldo[a,*], Rodolfo Zunino[a], Ingrid Heynderickx[b], Elena Vicario[b]

[a] *DIBE, University of Genoa, Via all'Opera Pia 11a, 16145 Genoa, Italy*
[b] *Philips Research Laboratories, Prof.Holstlaan 4, WY 8.08, 5656AA Eindhoven, The Netherlands*

Received 29 June 2004; accepted 18 March 2005

Abstract

Considerable research effort is being devoted to the development of image-enhancement algorithms, which improve the quality of displayed digital pictures. Reliable methods for measuring perceived image quality are needed to evaluate the performances of those algorithms, and such measurements require a univariant (i.e., no-reference) approach. The system presented in this paper applies concepts derived from computational intelligence, and supports an objective quality-assessment method based on a circular back-propagation (CBP) neural model. The network is trained to predict quality ratings, as scored by human assessors, from numerical features that characterize images. As such, the method aims at reproducing perceived image quality, rather than defining a comprehensive model of the human visual system. The connectionist approach allows one to decouple the task of feature selection from the consequent mapping of features into an objective quality score. Experimental results on the perceptual effects of a family of contrast-enhancement algorithms confirm the method effectiveness, as the system renders quite accurately the image quality perceived by human assessors.

Keywords: Perceptual quality; Objective image quality; Neural networks

1. Introduction

Algorithms for digital picture enhancement aim at improving the overall quality of displayed images. The effectiveness of those algorithms is determined by their impact on image quality as perceived by consumers; hence, reliable methods for assessing perceived image quality are needed.

Subjective testing [1–3] is the conventional approach for that purpose; it is essentially based on asking human assessors to judge the overall quality of a set of images. When properly implemented, subjective methods yield accurate

*Corresponding author. Tel. +39 0103532268; fax: +39 0103532175.
E-mail addresses: gastaldo@dibe.unige.it (P. Gastaldo), zunino@dibe.unige.it (R. Zunino), Ingrid.Heynderickx@philips.com (I. Heynderickx).

0923-5965/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2005.03.013

results but are time-consuming and therefore expensive.

Objective models of image quality [4–21], instead, estimate perceived quality while bypassing human assessors. These models predict image quality by processing numerical quantities (''objective features'') extracted from images. To be both consistent and effective, objective methods must match with perceived image quality as measured by subjective testing. Most objective models [4–18] aim at predicting image fidelity or image dissimilarity: the quality measure is based on the difference between a ''distorted'' image (e.g., by noise or compression) and the original image (as a reference). Thus, these models can follow a ''bivariant'' approach. Some methods [4–11,14,17,18] quantify the difference in quality between compared pictures by using a Minkowski metric, whereas others [4–13,15,16] simulate the human visual system by taking into account aspects such as light adaptation, the contrast-sensitivity function, masking, etc.

When, however, one aims at predicting the quality of enhanced images, a bivariant approach is often ineffective. The difference between ''processed'' and original images is expected to improve, rather than degrade, image quality; thus, this difference is not necessarily a good measure of the resulting perceived quality. In the specific case of picture-enhancement algorithms, one has to rely on a ''univariant'' approach, i.e., to assess the perceived quality from processed images only. Univariant approaches have recently been proposed for that purpose [10,11,18–21], and are based on a non-linear function of image features. However, existing univariant approaches have been designed to assess the quality of compressed images rather than enhanced images.

This paper presents a method that uses neural networks [22] for the objective assessment of enhanced pictures. A circular back-propagation (CBP) feedforward network [22] processes objective features extracted from an enhanced image, and returns the associated quality score. As the proposed method does not require information from the original image, it should be considered as a univariant method. The model exhibits several similarities to the neural-network system effectively used for the quality assessment of MPEG video streams [23]. In both cases, the approaches exploit the ability of feedforward neural structures to support a general paradigm for complex mathematical models. The overall goal of both methods is to mimic perceived image quality, rather than design an explicit model of the human visual system. This reduces the number of assumptions that are typically needed to model perceived image quality analytically; moreover, the problem of selecting objective features can be decoupled from the design of the function that maps those features into quality rates. With respect to [23], this paper shows that the neural-based framework can be effectively applied to quality assessment of still, uncompressed images. To achieve this goal, the present research proposes a non-parametric feature-selection criterion based on the Kolmogorov–Smirnov test to define an effective objective metric; furthermore, the introduction of an ensemble strategy into the neural-network architecture allows the model to reduce the variance in the estimated quality values.

Section 2 justifies the choice of a univariant paradigm. Section 3 describes the feature-selection criteria and the design of the set of objective metrics. Section 4 presents the CBP neural model and discusses its specific advantages in imaging applications. Section 5 reports on experimental results supporting the proposed approach. Some concluding remarks are made in Section 6.

2. Univariant approach to image-quality assessment

The lack of universally reliable objective models of image quality motivates the search for new methods that can automatically measure the quality of enhanced pictures. The overall problem can be set formally as follows. Denote by $\mathbf{I}$ the space of all possible images; a filter, $\gamma()$, can be viewed as an isomorphism that maps each image into the associate enhanced image: $\gamma : (\mathbf{I} \rightarrow \mathbf{I})$. Thus, the perceptual phenomenon is represented by a function, $Q(I, \gamma(I))$, which associates the pair of images $\{I, \gamma(I)\}$ (original and enhanced images, respectively) with a scalar measure of the

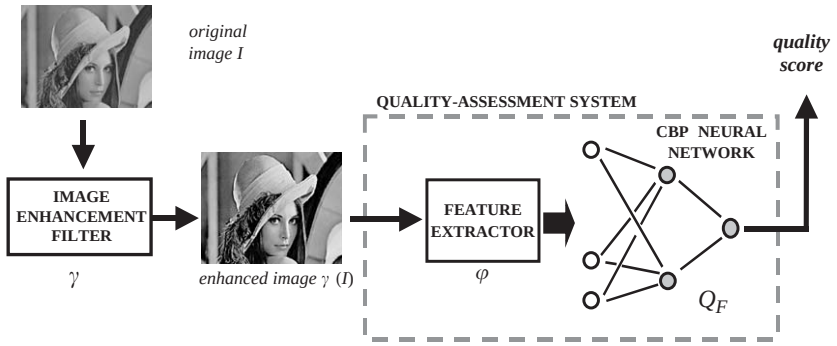

Fig. 1. The neural network system for assessing image quality.

perceived quality:

$$Q : \mathbf{I} \times \mathbf{I} \rightarrow [-1, +1]. \tag{1}$$

Here the quality score is normalized to the range $[-1, +1]$ without loss of generality.

The univariant paradigm assumes that only the subset of filtered (processed) images of Q is actually available to the end user. Thus, although the mapping to be modeled still remains (1), the perceptual phenomenon can be studied by projecting the input domain onto the subspace, $\mathbf{I}'$, of filtered images only:

$$Q : \mathbf{I}' \rightarrow [-1, +1]. \tag{2}$$

The problem setting (2) gives the scientific basis for univariant approaches [10,11,18–21]. The performance requirement implies that any approach to the implementation of (2) must consistently reproduce human perception, which, in any case, has to be measured experimentally. This gives rise to a major problem of statistical inference, as the virtually infinite size of the input space, $\mathbf{I}'$, makes an exhaustive-search approach unfeasible. The only reasonable way of limiting the statistical complexity of the problem (2) seems to be a reduction in the data-space dimensionality. In the area of image processing, this can be done by exploiting a *feature-based* representation of images. If $\varphi(I)$ denotes the feature-extracting operator that maps the image space, $\mathbf{I}'$, into the (lower-dimensional) feature space, $\mathbf{F}$, the original problem (2) is turned into the lower-dimensional formulation:

$$Q = Q_F \circ \varphi, \tag{3}$$

where the quality-assessment phenomenon is mapped by the operator $Q_F : \mathbf{F} \rightarrow [-1, +1]$.

The present study of objective quality assessment is based on the problem setting (3) and is illustrated in Fig. 1. The formulation (3) highlights the advantages of the divide-and-conquer strategy implemented by a feature-based representation. It allows one to reduce complexity by decoupling the prediction of image quality into two tasks: (1) selection of the features $\varphi(I)$, which, when chosen properly, result in an effective descriptive basis for the images, and (2) design of the mapping function Q_F, which may be highly non-linear and even mimic unknown perceptual mechanisms.

To accomplish the first task, a statistical analysis of possible descriptors was made in order to identify those able to provide useful information about the quality-related image characteristics. The second task takes advantage of the ability of CBP structures to deal with multidimensional data characterized by complex relationships, which are learned from examples by using a training algorithm. The filter chosen for this study is a contrast-enhancement filter. Hence, the set of natural images are limited to gray-scale pictures represented by 8 bits per pixel.

3. Feature-based description of images

3.1. Block-based description of images

The goal of the present research is to evaluate the validity of the approach described in Fig. 1 to predicting the quality of enhanced images. The

objective features characterizing the images are determined at signal level, i.e., they are based on pixel values. In principle, one can define features characterizing an image at a global level, such that each feature gives a single value for the whole picture. In practice, natural images are often too complex to be analyzed at a global level: the perceived quality of a picture may be conditioned by detail-related issues, which might not be considered by global image descriptors. Therefore, in this work, objective features are extracted from an image on a block-by-block, local basis for the purpose of characterizing effectively the space-variant nature of perceptual mechanisms. For the block size, a value of 32×32 pixels is chosen mainly because this value gives the best tradeoff between the rendering of local details and the need for reducing space dimensionality. The following notation indicates the basic quantities used throughout the paper.

Notation and conventions

- $\Psi = \{I^{(s)};\ s = 1, \ldots, n_p\}$ is the set of original images.
- $\gamma_1(\cdot)$, $(l = 1, \ldots, n_e)$ is the family of n_e enhancement filters, whose quality effects must be evaluated.
- $B^{(I)} = \{b_q^{(I)}; q = 1, \ldots, n_b\}$ is the set of blocks obtained by splitting the image I into n_b squares.
- $\Phi = \{f_k; k = 1, \ldots, n_f\}$ is the set of n_f objective features describing an image.
- $f_{kq}^{(s)}$ is the value of the feature $f_k \in \Phi$ for the qth block of the sth image $I^{(s)} \in \Psi$.
- $\widehat{I}^{(s,l)} = \gamma_l(I^{(s)})$ is the enhanced image obtained by processing $I^{(s)} \in \Psi$ by the th filter.
- $\widehat{f}_{kq}^{(s,l)}$ is the value of the feature $f_k \in \Phi$ for the qth block of the image $\widehat{I}^{(s,l)}$.
- Throughout the paper, as a general convention, the rounded-hat symbol $\widehat{\ }$ will denote quantities measured after the application of an enhancement filter.

From a modeling perspective, one must take into account that human assessors usually generate one overall quality score per image. Hence, somehow the block-based information has to be transferred into one vector per image, which has to be associated with this single score. To achieve this goal, the framework assembles such a vector by global-level statistical descriptors of the f_k as follows.

The construction algorithm for neural-network inputs

(1) Given an image $\widehat{I}^{(s,l)}$ enhanced by the lth filter, compute the following quantities:

$$f_k^M = \text{median}\left(\widehat{f}_{k1}^{(s,l)}, \ldots, \widehat{f}_{kn_b}^{(s,l)}\right),$$
$$f_k^S = \text{stdev}\left(\widehat{f}_{k1}^{(s,l)}, \ldots, \widehat{f}_{kn_b}^{(s,l)}\right),\ k = 1, \ldots, n_y, \tag{4}$$

where n_y is the number of active features that have been selected for the set Y.

(2) Assemble the vector $\vec{x}^{(s,l)}$ for the image $\widehat{I}^{(s,l)}$ as

$$\vec{x}^{(s,l)} = \{f_k^M, f_k^S; k = 1, \ldots, n_y\}. \tag{5}$$

The median operator is adopted because of its inherent robust statistical behavior. As a result, in the neural-network quality-assessment system, each image is represented by a unique input pattern, which is obtained by applying a two-step methodology: first, the image $\widehat{I}^{(s,l)}$ is analyzed on a block-by-block basis; secondly, statistical descriptors are used to generate the objective vector $\vec{x}^{(s,l)}$, whose dimensionality $d = 2n_y$ is twice the number of features included in the eventual objective metric.

The block-based description $B^{(I)}$ of the image can be designed according to two different approaches. The first approach uses overlapping blocks, whereas the second is designed to split the image into non-overlapping squares. The last methodology might be suboptimal as the block edges may interfere with the extraction of appropriate features, but it has the advantage of lower computational complexity. In view of this, the approach based on overlapping blocks should only be preferred in the case where it allows a more effective statistical description of the image.

3.2. Features definition

The features in Φ describe the image content in terms of luminance distribution, spatial orientation, frequency energy distribution, etc. They can be grouped into three families:

1. Features derived from the first-order histogram of image blocks, which describe the probabilistic distribution of gray levels within a picture.
2. Features derived from the co-occurrence matrix (also called "second-order histogram").
   $C(g_i, g_j, r, \omega)$ denotes the co-occurrence matrix [24] associated with the probability distribution of pairs of pixels with gray levels g_i and g_j, respectively, and are separated by r radial units at angle ω to the horizontal axis. Features derived from $C(g_i, g_j, r, \omega)$ are effective for the characterization of textural properties [24,25].
3. Features derived from a frequency-based representation.
   Information on image complexity can also be derived from the image's frequency energy content, which is described by the discrete cosine transform (DCT) of each block [26].

The complete set of features is listed in Appendix A.

3.3. A statistical approach to feature selection

A subset of extracted features may be insignificant or redundant to describe the part of an image that determines its perceived quality. The present study uses a statistical approach to selecting only those features that seem to carry most of the information about the effects of image-enhancement filters on the perceived quality.

The analysis starts from the complete feature set, Φ, and selects only the subset of statistically 'active' features. A feature is active if its statistical properties differ significantly from their original values after the application of an enhancement filter. Thus, for each objective feature $f_k \in \Phi$, the analysis compares the statistical properties of two samples: one contains the values of f_k for a set of original, unfiltered images, the other holds the values of f_k for a set of enhanced images. To guarantee the statistical independence of the two samples, the two sets of images are disjoint. This means that the original images used to create the (latter) set of enhanced pictures cannot be included in the (former) set of unfiltered pictures. The feature values are worked out on non-overlapping blocks of pixels randomly extracted from each image.

If the two data sets for f_k do not appear to have been drawn from the same distribution, then f_k is selected as an 'active' feature. Toward this end, the mutual independence of the data sets allows one to use the Kolmogorov–Smirnov test (denoted by KS for short) [27]. KS is the most widely accepted test for evaluating differences between continuous distributions, and has been preferred to parametric tests such as Student's t-test because one usually cannot assume a normal distribution of the data sets involved. In this case, KS is used to disprove the null hypothesis, i.e., the two data sets are drawn from the same population. The feature selection algorithm can be outlined as follows.

The feature-selection algorithm

0. (*Set-up: normalization factors*)
   For each objective feature $f_k \in \Phi$, $k = 1, \ldots, n_f$;
   for each image $I^{(s)} \in \Psi$, $s = 1, \ldots, n_p$
   0.a. Apply the n_e filters, split each resulting image (including the original ones) into n_b blocks, and compute the exhaustive set of feature values, Ω_k:

$$\Omega_k = \bigcup_{s=1}^{n_p} \left\{ \left\{ f_{kq}^{(s)}; q = 1, \ldots, n_b \right\} \cup \left\{ \widehat{f}_{kq}^{(s,l)}; l = 1, \ldots, n_e; q = 1, \ldots, n_b \right\} \right\}, \in k = 1, \ldots, n_f. \quad (6)$$

   0.b. Calculate the .05 and the .95 percentiles ($x_{.05}^{(k)}$ and $x_{.95}^{(k)}$, respectively) for the values in Ω_k.
1. (*Data set construction*)
   For each enhancement filter ($l = 1, \ldots, n_e$)
   1.a. Create two disjoint sets, $\Psi_1^{(l)} \cap \Psi_2^{(l)} = \varnothing$, each resulting from randomly extracting $n_d \leqslant n_p/2$ images from Ψ.

1.b. Apply the filter γ_l to every element of $\Psi_2^{(l)}$ to obtain $\widehat{\Psi}_2^{(l)} = \left\{ \widehat{I}^{(m,l)}; \forall I^{(m)} \in \Psi_2^{(l)} \right\}$.

1.c. Compute each feature $f_k \in \Phi$ $(k = 1, \ldots, n_f)$ for each image, and generate the sets $\Lambda_{1k}^{(l)}$ and $\Lambda_{2k}^{(l)}$:

$$\Lambda_{1k}^{(l)} = \left\{ f_{kq}^{(m)}; \forall I^{(m)} \in \Psi_1^{(l)}; q \in B^{(I^{(m)})} \right\},$$
$$\widehat{\Lambda}_{2k}^{(l)} = \left\{ \widehat{f}_{kq}^{(m,l)}; \forall \widehat{I}^{(m,l)} \in \widehat{\Psi}_2^{(l)}; q \in B^{(\widehat{I}^{(m,l)})} \right\}. \quad (7)$$

We recall that $B^{(X)}$ is the set of n_b non-overlapping blocks extracted from the image X.

1.d. Normalize each element of $\Lambda_{1k}^{(l)}$ and $\widehat{\Lambda}_{2k}^{(l)}$ to the range $[-1,1]$

$$\underline{f}_{kq}^{(m)} \stackrel{\text{def}}{=} 2\frac{\left(f_{kq}^{(m)} - x_{.05}^{(k)}\right)}{\left(x_{.95}^{(k)} - x_{.05}^{(k)}\right)} - 1,$$
$$\underline{\widehat{f}}_{kq}^{(m,l)} \stackrel{\text{def}}{=} 2\frac{\left(\widehat{f}_{kq}^{(m,l)} - x_{.05}^{(k)}\right)}{\left(x_{.95}^{(k)} - x_{.05}^{(k)}\right)} - 1. \quad (8)$$

Let $\underline{\Lambda}_{1k}^{(l)}$ and $\underline{\widehat{\Lambda}}_{2k}^{(l)}$ be the normalized sets of (7), respectively, including the values calculated in (8).

2. (*Kolmogorov–Smirnov test*)
Assemble a probability vector, $\vec{p}$,defined as

$$p[k,l] = p_{\text{KS}}(\underline{\Lambda}_{1k}^{(l)}, \underline{\widehat{\Lambda}}_{2k}^{(l)});$$
$$k = 1, \ldots, n_f; \quad l = 1, \ldots, n_e \quad (9)$$

where $p_{\text{KS}}(\cdot,\cdot)$ is the significance result of the KS test under the null hypothesis that the data sets $\underline{\Lambda}_{1k}^{(l)}$ and $\underline{\widehat{\Lambda}}_{2k}^{(l)}$ have been drawn from the same distribution.

3. (*Feature ranking*)

3.a. Set a reference confidence threshold, e.g. $p^* = 0.1$

3.b. Compute the indicator vector, $\vec{t}$, as

$$t[k,l] = \begin{cases} 1 & p[k,l] \leqslant p^* \\ 0 & p[k,l] > p^* \end{cases} \quad k = 1, \ldots, n_f; \quad l = 1, \ldots, n_e, \quad (10)$$

3.c. Assemble the occurrence vector, $\vec{o}$, whose kth element counts, over all possible filters, the event "the data sets $\underline{\Lambda}_{1k}^{(l)}$ and $\underline{\widehat{\Lambda}}_{2k}^{(l)}$ are not drawn from the same distribution"; thus, $0 \leqslant o[k] \leqslant n_e$, and

$$o[k] = \sum_{l=1}^{n_e} t[k,l], \quad k = 1, \ldots, n_f. \quad (11)$$

4. (*Output*)
Assemble the final feature set by including those features f_k for which $o[k]$ exceeds a threshold, $o^* \leqslant n_e$:

$$f_k \in Y \Leftrightarrow o[k] \geqslant o^*. \quad (12)$$

The above algorithm gathers in the set Y the features whose statistical properties are significantly altered by the enhancement filters. The threshold value o^* determines the number of filters for which the event "the statistical properties of the feature f_k are altered (as per the KS result) by the enhancement filter" must occur in order to allow one to select f_k from Φ. Its value is set empirically because it depends (1) on the shape of $o[k]$ and (2) on the expected dimension supported by Y. The final validation of the set Y comes from the performance of the whole system that assesses the objective image quality.

The feature-selection algorithm uses the information about the original images, yet the overall objective approach can still be considered as univariant. Indeed, the original images only help select the important features for the whole set of enhancement filters, and such a selection should be done only once at start-up. At run time, those features are evaluated on the enhanced images only, and enter the CBP neural network to estimate the related quality score.

4. Neural networks for image-quality estimation

4.1. The circular back-propagation model

The role of a feedforward neural network is to map feature-based image descriptions into scalar

values, which should represent the perceived image quality. Efficiency requirements (i.e., the storage size of the parameters) and generalization issues (i.e., the NN performance over data not used for training) drive the design of the neural-network model. The computational paradigm of feedforward neural networks [28] aims at implementing a stimulus-response behavior by properly combining several layers of elementary units ('neurons'); the resulting structure supports a unidirectional flow of information. Each unit involves a simple, nonlinear transformation of weighted inputs, and theory proves that feedforward networks embedding a sigmoidal non-linearity can support arbitrary mappings. The MultiLayer Perceptron (MLP) model [28] belongs to this class of networks, and applied research shows that MLP performs effectively whenever few computing units with a global scope can determine the target-mapping function.

The "Circular Back Propagation" (CBP) network [22] extends the conventional MLP model by adding one input, which is the sum of the squared values of all the network inputs. The quadratic augmentation does not affect the fruitful properties of the MLP structure. CBP networks can map both linear and circular separation boundaries. Moreover, the selection of either model is entirely data-driven and comes from the empirical training process: the selection of a model does not require any a priori assumption. Such an adaptive behavior makes CBP networks suitable for application to perceptual problems, whose domain structure is often obscure.

The CBP architecture involves two layers of neurons, as illustrated in Fig. 2. A d-dimensional vector, $\vec{x}$, supplies the input feature values, computed as described in Section 3. Those quantities connect to an intermediate *hidden* layer, including n_{h} neurons. First, each hidden neuron weights the input values by a specific set of coefficients; then, it applies a sigmoidal non-linearity:

$$a_u(\vec{x}) = \mathrm{sigm}\left(w_{u,0} + \sum_{k=1}^{d} w_{u,k} x_k + w_{u,d+1} \sum_{k=1}^{d} x_k^2 \right), \quad u = 1, \ldots, n_{\mathrm{h}}, \tag{13}$$

where $\mathrm{sigm}(r_u) = (1 + \mathrm{e}^{-r_u})^{-1}$, $\{w_{u,k}\}$ is the set of coefficients ("weights") and $w_{u,0}$ is a bias term. The last, quadratic term in the argument of the sigmoid represents the additional input to the conventional MLP; the notations r_u and a_u conventionally stand for the stimulus and activation of the uth neuron, respectively. The *output* layer provides the final response y (i.e., the assessment of perceived quality) by a similar transformation

$$y(\vec{x}) = \mathrm{sigm}\left(w'_0 + \sum_{u=1}^{n_h} w'_u a_u(\vec{x}) \right), \tag{14}$$

where $\{w'_u\}$ and w'_0 represent the output coefficients and the output bias, respectively.

A neural network can be regarded as a non-linear computing device having the set of weights as its own degrees of freedom; the training process adjusts those coefficients in such a way that the network is able to reproduce the desired input/output mapping. Toward this end, except for trivial cases, one has a sample-based formulation of the input/output behavior, which is described by a training set of input patterns with their expected responses. The empirical nature of the training sample makes the adjustment process data-driven, and is also the main reason for the remarkable flexibility of neural-network models.

For a given setting of the weights, W, that characterize a neural network, a performance cost measures its mapping accuracy on the training set. In the case of MLPs, the usual cost function is the mean square error, E_W, between the expected responses (the image quality scores from human assessors) and the actual network outputs. Thus, the network-training process is regarded as an optimization problem, which can be expressed as

$$\min_W E_W = \min_W \frac{1}{n_p} \sum_{s=1}^{n_p} \left[t^{(s)} - y(\vec{x}^{(s)}) \right]^2, \tag{15}$$

where n_p is the cardinality of the training set (one pattern per image), and $t^{(s)}$ and $y(\vec{x}^{(s)})$ denote the desired and actual network outputs, respectively, for the training pattern, $\vec{x}^{(s)}$.

In practice, the learning problem (15) is tackled efficiently and effectively by the back-propagation algorithm [28] (BP), which uses a stochastic gradient-descent strategy over the weight space.

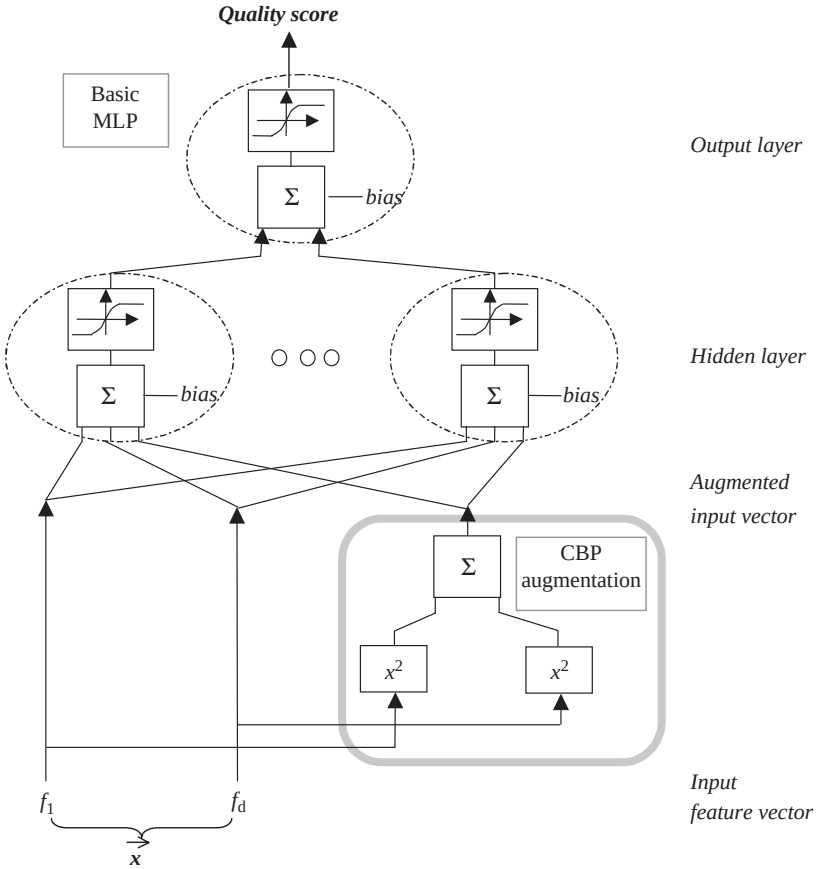

Fig. 2. The CBP additional input augments the MLP feedforward architecture.

The availability of this powerful tool represented the boosting factor to the practical impact of the MLP neural model. The research presented in this paper adopted an accelerated version [29] of classical BP in order to further increase convergence speed.

4.2. Using ensembles of CBP networks

A neural network designed and trained as described in Section 4.1 operates as an estimator whose predictions are always subject to some error, due to statistical fluctuations of the empirical sample drawn to form the training set. A typical approach to increasing the reliability of the neural stage is to replace the single network with an "ensemble" of different estimators [30] trained on the same problem. The statistical reason for this procedure is that the error, ε^2, on the estimate with respect to the sample can be described by a bias/variance decomposition [31]:

$$\varepsilon^2 = \beta^2 + \sigma^2. \tag{16}$$

The bias term, β^2, comes from using a possibly inappropriate estimation model, hence it results in a fixed offset and can be measured. The second term, the variance (σ^2), is due to statistical noise in training data and can be minimized by simply averaging over many independent realizations [30]. As such, the parallel use of several networks (Fig. 3a) can help reduce the variance of the overall quality assessment. In principle, a set of N statistically independent elements can ideally reduce the estimation variance to $\bar{\sigma}^2 = \sigma^2/N$. Such a straightforward approach has been successfully

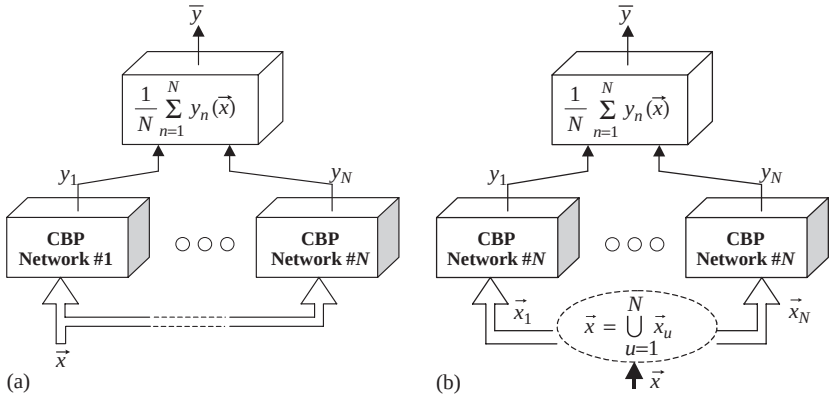

Fig. 3. Ensemble structures for the CBP-based neural estimator: (a) an ensemble of parallel independent networks, (b) an ensemble of specialized estimators.

exploited in various real applications of neural networks [32,33].

The crucial issue, however, is to obtain independent estimators. The simplest approach consists in randomly splitting the training data into as many disjoint subsets as ensemble elements in order to ensure the absence of correlation. As this requires a huge number of empirical samples, a more frequent solution is to train different networks on the same data set, but starting the optimization process under different initial conditions. This method suffers from the drawback that the obtained estimators may usually prove to be highly correlated.

In those cases where few patterns are available as compared with the data dimensionality, an alternative approach consists in partitioning the input space into several subspaces and in training a specialized neural network for each subspace. The subspaces are typically disjoint; averaging the outcomes of the local estimators provides the overall estimate. The following section will show that the particular nature of the image features that compose the input vector, $\vec{x}$, allows one to split $\vec{x}$ into N subvectors of lower dimensionality. These subvectors form the training sets for each network in the ensemble (Fig. 3(b)). The cognitive rationale is that the data-space partitions contribute to the global estimation task in a coordinate, but (ideally) independent, fashion.

4.3. Generalization issues in feature selection and network design

The problem of designing an estimator that proves effective at run time ultimately involves the complexity of the trained model, which should be kept to a minimum. In fact, taking into account generalization issues while designing a system is very difficult. Except for some empirical criteria [34] that prove effective in real applications, the literature does not provide established, theoretical guidelines that also exhibit practical applicability. In the present context, complexity is determined by the dimensionality, d, of the data space and by the number, N_h, of neurons in a network.

The former parameter calls for effective feature selection: in principle, the features should be chosen such that they do not carry too redundant information. Algorithms such as mutual information feature selection (MIFS) [35] can be used to select features with minimal redundancies among them. This will also increase the likelihood that the overall performance obtained by combining the outputs of the neural nets in the ensemble is better than the individual performance. When the features are redundant, the obtained estimates of the quality are not independent, and thus the variance does not decay as σ^2/N. On the other hand, one should consider that this family of methods supports a supervised paradigm, as the selection principle somehow takes into account the

desired estimator output. Choosing the features in compliance with their relevance to the specific quality-mapping task exposes the overall estimation system to the risk of overfitting (poor performance when processing unseen data).

In this respect, recent research [36,37] proved that a system's generalization error can be sharply bounded if the representation of input data is developed independently of the specific mapping task that is expected from the network. In compliance with the formalism (2), setting up the mapping function Q_F in (3) by an unsupervised analysis notably reduces overall complexity; this is the main rationale behind the statistical feature-selection criterion adopted in the present research. As the subjective judgments associated with images are never taken into account while sifting the overall feature set Φ, the eventual subset, Y, of objective features is assembled independently of the actual quality-estimation application.

5. Experimental results

5.1. Experiment setup

The neural-based model for image quality assessment was experimentally tested by using a family of contrast-enhancement filters. The filters adjusted the luminance levels of an image by enhancing only the pixels that belonged to a region containing noticeable details. The algorithm that modified the luminance value L^{old} of a pixel adopted the following procedure.

1. From the 5×5 region surrounding the considered pixel, determine the local contrast level ΔL (which estimates the luminance variation around the pixel) and the local luminance variance S (which is an estimator for high spatial frequencies). The local contrast level is given by $\Delta L = |L - L_{\text{m}}|$where L is the pixel luminance and L_{m} is the mean luminance of the surrounding area.
2. Set the luminance value of the pixel to

$$L^{\text{new}} = L_{\text{m}} + (L^{\text{old}} - L_{\text{m}})G, \quad G = f(\Delta L, S, \tau, \lambda), \tag{17}$$

where the gain G is either α or 1. The value of G depends on two parameters: τ, a threshold for the local contrast level, and λ, a threshold for the local luminance variance.

The latter quantities prevented lowly detailed, high-contrast regions in the image from being further enhanced, whereas mainly highly detailed, low contrast regions needed contrast enhancement. In summary, the filters relied on three parameters: two thresholds, τ and λ, and a gain value, α. In the present research, τ varied over three values, $\{\tau_n;\ n = 1, \ldots, 3\}$, λ over two values $\{\lambda_m;\ m = 1, 2\}$, and α spanned five values $\{\alpha_k;\ k = 1, \ldots, 5\}$. As a result, the family γ_l $(l = 1, \ldots, n_{\text{e}})$ of enhancement filters to be assessed in the quality-evaluation experiments comprised a set of $n_{\text{e}} = 30$ members.

The enhancement filters processed a library of $n_i = 16$ gray-scale images (252×189 pixels in size), whose contents varied from natural images to texture-like patterns (Fig. 4). Applying the family of $n_{\text{e}} = 30$ contrast-enhancement filters, as defined in (17), to the set of $n_i = 16$ original images yielded a sample of 480 enhanced images. Subjective evaluations of the enhanced images were collected with a double-stimulus setup: human assessors were asked to score the quality of each enhanced image against that of the original one. Since enhancement could result in a perceived quality being higher as well as lower (e.g. in the case of over-enhancement) for the enhanced image as compared with the original one, a double-ended scoring scale was used: the negative side reflected enhanced images being worse than the original image, whereas the positive side reflected enhanced images being better than the original one. To offer the subjects sufficient granularity at both ends of the discrete scale, an 11-point numerical scale ranging from "−5" to "+5" was used. Subsequent normalization of that scale to the range [−1;1] made these scores compatible with the neural-network output representation. The evaluation of the total set of 480 enhanced images was done according to a subgroup-based design [38], in which eighty people, mostly naïve viewers, participated. The design was such that the total group of participants provided 10 quality scores per enhanced image.

Fig. 4. The set of original images used for subjective testing.

Because of the large number of subjects required, the experiment was run on the Internet. Users were requested to adapt the spatial resolution and number of colors of their monitors to standard settings and to judge the images at a viewing distance usual in computer monitor applications, i.e., .5 m. The disadvantage of running a subjective experiment on the Internet is that there is no control of the type of monitor used to display the images, nor of the ambient illumination in a room. The practical advantage, however, is that a large group of subjects can be assessed in a relatively short time, and that the group of subjects better represents a random sample of the population.

5.2. Block-based description of an image

The procedure described in Section 3 requires the objective features to be first extracted from an image on a block-by-block, local basis, then assembled into global-level statistical descriptors that feed the neural network. The strategy to define blocks, $B^{(I)}$, may involve either non-overlapping or overlapping pixel square regions. The latter option notably increases the computational complexity of the process; hence, it should be chosen only if it leads to a more effective image representation of the quality phenomenon than the former option. Therefore, we compared both options to ascertain which choice would support the block-extraction mechanism properly.

The comparison followed a statistical approach, and observed the behavior of the objective features $f_k \in \Phi$ over a set of images, $\tilde{I}^{(s,l)}$. For each image, the analysis compared the statistical properties of two samples: the first sample held the values of features computed on non-overlapping blocks, whereas the second included the values of the same features for overlapping blocks. As usual, the statistical KS test was used to check the null hypothesis, namely, whether the two data sets had been drawn from the same distribution. The experiment involved the entire collection of objective features in the set Φ (see Appendix A).

For the analysis, a subset of 160 of the 480 enhanced images was used (the subset was created by processing the 16 original images by ten of the 30 available filter settings.) For each feature $f_k \in \Phi$ of each image of the subset, the pair of values calculated on overlapping and non-overlapping blocks underwent the KS test. Table 1 presents an excerpt of the results, and gives the KS-test significance levels for a subset of four images. The ten filter settings are indexed by capital letters in the table columns. The rows refer to the features $f_k \in \Phi$ as follows:

- *f1–f6*: features derived from the first-order normalized histogram;
- *f7–f16*: features derived from the co-occurrence matrix ($r = 2$, $\omega = 0^\circ$);
- *f17–f20*: features derived from the DCT.

Empirical findings provided strong evidence that, apart from minor exceptions, the null hypothesis could not be disproved. Hence, one could reasonably assert that, in most cases, the two samples involving non-overlapping and overlapping blocks, respectively, seem to be drawn from the same population. As a consequence, for the feature-extraction process of the quality-assessment system, a non-overlapping block strategy was chosen, as this option required a smaller computational overhead.

5.3. Quality assessment setup

The feature-selection procedure described in Section 3 defined the input vector for the neural-network system. Table 2 lists the resulting set of (four) descriptive features, all derived from the co-occurrence matrix, whose formal definitions are given in Appendix A. Those descriptors depended on two parameters, i.e., r and ω; the experiments adopted a fixed value of $r = 2$, meaning that the co-occurrence matrix was always computed within a neighborhood radius of two pixels. This setting was assessed empirically by measuring the relative advantages of using larger radius values; experiments showed that settings $r > 2$ did not bring significant benefits, but only implied an increased computational burden. The angular orientation parameter, ω, spanned the four principal directions, i.e., 0°, 45°, 90°, 135°. This led to a number $n_y = (4 \text{ descriptors} \times 4\ \omega \text{ settings}) = 16$ of selected features. As a consequence of (5), the eventual input-space dimension for the neural network amounted to $d = 2n_y = 32$.

A cross-validation approach [39] measured the performance of the quality-assessment system. The available sample was randomly divided into a training set and a test set, including 360 and 120 enhanced images, respectively. During the system setup and the training process, only the training set was used, whereas the test set was applied exclusively to measure the system generalization ability.

As the input space had a considerable dimensionality, as compared with the sample size of 360 enhanced images in the training set, an ensemble strategy using a partitioning of the input space and specialized networks was justified. The design of the assessment system led to $N = 4$ different neural networks; thus each ensemble element was entrusted with a specific angular orientation of the co-occurrence matrix. Table 2 illustrates the input-space partitioning, showing that the input vector to each ensemble element z_i includes the four features parametrized by $\omega = \omega_i$. This reduced the input-space dimensionality of each neural network to $d_i = 8$. For each ensemble element, the number of neurons in the hidden layer was empirically set to $N_h = 10$. No significant advantage resulted from increasing the number of neurons.

5.4. Quality-assessment measurements

The method effectiveness is based on the performance of each ensemble estimator, $\{z_i, i = 1, \ldots, 4\}$,

Table 1
Statistical KS-test results on four images for a comparison between overlapping and non-overlapping blocking strategies

	Image 1										Image 2									
	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
(a)																				
$f1$	.92	.92	.96	.92	.92	.92	.92	.92	.92	.96	.76	.82	.88	.77	.71	.76	.76	.88	.88	.77
$f2$	.92	.96	.98	.98	.98	.92	.96	.96	.92	.96	.96	.92	.83	.83	.83	.98	.96	.83	.71	.88
$f3$	.88	.88	.88	.77	.83	.88	.88	.88	.92	.98	.58	.88	.75	.73	.88	.60	.81	.73	.78	.82
$f4$	.77	.52	.88	.88	.88	.83	.77	.98	.99	.98	.71	.58	.92	.77	.77	.64	.64	.64	.71	.92
$f5$	.46	.77	.78	.74	.74	.64	.70	.70	.70	.70	.77	.52	.64	.98	.88	.58	.40	.78	.70	.58
$f6$	.88	.98	.77	.64	.52	.96	.92	.92	.92	.92	.83	.92	.92	.92	.96	.71	.98	.77	.58	.64
$f7$	.92	.96	.99	.99	.99	.92	.96	.98	.98	.99	.88	.98	.96	.96	.96	.83	.96	.96	.98	.88
$f8$	.96	.88	.99	.96	.96	.92	.96	.88	.96	.96	.88	.83	.98	.98	.99	.64	.83	.88	.92	.96
$f9$	.83	.71	.88	.88	.88	.83	.92	.83	.83	.83	.96	.92	.99	.96	.88	.96	.98	.99	.96	.98
$f10$	.70	.88	.71	.77	.83	.70	.83	.70	.70	.71	.88	.99	.98	.98	.96	.88	.98	.98	.96	.96
$f11$	.71	.77	.96	.88	.92	.83	.64	.98	.98	.83	.83	.98	.98	.88	.76	.58	.92	.83	.58	.77
$f12$	.64	.96	.88	.77	.77	.83	.71	.96	.92	.92	.77	.96	.58	.77	.71	.83	.88	.88	.83	.58
$f13$	.77	.77	.83	.99	.96	.83	.71	.83	.96	.92	.88	.52	.74	.99	.92	.77	.52	.76	.70	.77
$f14$	.71	.98	.88	.96	.96	.71	.88	.99	.99	.99	.92	.99	.98	.99	.98	.98	.99	.99	.98	.98
$f15$	.88	.99	.98	.98	.96	.88	.83	.92	.99	.98	.96	.58	.77	.71	.71	.99	.98	.96	.92	.92
$f16$	.96	.99	.99	.99	.99	.92	.92	.96	.96	.98	.99	.99	.99	.83	.83	.99	.99	.96	.92	.77
$f17$	.92	.92	.95	.92	.92	.96	.92	.92	.92	.92	.92	.92	.96	.99	.98	.88	.88	.88	.96	.96
$f18$	.96	.96	.96	.96	.88	.96	.96	.96	.92	.88	.88	.92	.92	.89	.89	.98	.96	.96	.96	.96
$f19$	.88	.92	.92	.92	.92	.88	.92	.92	.92	.99	.92	.92	.95	.92	.92	.96	.92	.92	.92	.92
$f20$	.92	.96	.99	.98	.99	.96	.99	.96	.96	.96	.98	.98	.92	.99	.98	.98	.98	.98	.98	.99
	Image 3										Image 4									
	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
(b)																				
$f1$	.77	.71	.71	.71	.71	.77	.77	.77	.71	.71	.52	.64	.83	.92	.92	.52	.64	.64	.71	.64
$f2$	.88	.92	.92	.83	.83	.98	.96	.96	.96	.96	.88	.98	.98	.98	.96	.99	.99	.98	.99	.99
$f3$	.88	.92	.83	.77	.77	.92	.96	.96	.88	.88	.96	.83	.98	.99	.98	.83	.88	.77	.64	.92
$f4$	.76	.64	.72	.78	.72	.70	.52	.76	.76	.72	.71	.88	.98	.83	.52	.92	.92	.96	.96	.92
$f5$	.71	.83	.83	.46	.58	.71	.71	.88	.83	.58	.99	.99	.99	.58	.71	.98	.96	.77	.96	.83
$f6$	.77	.66	.78	.76	.78	.77	.60	.76	.73	.75	.99	.64	.77	.83	.64	.92	.52	.60	.58	.64
$f7$	.99	.98	.96	.92	.92	.96	.92	.88	.96	.98	.92	.92	.96	.99	.98	.88	.88	.88	.88	.96
$f8$	.92	.77	.83	.98	.88	.71	.83	.71	.64	.64	.60	.64	.58	.52	.88	.60	.64	.46	.92	.96
$f9$	.64	.64	.64	.71	.71	.64	.64	.64	.64	.71	.88	.88	.92	.77	.92	.83	.88	.88	.88	.77
$f10$	.98	.98	.92	.99	.98	.98	.98	.98	.98	.99	.96	.88	.92	.83	.88	.92	.92	.77	.71	.83
$f11$	.71	.88	.96	.71	.92	.88	.83	.98	.96	.92	.88	.71	.92	.98	.96	.71	.98	.83	.88	.58
$f12$	.71	.98	.99	.99	.99	.58	.96	.92	.99	.99	.77	.77	.83	.71	.88	.96	.88	.64	.88	.58
$f13$	.83	.92	.77	.64	.71	.92	.96	.92	.83	.64	.83	.92	.64	.71	.71	.64	.88	.83	.96	.92
$f14$	.99	.99	.99	.99	.99	.99	.98	.98	.99	.99	.58	.58	.77	.77	.83	.64	.58	.76	.76	.76
$f15$	.99	.98	.99	.99	.99	.99	.99	.99	.99	.99	.64	.64	.58	.96	.98	.77	.71	.71	.71	.71
$f16$	.88	.96	.96	.96	.99	.92	.96	.98	.92	.96	.77	.77	.77	.77	.71	.77	.77	.64	.83	.64
$f17$	.88	.92	.92	.92	.92	.88	.92	.92	.88	.89	.99	.99	.99	.83	.83	.99	.99	.96	.92	.87
$f18$	.92	.96	.99	.99	.99	.92	.96	.96	.99	.96	.92	.96	.98	.98	.98	.92	.96	.96	.92	.96
$f19$	.96	.88	.91	.96	.96	.92	.91	.88	.96	.96	.88	.96	.98	.98	.96	.88	.83	.92	.99	.98
$f20$	.88	.98	.98	.98	.96	.99	.99	.98	.99	.99	.92	.96	.92	.98	.99	.96	.99	.96	.99	.99

Table 2
Parameter settings and feature grouping for the ensemble elements

Selected feature	r		ω		
co_absv(r,ω)	2	0°	45°	90°	135°
co_cont(r,ω)	2	0°	45°	90°	135°
co_diffVar(r,ω)	2	0°	45°	90°	135°
co_diffEnt(r,ω)	2	0°	45°	90°	135°
Ensemble element		z_1	z_2	z_3	z_4

Table 3
Test results for the four independent ensemble estimators

	z_1	z_2	z_3	z_4	Ensemble
ρ	.91	.91	.91	.91	.92
$\hat{\mu}_{\text{err}}$	−.012	−.002	−.007	−.011	−.008
s_{err}	.116	.121	.118	.117	.111
$\hat{\mu}_{\lvert\text{err}\rvert}$	.093	.098	.095	.093	.090
$s_{\lvert\text{err}\rvert}$	.07	.071	.069	.072	.064

individually considered. The values returned as quality scores by the neural network, y, were compared with the actual quality scores, t, collected from human assessors. As anticipated, all the comparisons were performed on *test* sets for cross-validation. Table 3 shows, for each estimator and for the ensemble, their performances in terms of the following values:

- Pearson's correlation coefficient, ρ, between y and t;
- the mean prediction error, $\hat{\mu}_{\text{err}}$, and the associate sample standard deviation, s_{err};
- the mean value of the absolute prediction error, $\hat{\mu}_{\lvert\text{err}\rvert}$, and its standard deviation, $s_{\lvert\text{err}\rvert}$.

Remarkably, all the four individual estimators exhibit correlation coefficients larger than .9. Likewise, the average absolute errors are always smaller than .1, which corresponds to half a unit when expressed in terms of the original 11-point quality scale used by human assessors. As expected, the integrated estimator scored, at the same time, a higher correlation coefficient and smaller errors. Thus, one can assert that the ensemble approach did enhance the performance of the overall objective-assessment system.

5.5. Generalization performance by using a new set of original images

In the experiments described so far, the stimuli of both the training and test sets were obtained by applying the same set of enhancement filters γ_l to the same sample of original images. Hence, there is some correlation between the two data sets. Therefore, one might question whether this level of correspondence between the training and test sets might have affected the reliability of the results.

To check this, the generalization performance of the neural-network assessment system was also evaluated by measuring its estimation performance on a novel set of pictures. Four new original (still gray-scale) images (Fig. 5) were processed by using the same family of contrast-enhancement filters γ_l ($l = 1,\ldots,n_e$) (17). Without loss of generality, the set of parameters $\{\alpha, \tau, \lambda\}$ was slightly simplified, keeping the parameter τ constant to $\tau = \tau_2$. As a result, the family of filters for this experiment consisted of $n_e = 10$ members, and after filtering, the set of enhanced images amounted to $n_p = n_i \times n_e = 40$ images. The enhanced images were used to perform both a subjective test (according

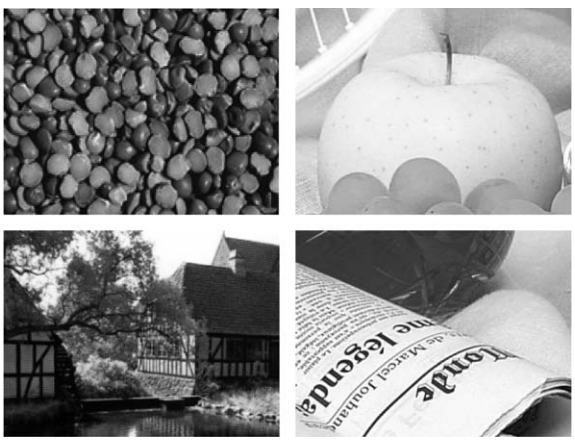

Fig. 5. The validation set of gray-scale images used to assess the system generalization ability.

to the same protocol as described in Section 5.1) and an objective estimation of perceived image quality by the neural-network assessment system.

Table 4 summarizes the performance results of each individual ensemble estimator (z_1-z_4) and of the overall ensemble on the validation set. It confirms that the ensemble estimator did improve the system performance, as it featured smaller errors than the individual estimators.

The graphs in Fig. 6 show the error distributions of the quality scores as measured by the ensemble estimator on the validation set. The scatter plot in Fig. 6(a) gives the estimated objective quality (x-axis) as a function of the subjective scores (on the y-axis). The concentration of data points around the diagonal line of this plot confirms the good generalization performance of the neural-network assessment system on a (random) validation set of images. The ensemble system reached a Pearson's correlation coefficient equal to .85, whereas the average prediction error was $\hat{\mu}_{\text{err}}^{(\text{Ens})} = .007\left(s_{\text{err}}^{(\text{Ens})} = .1\right)$ and the mean absolute prediction error was $\hat{\mu}_{|\text{err}|}^{(\text{Ens})} = .008\left(s_{|\text{err}|}^{(\text{Ens})} = .078\right)$. The quantile–quantile (Q–Q) plot in Fig. 6(b) compares the quantiles of the experimental error distribution (x-axis) with the corresponding quantiles of an ideal Gaussian distribution, $N\left(\hat{\mu}_{\text{err}}^{(\text{Ens})}, s_{\text{err}}^{(\text{Ens})}\right)$ (on the y axis), where the Gaussian parameters were the average and variance values empirically measured. Both axes in the plot are expressed in units of their respective data sets, and for each point in the Q–Q plot, the quantile level is the same for both distributions. The graph shows that the error distribution of the estimated quality scores can be modeled by a normal distribution, as the dots lie approximately along the dashed line.

Table 4
Validation results for the four independent ensemble estimators

	z_1	z_2	z_3	z_4	Ensemble
ρ	.73	.86	.7	.81	.85
$\hat{\mu}_{\text{err}}$	−.001	.02	.006	.001	.007
s_{err}	.16	.12	.16	.14	.1
$\hat{\mu}_{\|\text{err}\|}$	.12	.1	.12	.11	.088
$s_{\|\text{err}\|}$	.1	.07	.09	.08	.078

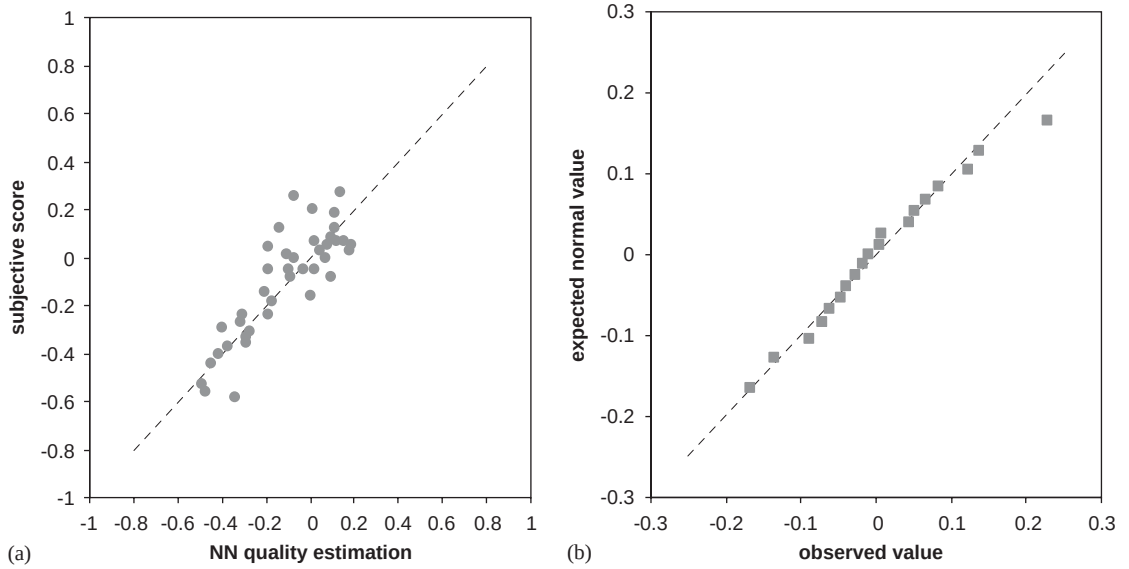

Fig. 6. Generalization performance of the ensemble neural-network estimator on a new set of original images: (a) scatter plot of estimated quality values versus subjective scores, (b) Q–Q plot comparing the error distribution of the estimated scores with a distribution $\left(\hat{\mu}_{\text{err}}^{(\text{Ens})}, s_{\text{err}}^{(\text{Ens})}\right)$.

5.6. Generalization performance by using a new set of enhancement filters

The experiments described in Section 5.5 showed that the quality-assessment method could attain a satisfactory generalization performance on a set of original images that were not used in the neural-network setup. Similarly, it seemed interesting to check the generalization performance of the neural-based assessment system when using a different set of enhancement filters. These tests aimed at verifying whether the trained network actually measured the perceived quality of contrast-enhanced images, or otherwise, whether the estimation system was just fitted to the family γ_l of enhancement filters (17) introduced in Section 5.1.

This new experimental session used the same library of $n_i = 16$ gray-scale images as presented in Fig. 4, but now processed by a different family, $\bar{\gamma}_i$, of contrast-enhancement filters. As compared with the original family of filters γ_l, the family $\bar{\gamma}_i$ was parameterized by different settings of the gain parameter, α; which spanned a new range $\{\bar{\alpha}_k; k = 1, \ldots, 4\}$. The other, less critical parameters, τ and λ took on again three values $\{\tau_n; n = 1, \ldots, 3\}$ and two values $\{\lambda_m; m = 1, 2\}$, respectively. As a result, the family $\bar{\gamma}_i$ $(i = 1, \ldots, n_e)$ of enhancement filters for this new experiment comprised a set of $n_e = 24$ members. Applying this set of filters to the library of n_i original images yielded a new sample of $n_e n_i = 24 \times 16 = 384$ enhanced images, which underwent an additional panel test of human assessors to collect subjective ratings. The neural quality-assessment system, which was not changed with respect to that developed in Section 5.3, was used to predict the objective quality scores.

The graphs in Fig. 7 show the error distributions of the quality scores as measured by the ensemble estimator on the new set of enhanced images. The scatter plot in Fig. 7(a) gives the estimated objective quality (x-axis) as a function of the subjective scores (on the y-axis). The concentration of data points around the diagonal line of this plot confirms the satisfactory generalization performance of the neural-network assessment system on the set of images obtained by applying the new set of enhancement filters $\bar{\gamma}_i$. The ensemble system reached a Pearson's correlation coefficient equal to .93, whereas the average prediction error was $\hat{\mu}_{\text{err}}^{(\text{Ens})} = -.001\left(s_{\text{err}}^{(\text{Ens})} = .09\right)$ and the mean absolute prediction error was $\hat{\mu}_{|\text{err}|}^{(\text{Ens})} = .07$. The quantile-quantile (Q–Q) plot in Fig. 7(b) compares the quantiles of the experimental error distribution

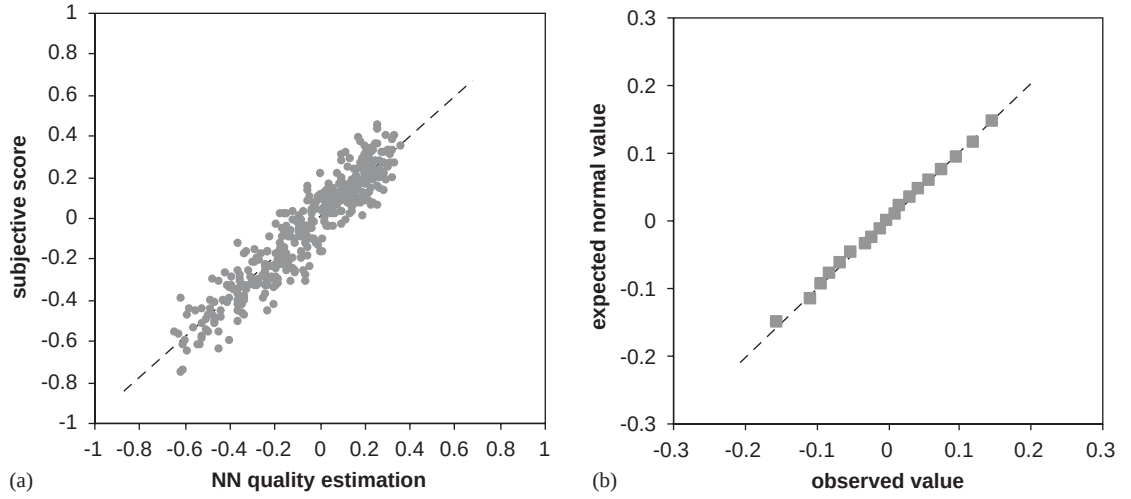

Fig. 7. Generalization performance of the ensemble neural-network estimator on the set of pictures processed by the family $\bar{\gamma}_i$ of contrast-enhancement filters: (a) scatter plot of estimated quality values versus subjective scores, (b) Q–Q plot comparing the error distribution of the estimated scores with a distribution $\left(\hat{\mu}_{\text{err}}^{(\text{Ens})}, s_{\text{err}}^{(\text{Ens})}\right)$.

(x-axis) with the corresponding quantiles of an ideal Gaussian distribution, $N\left(\hat{\mu}_{\mathrm{err}}^{(\mathrm{Ens})}, s_{\mathrm{err}}^{(\mathrm{Ens})}\right)$ (on the y-axis), where the Gaussian parameters were the average and variance values empirically measured. Both axes in the plot are expressed in units of their respective data sets, and for each point in the Q–Q plot, the quantile level is the same for both distributions. The graph shows that the error distribution of the estimated quality scores can be modeled by a normal distribution, as the dots lie along the dashed line.

6. Conclusions

For the evaluation of the effects of image-enhancement filters on image quality, one cannot rely on the bivariant approach based on image fidelity or image dissimilarity, as used in the majority of objective image quality models developed so far. Indeed, most of the dissimilarity between the original and enhanced images is expected to improve rather than degrade the overall quality. Hence, a univariant approach is the prerequisite for the evaluation of the performance of image-enhancement filters. In this paper, an objective quality assessment system based on a CBP neural network has been presented as a univariant approach to estimating the image quality that results from using a contrast-enhancement filter. The approach consists of two steps: (1) finding features, which reliably describe the enhanced images, and (2) defining a neural network, which maps the image features into image quality scores. As to feature selection, the present paper has proven that a non-parametric criterion based on the Kolmogorov–Smirnov test can effectively define image features that are able to determine the effect of an enhancement filter on an image. A crucial issue of the proposed method is that the set of (small number of) features is used for all the different images. In principle, it might be necessary to extract different types of features for different images corresponding to the nature of content in the image; in fact it might even be necessary to extract different sets of features for different image blocks. This would also correspond to the way in which humans perceive images. On the other hand, an adaptive strategy in (either block-based or image-based) feature extraction might affect the method's generality and expose the overall research to the risk of overfitting. The complexity of these issues will constitute future lines of research in this area.

As to the neural-network architecture, it has been shown that the CBP network (which is an extended version of the well-established Multi-Layer Perceptron) allows more flexibility of the underlying functional behavior, and thus is able to mimic the complex process of image quality perception. Moreover, it has been proven that the introduction of an ensemble strategy into the neural-network architecture reduces the variance in the estimated quality values.

The approach has been validated by two additional experiments: (1) using new (original) images, which have been randomly taken from a database of (gray-scale) images and processed by the same contrast-enhancement filter, and (2) using the same original images, but processed according to different settings of the contrast-enhancement filter. For both new sets of enhanced images, the neural-network assessment system was able to predict to a sufficiently high accuracy the image quality as perceived by human viewers. These results motivate further developments of the present research towards colored images, other image-enhancement filters, and possibly the use of more refined statistical criteria for the feature-selection procedure.

Acknowledgements

The authors wish to thank Dr. P. Carrai and Dr. G. Ferretti, Philips Research, Monza, for their assistance during the first phase of the research.

Appendix A. Pixel-based features

- derived from the first-order normalized histogram, $H_q(g)$ of block $b_b^{(I)}$ measuring $D \times D$ pixels; $H_q(g)$ is calculated as

$$H_q(g) = N_{qg}/D^2, \quad g = 0, \ldots, n_l - 1,$$

where g denotes a generic gray-level value of a pixel, n_l indicates the number of such levels (for 8-bit images $n_l = 256$), and N_{qg} is the number of pixels in $b_b^{(I)}$ having a gray level g. From this histogram the following features are calculated:

$$\text{mean} = \mu_g = \sum_g g H_q(g),$$

$$\text{stdev} = \sigma_g = \left[\sum_g (g-\mu_g)^2 H_q(g)\right]^{1/2},$$

$$\text{entropy} = -\sum_g H_q(g)\log_2 H_q(g),$$

$$\text{energy} = \sum_g \left[H_q(g)\right]^2,$$

$$\text{skew} = \frac{1}{\sigma_g^3}\sum_g (g-\mu_g)^3 H_q(g),$$

$$\text{kurt} = \frac{1}{\sigma_g^4}\sum_g (g-\mu_g)^4 H_q(g) - 3.$$

- Features derived from the co-occurrence matrix ($0 \leqslant g_i, g_j < n_l - 1$):

$$\text{co\_autoc} = \sum_{g_i,g_j} g_i g_j C_q(g_i,g_j,r,\omega),$$

$$\text{co\_invd} = \sum_{g_i,g_j} \frac{C_q(g_i,g_j,r,\omega)}{1+(g_i-g_j)^2},$$

$$\text{co\_energy} = \sum_{g_i,g_j} \left[C_q(g_i,g_j,r,\omega)\right]^2,$$

$$\text{co\_entropy} = -\sum_{g_i,g_j} C_q(g_i,g_j,r,\omega)\log_2 C_q(g_i,g_j,r,\omega),$$

$$\text{co\_cov} = \sum_{g_i,g_j} (g_i-\mu_i)(g_j-\mu_j) C_q(g_i,g_j,r,\omega)$$

$$\left(\mu_i = \sum_{g_i,g_j} g_i C_q(g_i,g_j,r,\omega),\ \mu_j = \sum_{g_i,g_j} g_j C_q(g_i,g_j,r,\omega)\right).$$

The remaining features need the following quantity:

$$P_z(r,\omega) = \sum_{\substack{g_i,g_j \\ |g_i-g_j|=z}} C_q(g_i,g_j,r,\omega);\ (0 \leqslant z < n_l - 1).$$

Based on $P_z(r,\omega)$, four features are defined as follows:

$$\text{co\_absv} = \sum_z z P_z(r,\omega),$$

$$\text{co\_diffVar} = \left[\sum_z (z-\text{co\_absv})^2 P_z(r,\omega)\right]^{1/2},$$

$$\text{co\_cont} = \sum_z z^2 P_z(r,\omega),$$

$$\text{co\_diffEnt} = -\sum_z P_z(r,\omega)\log_2 P_z(r,\omega).$$

- Features derived from the DCT ($0 \leqslant m, n < D$):

$$\text{f\_dcEn} = B_q[0,0] \Big/ \sum_{m,n} B_q[m,n],$$

$$\text{f\_horEn} = \sum_n B_q[0,n] \Big/ \sum_{m,n} B_q[m,n],$$

$$\text{f\_verEn} = \sum_m B_q[m,0] \Big/ \sum_{m,n} B_q[m,n],$$

$$\text{f\_diagEn} = \sum_{\substack{m,n \\ m=n}} B_q[m,n] \Big/ \sum_{m,n} B_q[m,n],$$

where $b_q(m,n)$ is the pixel value at position (m,n) within block $b_q(I)$, and $B_q[m,n]$ is the DCT component at the angular frequencies m, n.

References

[1] Methodology for the subjective assessment of the quality of television pictures, International Telecomunication Union, Geneva, Switzerland, ITU-R BT.500, 1995.
[2] P. Engeldrum, Psychometric Scaling: A Toolkit for Imaging Systems Development, Imcotek Press, Winchester, 2000.
[3] H. de Ridder, Cognitive issues in image quality measurement, J. Electronic Imaging 10 (1) (2001) 47–55.
[4] A. Ahumada, Computational image quality metrics: a review, SID Digest 24 (1993) 305–308.
[5] A.M. Eskicioglu, P.S. Fisher, Image quality and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.

[6] M.P. Eckert, A.P. Bradley, Perceptual quality metrics applied to still image compression, Signal Processing 70 (3) (1998) 177–200.
[7] J.-B. Martens, L. Meesters, Image dissimilarity, Signal Processing 70 (3) (1998) 155–176.
[8] T.N. Pappas, R.J. Safranek, Perceptual criteria for image quality evaluation, in: A. Bovik (Ed.), Handbook of Image and Video Processing, Academic Press, San Diego, 2000, pp. 669–684.
[9] B.E. Rogowitz, T.N. Pappas, J.P. Allebach, Human vision and electronic imaging, J. Electronic Imaging 10 (1) (2001) 10–19.
[10] J.-B. Martens, Image Technology Design—A Perceptual Approach, Kluwer, Norwell, MA, 2003.
[11] Z. Wang, H.R. Sheikh, A.C. Bovik, Objective video quality assessment, in: B. Furth, O. Marques (Eds.), The Handbook of Video Databases: Design and Applications, CRC Press, Boca Raton, FL, 2003.
[12] S. Daly, The visible differences predictor: an algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 179–206.
[13] J. Lubin, The use of psychophysical data and models in the analysis of display system performance, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 163–178.
[14] M. Miyahara, K. Kotani, V.R. Algazi, Objective picture quality scale (PQS) for image coding, IEEE Trans. Commun. 46 (9) (1998) 1215–1225.
[15] Y.K. Lai, C.C.J. Kuo, Haar wavelet approach to compressed image quality measurement, J. Visual Commun. Image Representation 11 (1) (2000) 17–40.
[16] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, A.C. Bovik, Image quality assessment based on a degradation model, IEEE Trans. Image Process. 9 (4) (2000) 636–649.
[17] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Processing Lett. 9 (3) (2002) 81–84.
[18] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, Perceptual blur and ringing metrics: application to JPEG2000, Signal Processing: Image Commun. 19 (2) (2004) 163–172.
[19] M. Jung, D. Léger, M. Gazalet, Univariant assessment of the quality of images, J. Electronic Imaging 11 (3) (2002) 354–364.
[20] J.E. Caviedes, F. Oberti, No-reference quality metric for degraded and enhanced video, Proceedings of the VCIP, Lugano, Switzerland, 2003.
[21] D.S. Turaga, Y. Chen, J. Caviedes, No reference PSNR estimation for compressed pictures, Signal Processing: Image Commun. 19 (2) (2004) 173–184.
[22] S. Ridella, S. Rovetta, R. Zunino, Circular back-propagation networks for classification, IEEE Trans. Neural Networks 8 (1) (1997) 84–97.
[23] P. Gastaldo, R. Zunino, S. Rovetta, Objective assessment of MPEG-2 video quality, J. Electronic Imaging 11 (3) (2002) 365–374.
[24] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Systems, Man Cybern. SMC-3 (6) (1973) 610–621.
[25] R.M. Haralick, Statistical and structural approaches to texture, Proceedings of the IEEE 67 (5) (1979) 786–804.
[26] A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[27] I.M. Chakravarti, R.G. Laha, J. Roy, Handbook of Methods of Applied Statistics, vol. I, Wiley, New York, 1967.
[28] D.E. Rumelhart, J.L. McClelland, Parallel Distributed Processing, MIT Press, Cambridge, MA, 1986.
[29] T.P. Vogl, J.K. Mangis, A.K. Rigler, W.T. Zink, D.L. Alkon, Accelerating the convergence of the back propagation method, Biol. Cybern. 59 (1988) 257–263.
[30] M. Perrone, Improving regression estimates: averaging methods for variance reduction with extension to general convex measure optimization, Ph.D. Dissertation, Physics Department, Brown University, Providence, RI, 1993.
[31] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. 4 (1) (1992) 1–48.
[32] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 226–239.
[33] S. Rovetta, R. Zunino, A multiprocessor-oriented visual tracking system, IEEE Trans. Ind. Electronics 46 (4) (1999) 842–850.
[34] B. Widrow, M.A. Lehr, 30 Years of adaptive neural networks: perceptron. Madaline and back propagation, Proc. IEEE 78 (9) (1990) 1415–1442.
[35] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5 (4) (1994) 537–550.
[36] S. Ridella, S. Rovetta, R. Zunino, K-winner machines for pattern classification, IEEE Trans. Neural Networks 12 (2) (2001) 371–385.
[37] S. Ridella, R. Zunino, Empirical measure of multiclass generalization performance: the K-winner machine case, IEEE Trans. Neural Networks 12 (6) (2001) 1525–1529.
[38] W.G. Cochran, G.M. Cox, Experimental Designs, Wiley, New York, 1957.
[39] R. Hecht-Nielsen, Neurocomputing, Addison-Wesley, Reading, MA, 1990.