A Locally Adaptive System for the Fusion of Objective Quality Measures

Adriaan Barri, Ann Dooms, *Member, IEEE*, Bart Jansen, and Peter Schelkens, *Member, IEEE*

***Abstract*— Objective measures to automatically predict the perceptual quality of images or videos can reduce the time and cost requirements of end-to-end quality monitoring. For reliable quality predictions, these objective quality measures need to respond consistently with the behavior of the human visual system (HVS). In practice, many important HVS mechanisms are too complex to be modeled directly. Instead, they can be mimicked by machine learning systems, trained on subjective quality assessment databases, and applied on predefined objective quality measures for specific content or distortion classes. On the downside, machine learning systems are often difficult to interpret and may even contradict the input objective quality measures, leading to unreliable quality predictions. To address this problem, we developed an interpretable machine learning system for objective quality assessment, namely the locally adaptive fusion (LAF). This paper describes the LAF system and compares its performance with traditional machine learning. As it turns out, the LAF system is more consistent with the input measures and can better handle heteroscedastic training data.**

***Index Terms*— Objective quality assessment, machine learning, measure fusion.**

I. Introduction

RECENT advances in information technology have increased user expectations regarding the visual quality of multimedia services. However, during distribution, the perceived visual quality may decrease, mainly due to compression or transmission errors. In order to satisfy the high demands of the end-user, the visual quality needs to be continuously monitored.

Subjective quality experiments currently provide the most accurate way to measure and monitor the perceptual quality, in which a representative group of test subjects is asked to rate the quality of distorted signals [1], [2]. However, subjective experiments are not popular, because they are expensive, time-consuming and unsuitable for real-time quality monitoring.

The drawbacks of subjective experiments triggered the design of *objective quality measures* to automatically predict the visual quality as it is perceived by human viewers [3], [4]. Traditional objective quality measures attempt to model the behavior of the *Human Visual System* (HVS). However, many important mechanisms of the HVS are discarded, because they are either too complex or not sufficiently understood.

To improve the quality prediction, objective quality measures based on *Machine Learning* (ML) have been introduced. These ML-based objective quality measures try to mimic the HVS mechanisms. As a consequence, they do not require explicit mathematical models of the HVS [5], [6].

Although many ML-based objective quality measures have already been proposed in literature, there is still quite some room for improvement. While ML systems with a linear response cannot handle the complex behavior of the HVS, current ML systems with a nonlinear response are often difficult to interpret and may even contradict the input objective quality measures. Hence, new forms of ML are needed that can provide more reliable quality predictions.

This paper introduces the *Locally Adaptive Fusion* (LAF) system, an extension of our previous work in [7]. The LAF system is specifically designed for the perceptual quality prediction of images or videos. A LAF-based objective quality measure is constructed in two steps. The first step comprises a selection of *limited-scope* objective quality measures, i.e. they are only reliable for specific content and distortion classes. The second step comprises a combination of the selected limited-scope objective quality measures through adaptive weighting, where the weighting factors are determined by training on a subjective quality assessment database. In this way, the composite objective quality measure is suitable for perceptual quality predictions on a broad scope of content and distortion classes.

The remainder of this paper is structured as follows. Section II introduces some basic notations and gives a global view of the state-of-the-art. Section III describes the LAF system. Section IV presents a concrete implementation of the LAF system for the quality prediction of images. Section V validates the LAF system and compares its prediction performance with the most prominent ML systems in the field of objective quality assessment. Section VI summarizes the obtained results.

II. Preliminaries

In this paper, we represent perceptual quality by an operator Q that assigns a score $Q(x, y)$ between 0 and 1 to each distorted signal x, relative to its original, undistorted reference signal y. The higher the value of $Q(x, y)$, the better the

Manuscript received April 3, 2013; revised October 6, 2013 and January 24, 2014; accepted March 24, 2014. Date of publication April 10, 2014; date of current version April 28, 2014. This work was supported in part by the Flemish Institute for the Promotion of Innovation by Science and Technology and in part by the Vrije Universiteit Brussel Strategic Research Program Processing of Large Scale Multidimensional, Multi-spectral, Multi-sensorial and Distributed Data. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sanghoon Lee.

The authors are with the Department of Electronics and Informatics, Vrije Universiteit Brussel, Brussels B-1050, Belgium, and also with the iMinds-Department of Future Media and Imaging, Ghent B-9050, Belgium (e-mail: abarri@etro.vub.ac.be; adooms@etro.vub.ac.be; bjansen@etro.vub.ac.be; pschelke@etro.vub.ac.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2316379

perceptual quality. We assume the *perfect reference hypothesis*, which states that only the reference signals have the highest perceptual quality [4]. The perfect reference hypothesis is formulated by the *strong reflexivity property* [8]:

$$Q(x, y) = 1 \text{ if and only if } x = y \tag{1}$$

Note that, under this hypothesis, every distorted signal receives a quality score that is strictly smaller than one, even when the impairments are perceptually invisible. The notion of *perceptual similarity* can be formalized as follows: given a perceptibility threshold T, two signal pairs (x_1, y_1) and (x_2, y_2) are said to be perceptual similar if and only if

$$-T \leq Q(x_1, y_1) - Q(x_2, y_2) \leq T.$$

The perceptibility threshold T depends on several factors, such as the display device, the viewing conditions, and the lighting conditions. One possible approach to determine T from subjective experiments is described in [9].

The perceptual quality operator Q is determined through subjective experiments. Each test subject individually rates the experienced quality of the sample signals. The obtained raw opinion scores are processed and averaged to yield a *Mean Opinion Score* (MOS) for all considered signals. We estimate the perceptual quality score $Q(x, y)$ w.r.t. the *Difference Mean Opinion Score* (DMOS) scale, i.e. the difference between the MOS of the reference signal and the MOS of the distorted signal:

$$\text{DMOS}(x, y) = \text{MOS}(y) - \text{MOS}(x). \tag{2}$$

We then set

$$Q(x, y) = 1 - \frac{\text{DMOS}(x, y) - a}{b - a} \tag{3}$$

where a and b are respectively the minimum and maximum DMOS value. The perfect reference hypothesis is satisfied when $\text{DMOS}(x, y) > 0$ for all considered distorted signals x that differ from the reference signal y. Note that the DMOS value can be negative for certain distorted signals with a higher perceptual quality than their reference. The quality prediction of signals with a negative DMOS is not covered by the present paper.

We denote by $\mathcal{M}(x, y)$ the score assigned by an objective quality measure $\mathcal{M}$ to a distorted signal x relative to the reference signal y. We assume that all objective quality measures have a monotonically increasing *conditional mean* w.r.t perceptual quality, that is, the measured scores tend to increase as the perceptual quality increases. Objective quality measures with a monotonically decreasing conditional mean can simply be replaced by their additive inverse.

The remainder of this section is divided into three parts. Part A focuses on subjective quality assessment databases. Part B lists the state-of-the-art objective visual quality measures. Part C describes the ML systems that are currently employed in the field of objective visual quality assessment.

A. Subjective Quality Assessment Databases

Subjective quality assessment databases are essential for the training and validation of ML-based objective quality measures. These databases consist of distorted signals that are annotated with MOS or DMOS scores. An extensive list of publicly available subjective quality assessment databases can be found in [10] and [11]. By means of example, we describe three popular subjective quality assessment databases for images, namely the LIVE, the CSIQ, and the TID database.

The LIVE subjective image quality assessment database consists of 29 reference images and 779 distorted images that are annotated with DMOS scores [12], [13]. The LIVE database focuses on five distortion types, namely Gaussian blur, JPEG compression, JPEG2000 compression, white noise, and bit errors induced by a Rayleigh fading channel.

The CSIQ database consists of 30 reference images and 866 distorted images that are annotated with DMOS [14], [15]. The considered distortion types are Gaussian blur, JPEG compression, JPEG2000 compression, white noise, global contrast decrements, and additive pink Gaussian noise.

The TID database contains 25 reference images, which largely overlap with those of the LIVE database [16], [17]. The TID database considers no less than 17 different distortion types to yield a total amount of 1700 images, all annotated with MOS scores.

B. Objective Quality Measures

Objective quality measures are classified into three categories according to the need for the reference signal [3], [4].

Most present-day objective quality measures are *full-reference* (FR), meaning that they require the entire reference signal. The simplest FR objective quality measure is the *peak signal-to-noise ratio* (PSNR). However, the PSNR does not have a high correlation with perceptual quality [18]. As a result, FR objective quality measures that incorporate important HVS characteristics have been proposed, such as the *Structural Similarity Index* (SSIM) [19]–[21].

In most applications the reference signal is not available, so that FR objective quality measures cannot be used. On the other hand, it is often possible to send a limited amount of information on the reference signal through a side-channel. In this scenario, *reduced-reference* (RR) quality measures can be employed. A popular RR objective quality measure is the *Generalized Video Quality Model* (VQM), standardized by the American National Standards Institute (ANSI) and the International Telecommunication Union (ITU) [22].

No-reference (NR) objective quality measures are completely independent of the reference signal, that is, $\mathcal{M}(x, y) = \mathcal{M}(x)$. Most NR objective quality measures in literature are limited to detect a single distortion type, such as blurriness [23], [24], JPEG compression artifacts [25]–[27], JPEG2000 compression artifacts [28], [29], and MPEG compression artifacts [30], [31]. Following the ML approach, three distortion-generic NR objective image quality measures have been recently constructed in [32]–[34].

C. Machine Learning

The purpose of ML is to predict a certain variable from statistical data [35]. Within the broad field of ML, this paper focuses on systems that are trained on a subjective quality

assessment database and applied on a selection of objective quality measures $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$. We assume that the training database is annotated with DMOS scores. We denote the pairs of all distorted and corresponding reference signals in the training set by (x_i, y_i), $i = 1, 2, \ldots, p$. We put $q_i = Q(x_i, y_i)$ where Q is the perceptual quality operator defined in (3). For the task of objective quality assessment, a ML system determines a *regression function* γ that approximates perceptual quality. If $\mathbf{M}$ is the m-dimensional vector of all input quality measures $\mathcal{M}_i$ then

$$\gamma\left(\mathbf{M}(x_i, y_i)\right) \approx q_i.$$

Based on the regression function, ML systems can be divided into two categories, namely *nonparametric* and *parametric*. In nonparametric systems, the regression function γ is directly estimated from the training data. In parametric systems, the regression function γ has a specific functional form ϕ_β that depends on a vector of *regression parameters* β. Usually, β is determined by the method of *Ordinary Least Squares* (OLS):

$$\beta = \arg\,\min_b \sum_{i=1}^{p} \Big(q_i - \phi_b\big(\mathbf{M}(x_i, y_i)\big)\Big)^2. \tag{4}$$

Parametric ML systems are further classified into *linear regression systems* and *nonlinear regression systems*. The latter is more flexible and may achieve a higher prediction performance. The next paragraphs describe the most common ML systems in the field of objective quality assessment. A more elaborate overview on machine learning in quality assessment can be found in the excellent survey papers [3], [5], and [6].

The *General Regression Neural Network* (GRNN) is a nonparametric ML system, introduced by Specht in [36]. It has been used to construct a generic NR objective quality measure for images [34]. The GRNN depends on a *spread factor* h, which is empirically determined. The higher the value of h, the smoother the functional approximation of the target perceptual quality score. The GRNN output is given by

$$\mathrm{GRNN}(x, y) = \frac{1}{C} \sum_{i=1}^{p} q_i e^{\frac{-d_i^2}{2h^2}} \tag{5}$$

with $C = \sum_{i=1}^{p} e^{\frac{-d_i^2}{2h^2}}$ and $d_i^2 = \sum_{j=1}^{m} (\mathcal{M}_j(x, y) - \mathcal{M}_j(x_i, y_i))^2$. The GRNN is computationally inefficient, because the number of calculations increases with both the size of the training set and the number of input quality measures.

The *Principal Component Regression* (PCR) is a linear regression system that combines the principal components (PCs) of the input objective quality measures [37]. The variances of the selected PCs approximate the variances of the measure scores. The PCR has been further optimized for video objective quality assessment in [38]. An alternative to the PCR is the *Partial Least Squares Regression* (PLSR), which takes the variances of both the perceptual quality scores and the input quality measure scores into account. The PLSR has been employed for the NR quality assessment of video streams compressed with the H.264/AVC codec [31].

The standard *Feed Forward Neural Network* (FFNN) with one hidden layer is a nonlinear regression system of the form

$$\mathrm{FFNN}(x, y) = g\Big(w_0^{(1)} + \sum_{k=1}^{n} w_k^{(1)} N_k(x, y)\Big), \tag{6}$$

where g is the *output transfer function* and N_k, $i = 1, 2, \ldots, n$, are the *hidden neurons* of the neural network defined by

$$N_k(x, y) = h\Big(w_{k,0}^{(2)} + \sum_{j=1}^{m} w_{k,j}^{(2)} \mathcal{M}_j(x, y)\Big). \tag{7}$$

In the above equation, the *hidden transfer function* h is typically a sigmoid function. Note that the FFNN can be extended to multiple hidden layers, but this generally increases the risk of overfitting [39]. The network weights $w_k^{(1)}$ and $w_{k,j}^{(2)}$ are numerically determined by the OLS cost function (4). As the OLS minimization problem, adopted for the FFNN, is typically non-convex, the numerical solution can get trapped in one of the local minima. To increase the probability that the optimal network weights are found, the FFNN needs to be re-trained several times with random weight initializations. However, these random initializations do not allow for perfectly reproducible quality predictions.

Several variations of the standard FFNN have been proposed in literature. For instance, a *Radial Basis Function Network* (RBFN) based on a growing and pruning algorithm has been used to design a NR image quality measure for JPEG compression [27]. In contrast to the standard FFNN, the RBFN does not suffer from local minima, but it typically requires much more training data to achieve the same prediction accuracy [40, Ch. 6.1.1]. The FFNN and the RBFN can be generalized to a *Circular Back Propagation Network* (CBPN) [41]. The CBPN has been implemented for the objective quality assessment of MPEG-2 compressed video content [42]. Recently, another generalization of the FFNN, the *Extreme Learning Machine* (ELM), has been applied to design a NR objective quality measure for JPEG compressed images [26]. For video quality assessment, a *Convolutional Neural Network* (CNN) has been proposed to integrate the frame-by-frame signal descriptors into a *Time Delay Neural Network* (TDNN) [43]. By sharing network weights, the CNN greatly reduces the number of free parameters to learn. The *Support Vector Regression* (SVR) system provides an alternative to neural networks for high-dimensional regression problems [44]. The SVR has been implemented to combine 88 signal descriptors for the generic NR quality assessment of images [33]. More recently, the SVR has been employed to combine several state-of-the art quality measures [45].

The above described ML systems, although frequently used, are not optimized for objective quality assessment. Linear regression systems, such as the PCR and PLSR, cannot handle the complex behavior of the HVS. The learning process of the existing nonlinear regression systems is often difficult to analyze and interpret. Nonparametric regression systems, such as the GRNN, are often computationally or memory intensive. Our proposed *Locally Adaptive Fusion* (LAF) system does not suffer from any of the previously mentioned issues, as will be shown in the next section.

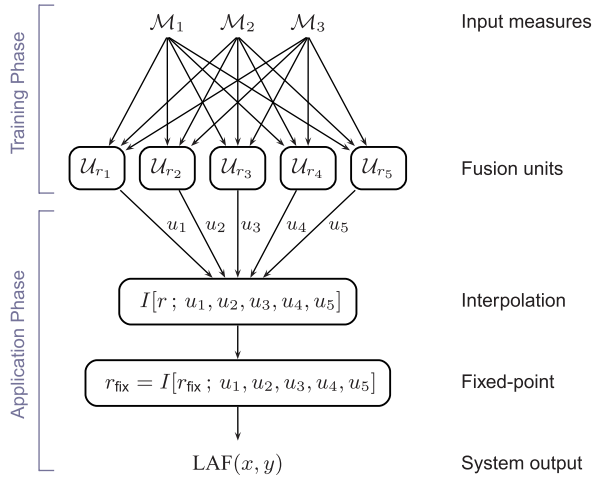

Fig. 1. In the above block diagram, the LAF system is used for three objective quality measures $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. During the training phase, the input measures are combined into five fusion units $\mathcal{U}_{r_1}, \mathcal{U}_{r_2}, \ldots, \mathcal{U}_{r_5}$, which are associated with certain target values $r_1, r_2, \ldots, r_5$ (18). During the application phase, the fusion units are evaluated on the received signal pair (x, y). By construction through the use of the local prediction accuracy, the five fusion responses $u_i = \mathcal{U}_{r_i}(x, y)$ yield more reliable quality indications when they are closer to their target values r_i. The fusion responses $\mathcal{U}_r(x, y)$ for other target values are approximated by interpolating the five points (r_i, u_i) (Fig. 3). This interpolation line typically contains a fixed-point $(r_{\text{fix}}, r_{\text{fix}})$. The value of r_{fix} is the final LAF-based quality prediction of the distorted signal x, relative to the reference signal y.

III. The Locally Adaptive Fusion System

This section introduces the *Locally Adaptive Fusion* (LAF) system for the combination of objective quality measures $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ (Fig. 1). The LAF system relies on multiple *fusion units* $\mathcal{U}_{r_1}, \mathcal{U}_{r_2}, \ldots, \mathcal{U}_{r_n}$, which are normalized weighted sums of the input objective quality measures. The fusion units $\mathcal{U}_{r_i}$ are each associated with a certain value r_i of perceptual quality, called the *target value*. The fusion units are constructed in such a way that the *local prediction accuracy* is optimized near their target value. To this end, the weights of the fusion units need to be *trained* on a subjective quality assessment database.

The LAF-based quality prediction of a signal x with reference y is a weighted sum of the fusion unit responses. More precisely, the fusion unit responses are weighted according to the distance to their target value. By construction, the weights used to combine the fusion units are *adaptive*, i.e. their values change with the received signal. For example, when the true (unknown) perceptual quality is equal to r_i, the distance between r_i and $U_{r_i}(x, y)$ will typically be smaller than the distance between r_i and $U_{r_j}(x, y)$ for $j \neq i$. In this case, the i-th fusion unit has the best prediction accuracy, and will therefore receive the highest weight.

Accordingly, the LAF system predicts the perceptual quality of a signal by cleverly weighting the locally optimized fusion units. In this way, the LAF system can achieve an improved prediction accuracy on a broader scope of content and distortion classes.

The remainder of this section is divided into four parts. Part A provides a closed-form formula to determine the local prediction accuracy of an objective quality measure. Part B explains the construction of the locally optimized fusion units. Part C describes the LAF system. Finally, Part D states some important properties of the LAF system.

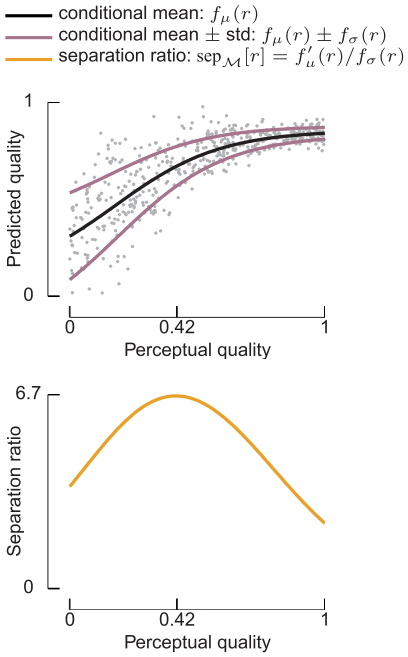

Fig. 2. The behavior of the separation ratio is visualized for a hypothetical objective quality measure $\mathcal{M}$. The higher the value of the separation ratio, the higher the local prediction accuracy. In this example, the highest local prediction accuracy is obtained in 0.42. The data points are of the form (q, s), where q is uniformly sampled from the interval $[0, 1]$ and s is sampled from a normal distribution with mean $f_\mu(q)$ and standard deviation $f_\sigma(q)$.

A. Local Prediction Accuracy

Consider two signal pairs (x_1, y_1) and (x_2, y_2) with a perceptual quality slightly higher and lower than a fixed quality score r, say $Q(x_1, y_1) = r + \epsilon$ and $Q(x_2, y_2) = r - \epsilon$ for some small $\epsilon > 0$. The probability that $\mathcal{M}(x_1, y_1) > \mathcal{M}(x_2, y_2)$ indicates the *local prediction accuracy* of an objective quality measure $\mathcal{M}$ at the distance of ϵ. We denote this probability by $\text{acc}_{\mathcal{M}}[r, \epsilon]$.

When we represent the set of all considered signal pairs (x, y) by a continuous random vector X we obtain the identity

$$\text{acc}_{\mathcal{M}}[r, \epsilon] = \mathbb{P}\big[\mathcal{M}(X_{r+\epsilon}) > \mathcal{M}(X_{r-\epsilon})\big], \tag{8}$$

where the random vectors X_q with q in $[0, 1]$ are independent and distributed according to the conditional probability of X given $Q(X) = q$.

The value of $\text{acc}_{\mathcal{M}}[r, \epsilon]$ is hard to estimate, because the distributions of $\mathcal{M}(X_{r+\epsilon})$ and $\mathcal{M}(X_{r-\epsilon})$ are not known in practice. We therefore introduce an alternative characterization of the local prediction accuracy, called *separation ratio*:

$$\text{sep}_{\mathcal{M}}[r] = \frac{\frac{d}{dq}\mathbb{E}[\mathcal{M}(X_q)]\Big|_{q=r}}{\text{std}[\mathcal{M}(X_r)]}. \tag{9}$$

The separation ratio measures the rate of increase of the mean of $\mathcal{M}(X_q)$ as q approaches r, weighted according to the standard deviation of $\mathcal{M}(X_r)$. In this way, the separation ratio can approximate the local prediction accuracy of an objective quality measure (Fig. 2). The separation ratio is scale invariant,

that is, $\text{sep}_{\mathcal{M}}[r] = \text{sep}_{a\mathcal{M}}[r]$ for all $a > 0$. As stated in the next theorem, the separation ratio $\text{sep}_{\mathcal{M}}[r]$ is consistent with the local accuracy operator $\text{acc}_{\mathcal{M}}[r, \epsilon]$ when $\mathcal{M}(X_q)$ is normally distributed. For a proof, see the Appendix.

Theorem 1: Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two objective quality measures with $\text{sep}_{\mathcal{M}_1}[r] \neq \text{sep}_{\mathcal{M}_2}[r]$. Suppose that $\mathcal{M}_1(X_q)$ and $\mathcal{M}_2(X_q)$ are normally distributed for every q in a neighborhood around r. The following expressions are equivalent:

1. $\text{sep}_{\mathcal{M}_1}[r] > \text{sep}_{\mathcal{M}_2}[r]$;
2. $\text{acc}_{\mathcal{M}_1}[r, \epsilon] > \text{acc}_{\mathcal{M}_2}[r, \epsilon]$ *for all sufficiently small* $\epsilon > 0$.

The separation ratio of an objective quality measure $\mathcal{M}$ can be determined through regression analysis on a subjective quality assessment database. Observe that $\text{sep}_{\mathcal{M}}[r] = f'_\mu(r)/f_\sigma(r)$, where f_μ and f_σ are the real functions defined by

$$f_\mu(q) = \mathbb{E}[\mathcal{M}(X_q)] \quad \text{and} \quad f_\sigma(q) = \text{std}(\mathcal{M}(X_q)). \tag{10}$$

The values of $f_\mu(q)$ and $f_\sigma(q)$ represent the conditional mean and standard deviation of the random variable $\mathcal{M}(X)$, given that $Q(X) = q$. The conditional mean $f_\mu(q)$ and the lower confidence bound $f_\ell(q) = f_\mu(q) - f_\sigma(q)$ are closely approximated by a 4-parameter generalized logistic function, described by the *Video Quality Experts Group* (VQEG) in [46]:

$$\text{lgst}_\beta(q) = \beta(1) + \frac{\beta(2)}{1 + \exp\left(-\frac{q-\beta(3)}{\beta(4)}\right)}. \tag{11}$$

Hence, when $f_\mu(q) = \text{lgst}_{\beta_\mu}(q)$ and $f_\ell(q) = \text{lgst}_{\beta_\ell}(q)$, then

$$\text{sep}_{\mathcal{M}}[r] = \frac{\text{lgst}'_{\beta_\mu}(r)}{\text{lgst}_{\beta_\mu}(r) - \text{lgst}_{\beta_\ell}(r)}. \tag{12}$$

The 4-dimensional parameter vectors β_μ and β_ℓ are traditionally obtained by the OLS method described in (4). The OLS method assumes that the data to be fitted is *homoscedastic*, meaning that the conditional standard deviation $f_\sigma(q)$ is constant in q. However, this assumption is rarely satisfied for objective quality measures. As a result, the OLS method often produces suboptimal fittings (see for example Fig. 8 in Section IV). The heteroscedasticity of objective quality measures can be taken into account by weighting the individual error terms in (4) by $w_i = 1/f_\sigma^2(q_i)$. In this way, the data points with a smaller conditional variance get a relatively larger weight in the estimation of β. This extension of OLS is called the *Weighted Least Squares* (WLS) method [47].

In order to compute the weights w_i, the conditional standard deviation $f_\sigma(q)$ must be approximated independently of the conditional mean $f_\mu(q)$, which is addressed with some success in [48] and [49]. To improve the accuracy of the weight estimations, we developed a new algorithm for the nonparametric regression of objective quality measures. This algorithm is used to obtain initial estimates $\hat{f}_\mu$ and $\hat{f}_\sigma$ of the conditional mean and standard deviation. We then define

$$\beta_\mu = \arg\min_b \sum_{q \in I_\delta} \frac{\left(\hat{f}_\mu(q) - \text{lgst}_b(q)\right)^2}{\hat{f}_\sigma^2(q)} \tag{13}$$

and

$$\beta_\ell = \arg\min_b \sum_{q \in I_\delta} \frac{\left(\hat{f}_\mu(q) - \hat{f}_\sigma(q) - \text{lgst}_b(q)\right)^2}{\hat{f}_\sigma^2(q)}, \tag{14}$$

where I_δ is a discretization of the interval $[0, 1]$ with a small step size δ. In our experiments, we have put $\delta = 0.01$.

The initial estimates $\hat{f}_\mu(q)$ and $\hat{f}_\sigma(q)$ are calculated in three steps. First, the subjective quality assessment database is divided into sequences $\mathbf{x}^{y,t}$ of signals $x_k^{y,t}$ with the same reference signal y and distortion type t, ordered by increasing perceptual quality. Second, the data points $\left(q_k^{y,t}, s_k^{y,t}\right)$, with $q_k^{y,t} = Q\left(x_k^{y,t}, y\right)$ and $s_k^{y,t} = \mathcal{M}\left(x_k^{y,t}, y\right)$ are linearly interpolated to yield a continuous function $\hat{f}_\mu^{y,t}$. This function approximates the conditional mean of $\mathcal{M}$ restricted to the signal sequence $\mathbf{x}^{y,t}$. Finally, the values of $\hat{f}_\mu(q)$ and $\hat{f}_\sigma(q)$ are respectively defined as the sample mean and sample standard deviation of $\hat{f}^{y,t}(q)$ over all y and t. More precisely,

$$\hat{f}_\mu(q) = \frac{1}{\nu_q} \sum_{y,t} \hat{f}_\mu^{y,t}(q) \tag{15}$$

and

$$\hat{f}_\sigma(q) = \sqrt{\frac{1}{\nu_q - 1} \sum_{y,t} \left(\hat{f}_\mu(q) - \hat{f}_\mu^{y,t}(q)\right)^2} \tag{16}$$

where ν_q is the total number of signal sequences $\mathbf{x}^{y,t}$ for which $\hat{f}_\mu^{y,t}(q)$ is defined.

To summarize, the separation ratio provides a measure for the local prediction accuracy of an objective quality measure $\mathcal{M}$. It can be calculated through logistic regression on a subjective database. First, the initial estimates $\hat{f}_\mu$ and $\hat{f}_\sigma$ are computed as in (15) and (16). Second, the logistic parameter vectors β_μ and β_ℓ are determined by the WLS minimization problems in (13) and (14). Finally, the separation ratio is calculated as in (12).

B. Fusion Units

In order to construct the fusion unit $\mathcal{U}_r$, relative to some target quality score r in the interval $[0, 1]$, the objective quality measures $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_m$ are first linearly combined to yield a weighted quality measure

$$\mathcal{W}_r = \mathbf{w}_r^T \mathbf{M},$$

where $\mathbf{M}$ is the vector of all input quality measures. We impose that all weights are positive, so that the behavior of the fusion units is easier to interpret. The weight vector $\mathbf{w}_r$ is the solution of the optimization problem

$$\mathbf{w}_r = \arg\max_{\mathbf{w} \geq 0} \left(\text{sep}_{\mathbf{w}^T\mathbf{M}}[r]\right). \tag{17}$$

Hence, the weighted quality measure $\mathcal{W}_r$ is optimized to predict the perceptual quality of signals near its target quality score r. Note that the optimization problem in (17) is not solvable when one or more input quality measures have a zero conditional standard deviation in q. In that case, we set $\mathcal{W}_r = \mathcal{M}_j$, where $\mathcal{M}_j$ is the objective quality measure that has the highest rate of increase among all input quality measures with a zero conditional standard deviation in q.

The *fusion unit* $\mathcal{U}_r$ is then a normalization of the weighted quality measure $\mathcal{W}_r$ by a transfer function g_r. It is defined by

$$\mathcal{U}_r = g_r \circ \mathcal{W}_r = g_r \circ \mathbf{w}_r^T \mathbf{M}. \tag{18}$$

where g_r ensures that the average of the fusion responses $\mathcal{U}_r(x, y)$ of images with perceptual quality q is approximately equal to q. To this end, we need that $g_r\big(\mathbb{E}[\mathcal{W}_r(X_q)]\big) = q$ for every q in $[0, 1]$. Then $g_r = f^{-1}_{\mu,\mathbf{w}_r}$, where $f_{\mu,\mathbf{w}_r}$ is the conditional mean function of $\mathcal{W}_r$. The function $f_{\mu,\mathbf{w}_r}$ can again be approximated by a generalized logistic function with parameter vector β_r. The inverse of this logistic function is

$$\text{logit}_{\beta_r}(s) = \beta_r(3) + \beta_r(4) \log\left(\frac{s - \beta_r(1)}{\beta_r(2) - s + \beta_r(1)}\right).$$

Hence, the transfer function g_r is given by

$$g_r(s) = f^{-1}_{\mu,\mathbf{w}_r}(s) = \text{logit}_{\beta_r}(s). \tag{19}$$

The optimization problem in (17) corresponds to a *convex quadratic programming problem* and can therefore be efficiently solved [50]. Consider the m-dimensional vector $\mathbf{v}_r = \frac{d}{dq}\mathbb{E}[\mathbf{M}(X_q)]\Big|_{q=r}$ and the $m \times m$ covariance matrix Σ_r with $\Sigma_r(i, j) = \text{cov}\big(\mathcal{M}_i(X_r), \mathcal{M}_j(X_r)\big)$. Then

$$\text{sep}_{\mathbf{w}^T\mathbf{M}}[r] = \frac{\mathbf{w}^T \mathbf{v}_r}{\sqrt{\mathbf{w}^T \Sigma_r \mathbf{w}}}. \tag{20}$$

Hence, the weight vector $\mathbf{w}_r$ is obtained by minimizing $\mathbf{w}^T \Sigma_r \mathbf{w}$ given the constraints $\mathbf{w}^T \mathbf{v}_r = 1$ and $\mathbf{w} \geq 0$. The weight vector $\mathbf{w}_r$ is then rescaled, so that it sums to 1. This does not alter the value of the scale-invariant separation ratio.

The values of $\mathbf{v}_r$ and Σ_r, required in Formula (20), are calculated as follows. Let $f_{\mu,i}$ and $f_{\sigma,i}$ the conditional mean and standard deviation functions of the objective quality measure $\mathcal{M}_i$ as defined in (10). Then we have $\mathbf{v}_r = \big(f'_{\mu,1}(r), f'_{\mu,2}(r), \ldots, f'_{\mu,m}(r)\big)$ and $\Sigma_r(i, j) = 2f^2_{\sigma,i,j}(r) - \frac{1}{2}\big(f^2_{\sigma,i}(r) + f^2_{\sigma,j}(r)\big)$, where $f_{\sigma,i,j}$ is the conditional standard deviation of the objective quality measure $\mathcal{M}_{i,j} = (\mathcal{M}_i + \mathcal{M}_j)/2$. The values of $f_{\mu,i}(r)$, $f_{\sigma,i}(r)$, $f_{\sigma,j}(r)$, and $f_{\sigma,i,j}(r)$ can be estimated by generalized logistic functions, as previously described in Part A of this section.

C. Description of the LAF System

The LAF systems predicts the perceptual quality of a signal pair (x, y) based on the response of the fusion units. The optimal quality prediction is the fusion response $\mathcal{U}_\rho(x, y)$ that is the closest to its target quality score:

$$\rho = \arg\min_r \big|\mathcal{U}_r(x, y) - r\big|. \tag{21}$$

To numerically solve the above minimization problem, select n fusion units $\mathcal{U}_{r_i}$ with equidistantly spaced target quality scores $r_i = \frac{i-1}{n-1}$, for $i = 1, 2, \ldots, n$. The response of any other fusion unit is approximated by linearly interpolating the responses of $\mathcal{U}_{r_i}$ with respect to r_i. More precisely, $\mathcal{U}_r(x, y) \approx I\big[r;\ u_1, u_2, \ldots, u_n\big]$, where $u_i = \mathcal{U}_{r_i}(x, y)$, $\omega_r = \frac{r - r_i}{r_{i+1} - r_i}$, and

$$I[r\,;\ u_1, u_2, \ldots u_n] = \begin{cases} u_1 & \text{if } r = 0, \\ u_i + \omega_r(u_{i+1} - u_i) & \text{if } r \in [r_i, r_{i+1}], \\ u_n & \text{if } r = 1. \end{cases}$$

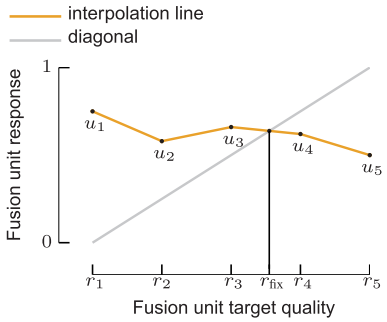

Fig. 3. In this example, the LAF system is applied on a specific signal pair (x, y). The calculated fusion responses $u_i = \mathcal{U}_{r_i}(x, y)$ all deviate from their target quality scores r_i, in which their local prediction accuracies were optimized. The fusion responses u_3 and u_4 are closer to their respective target scores, and thus provide more reliable quality indications than the other three responses. The fusion responses $\mathcal{U}_r(x, y)$ that were not calculated are approximated by linearly interpolating the five points (r_i, u_i). The interpolation line crosses the diagonal in the fixed-point $(r_{\text{fix}}, r_{\text{fix}})$. This fixed-point provides the most reliable quality indication, because the fusion response $\mathcal{U}_{r_{\text{fix}}}(x, y)$ is (approximately) equal to its target score r_{fix}.

The LAF system predicts the quality of the distorted signal x by the fixed-point r_{fix} of the interpolated function (Fig. 3). In other words, the output is given by $\text{LAF}(x, y) = r_{\text{fix}}$ where $r_{\text{fix}} = I\big[r_{\text{fix}}\,;\ u_1, u_2, \ldots, u_n\big]$. By construction,

$$\text{LAF}\ (x, y) = \arg\min_r \big|I[r\,;\ u_1, u_2, \ldots, u_n] - r\big|. \tag{22}$$

Hence, LAF is an approximate solution of the minimization problem in (21).

When there is no fixed-point, then either $\mathcal{U}_{r_i}(x, y) < r_i$ for every i or $\mathcal{U}_{r_i}(x, y > r_i$ for every i. The LAF output is set to $\mathcal{U}_0(x, y)$ in the former and $\mathcal{U}_1(x, y)$ in the latter case. Multiple fixed-points indicate that the distorted signal x differs significantly from the training signals in the subjective quality assessment database. This is due to e.g. an unknown distortion type or unexpected visual content. In such situations, the LAF system has the option to make no quality prediction.

The LAF system depends on only one tuning parameter, namely n, the number of fusion units. Increasing the value of n generally improves the prediction performance, but also increases the computational complexity of the quality prediction. The number of computations is linear in n.

D. Corollaries

By construction, the LAF system satisfies the *consistency* and *preservation of strong reflexivity* properties, which are stated below. Other ML systems often violate these properties and are therefore more prone to unreliable quality predictions, as will be shown in Section V.

Corollary 1 (Consistency): If the signal pair (x_1, y_1) *is assigned a lower score than* (x_2, y_2) *by all input quality measures, i.e.* $\mathcal{M}_i(x_1, y_1) \leq \mathcal{M}_i(x_2, y_2)$ *for all* $i = 1, 2, \ldots, m$, *then the* LAF *output satisfies* $\text{LAF}(x_1, y_1) \leq \text{LAF}(x_2, y_2)$.

Corollary 2 (Preservation of Strong Reflexivity): Suppose that at least one input quality measure $\mathcal{M}_i$ *is reflexive, that is,* $\mathcal{M}_i(x, y) = 1$ *if and only if* $x = y$. *If the* LAF *output has one fixed-point, then also* $\text{LAF}(x, y) = 1$ *if and only if* $x = y$.

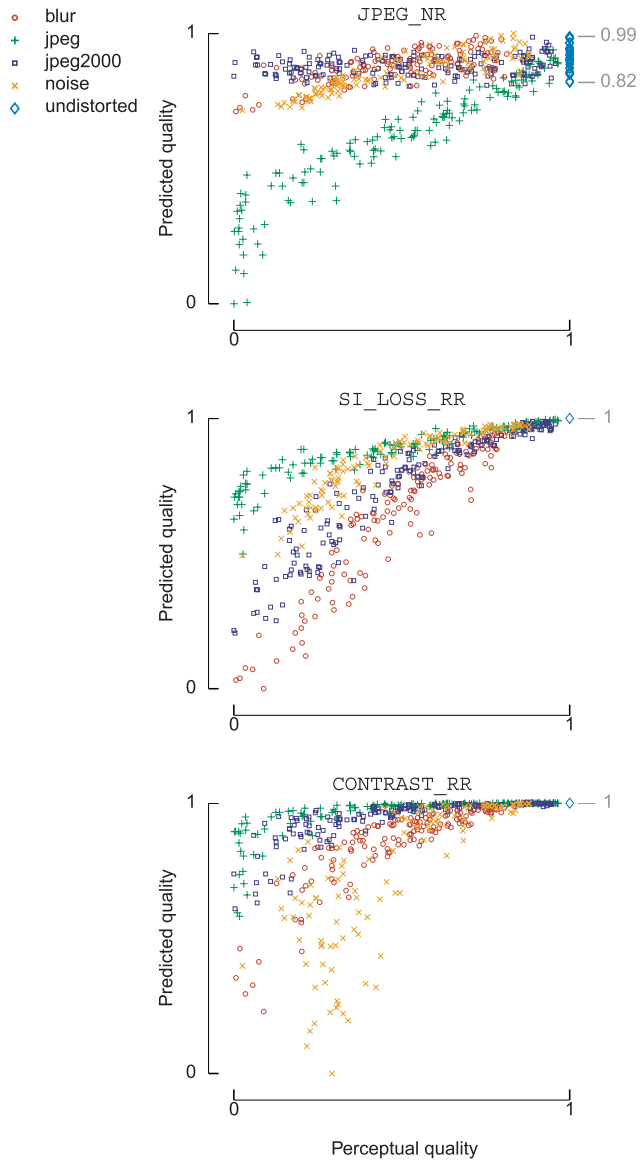

Fig. 4. The selected input objective quality measures, evaluated on a subset of the LIVE image quality assessment database [12], [13], respond very differently to the considered distortion types. For example, the JPEG_NR measure is only reliable for JPEG compression, while the CONTRAST_RR measure is more sensitive to white noise than the SI_LOSS_RR measure. In contrast to the CONTRAST_RR and SI_LOSS_RR measures, the JPEG_NR measure fails to assign a perfect quality score to the reference signals.

IV. System Implementation

This section provides a possible implementation of the LAF system, which involves one NR and two RR objective quality measures for images. We focused on four standard distortion types that are considered in most subjective image quality assessment databases: Gaussian blur, JPEG compression, JPEG2000 compression and white noise.

The input quality measures, denoted by JPEG_NR, SI_LOSS_RR, and CONTRAST_RR, are each designed to measure a specific aspect of perceptual quality. They all respond differently to the selected distortion types (Fig. 4). For consistency reasons, the input quality measures are linearly scaled to the interval $[0, 1]$.

The first input quality measure, JPEG_NR, is developed in [25] for the NR quality assessment of JPEG compressed images. The quality estimation is based on three image features that measure the intensity of the blocking artifacts and the attenuation of detail in the DCT blocks. The second input quality measure, SI_LOSS_RR, which is part of the NTIA General Video Quality Model [22], detects a decrease or loss of spatial information (e.g. blurring). We applied the SI_LOSS_RR measure on disjoint image patches of 24×24 pixels. The third input quality measure, CONTRAST_RR, is the contrast component of the popular SSIM index [19], [20]. By construction, the CONTRAST_RR measure is highly effective to detect noise contamination in an image. Like with SI_LOSS_RR, we applied the CONTRAST_RR measure on disjoint image patches of 24×24 pixels.

We trained the LAF system on the LIVE subjective quality assessment database (see Section II-A). The LIVE database includes the four distortion types that are listed above. The fading distortion type is excluded. To better align the maximum DMOS scores for each distortion type (resp. 93, 109, 91 and 112 for blur JPEG, JPEG2000 and noise), we removed all images with a DMOS score higher than 91. We also removed the three JPEG-compressed images with a negative DMOS score. In this way, the corresponding perceptual quality operator Q as defined in (3) satisfies the strong reflexivity property (1). The adjusted LIVE image quality assessment database contains 580 distorted images and 29 reference images in total.

Note that a linear regression system, applied on the three selected quality measures, cannot simultaneously optimize the local prediction accuracy for all quality scores. On one hand, it should assign a positive weight to the JPEG_NR measure to better predict the perceptual quality of JPEG compressed images. On the other hand, any nonzero weight for the JPEG_NR measure reduces the local prediction performance in the high quality range, as this measure is not suitable for nearly imperceptible distortions (Fig. 4).

The LAF system adapts the weights of the input quality measures, depending on the responses of the fusion units (Fig. 7). These fusion units are determined by fitting the conditional mean and standard deviation of the objective quality measures $\mathcal{M}_{i,j} = (\mathcal{M}_i + \mathcal{M}_j)/2$ for $1 \leq i \leq j \leq 3$. To this end, we proposed a logistic fitting method based on an initial sequence-based nonparametric fitting (see Section III-B). This method can better handle heteroscedastic and non-uniformly distributed data (Fig. 8). The sequence-based nonparametric fittings are closely approximated by the generalized logistic functions (Fig. 5). The fusion units and the final LAF-based quality measure for $n = 5$, are visualized in Fig. 6.

V. Experimental Verification

To verify the prediction performance, we evaluated the LAF system on three subjective quality assessment databases (part A) and performed additional stress tests on a large, unannotated image database (part B).

We compared the performance of the LAF system with one nonparametric, one linear regression, and one nonlinear

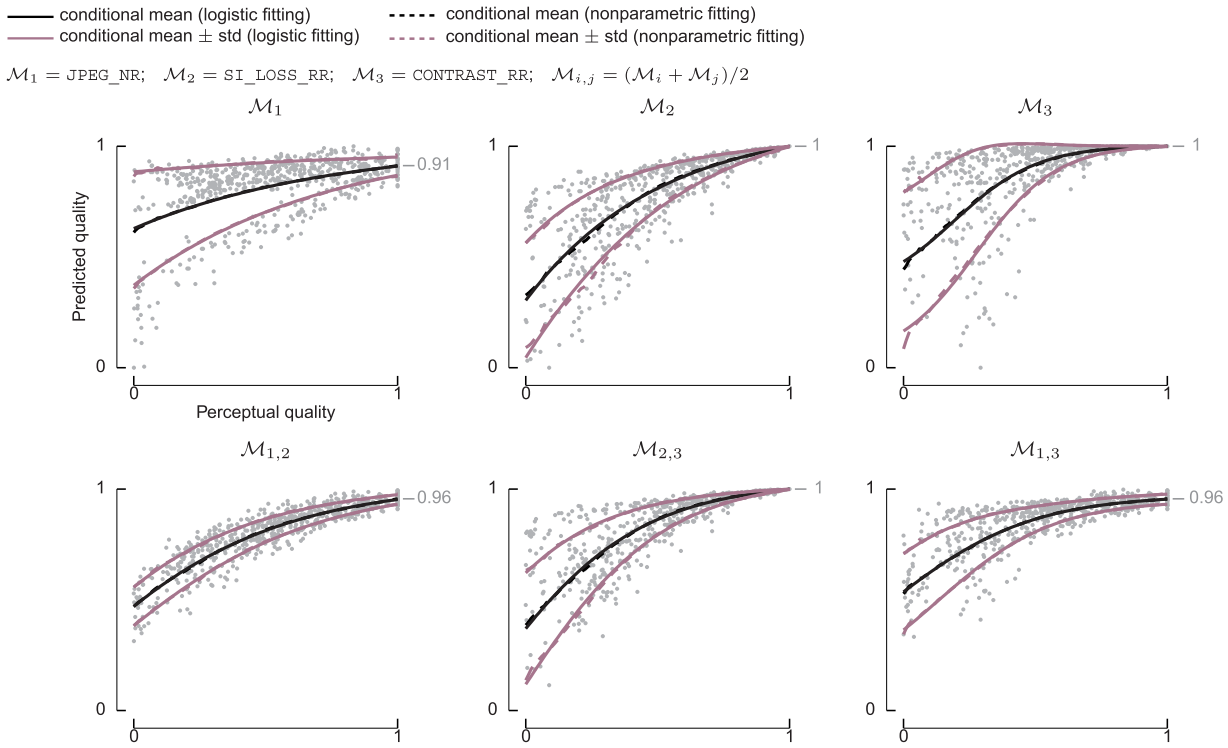

Fig. 5. The fusion units are determined by fitting the conditional mean and standard deviation of the above six objective quality measures. The initial (sequence-based) nonparametric fittings are closely approximated by logistic fittings.

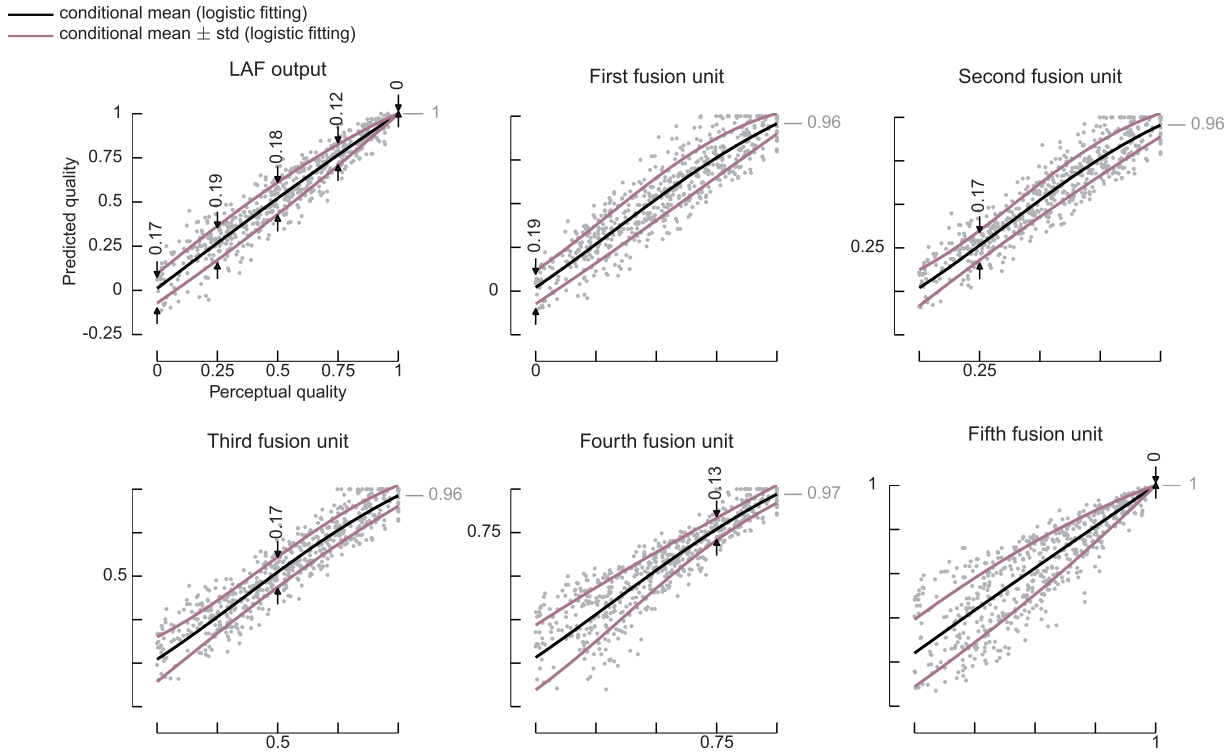

Fig. 6. In the above plots, the LAF system and the five employed fusion units are evaluated on the selected subset of the LIVE quality assessment database. The LAF system achieves a high prediction accuracy on the entire quality range by cleverly weighting the locally optimized fusion units. By construction, the fusion units maximize the local prediction accuracy near their target values (resp. 0, 0.25, 0.5, 0.75, and 1).

regression system. Among the various ML systems available, we chose the *General Regression Neural Network* (GRNN), the *Principal Component Regression* (PCR) and the *Feed Forward Neural Network* (FFNN) with one hidden layer, all described in Section II-C. The correlation with perceptual quality is measured using the *Pearson Linear Correlation Coefficient* (PLCC) and the *Spearman Rank Correlation Coefficient* (SRCC).

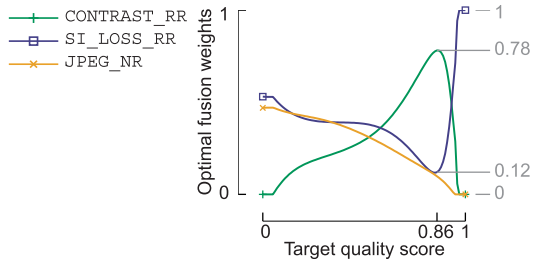

Fig. 7. The LAF system assigns different weights to the input quality measures depending on the target quality score of the fusion units. These weight variations reveal the relative importance of the input quality measures in the different quality ranges. This provides a way to interpret the learned weights. In the current system setup, the weights assigned to the JPEG_NR measure gradually decrease. Indeed, as can be observed in Figure 4, JPEG_NR is less reliable in the high quality range than the other two quality measures. The SI_LOSS_RR and CONTRAST_RR measures receive higher weights in the high quality range. However, only the SI_LOSS_RR measure is used to predict the quality of the reference images, because its conditional mean is steeper in the near-perfect quality range.

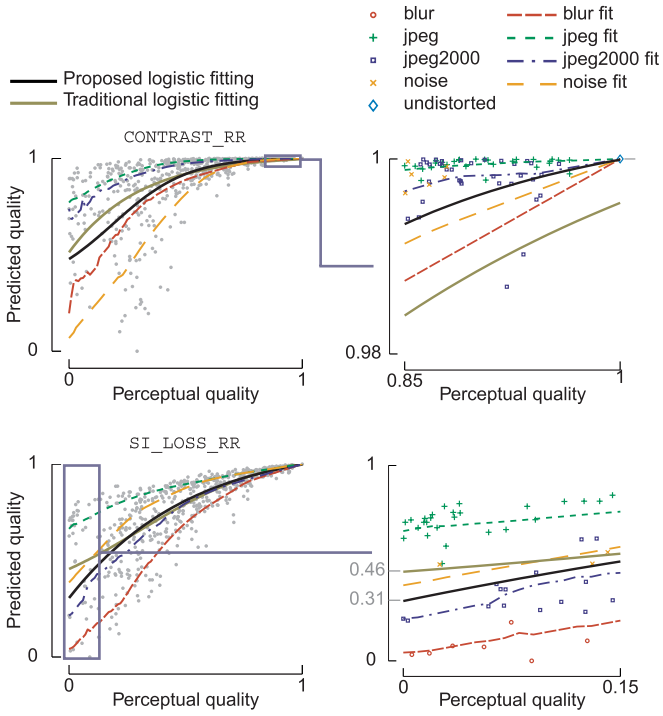

Fig. 8. The proposed logistic fitting, based on an initial sequence-based nonparametric fitting, can better handle heteroscedastic and non-uniformly distributed data. In particular, the fitting of the CONTRAST_RR measure is more accurate in the high quality range and the fitting of the SI_LOSS_RR measure is not biased towards the JPEG distorted images in the low quality range. To improve visualization, we also fitted the individual distortion types.

The computational complexity of the measurement fusion of a previously unseen signal depends on one or more of the following factors: m, the number of input quality measures, n, the number of fusion units or neurons, and p, the size of the training database. For the PCR, the computational complexity grows linearly with m. For the LAF and the one-layered FFNN, the computational complexity grows linearly with both m and n, and for these two ML systems, the total number of computations is approximately equal. For the GRNN, the computational complexity grows linearly with p. Since the value of p is much larger than the value of m or n, the computational complexity of the GRNN is typically the highest.

TABLE I
INFLUENCE OF THE LAF TUNING PARAMETER ON THE PERFORMANCE THROUGH REPEATED CROSS-VALIDATION ON THE LIVE DATABASE

# units	median PLCC	median SRCC
$n = 2$	0.9503	0.9499
$n = 5$	0.9602	0.9573
$n = 10$	0.9604	0.9582
$n = 100$	0.9605	0.9587

We implemented the ML systems as follows. For the LAF system, we set the number of fusion units to $n = 5$. More fusion units do not significantly improve the correlation with perceptual quality, determined through a repeated cross-validation procedure on the LIVE database (Table I). The PCR implementation is based on the MATLAB functions princomp and regress, and the output is normalized using a logistic function [46]. The GRNN is based on the MATLAB function newgrnn. We obtained the best results for the spread factor $h = 0.04$, the same optimal value found in [34]. The FFNN is implemented using the neural network toolbox, adopted to the Levenberg-Marquardt algorithm [51]. It includes an *early-stopping* method to improve the neural network generalization performance. The hidden and output transfer functions are $h(x) = \tanh(x)$ and $g(x) = x$, respectively. We empirically set the number of hidden neurons to $n = 3$. To avoid local minima, we re-initialized the FFNN training 10 times and selected the weights with the smallest mean squared error.

A. Evaluation on Annotated Quality Assessment Databases

We conducted three tests that validate and compare the prediction accuracy of the selected ML systems. To this end, we employed the three subjective quality assessment databases described in Section II. The resulting PLCC and SRCC correlation coefficients are listed in Table II. As a first test, we performed a repeated cross-validation on the adjusted LIVE database, as described in Section IV. We iteratively split the images in the database in a training and a test set according to the reference images. At each iteration step, the training and test set consist of 26 and 3 reference images, respectively, and their associated distorted images. We evaluated the prediction performance of the selected ML systems for all 3654 possible database divisions. The FFNN and GRNN perform slightly better than the proposed LAF system in the cross-validation test. However, the results of a cross-validation procedure are often biased by the peculiarities of the subjective quality assessment database, such as the recurrence of the same degradation levels [52].

As a second test, we verified the database independence of the ML systems by training on the adjusted LIVE database and testing on the CSIQ database, restricted to the four standard distortion types. The proposed LAF system

TABLE II
PREDICTION PERFORMANCE OF THE ML SYSTEMS

		LAF	FFNN	GRNN	PCR
Test 1	Repeated cross-validation on the LIVE database [12], [13]				
	median PLCC	0.960	0.965	0.960	0.944
	median SRCC	0.957	0.963	0.961	0.944
	std PLCC	0.009	0.009	0.010	0.012
	std SRCC	0.011	0.011	0.012	0.012
Test 2	Database independence *Training set:* LIVE [12], [13] - *Test set:* CSIQ [14], [15]				
		LAF	FFNN	GRNN	PCR
	PLCC	0.967	0.959	0.955	0.953
	SRCC	0.963	0.961	0.956	0.937
Test3	Robustness for unknown distortions *Training set:* LIVE [12], [13] - *Test set:* TID [16], [17]				
		LAF	FFNN	GRNN	PCR
	PLCC	0.822	0.790	0.794	0.808
	SRCC	0.810	0.793	0.786	0.797

achieves the highest correlation with the perceptual quality scores.

As a third test, we analyzed the robustness of the ML systems to unknown distortion types. The ML systems, trained on the adjusted LIVE database, are now evaluated on the TID database, containing no less than 17 different distortion types. In this last test, the nonlinear FFNN and GRNN systems even perform worse than the linear PCR system. The proposed LAF system has again the best performance.

According to the test results, the proposed LAF system can compete with the other ML-systems in terms of prediction accuracy. Part B of this section further investigates the *prediction robustness* of the LAF system by means of a stress testing methodology for objective quality measures.

B. Complementary Stress Tests

Subjective quality assessment databases constitute essential ground truth for the validation of objective quality measures. However, due to the high time and cost requirements, they are typically very limited in size and cannot cover the various content available in actual applications.

Recently, Ciaramello and Reibman developed a stress testing model to expose potential vulnerabilities in the design of objective quality measures [52]. Based on this approach, we performed three stress tests to validate the robustness of the considered ML systems. These stress tests do not require any time-consuming subjective experiments, so that much larger databases can be employed.

Our stress test database consists of 650 reference images from the publicly available quality image collection of Wikimedia Commons [53]. Together, these images cover a great variety in content, including animals, buildings, natural scenes, people and sport events (Fig. 9). All images follow the strict quality guidelines described in [54]. The reference images are systematically degraded at 10 different distortion levels for each of the four standard distortion types. The resulting stress test database contains 26 000 distorted images and is therefore more than 40 times larger than the previously employed LIVE database.

Fig. 9. The constructed stress test database with 650 reference images and 26 000 distorted images covers a great variety in content, including animals, buildings, natural scenes, people, and sport events.

We conducted three stress tests that verify the reliability of the ML-based quality predictions based on the following rules:

1) the ML-based quality predictions of some distorted images should be monotonically increasing when the quality predictions from the input quality measures are monotonically increasing;
2) the ML-based quality predictions of the undistorted reference images should be the highest;
3) the ML-based quality predictions should increase when the degradation level decreases, while the reference image and distortion type are kept fixed.

The first two stress tests verify the *consistency* and the *preservation of strong reflexivity* rules, which are always satisfied by the proposed LAF system (see Section III-D). However, these two rules are not always satisfied by the traditional ML systems. On the stress test database, the consistency rule is often violated by the FFNN and the GRNN systems, which respectively produce a total of 248 294 and 1 583 722 inconsistencies (Fig. 10). The strong reflecivity rule is not preserved by the FFNN, GRNN, and PCR systems (Fig. 11). The FFNN and GRNN systematically underrate the perceptual quality of the reference images in the stress test database. The PCR output scores vary from 0.75 to 1.13. Only the LAF system correctly assigns a score of 1 to all reference images.

The third stress test investigates the number of *false orderings* (FOs) per image sequence. Here, each image sequence $\mathbf{x}^{y,t}$ consists of all images with the same distortion type t and reference image y, ordered at decreasing distortion levels. A false ordering of an objective quality measure $\mathcal{M}$ is an image pair $(x_{k_1}^{y,t}, x_{k_2}^{y,t})$ with $k_1 < k_2$, for which $\mathcal{M}(x_{k_1}^{y,\ell}, y) > \mathcal{M}(x_{k_2}^{y,l}, y)$. The proposed LAF system produces significantly less false orderings than all other ML systems: it produces no more 6 false orderings with a maximum of 1 false ordering per sequence (Fig. 12). The FFNN, GRNN, and PCR systems respectively produce a total of 119, 2383 and 342 false orderings.

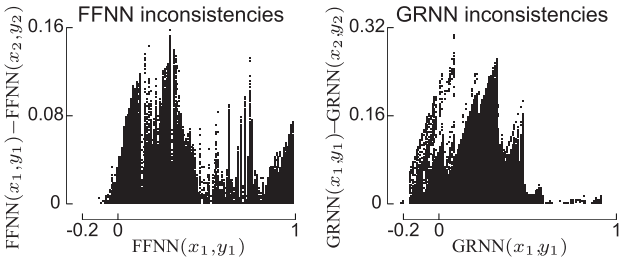

Fig. 10. The above graphs visualize every two image pairs (x_1, y_1) and (x_2, y_2) from the stress test database for which the FFNN and GRNN systems are inconsistent. The first image pair is given a higher score by the ML system, but a lower score by all input quality measures. The score differences between inconsistent pairs reach values close to 0.16 for FFNN and 0.32 for GRNN. The PCR and the proposed LAF system are always consistent.

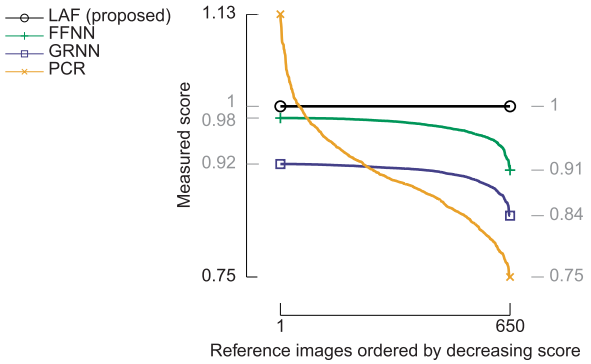

Fig. 11. The FFNN and GRNN systematically underrate the perceptual quality of the reference images in the stress test database. The PCR output scores vary from 0.75 to 1.13. Only the LAF system correctly assigns a score of 1 to all reference images.

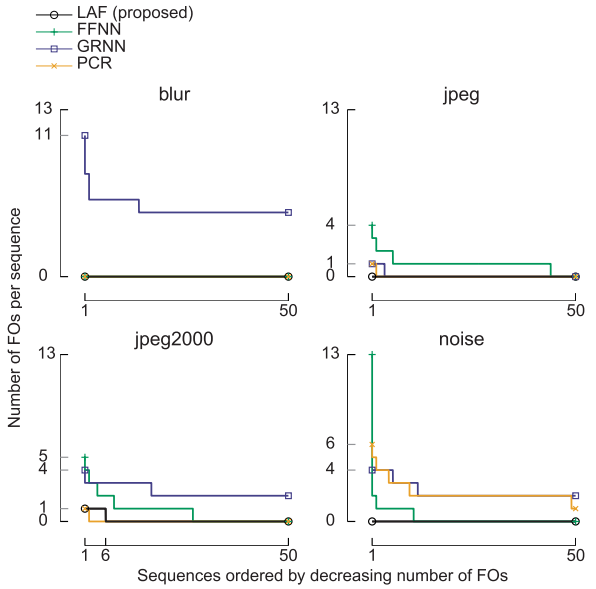

Fig. 12. The above figures show the number of false orderings (FOs) per sequence in the stress test database. The FFNN, GRNN and PCR systems respectively produces a total of 119, 2383 and 342 FOs. The proposed LAF systems produces no more than 6 FOs with a maximum of 1 FO per sequence.

VI. Conclusion

The presented *Locally Adaptive Measure Fusion* (LAF) addresses important issues of *machine learning* (ML) inherent to objective quality assessment, which are summarized in Table III. The LAF system involves a training phase and an application phase (Fig. 1). The training phase is based on three new concepts: the *separation ratio* (Fig. 2), the *fusion units* (Fig. 6), and the *initial nonparametric fittings* (Figs. 5, 8). The application phase consists of two steps (Fig. 3). First, the *interpolation* of the fusion responses is calculated. Second, the *fixed-point* of the interpolation line is determined. This fixed-point is the final quality prediction of the distorted signal.

TABLE III
Main Characteristics of the Machine Learning Systems

	LAF	FFNN	GRNN	PCR
Nonlinear Response	✓	✓	✓	×
Interpretability	✓	×	✓	✓
Reproducibility	✓	×	✓	✓
Computational Efficiency	✓	✓	×	✓
Consistency	✓	×	×	✓
Strong Reflexivity	✓	×	×	×

Nonlinear Response and Interpretability. The LAF system adapts the weights of the input quality measures in a comprehensive manner (Fig. 7). As a result, the LAF system is more flexible than linear regression systems (e.g. PCR).

Reproducibility. Unlike most neural network implementations, the LAF training process does not require a random initialization. Hence, retraining the LAF system on the same data always produces the same weights. The training of the LAF system relies on a set of convex quadratic programming problems that can be efficiently solved. The solution of such an optimization problem is always a global minimum.

Computational Efficiency. The computational complexity of the LAF system for the quality prediction of an unknown signal increases linearly with n, the number of fusion units. Increasing the value of n generally increases the prediction accuracy and converges very rapidly to the optimal solution (Table I). We already obtained good results for $n = 5$.

Consistency and Strong Reflexivity. As shown by our experimental results on the LIVE, CSIQ, and TID databases, the LAF system can compete with the traditional ML systems (Table II). According to the complementary stress tests, the traditional ML systems are often inconsistent with the input quality measures and do not preserve strong reflexivity (Figs. 10 and 11). The LAF system does not suffer from these problems, and is therefore less vulnerable to unreliable quality predictions (Corollaries 1 and 2). In fact, the LAF system produces less false orderings than the other ML systems when the degradation level is systematically decreased (Fig. 12).

According to the performed experiments, the proposed LAF system yields a higher robustness compared to the traditional ML systems for objective quality assessment. As a result, the LAF system may lead to more reliable objective quality measures for images and videos. In addition, the LAF system may also be applied for the objective quality assessment of other media types (e.g. stereoscopic or audiovisual content).

APPENDIX: PROOF OF THEOREM 1

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two objective quality measures with $\text{sep}_{\mathcal{M}_1}[r] \neq \text{sep}_{\mathcal{M}_2}[r]$ and assume that $\mathcal{M}_1(X_q)$ and $\mathcal{M}_2(X_q)$ are normally distributed. Theorem 1 states that $\text{sep}_{\mathcal{M}_1}[r] > \text{sep}_{\mathcal{M}_2}[r]$ if and only if $\text{acc}_{\mathcal{M}_1}[r,\epsilon] > \text{acc}_{\mathcal{M}_2}[r,\epsilon]$ for every sufficiently small, positive ϵ.

The expression $\text{acc}_{\mathcal{M}_1}[r,\epsilon] > \text{acc}_{\mathcal{M}_2}[r,\epsilon]$ is equivalent to

$$\mathbb{P}[\Delta_{1,\epsilon} > 0] > \mathbb{P}[\Delta_{2,\epsilon} > 0] \tag{23}$$

where $\Delta_{i,\epsilon} = \mathcal{M}_i(X_{r+\epsilon}) - \mathcal{M}_i(X_{r-\epsilon})$.

We now show that the separation ratio can be expressed in terms of the difference variable $\Delta_{i,\epsilon}$. To simplify notations, set $f_{\mu,i}(q) = \mathbb{E}[\mathcal{M}_i(X_q)]$, $f_{\sigma,i}(q) = \text{std}(\mathcal{M}_i(X_q))$ and $\lambda_i(q) = \text{sep}_{\mathcal{M}_i}[q]$. By definition,

$$f'_{\mu,i}(r) - \lambda_i(r) f_{\sigma,i}(r) = 0. \tag{24}$$

By Taylor's theorem, there exists $\eta_{i,\epsilon}$, with $\eta_{i,\epsilon} = O(\epsilon^2)$ as $\epsilon \to 0$, such that

$$f'_{\mu,i}(r) = \frac{\mathbb{E}[\Delta_{i,\epsilon}]}{2\epsilon} + \eta_{i,\epsilon} \tag{25}$$

On the other hand,

$$f_{\sigma,i}(r) \ = \frac{1}{\sqrt{2}} K_{i,\epsilon} \text{std}[\Delta_{i,\epsilon}] \tag{26}$$

where $K_{i,\epsilon} = \sqrt{2} f_{\sigma,i}(r)/\text{std}[\Delta_{i,\epsilon}]$. As $X_{r+\epsilon}$ and $X_{r-\epsilon}$ are independent values, we have

$$\begin{aligned} \text{var}\big[\Delta_{i,\epsilon}\big] &= \text{var}\big[\mathcal{M}_i(X_{r+\epsilon})\big] + \text{var}\big[\mathcal{M}_i(X_{r-\epsilon})\big] \\ &= f^2_{\sigma,i}(r+\epsilon) + f^2_{\sigma,i}(r-\epsilon). \end{aligned}$$

Hence,

$$K_{i,\epsilon} = \left(\frac{2 f^2_{\sigma,i}(r)}{f^2_{\sigma,i}(r+\epsilon) + f^2_{\sigma,i}(r-\epsilon)} \right)^{1/2},$$

so $K_{i,\epsilon}$ converges to 1 as $\epsilon \to 0$.

Substituting (25) and (26) in Eq. (24) and multiplying both sides with 2ϵ gives

$$\mathbb{E}[\Delta_{i,\epsilon}] - \sqrt{2}\epsilon \lambda_i(r) \tilde{K}_{i,\epsilon} \text{std}[\Delta_{i,\epsilon}] = 0, \tag{27}$$

where we put $\tilde{K}_{i,\epsilon} = K_{i,\epsilon} - \sqrt{2}\eta_{i,\epsilon}/(\lambda_i(r)\text{std}[\Delta_{i,\epsilon}])$. By construction, $\tilde{K}_{i,\epsilon}$ also converges to 1 as $\epsilon \to 0$. Eq. (27) implies that

$$\mathbb{P}[\Delta_{i,\epsilon} > 0] = \mathbb{P}\left[\frac{\Delta_{i,\epsilon} - \mathbb{E}[\Delta_{i,\epsilon}]}{\text{std}[\Delta_{i,\epsilon}]} > -\sqrt{2}\epsilon \lambda_i(r) \tilde{K}_{i,\epsilon} \right].$$

The difference variable $\Delta_{i,\epsilon}$ is normally distributed, because $\mathcal{M}_i(X_{r+\epsilon})$ and $\mathcal{M}_i(X_{r-\epsilon})$ are independent and normally distributed. We obtain that

$$\mathbb{P}[\Delta_{i,\epsilon} > 0] = \int_{-\sqrt{2}\epsilon\lambda_i(r)\tilde{K}_{i,\epsilon}}^{\infty} \phi(x)\, dx \tag{28}$$

where ϕ is the standard normal probability density function. Theorem 1 immediately follows from Eq. (28). Indeed, suppose that $\text{sep}_{\mathcal{M}_1}[r] > \text{sep}_{\mathcal{M}_2}[r]$. When ϵ is sufficiently small, the lower bound of the integral in Eq. (28) is smaller for $i = 1$ than for $i = 2$ and therefore, Eq. (23) is satisfied. This shows that $\text{acc}_{\mathcal{M}_1}[r,\epsilon] > \text{acc}_{\mathcal{M}_2}[r,\epsilon]$. Similarly, if $\text{sep}_{\mathcal{M}_1}[r] < \text{sep}_{\mathcal{M}_2}[r]$, then $\text{acc}_{\mathcal{M}_1}[r,\epsilon] < \text{acc}_{\mathcal{M}_2}[r,\epsilon]$.

REFERENCES

[1] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU Rec. BT.500-11, Geneva, Switzerland, 2002.

[2] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. New York, NY, USA: Wiley, 2005.

[3] W. Lin and C. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011.

[4] Z. Wang, H. R. Sheikh, and A. C. Bovik, "Objective video quality assessment," in *The Handbook of Video Databases: Design and Applications*. Boca Raton, FL, USA: CRC Press, 2003, ch. 41, pp. 1041–1078.

[5] P. Gastaldo and J. Redi, "Machine learning solutions for objective quality assessment," in *Proc. 6th Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, 2012.

[6] C. Keimel, M. Rothbucher, S. Hao, and K. Diepold, "Video is a cube," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 41–50, Nov. 2011.

[7] A. Barri, A. Dooms, and P. Schelkens, "Combining the best of perceptual quality metrics," in *Proc. 6th Int. Workshop Video Process. Qual. Metrics Consum. Electron.*, 2012.

[8] H. D. Burkhard and M. M. Richter, "On the notion of similarity in case-based reasoning and fuzzy theory," in *Soft Computing in Case Based Reasoning*. Berlin, Germany: Springer-Verlag, 2000, pp. 29–46.

[9] *Method for Specifying Accuracy and Cross-Calibration of Video Quality Metrics*, document ITU Rec. J.149, Geneva, Switzerland, 2004.

[10] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[11] K. Fliegel and C. Timmerer, *WG4 Databases White Paper v1.5: QUALINET Multimedia Database Enabling QoE Evaluations and Benchmarking*. Munich, Germany: Prague/Klagenfurt, Mar. 2013.

[12] H. R. Sheikh, Z. Wang, L. K. Cormack, and A. C. Bovik. (2003). *LIVE Image Quality Assessment Database Release 2* [Online]. Available: http://live.ece.utexas.edu/research/quality

[13] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[14] E. C. Larson and D. M. Chandler. (2009). *Consumer Subjective Image Quality Database*. [Online]. Available: http://vision.okstate.edu/index.php?loc=csiq

[15] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, p. 011006, 2010.

[16] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. (2008). *Tampere Image Quality Database* [Online]. Available: http://www.ponomarenko.info/tid2008.htm

[17] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 1, pp. 30–45, 2009.

[18] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[20] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 1998.

[21] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[22] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[23] S. Wu, W. Lin, S. Xie, Z. Lu, E. Ong, and S. Yao, "Blind blur assessment for vision-based applications," *J. Vis. Commun. Image Represent.*, vol. 20, no. 4, pp. 231–241, 2009.

[24] P. Marzilano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. IEEE ICIP*, vol. 3. Jan. 2002, pp. 57–60.

[25] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1. Jan. 2002, pp. 1–477.

[26] S. Suresh, R. Venkatesh Babu, and H. J. Kim, "No-reference image quality assessment using modified extreme learning machine classifier," *Appl. Soft Comput.*, vol. 9, no. 2, pp. 541–552, Mar. 2009.

[27] R. V. Babu, S. Suresh, and A. Perkis, "No-reference JPEG-image quality assessment using GAP-RBF," *Signal Process.*, vol. 87, no. 6, pp. 1493–1503, 2007.
[28] Z. M. P. Sazzad, Y. Kawayoke, and Y. Horita, "No reference image quality assessment for JPEG2000 based on spatial features," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 257–268, Apr. 2008.
[29] H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.
[30] T. Brandao and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.
[31] T. Oelbaum, C. Keimel, and K. Diepold, "Rule-based no-reference video quality evaluation using additionally coded videos," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 294–303, Apr. 2009.
[32] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.
[33] A. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
[34] C. Li, A. C. Bovik, and W. Xiaojun, "Blind image quality assessment using a general regression neural network," *IEEE Trans. Neural Netw.*, vol. 22, no. 5, pp. 793–799, May 2011.
[35] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY, USA: Springer-Verlag, 2009.
[36] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, no. 6, pp. 568–576, Nov. 1991.
[37] H. Martens and M. Martens, *Multivariate Analysis of Quality*. New York, NY, USA: Wiley, 2001.
[38] C. Keimel, M. Rothbucher, and K. Diepold, "Extending video quality metrics to the temporal dimension with 2D-PCR," *Proc. SPIE*, vol. 7867, pp. 786713-1–786713-10, Jan. 2011.
[39] E. Moody, "The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems," *Adv. Neural Inf. Process. Syst.*, vol. 4, pp. 847–854, Jun. 1992.
[40] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*. Cambridge, MA, USA: MIT Press, 1995.
[41] S. Ridella, S. Rovetta, and R. Zunino, "Circular back-propagation networks for classification," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 84–97, Jan. 1997.
[42] P. Gastaldo, S. Rovetta, and R. Zunino, "Objective quality assessment of MPEG-2 video streams by using CBP neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 939–947, Jul. 2002.
[43] P. Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.
[44] M. Narwaria and W. Lin, "Objective image quality assessment based on support vector regression," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 515–519, Mar. 2010.
[45] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion (MMF)," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1793–1807, May 2013.
[46] (2000, Mar.). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality* [Online]. Available: http://www.vqeg.org/
[47] C. Heij, P. de Boer, P. Franses, T. Kloek, and H. K. van Dijk, *Econometric Methods With Applications in Business and Economics*. Oxford, U.K.: Oxford Univ. Press, 2004.
[48] L. Chen, M. Cheng, and L. Peng, "Conditional variance estimation in heteroscedastic regression models," *J. Statist. Planning Inference*, vol. 139, no. 2, pp. 236–245, 2009.
[49] J. Fan and Q. Yao, "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, vol. 85, no. 3, pp. 645–660, 1998.
[50] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Berlin, Germany: Springer-Verlag, 2006.
[51] M. Beale, M. Hagan, and H. Demuth, *Neural Network Toolbox User's Guide (R2012a)*. Natick, MA, USA: Mathworks, 2012.
[52] F. Ciaramello and A. Reibman, "Systematic stress testing of image quality estimators," in *Proc. 18th IEEE ICIP*, Sep. 2011, pp. 3101–3104.
[53] Wikimedia Commons. (2013). *Quality Image Collection* [Online]. Available: http://commons.wikimedia.org/wiki/Commons:QI
[54] Wikimedia Commons. (2008). *Image Guidelines* [Online]. Available: http:// commons.wikimedia.org/wiki/Commons:Quality_images_guidelines

Adriaan Barri received the M.Sc. degree in mathematics from Vrije Universiteit Brussel (VUB), Belgium, in 2010, where he is currently pursuing the Ph.D. degree in electrical engineering. He holds a Ph.D. bursary from the Flemish Institute for the promotion of Science and Technology (IWT). He is a member of iMinds and the Department of Electronics and Informatics at VUB. His research interests include statistical data analysis and perceptual image and video quality assessment algorithms.

Ann Dooms (M'07) received the master's and Ph.D. degrees in mathematics from Vrije Universiteit Brussel, Brussels, Belgium, in 2000 and 2004, respectively, where she is currently a Professor with the Department of Electronics and Informatics, which is part of iMinds. She leads a research team in Multimedia Forensics, which studies the lifecycle of a multimedia item or its source solely through digital image processing to provide answers to forensic questions ranging from authenticity and traitor tracing over perceptual quality and compressed sensing to digital painting analysis.

Bart Jansen was born in 1979. He received the master's degree in computer science, the master's degree in computer science in artificial intelligence, and the Ph.D. degree in computer science in 2001, 2003, and 2005, respectively. During the Ph.D. degree with the Artificial Intelligence Laboratory, Vrije Universiteit Brussel, he was headed by Prof. L. Steels, where he was a Post-Doctoral Researcher with the Electronics and Engineering Department, headed by Prof. J. Cornelis and Prof. R. Vounckx, and where he was involved in building computer systems, which can detect, classify, and understand aspects of human behavior. During his Ph.D. research, the specific research domain was robot imitation, which is a part of cognitive robotics and developmental robotics. Since 2006, his research has focused on e-health and recognizing human behavior for medical applications for instance targeting the monitoring of physical activity of geriatric patients. His research combines aspects from computer vision (e.g., color segmentation, stereo vision, and 3-D vision) with artificial intelligence (e.g., pattern recognition and machine learning) applied to the domain of biomedical engineering.

Peter Schelkens is currently a Professor with the Department of Electronics and Informatics, Vrije Universiteit Brussel, and a Research Director with the iMinds Research Institute. In 2010, he became a Board of Councilors Member of the Interuniversity Microelectronics Institute, Belgium.

His research interests are situated in the field of multidimensional signal processing, in particular, focusing on cross-disciplinary research. In 2013, he received an ERC Consolidator Grant focusing on digital holography.

Dr. Schelkens has co-edited the books *The JPEG 2000 Suite* (Wiley, 2009) and *Optical and Digital Image Processing* (Wiley, 2011). He is an Elected Committee Member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee and the IEEE Multimedia Signal Processing Technical Committee. He is participating in the ISO/IEC JTC1/SC29/WG1 (JPEG) and WG11 (MPEG) standardization activities.