PAPER

Revisiting the Regression between Raw Outputs of Image Quality Metrics and Ground Truth Measurements

Chanho JUNG^{†a)}, Member, Sanghyun JOO^{††b)}, Do-Won NAM^{††c)}, Nonmembers, and Wonjun KIM^{†††d)}, Member

SUMMARY In this paper, we aim to investigate the potential usefulness of machine learning in image quality assessment (IQA). Most previous studies have focused on designing effective image quality metrics (IQMs), and significant advances have been made in the development of IQMs over the last decade. Here, our goal is to improve prediction outcomes of “any” given image quality metric. We call this the “IQM’s Outcome Improvement” problem, in order to distinguish the proposed approach from the existing IQA approaches. We propose a method that focuses on the underlying IQM and improves its prediction results by using machine learning techniques. Extensive experiments have been conducted on three different publicly available image databases. Particularly, through both 1) in-database and 2) cross-database validations, the generality and technological feasibility (in real-world applications) of our machine-learning-based algorithm have been evaluated. Our results demonstrate that the proposed framework improves prediction outcomes of various existing commonly used IQMs (e.g., MSE, PSNR, SSIM-based IQMs, etc.) in terms of not only prediction accuracy, but also prediction monotonicity.

key words: machine learning, image quality assessment (IQA), image quality metric (IQM), DMOS, human visual system

1. Introduction

Developing objective image quality assessment (IQA) algorithms has attracted much research interest during the past decade. Such a machine evaluation of picture quality plays an essential role in a wide range of applications, including image acquisition, processing, compression, transmission, reproduction, display, in-service quality assessment in visual communication networks [1], [2]. The most accurate and reliable method of measuring image quality is “subjective” evaluation by human observers [3], [4]. However, it is time-consuming, cumbersome, expensive, and cannot be used in real-world applications. For this reason, the development of expert systems to predict image quality in a way that is “consistent” with subjective human evaluation is the goal of objective IQA research [4], [5]. Namely, the ultimate objective (of quality assessment research) is to predict subjective

quality rating (ground truth measurement, i.e., DMOS*) of a visual signal. Note that the closer to the subjective quality ratings the system’s assessments of quality are, the better the system is. A large number of IQA systems have been reported in the literature so far. Depending upon the availability of reference image, they are mainly classified as full-reference (FR), reduced-reference (RR), and no-reference (NR) systems. Each of these three IQA methods has potential in different applications, and this paper gives particular attention to the problem of “FR” IQA.

Most previous FR IQA approaches have focused on designing effective image quality metrics (IQMs) [5]–[12]. Here, our purpose is to improve prediction outcomes of “any” given image quality metric by using machine learning techniques. We refer to this problem as “IQM’s Outcome Improvement”, to distinguish our work from the existing IQA approaches. Our proposed approach is heavily influenced by recent promising results demonstrating the usefulness of machine learning techniques to predict perceptual image quality [13]–[19]. We propose a method for improving prediction results of a given IQM by using machine learning methods. In this paper, we revisit the problem of correlating (regressing) the outputs of IQM (e.g., MSE, PSNR, and SSIM [5]) with subjective scores, which are “ground truth” measurements. Note that, until now, the regression between the rating values of IQM and subjective ratings has been mainly performed as a part of the performance assessment (PA) of IQM [20], [21] (refer to Sect. 2.2). As a result, after careful analysis, we find that a machine-learning-based correlation between the two ratings has great ability of improving prediction outcomes of existing IQMs. We show how such machine learning techniques can be effectively used in conjunction with the underlying IQM, and prove that the proposed approach provides a means of “extending” existing IQMs: we will demonstrate that our proposed framework enhances prediction results of IQMs with respect to prediction accuracy and monotonicity (refer to Sect. 3). Figure 1 shows an example of comparison between naive PSNR and machine-learning-based PSNR (the proposed method) for TID2008 database [22]. As shown in Fig. 1, the naive PSNR fails to predict the visual quality of distorted images, whereas our machine-learning-based PSNR correctly predicts the visual quality.

The rest of this paper is organized as follows. Section 2 describes the technical details for IQA. Extensive experi-

Manuscript received March 24, 2015.

Manuscript revised May 19, 2016.

Manuscript publicized August 8, 2016.

[†]The author is with the Department of Electrical Engineering, Hanbat National University, Daejeon, Republic of Korea.

^{††}The authors are with the SWContent Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), Yuseong-Gu, Daejeon 305-700, Republic of Korea.

^{†††}The author is with the Department of Electronics Engineering, Konkuk University, Seoul 05029, Republic of Korea.

a) E-mail: peterjung@hanbat.ac.kr

b) E-mail: joos@etri.re.kr

c) E-mail: dwnam@etri.re.kr

d) E-mail: wonjkim@konkuk.ac.kr (Corresponding author)

DOI: 10.1587/transinf.2015EDP7099

*Differential mean opinion score.

(a) Reference image

(b) Distorted image 1

(c) Distorted image 2

Comparison

	Distorted image 1		Distorted image 2
Ground truth (MOS)	3.4839	<	6.1935
PSNR	24.0143	>	18.6483
Machine-learning-based PSNR (Our method)	3.7198	<	5.2174

Fig. 1 Example of comparison between naive PSNR and machine-learning-based PSNR (the proposed method) for TID2008 database [22]. Higher values of MOS (0 - minimal, 9 - maximal) and PSNR correspond to higher visual quality of the image. In this example, the MOS for distorted image 2 is higher than the one for distorted image 1. Note that the naive PSNR fails to predict the visual quality of the distorted images, whereas our machine-learning-based PSNR correctly predicts the visual quality. In the testing, training examples are not included (refer to Sect. 3).

mental results are reported in Sect. 3, followed by concluding remarks in Sect. 4.

2. Methodologies

In this section, we first briefly review some representative and widely used FR IQMs (such as MSE, PSNR, and SSIM-based IQMs [5], [6], [11]) and the scheme of evaluating the performance of IQMs. Then, the fundamentals (which are the motivation of the proposed method) of recent machine-learning-based IQA systems are discussed. Finally, our proposed algorithm is presented.

2.1 Full-Reference Image Quality Metrics

The purpose of FR IQMs is to quantify the visual quality of a distorted image under the circumstances where its reference counterpart is “fully” available. For example, the mean

squared error (MSE) and peak signal-to-noise ratio (PSNR) between the two images are computed as follows:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2, \quad (1)$$

$$\text{PSNR} = 10 \log_{10} \left(\frac{255^2}{\text{MSE}} \right), \quad (2)$$

where x_i and y_i denote the i th pixel in the reference image $\mathbf{x}$ and the distorted image $\mathbf{y}$, respectively, and N denotes the total number of pixels in the image. The structural similarity (SSIM) index [5] and its derivatives (e.g., MS-SSIM [6] and IW-SSIM [11]) basically measure the similarity between the two signals based on local comparison of luminance, contrast, and structure.

2.2 Performance Assessment of Image Quality Metric

In this subsection, we briefly discuss the principle of assessing the performance of IQM, in order to avoid confusion with the proposed method. The PA scheme, which is first proposed by Video Quality Expert Group (VQEG) [21], has been adopted as a “gold standard” process in the IQA research community [20]. As pointed out in [20], the IQA research unfolds into three major steps: 1) IQM computation, 2) regression (correlation), and 3) index (such as Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SROCC)) computation. We note here that the task of PA consists of both steps 2 and 3. Particularly, as mentioned earlier, the regression between subjective (DMOSs) and objective image quality ratings is performed in step 2. Notice that, in this step, VQEG puts “two constraints” for the regression function: 1) it should be kept monotonic over the full range of data and 2) it should be best fitted in a least-square sense [20], [21]. In accordance with such constraints, VQEG has recommended some non-linear regressions such as logistic regression and polynomial regression, and they have been universally adopted by researchers [20]. Recently, Han et al. [20] have proposed a more advanced regression function satisfying the two constraints for the use of it in the PA. Note that it is not the subject of comparison with the proposed approach: in our method, (a machine-learning-based) regression between the two ratings is performed “in the quality assessment stage” (refer to Sect. 2.4). In step 3, performance indexes between the DMOSs and the predicted DMOSs (which are obtained through the regression function) are computed. Figure 2 illustrates the general framework of IQA research in the case of FR assessment: note that, in Figs. 2, 3, and 4, what is actually used in real-life application scenarios is the objective image quality score obtained from the quality assessment stage.

2.3 Machine-Learning-Based Image Quality Assessments

Thanks to the proliferation of DMOS data (ground truth measurements, e.g., [22]–[24]), many technical advances

Fig. 2 Generic framework of IQA research in the case of FR assessment. Note that the framework is largely divided into two stages: 1) objective image quality measurement (i.e., IQM computation) and 2) performance assessment. The second stage is subdivided into two steps: 1) regression and 2) index computation.

Fig. 3 Generic framework of machine-learning-based IQA systems. Note that the performance assessment stage is identical to that in Fig. 2, and thus is omitted in this figure. The inputs to the feature extraction step depend on the purpose (FR, RR, or NR assessment) of the system.

Fig. 4 Diagram of the proposed algorithm. Note that, as in the studies of machine-learning-based IQA, the DMOS data is used only in the “training” phase.

have recently been made in the development of machine-learning-based IQA systems. These systems use “supervised” learning (such as general regression neural network (GRNN) [25], support vector regression (SVR) [26], etc.) to estimate subjective quality of an image based on a set of training examples (i.e., pairs of feature vectors and DMOSs measured by human beings beforehand). More specifically, they essentially attempt to learn the functional mapping from image features (e.g., wavelet coefficients, image gradient, image entropy, phase congruency, 2D mel-cepstrum coefficients, singular value decomposition coefficients) to a desired score (i.e., “subjective” quality rating): indeed, it allows reliable prediction of subjective human evaluation [13]–[19]. For example, in [17], the relationship between visual features and DMOS is modeled by the GRNN. Moorthy and Bovik [14] also identify such a relationship using the SVR. For details, we refer the reader to [17] and [14]. Once trained, such systems do not require any prior knowledge. The general framework of machine-learning-based IQA methods in the literature is shown in Fig. 3. In this figure, the PA stage is omitted since it is identical to the one in Fig. 2. Here, it is important to note that the machine-learning-based mapping in the objective quality assessment stage is also a regression (where the outcome is a predicted DMOS) related to the subjective ratings. However, it differs from that in the PA stage, i.e., the regression in the quality assessment stage is entirely free from the aforementioned constraints in the previous subsection. In the quality assessment stage, the DMOS data is of course involved only in the “training” phase.

2.4 Proposed Method: Revisiting the Regression between the Raw Outputs of IQM and Subjective Image Quality Ratings (Ground Truth Measurements)

As described in Sect. 1, the ultimate goal of quality assessment research is to predict subjective quality rating (i.e., ground truth measurement) of visual signal. To improve prediction outcomes of the given IQM, we follow the idea of learning-based IQA approaches discussed in the previous subsection: “in the quality assessment stage”, we involve the DMOS data (quantified by human observers in advance), and take advantage of the constraint-free regression related to it. More specifically, with supervised learning techniques (e.g., GRNN, SVR, etc.), we derive a one-to-one functional relationship between the resultant response from the underlying IQM (e.g., MSE, PSNR, SSIM-based IQMs, etc.) and “subjective” quality rating (i.e., DMOS) in the quality assessment stage: we find that such a modeling has great potential of enhancing prediction results of any given IQM and providing a means of extending existing IQMs. In this paper, the general applicability of the machine-learning-based mapping between the “raw” outputs of IQM and subjective image quality scores in the quality assessment stage is demonstrated through specific examples where the GRNN [25] and SVR [26] are successfully applied. The system diagram of the proposed approach is

Fig. 5 Example of quality assessment in the proposed method. Note that, in this example, the PSNR and GRNN are used.

Table 1 Summarization of differences between existing methods and the proposed method.

	Conventional methods	Proposed method
Goal	Designing a new image quality metric	Enhancing prediction results of widely used well-established existing image quality metrics (e.g., MSE, PSNR, SSIM, etc.)
Mapping type	Many-to-one mapping	One-to-one mapping
Input data type to the mapping	Vector	Scalar
Input data content to the mapping	Feature vector	Initial quality score

shown in Fig. 4. The PA stage is also omitted in this figure (because it is identical to that in Fig. 2). Note that we employ the trained constraint-free regression function in the “testing” phase.

Figure 5 shows an example of image quality assessment in the proposed method. Noting that the PSNR and GRNN are used in this example. For an initial quality score PSNR, the output $f(\text{PSNR})$ (i.e., final quality score) of the GRNN is defined as follows:

$$f(\text{PSNR}) = \frac{\sum_{i=1}^n \text{DMOS}_i \exp(-D_i^2 / 2\sigma^2)}{\sum_{i=1}^n \exp(-D_i^2 / 2\sigma^2)}, \quad (3)$$

where n represents the number of training samples; $D_i^2 = (\text{PSNR} - \text{PSNR}_i)^2$; PSNR_i and DMOS_i are i th training input and output values, respectively; and σ is the spread parameter. The larger the value of σ , the smoother the functional approximation. In this paper, as was done in [14], a five-fold cross validation on the training set is used to select the parameter. It is worth noting here that, as mentioned above, the goal of our proposed machine-learning-based method is to enhance prediction results of widely used well-established existing image quality metrics (e.g., MSE, PSNR, SSIM, etc.), whereas the conventional machine-

learning-based methods aim to develop a new image quality metric. From the point of view of machine-learning-based mapping, the proposed machine-learning-based method is different from the previous machine-learning-based methods in three regards (see an example in Fig. 5): 1) the mapping type is a one-to-one mapping; 2) the input data type to the mapping is a scalar; and 3) the input data to the mapping is an initial quality score. The differences are summarized in Table 1. Note that, to the best of our knowledge, our work is the first to formally study the dynamic relationship between those pairs of measures (i.e., the “raw” outputs of IQM and subjective image quality scores) from the point of view of supervised machine learning in the quality assessment stage. The success of our method may be understood as a natural outcome of an effective combination of several promising approaches in the area of IQA. In the following section, to confirm the effectiveness of the proposed method, extensive experimental results on three large scale image databases are reported.

3. Experiments and Discussion

The focus of our proposed method is to improve prediction outcomes of any given IQM using machine learn-

Table 2 Prediction performance of different naive IQMs and IQMs with pretrained machine-learning-based mapping on TID2008 image database. Boldface indicates the better performer.

TID2008 Database [22] (1700 images, 17 distortions)

	Naive IQM					IQM with Pretrained Machine-Learning-Based Mapping (Our Method)				
	PLCC	RMSE	MAE	SROCC	KROCC	PLCC	RMSE	MAE	SROCC	KROCC
MSE	0.5163	1.1492	0.9308	0.5376	0.3810	0.6468	1.0234	0.8121	0.6471	0.4629
SNR	0.5272	1.1403	0.8731	0.5377	0.3843	0.5469	1.1234	0.8572	0.5483	0.3925
PSNR	0.5363	1.1326	0.9045	0.5354	0.3787	0.6303	1.0418	0.8251	0.6358	0.4534
PSNR-HVS-M [27]	0.5804	1.0927	0.7649	0.5968	0.4689	0.7123	0.9418	0.6528	0.7246	0.5604
NQM [28]	0.6096	1.0638	0.7807	0.6243	0.4608	0.6230	1.0497	0.7717	0.6282	0.4622
VSNR [9]	0.6523	1.0171	0.7151	0.7000	0.5306	0.6818	0.9817	0.6912	0.7046	0.5340
MSVD [7]	0.5709	1.1018	0.8413	0.5880	0.4288	0.5835	1.0921	0.8333	0.5957	0.4353
IFC [29]	0.5834	1.0899	0.8660	0.6323	0.4566	0.7630	0.8674	0.6941	0.7126	0.5221
VIF [8]	0.8101	0.7867	0.6025	0.7496	0.5863	0.8258	0.7568	0.5744	0.7603	0.5940
MAD [30]	0.8370	0.7343	0.5456	0.8420	0.6537	0.8389	0.7304	0.5431	0.8420	0.6537
SSIM [5]	0.8054	0.7953	0.6249	0.8081	0.6056	0.8145	0.7785	0.6072	0.8087	0.6070
MS-SSIM [6]	0.8359	0.7365	0.5786	0.8528	0.6543	0.8509	0.7049	0.5503	0.8530	0.6552
IW-SSIM [11]	0.8530	0.7003	0.5455	0.8559	0.6636	0.8676	0.6672	0.5116	0.8585	0.6665

Table 3 Prediction performance of different naive IQMs and IQMs with pretrained machine-learning-based mapping on CSIQ image database. Boldface indicates the better performer.

CSIQ Database [23] (866 images, 6 distortions)

	Naive IQM					IQM with Pretrained Machine-Learning-Based Mapping (Our Method)				
	PLCC	RMSE	MAE	SROCC	KROCC	PLCC	RMSE	MAE	SROCC	KROCC
MSE	0.7972	0.1585	0.1237	0.8279	0.6365	0.8731	0.1280	0.0939	0.8753	0.6900
SNR	0.7748	0.1660	0.1302	0.7995	0.5987	0.7915	0.1605	0.1212	0.8127	0.6175
PSNR	0.8049	0.1558	0.1202	0.8372	0.6413	0.8592	0.1343	0.0985	0.8680	0.6833
PSNR-HVS-M [27]	0.8054	0.1556	0.1139	0.8184	0.6447	0.8896	0.1199	0.0825	0.8947	0.7234
NQM [28]	0.7432	0.1756	0.1323	0.7437	0.5656	0.7463	0.1748	0.1314	0.7481	0.5695
VSNR [9]	0.8165	0.1516	0.1119	0.8275	0.6332	0.8227	0.1492	0.1098	0.8323	0.6419
MSVD [7]	0.7958	0.1590	0.1166	0.8109	0.6278	0.8228	0.1492	0.1093	0.8304	0.6461
IFC [29]	0.8390	0.1428	0.1118	0.7924	0.6061	0.8689	0.1299	0.0986	0.8370	0.6537
VIF [8]	0.9470	0.0843	0.0645	0.9381	0.7825	0.9481	0.0835	0.0642	0.9381	0.7826
MAD [30]	0.9502	0.0818	0.0635	0.9466	0.7970	0.9514	0.0808	0.0631	0.9467	0.7971
SSIM [5]	0.8396	0.1386	0.1055	0.8602	0.6641	0.8397	0.1385	0.1054	0.8603	0.6642
MS-SSIM [6]	0.8882	0.1206	0.0933	0.9082	0.7311	0.8978	0.1156	0.0892	0.9152	0.7418
IW-SSIM [11]	0.9002	0.1143	0.0916	0.9209	0.7519	0.9077	0.1102	0.0865	0.9216	0.7536

Table 4 Prediction performance of different naive IQMs and IQMs with pretrained machine-learning-based mapping on LIVE image database. Boldface indicates the better performer.

LIVE Database [24] (779 images, 5 distortions)

	Naive IQM					IQM with Pretrained Machine-Learning-Based Mapping (Our Method)				
	PLCC	RMSE	MAE	SROCC	KROCC	PLCC	RMSE	MAE	SROCC	KROCC
MSE	0.6334	18.2935	14.9319	0.8705	0.6808	0.8469	12.5709	10.3045	0.8761	0.6875
SNR	0.8571	12.1793	9.4853	0.8648	0.6735	0.8649	11.8681	9.2038	0.8650	0.6740
PSNR	0.8686	11.7163	9.1141	0.8755	0.6865	0.8739	11.4923	8.9247	0.8756	0.6866
PSNR-HVS-M [27]	0.9228	9.1072	6.9740	0.9288	0.7635	0.9275	8.8357	6.8048	0.9295	0.7659
NQM [28]	0.9130	9.6470	7.3591	0.9093	0.7431	0.9161	9.4782	7.1392	0.9107	0.7454
VSNR [9]	0.6863	17.1936	13.1309	0.6481	0.4879	0.8180	13.5975	10.2850	0.8120	0.6189
MSVD [7]	0.8732	11.5227	8.9817	0.8771	0.6956	0.8787	11.2856	8.6262	0.8772	0.6957
IFC [29]	0.9221	9.1465	7.5292	0.9247	0.7561	0.9312	8.6195	6.9008	0.9283	0.7622
VIF [8]	0.9604	6.5909	5.2880	0.9636	0.8282	0.9605	6.5787	5.2609	0.9636	0.8282
MAD [30]	0.9654	6.0878	4.5634	0.9619	0.8372	0.9655	6.0876	4.5613	0.9620	0.8375
SSIM [5]	0.9082	9.8925	7.7311	0.9104	0.7311	0.9157	9.4995	7.4517	0.9177	0.7440
MS-SSIM [6]	0.9280	8.8103	7.1596	0.9478	0.7967	0.9532	7.1443	5.5735	0.9544	0.8116
IW-SSIM [11]	0.9371	8.2532	6.7693	0.9560	0.8157	0.9547	7.0374	5.4141	0.9575	0.8204

ing techniques. Therefore, our primary aim is to evaluate the proposed approach based on the improvement in prediction results of baseline (i.e., naive) IQMs after applying supervised learning. Toward this end, we have used 13 different IQMs: MSE, SNR, PSNR, PSNR-HVS-M [27], NQM [28], VSNR [9], MSVD [7], IFC [29], VIF [8], MAD [30], SSIM [5], MS-SSIM [6], and IW-SSIM [11]. The three largest publicly accessible subject-rated image databases are employed in our demonstration: TID2008 database [22], CSIQ database [23], and LIVE database [24]. The TID2008 database [22] contains 1700 distorted images obtained from 25 reference images with 17 distortion types at four distortion levels. The CSIQ database [23] includes 866 distorted images generated from 30 original images with six types of distortions at four or five distortion levels. The LIVE database [24] includes 29 reference images from which 779 distorted images are obtained with five types of distortions. For the sake of completeness, in this paper, we show 1) in-database (Tables 2, 3, 4, 5, 6, 7, 8, and 10) and 2) cross-database (Table 9) performance validations of our machine-learning-based algorithm. We have employed five evaluation criteria [4], [21] (see Tables 2, 3, and 4): Pearson linear correlation coefficient (PLCC), root mean squared error (RMSE), mean absolute error (MAE), Spearman rank order correlation coefficient (SROCC), and Kendall rank order correlation coefficient (KROCC). PLCC, RMSE, and MAE are employed to evaluate prediction accuracy. SROCC and KROCC are used to assess prediction monotonicity. Note that the higher the PLCC, SROCC, and KROCC (the lower the RMSE and MAE), the better the performance.

To demonstrate the effectiveness of the proposed method, as mentioned above, we have used the cross validation scheme [14], [16], [17]: we have randomly split the set of images into five subsets (i.e., five-fold cross validation). In each fold, four subsets are used for training and the remaining subset is used for testing: in the testing phase,

training examples are not included. For fairness, we have followed the PA procedure recommended in [21]: we have fitted the objective quality scores to the subjective quality ratings via a four-parameter cubic polynomial function as follows [16], [20], [21]:

$$g(x) = a_0 + a_1x + a_2x^2 + a_3x^3, \quad (4)$$

where $a_0, \dots, a_3$ are coefficients. In this paper, for the purpose of assessing image quality, we have chosen the general regression neural network (GRNN) [17], [25] as a training and testing tool due to its good generalization performance through empirical experiments (see Table 5). There is a parameter σ (spread) to be determined in the GRNN [17], [25]. Note that, as was done in [14], a five-fold cross validation on the training set of images is used to select the parameter.

We first report the prediction performance (through in-database cross validation) of different 1) “naive IQMs” and 2) “IQMs with pretrained machine-learning-based mapping (the proposed method)” on the three image databases in Tables 2, 3, and 4. Note that, throughout this paper, indices (such as PLCC and SROCC) of both 1) “naive IQMs” and 2) “IQMs with pretrained machine-learning-based mapping (the proposed method)” are computed after the fitting using (4), except for Table 10. As shown in the tables, the proposed approach improves prediction outcomes of a wide range of IQMs. Even though not all prediction results of

Table 5 Database size-weighted average performance of some selected IQMs with pretrained machine-learning-based mapping over three databases. For this demonstration, the ν -SVR [26] with the radial basis function kernel is used. Note that, as was done in [14], a five-fold cross validation on the training set is used to determine the parameters of the SVR.

	SVR		GRNN	
	PLCC	SROCC	PLCC	SROCC
PSNR	0.7281	0.7333	0.7463	0.7518
MS-SSIM [6]	0.8792	0.8924	0.8869	0.8927
IW-SSIM [11]	0.8924	0.8975	0.8983	0.8979

Fig. 6 Example of comparison between naive PSNR and machine-learning-based PSNR (the proposed method) for TID2008 database [22]. Higher values of MOS (0 - minimal, 9 - maximal) and PSNR correspond to higher visual quality of the image. In this example, the MOS for distorted image 1 is higher than the one for distorted image 2. Note that the naive PSNR fails to predict the visual quality of the distorted images, whereas our machine-learning-based PSNR correctly predicts the visual quality.

Table 6 PLCC of some selected naive IQMs and IQMs with pretrained machine-learning-based mapping on TID2008 image database for individual distortion types.

TID2008 Database [22] (1700 images, 17 distortions)

	Naive IQM		IQM with Pretrained Machine-Learning-Based Mapping (Our Method)	
	MS-SSIM [6]	IW-SSIM [11]	MS-SSIM [6]	IW-SSIM [11]
AWGN	0.7846	0.7495	0.8000	0.7736
AWGN Color	0.7898	0.7654	0.8034	0.7760
Corr. Noise	0.8248	0.7733	0.8328	0.7775
Masked Noise	0.7968	0.8255	0.7635	0.8262
High Freq.	0.8789	0.8551	0.8842	0.8715
Impulse	0.6619	0.6145	0.6628	0.6164
Quantization	0.8158	0.7897	0.8345	0.8135
Gaussian Blur	0.9427	0.9387	0.9538	0.9506
Denoising	0.9632	0.9570	0.9642	0.9584
JPEG	0.9583	0.9523	0.9593	0.9539
JP2K	0.9706	0.9522	0.9713	0.9728
JPEG Trans.	0.8811	0.8669	0.8886	0.8673
JP2K Trans.	0.8266	0.8141	0.7980	0.8204
Non Eccen.	0.7008	0.7670	0.7146	0.7847
Block-Wise	0.7928	0.7643	0.7850	0.7442
Mean Shift	0.6909	0.6702	0.6765	0.6584
Contrast	0.7638	0.7698	0.7887	0.8215
# better ¹	4	2	13	15

¹The number of times that a naive IQM or IQM with pretrained machine-learning-based mapping performs better on PLCC.

Table 7 PLCC of some selected naive IQMs and IQMs with pretrained machine-learning-based mapping on CSIQ image database for individual distortion types.

CSIQ Database [23] (866 images, 6 distortions)

	Naive IQM		IQM with Pretrained Machine-Learning-Based Mapping (Our Method)	
	MS-SSIM [6]	IW-SSIM [11]	MS-SSIM [6]	IW-SSIM [11]
AWGN	0.9166	0.8848	0.9260	0.8921
JPEG	0.9764	0.9762	0.9804	0.9774
JP2K	0.9703	0.9711	0.9711	0.9725
Pink Noise	0.9141	0.8848	0.9216	0.9003
Gaussian Blur	0.9511	0.9544	0.9522	0.9628
Contrast	0.9258	0.9406	0.9325	0.9440
# better ¹	0	0	6	6

¹The number of times that a naive IQM or IQM with pretrained machine-learning-based mapping performs better on PLCC.

Table 8 PLCC of some selected naive IQMs and IQMs with pretrained machine-learning-based mapping on LIVE image database for individual distortion types.

LIVE Database [24] (779 images, 5 distortions)

	Naive IQM		IQM with Pretrained Machine-Learning-Based Mapping (Our Method)	
	MS-SSIM [6]	IW-SSIM [11]	MS-SSIM [6]	IW-SSIM [11]
JP2K	0.9478	0.9552	0.9695	0.9742
JPEG	0.9711	0.9687	0.9796	0.9786
AWGN	0.9409	0.9529	0.9745	0.9778
Gaussian Blur	0.9371	0.9500	0.9543	0.9719
JP2K Trans.	0.9268	0.9252	0.9468	0.9432
# better ¹	0	0	5	5

¹The number of times that a naive IQM or IQM with pretrained machine-learning-based mapping performs better on PLCC.

IQMs are improved in terms of prediction accuracy and monotonicity (e.g., MAD [30] for TID2008 database), the tables demonstrate the potential effectiveness of our pro-

posed method. Figure 6 shows an example of comparison between naive PSNR and machine-learning-based PSNR (our method) for TID2008 database [22]. We have also ex-

Table 9 PLCC of some selected IQMs with pretrained machine-learning-based mapping through cross-database validation. Boldface indicates the training database.

MS-SSIM [6]			
	TID2008	CSIQ	LIVE
TID2008	–	0.8960	0.9450
CSIQ	0.8363	–	0.9395
LIVE	0.8381	0.8975	–
IW-SSIM [11]			
	TID2008	CSIQ	LIVE
TID2008	–	0.9098	0.9486
CSIQ	0.8581	–	0.9400
LIVE	0.8538	0.9085	–

Table 10 Prediction performance of some selected naive IQMs and IQMs with pretrained machine-learning-based mapping on three image databases. Boldface indicates the better performer. Note that, in this table, the five-parameter logistic function in (5) is used in the PA stage.

TID2008 Database [22] (1700 images, 17 distortions)				
	Naive IQM		Our Method	
	PLCC	RMSE	PLCC	RMSE
MS-SSIM [6]	0.8451	0.7173	0.8578	0.6898
IW-SSIM [11]	0.8579	0.6895	0.8701	0.6593
CSIQ Database [23] (866 images, 6 distortions)				
	Naive IQM		Our Method	
	PLCC	RMSE	PLCC	RMSE
MS-SSIM [6]	0.8991	0.1149	0.9075	0.1112
IW-SSIM [11]	0.9144	0.1063	0.9205	0.1035
LIVE Database [24] (779 images, 5 distortions)				
	Naive IQM		Our Method	
	PLCC	RMSE	PLCC	RMSE
MS-SSIM [6]	0.9489	8.619	0.9611	7.258
IW-SSIM [11]	0.9522	8.347	0.9645	7.047

amined the effects of the proposed method on different image distortion types. Tables 6, 7, and 8 show the PLCC performance of some selected naive IQMs and IQMs with pretrained machine-learning-based mapping for individual distortion types. In Tables 6, 7, and 8, we can see that our proposed approach leads to consistent improvement for different image distortions, except for TID2008 database. Further, through cross-database validation, we have demonstrated the general applicability of the proposed scheme as shown in Table 9. Note that, for cross-database validation, the entire database is employed for training and testing. We can see from Tables 2, 3, 4, and 9 that “IQMs with pretrained machine-learning-based mapping” for each database consistently perform better than their naive counterparts (i.e., the IQMs without pretrained machine-learning-based mapping), even when the database is not involved in training. For example, for LIVE database, the MS-SSIM [6] with pretrained machine-learning-based mapping gives correlation coefficient of 0.9450 when the training is done only with TID2008 database, on the other hand, the naive MS-SSIM [6] gives correlation coefficient of 0.9280. These results again verify that our algorithm is independent of image contents and distortions. Lastly, the effects of the regression in the PA stage and the computational complexity of our ap-

Table 11 Comparison of running times for some selected naive IQMs and our proposed machine-learning-based mapping (Seconds/Image).

PSNR	MAD [30]	MS-SSIM [6]	IW-SSIM [11]	Ours
0.097	50.79	0.418	1.607	0.177

proach are examined. For the former, we have fitted the objective quality ratings to the subjective quality scores using a five-parameter logistic function [3], [11], which is another common regression model for the PA in the IQA research, as follows:

$$h(x) = a_0 \left(0.5 - \frac{1}{1 + \exp[a_1(x - a_2)]} \right) + a_3x + a_4, \quad (5)$$

where $a_0, \dots, a_4$ are coefficients. Table 10 reports the prediction performance of some selected “naive IQMs” and “IQMs with pretrained machine-learning-based mapping” with the five-parameter logistic regression in the PA stage. As shown in Table 10, our method enhances prediction outcomes of IQMs, regardless of the regression function adopted in the PA stage. Table 11 shows the comparison of (average) running times for some selected naive IQMs and our proposed machine-learning-based mapping. The experiments were performed on a 3.0 GHz Intel machine for a 512×512 image with MATLAB implementations. Note that the running time for the PA is not included in these experiments. This table confirms the feasibility of our method in real applications: our framework improves prediction results of IQMs, at the expense of a moderate increase in computation.

4. Conclusions and Future Work

In this paper, we have proposed a simple but effective framework for improving prediction outcomes of a given IQM by using machine learning techniques. The strength of the proposed approach lies in its ability to improve prediction results of any given naive IQM. Overall, the extensive experimental results on three different publicly available image databases reported in Sect. 3 demonstrate the feasibility of this approach and suggest that it can be useful in practice. In this paper, we have investigated specific examples of making use of machine-learning-based mapping between the outputs of IQM and subjective image-quality ratings in the quality assessment stage with general regression neural network [25] and support vector regression [26]. Therefore, our future work includes further research on more efficient machine learning algorithms for the task of correlating those pairs of measures in the quality assessment stage. We will also investigate the effectiveness of our proposed approach in vision-based sport video analysis systems involving multiple cameras for the purpose of quality monitoring. Image and video contents delivered over various wired and wireless networks inevitably suffer from visual quality degradations during transmission over error prone networks [1]. For this reason, it is required to monitor such quality deteriorations in real-time to meet system requirements.

Acknowledgments

This work was supported by MCST (Ministry of Culture, Sports & Tourism) / KOCCA (Korea Creative Content Agency) (R2016030044 - Development of Context-Based Sport Video Analysis, Summarization, and Retrieval Technologies).

References

- [1] Z. Wang, "Applications of Objective Image Quality Assessment Methods [Applications Corner]," *IEEE Signal Process. Mag.*, vol.28, no.6, pp.137–142, 2011.
- [2] X. Xie, G. Li, and Z. Wang, "A Low-Complexity and High-Quality Image Compression Method for Digital Cameras," *ETRI Journal*, vol.28, no.2, pp.260–263, 2006.
- [3] J. Zhu and N. Wang, "Image Quality Assessment by Visual Gradient Similarity," *IEEE Trans. Image Process.*, vol.21, no.3, pp.919–933, 2012.
- [4] 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, 2006.
- [5] 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, 2004.
- [6] Z. Wang, E.P. Simoncelli, and A.C. Bovik, "Multiscale Structural Similarity for Image Quality Assessment," *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, pp.1398–1402, 2003.
- [7] A. Shnayderman, A. Gusev, and A.M. Eskicioglu, "An SVD-Based Grayscale Image Quality Measure for Local and Global Assessment," *IEEE Trans. Image Process.*, vol.15, no.2, pp.422–429, 2006.
- [8] H.R. Sheikh and A.C. Bovik, "Image Information and Visual Quality," *IEEE Trans. Image Process.*, vol.15, no.2, pp.430–444, 2006.
- [9] D.M. Chandler and S.S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Trans. Image Process.*, vol.16, no.9, pp.2284–2298, 2007.
- [10] D.-O. Kim and R.-H. Park, "New Image Quality Metric Using the Harris Response," *IEEE Signal Process. Lett.*, vol.16, no.7, pp.616–619, 2009.
- [11] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, vol.20, no.5, pp.1185–1198, 2011.
- [12] A. Liu, W. Lin, and M. Narwaria, "Image Quality Assessment Based on Gradient Similarity," *IEEE Trans. Image Process.*, vol.21, no.4, pp.1500–1512, 2012.
- [13] R.V. Babu, S. Suresh, and A. Perkis, "No-reference JPEG-image quality assessment using GAP-RBF," *Signal Processing*, vol.87, no.6, pp.1493–1503, 2007.
- [14] A.K. Moorthy and A.C. Bovik, "A Two-Step Framework for Constructing Blind Image Quality Indices," *IEEE Signal Process. Lett.*, vol.17, no.5, pp.513–516, 2010.
- [15] A. Chetouani, A. Beghdadi, S. Chen, and G. Mostafaoui, "A Novel Free Reference Image Quality Metric Using Neural Network Approach," *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, pp.1–4, 2010.
- [16] 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, 2010.
- [17] C. Li, A.C. Bovik, and X. Wu, "Blind Image Quality Assessment Using a General Regression Neural Network," *IEEE Trans. Neural Netw.*, vol.22, no.5, pp.793–799, 2011.
- [18] M. Narwaria, W. Lin, and A.E. Cetin, "Scalable image quality assessment with 2D mel-cepstrum and machine learning approach," *Pattern Recognition*, vol.45, no.1, pp.299–313, 2012.
- [19] M. Narwaria and W. Lin, "SVD-Based Quality Metric for Image and Video Using Machine Learning," *IEEE Trans. Syst., Man, Cybern., Part B*, vol.42, no.2, pp.347–364, 2012.
- [20] Y. Han, Y. Cai, Y. Cao, and X. Xu, "Monotonic Regression: A New Way for Correlating Subjective and Objective Ratings in Image Quality Research," *IEEE Trans. Image Process.*, vol.21, no.4, pp.2309–2313, 2012.
- [21] VQEG, Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II Aug. 2003, URL <http://www.vqeg.org/>, 2003.
- [22] N. Ponomarenko and K. Egiazarian, Tampere image database 2008, URL <http://www.ponomarenko.info/tid2008.htm>, 2008.
- [23] E.C. Larson and D.M. Chandler, Categorical image quality (CSIQ) database, URL <http://vision.okstate.edu/csiq>, 2010.
- [24] H.R. Sheikh, K. Seshadrinathan, A.K. Moorthy, Z. Wang, A.C. Bovik, and L.K. Cormack, Image and video quality assessment research at LIVE, URL <http://live.ece.utexas.edu/research/quality/>, 2006.
- [25] D.F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol.2, no.6, pp.568–576, 1991.
- [26] B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett, "New support vector algorithms," *Neural Computation*, vol.12, no.5, pp.1207–1245, 2000.
- [27] A. Chetouani, A. Beghdadi, S. Chen, and G. Mostafaoui, "On between coefficient contrast masking of DCT basis functions," *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, CD-ROM, 2007.
- [28] N. Damara-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol.9, no.4, pp.636–650, 2000.
- [29] H.R. Sheikh, A.C. Bovik, and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," *IEEE Trans. Image Process.*, vol.14, no.12, pp.2117–2128, 2005.
- [30] E.C. Larson and D.M. Chandler, "Most apparent distortion: Full reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol.19, pp.011006:1–011006:21, 2010.

Chanho Jung received the B.S. and M.S. degrees in electronic engineering from Sogang University, Seoul, Korea, in 2004 and 2006, respectively, and the Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 2013. From 2006 to 2008, he was a Research Engineer with the Digital Television Research Laboratory, LG Electronics, Seoul. From 2013 to 2016, he was a Senior Researcher with Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea. Since 2016, he has been with the Department of Electrical Engineering, Hanbat National University, Daejeon, Korea, where he is currently an Assistant Professor. His current research interests are computer vision, machine learning, embedded system, pattern recognition, and image processing.

Sanghyun Joo received the B.S. and M.S. degrees in Computer Engineering from Dongguk University, Republic of Korea, in 1989 and 1994, respectively. He also received the Ph. D. degree in Natural Science Research from Niigata University, Japan, in 1999. He was an associated professor in the Department of Electronic and Electrical Engineering from 1999 to 2001. Since 2001 he has been a principal researcher in the Mobile Content Section, SW-Content Research Laboratory, Electronics and Telecommu-

nications Research Institute (ETRI), Republic of Korea. His research interests include image/video compression, user information description, and image/video communication.

Do-Won Nam received the B.S. degree in Computer Science from Korea Advanced Institute of Science and Technology (KAIST) in 1996 and M.S. degree in Information Technology from Pohang University of Science of Technology (POSTECH), Korea in 1998. He is working as a principal researcher in the Electronics and Telecommunications Research Institute (ETRI) since 2001. His research interests include data mining, digital rights management, digital cinema system and sports video analysis.

Wonjun Kim received the B.S. degree in electrical engineering from Sogang University, Seoul, Korea, the M.S. degree from the Department of Information and Communications, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, and the Ph.D. degree from the Computational Imaging Laboratory, Department of Electrical Engineering, KAIST, in 2006, 2008, and 2012, respectively. From Sept. 2012 to Feb. 2016, he was a Research Staff Member in Samsung Advanced Institute of Technology (SAIT). Since Mar. 2016, he has been with the Department of Electronics Engineering, Konkuk University, Seoul, Korea, where he is currently an Assistant Professor. His research interests include image and video understanding, computer vision, pattern recognition, and biometrics, with an emphasis on saliency detection, face and action recognition.