MEASUREMENT SCIENCE REVIEW, **16**, (2016), No. 6, 316–325

DE GRUYTER OPEN

MEASUREMENT SCIENCE REVIEW

Journal homepage: http://www.degruyter.com/view/j/msr

ISSN 1335-8871

A Regression-Based Family of Measures for Full-Reference Image Quality Assessment

Mariusz Oszust

Department of Computer and Control Engineering, Rzeszow University of Technology, W. Pola 2, 35-959 Rzeszow, Poland, marosz@kia.prz.edu.pl

The advances in the development of imaging devices resulted in the need of an automatic quality evaluation of displayed visual content in a way that is consistent with human visual perception. In this paper, an approach to full-reference image quality assessment (IQA) is proposed, in which several IQA measures, representing different approaches to modelling human visual perception, are efficiently combined in order to produce objective quality evaluation of examined images, which is highly correlated with evaluation provided by human subjects. In the paper, an optimisation problem of selection of several IQA measures for creating a regression-based IQA hybrid measure, or a multimeasure, is defined and solved using a genetic algorithm. Experimental evaluation on four largest IQA benchmarks reveals that the multimeasures obtained using the proposed approach outperform state-of-the-art full-reference IQA techniques, including other recently developed fusion approaches.

Keywords: Image quality assessment, quality measurement, feature selection, multiple linear regression, genetic algorithm.

1. Introduction

Different image processing schemas, as well as a variety of imaging devices may interfere with the content of the displayed image. Human evaluation of such images can be either inconvenient or expensive. Therefore, in order to measure image quality from the human perception point of view, many image quality assessment (IQA) measures have been developed [1],[2]. They are divided into three categories. Full-reference techniques evaluate the quality of distorted images based on their distortion-free equivalents; no-reference and reduced reference measures, in turn, do not use such information or it is partially provided [3],[4],[5].

This paper presents a full-reference IQA measure. One of the simplest techniques in this category is peak signal-to-noise ratio (PSNR); noise quality measure (NQM) [6], in turn, uses a linear frequency distortion and an additive noise injection. Other measures which have been introduced in the last decade use: luminance and contrast distortions [7], structural information [8], [9], [10], statistical properties [10], [11], [12], phase congruency and image gradient magnitude [13], visual saliency maps [14], [15], Riesz-transform features [16], speeded-up robust features [17], local binary patterns [18], structure and contrast changes [19], inter-patch and intra-patch similarities [20], or fuzzy gradient similarity deviation [21].

There are also IQA measures which aggregate several IQA techniques. For example, in [22], a canonical correlation analysis was used to combine SNR, SSIM, VIF, and VSNR. Larson and Chandler in [23] introduced the most apparent distortion algorithm (MAD) that adapts local luminance with contrast masking for assessment of high quality images and evaluates low quality images using local statistics of spatial-frequency components. A fusion of IQA measures with preservation of edge direction was introduced in [24]. Local and global measures of distortion were considered in [25]. In [26], [27], MSSIM, VIF, and R-SVD, were non-linearly combined. A linear combination of IQA measures can be found in [28]. In [29], a preliminary approach was presented with a regularized multi-dimensional polynomial estimator for combining seven IQA measures. A conditional Bayesian mixture of experts model with a support vector machines classifier was used in [30] for combining SSIM, VSNR, and VIF using k-nearest-neighbour regression. A support vector regression approach was shown in [31]. In [32], in turn, image blocks were first classified using decision trees and then FSIM [13], mean squared error, and different variations of PSNR [33] were combined. In [34], in turn, lasso regression models were obtained using pairwise scores differences. Six IQA measures were fused using neural network in [35]. In [36], the two-step approach was proposed in which local image patches were used for finding perceptually meaningful structures and local distortion measurements were combined into a multimeasure by kernel ridge regression.

A fusion measure introduced in this paper uses the multiple linear regression [37] of opinions provided by genetically selected IQA measures. Furthermore, the proposed approach is able to find a well-performing hybrid measure which con-

DOI: 10.1515/msr-2016-0040

Table 1. IQA benchmark image datasets.

Benchmark	**No. of ref. images**	**No. of dist. images**	**No. of distortions**
TID2013	25	3000	24
TID2008	25	1700	17
CSIQ	30	866	6
LIVE	29	779	5

sists of a small number of IQA measures. Although in the literature many works have used different types of regression for aggregating IQA measures, this paper introduces a very efficient conjunction of the genetic algorithm with multiple linear regression, able to outperform other, often more complex techniques. The multimeasures obtained using the proposed approach are compared with the state-of-the-art techniques using typical evaluation protocol on the four largest IQA benchmark image datasets.

The rest of this paper is organised as follows. Section 2 presents the formulation of the optimisation problem of IQA measures fusion. In section 3, the proposed family of regression-based multimeasures is obtained and discussed. Then, in section 4, the approach is compared with popular measures. Finally, section 5 concludes the paper.

2. METHOD

The desired IQA measure should return scores that are consistent with human subjective evaluation. In order to compare IQA approaches [1], specific benchmark image datasets can be used. They contain pristine, distortion-free images, their corresponding distorted equivalents and human scores. Human evaluation of images in benchmarks is provided in the form of mean opinion scores (MOS values) or differential mean opinion scores (DMOS values). In the approach proposed in this paper, it is assumed that the resulting measure can provide objective scores that are closer to the human evaluation than measures that were considered as the part of the fusion.

The considered problem is formulated as follows. Let $\hat{S}$ be the estimated response and Q_n is the one of N prediction variables in the multiple linear regression model [37], $n = 1,\ldots,N$. $\hat{S}$ can also be seen as the output of a joint decision of $k \in N$ IQA measures, or the objective score. In the model, vector B contains fitted coefficients estimated by minimising the mean squared difference between the prediction vector $B(Q)$ that contains objective scores and the vector of subjective scores S. The fitted linear function can be written as:

$$\hat{S} = B_0 + \sum_{l=1}^{k} B_l(Q_l), \tag{1}$$

where $l = 1,\ldots,k$ denotes a Q_n selected for regression.

The selection of IQA measures that are used in the regression can be considered as an optimisation problem which requires a formulation of an objective function. In order to find well-performing IQA multimeasures, a one of typically used IQA techniques performance index can be employed. There are four indices used for comparison of IQA measures [38], [39]: Spearman Rank order Correlation Coefficient (SRCC), Kendall Rank order Correlation Coefficient (KRCC), Pearson linear Correlation Coefficient (PCC), and Root Mean Square Error (RMSE). It is worth noticing that RMSE is widely used in the development of image processing algorithms, e.g., in [40]. SRCC and KRCC evaluate prediction monotonicity, while PCC and RMSE evaluate prediction accuracy. These indices are calculated after a non-linear mapping between a vector of objective scores $\hat{S}$, and MOS or DMOS, denoted by S, using the following mapping function for the non-linear regression [38]:

$$\hat{S_m} = \beta_1\left(\frac{1}{2} - \frac{1}{1+exp(\beta_2(\hat{S}-\beta_3))}\right) + \beta_4\hat{S} + \beta_5, \tag{2}$$

where $\beta = [\beta_1,\beta_2,\ldots,\beta_5]$ are parameters of the non-linear regression model [38], and $\hat{S_m}$ is the non-linearly mapped $\hat{S}$. SRCC is given as:

$$\mathrm{SRCC}(\hat{S},S) = 1 - \frac{6\sum_{i=1}^{m} d_i^2}{m(m^2-1)}, \tag{3}$$

where d_i is the difference between i^{th} image in $\hat{S}$ and S, m is the total number of images. In KRCC, the number of concordant pairs in the dataset, m_c, is used, as well as the number of discordant pairs, m_d,

$$\mathrm{KRCC}(\hat{S},S) = \frac{m_c - m_d}{0.5m(m-1)}. \tag{4}$$

PCC, in turn, is calculated as:

$$\mathrm{PCC}(\hat{S_m},S) = \frac{\bar{\hat{S}_m}^T \bar{S}}{\sqrt{\bar{\hat{S}_m}^T \bar{\hat{S}}_m \bar{S}^T \bar{S}}}, \tag{5}$$

where mean-removed vectors are denoted by $\bar{\hat{S}_m}$ and $\bar{S}$. RMSE is calculated as:

$$\mathrm{RMSE}(\hat{S_m},S) = \sqrt{\frac{(\hat{S_m}-S)^T(\hat{S_m}-S)}{m}}. \tag{6}$$

In the proposed approach, RMSE was used as the objective function in the optimisation problem of finding well-performing IQA joint measure. The problem can be written as:

$$\begin{aligned} \min_{x,\beta} \quad & \mathrm{RMSE}(\hat{S_m},S) \\ \text{s.t.} \quad & x_j \in \{0,1\}, \quad \sum_{j=1}^{N} x_j \le k, \quad k \le N, \quad \beta \ge 0, \end{aligned} \tag{7}$$

where x and β are optimised decision variables. In the problem, x is a vector of binary weights which indicate measures that are used in the regression. Given fitted regression model, β is used for computation of the objective function, RMSE.

Table 2. SRCC between objective scores of IQA measures on LIVE.

	VSI	**FSIMc**	**IW-SSIM**	**MAD**	**MSSIM**	**PSNR**	**SR-SIM**	**VIF**	**IFS**	**SFF**
VSI	1.0000	0.9866	0.9708	0.9714	0.9804	0.9407	0.9816	0.9549	**0.9812**	**0.9734**
FSIMc	0.9866	1.0000	0.9897	0.9823	0.9828	0.9096	0.9938	0.9745	0.9860	0.9855
IW-SSIM	0.9708	0.9897	1.0000	0.9764	0.9860	0.8801	0.9835	0.9781	0.9761	**0.9761**
MAD	0.9714	0.9823	0.9764	1.0000	0.9688	0.8947	0.9781	**0.9632**	**0.9716**	**0.9713**
MSSIM	0.9804	0.9828	0.9860	0.9688	1.0000	0.9096	0.9713	0.9595	0.9725	0.9627
PSNR	0.9407	0.9096	0.8801	0.8947	0.9096	1.0000	0.9156	0.8714	0.9102	0.8990
SR-SIM	0.9816	0.9938	0.9835	0.9781	0.9713	0.9156	1.0000	0.9747	**0.9824**	0.9858
VIF	0.9549	0.9745	0.9781	**0.9632**	0.9595	0.8714	0.9747	1.0000	0.9744	0.9807
IFS	**0.9812**	0.9860	0.9761	**0.9716**	0.9725	0.9102	**0.9824**	0.9744	1.0000	0.9928
SFF	**0.9734**	0.9855	**0.9761**	**0.9713**	0.9627	0.8990	0.9858	0.9807	0.9928	1.0000

Note: SRCC values between IQA measures aggregated in rSIM1-4$^{2\text{-}3}$ are written in boldface.

3. Obtained multimeasures

In this section, experiments that were carried out in order to obtain a family of multimeasures using the proposed approach are discussed, as well as its evaluation on benchmark image datasets. Moreover, the section considers a contribution of aggregated IQA measures to the performance.

3.1. Optimisation results

In experiments, the following 16 IQA measures with publicly available objective scores, or source-code, for used benchmarks took part in the optimisation: VSI [15], FSIM [13], FSIMc [13], GSM [19], IFC [12], IW-SSIM [10], MAD [23], MSSIM [9], NQM [6], PSNR [38], RFSIM [16], SR-SIM [14], SSIM [8], VIF [11], IFS [41], and SFF [42].

It is worth noting that MAD is a multimeasure, but it was used in the optimisation due to its popularity and availability of the source-code. In the approach, the vector of decision variables in the optimisation problem is obtained in a data-driven fashion using a part of images and their subjective scores from benchmark datasets. Objective scores of used measures, if needed, were scaled to be in the $[0;1]$ range. There are four largest widely used IQA benchmark image datasets; therefore, four IQA multimeasures are introduced in this paper. In the approach, 20% of reference images and their distorted equivalents from a dataset were used, as in [17]. In order to show dataset independent results, each introduced multimeasure was evaluated on all datasets. In the literature, some authors used different numbers of images from benchmarks for this purpose, e.g., 30% [13], [15], a one dataset [29], [24], [30], [18], [35], or even several datasets jointly, as in [20]. The following four IQA benchmarks were used: TID2013 [2], TID2008 [43], CSIQ [23], and LIVE [8], they are characterised in Table 1.

A genetic algorithm (GA) [44] was used to solve the optimisation problem proposed in this paper. The algorithm operates on a population of solutions, called individuals, and applies selection, crossover and mutation operators in order to obtain better solutions in emerging generations of individuals. Experiments were carried out using Matlab version R2012a with Genetic Algorithms and Statistics Toolboxes. The GA parameters were determined experimentally, observing the convergence of the objective function. A population of 100 individuals was used in the GA, which was run for 200 generations. The scattered crossover, Gaussian mutation and stochastic uniform selection rules were used [44]. In experiments, four IQA multimeasures, namely Regression-based SImilarity Measures (rSIMs), were obtained, and their fitted models can be written as:

$$\begin{aligned}\text{rSIM1} = &-43.51 + 8.90VSI - 26.60FSIM \\ &+ 23.18FSIMc + 43.77GSM - 0.52IFC \\ &+ 0.97IW\text{-}SSIM - 2.56MAD \\ &- 2.40MSSIM - 0.71NQM - 1.18PSNR \\ &+ 0.71RFSIM + 0.89SR\text{-}SIM - 1.39SSIM \\ &+ 1.40VIF + 3.55IFS - 2.48SFF,\end{aligned} \tag{8}$$

$$\begin{aligned}\text{rSIM2} = &-20.21 + 10.10VSI - 8.14FSIM \\ &+ 6.66FSIMc + 11.51GSM - 0.40IFC \\ &+ 2.33IW\text{-}SSIM - 1.98MAD \\ &- 2.24MSSIM - 0.84NQM \\ &+ 1.85PSNR + 0.70RFSIM \\ &+ 3.83SR\text{-}SIM - 3.49SSIM + 2.30VIF \\ &+ 3.09IFS + 0.10SFF,\end{aligned} \tag{9}$$

Table 3. Performance of obtained multimeasures on four benchmark datasets in terms of SRCC.

	rSIM1	**rSIM1^3**	**rSIM1^2**	**rSIM2**	**rSIM2^3**	**rSIM2^2**	**rSIM3**	**rSIM3^3**	**rSIM3^2**	**rSIM4**	**rSIM4^3**	**rSIM4^2**
TID2013	**0.9008**	**0.8994**	0.8951	0.8202	0.8329	0.8720	0.8225	0.8175	0.8027	0.8070	0.8036	0.8027
TID2008	**0.9124**	0.8962	0.9056	**0.9218**	0.9073	0.9038	0.9090	0.9014	0.8890	0.8881	0.8875	0.8890
CSIQ	0.9552	0.9339	0.9542	0.9609	0.9695	0.9577	**0.9738**	**0.9685**	0.9666	0.9601	0.9670	0.9666
LIVE	0.9643	0.9520	0.9582	0.9670	0.9730	0.9639	0.9701	**0.9736**	0.9731	**0.9754**	0.9731	0.9731
Overall direct	**0.9332**	0.9204	**0.9283**	0.9175	0.9207	0.9244	0.9189	0.9153	0.9079	0.9077	0.9078	0.9079
Overall weighted	**0.9194**	0.9099	**0.9140**	0.8851	0.8892	0.9039	0.8850	0.8803	0.8697	0.8709	0.8698	0.8697

Note: The best two measures for each dataset are written in bold.

$$\begin{aligned}\text{rSIM3} = &-1.78 - 1.60VSI - 1.75FSIM \\ &+ 1.72FSIMc + 3.46GSM - 0.12IW\text{-}SSIM \\ &+ 0.55MAD + 0.75MSSIM \\ &+ 0.06NQM - 0.01PSNR - 0.23RFSIM \\ &- 0.31SR\text{-}SIM + 0.02SSIM - 0.18VIF \\ &- 0.61IFS + 0.61SFF,\end{aligned} \tag{10}$$

$$\begin{aligned}\text{rSIM4} = &-287.26 - 10.35VSI - 131.16FSIM \\ &+ 84.83FSIMc + 433.46GSM + 12.00IFC \\ &+ 52.36IW\text{-}SSIM + 73.16MAD \\ &- 12.11MSSIM + 11.54NQM \\ &+ 9.79RFSIM - 42.10SR\text{-}SIM \\ &- 18.88SSIM - 51.12VIF - 41.30IFS \\ &+ 0.93SFF.\end{aligned} \tag{11}$$

Their corresponding β components are as follows: β_{rSIM1} = [4.9054, 6.6992, 5.8301, 2.5195, 8.7132], β_{rSIM2} = [7.2085, 5.9428, 5.5770, 1.3571, 8.1656], β_{rSIM3} = [6.7406, 9.0813, 1.7005, 2.9973, 6.6393], β_{rSIM4} = [4.3063, 2.4294, 4.4762, 0.7002, 6.0786].

3.2. Reduced multimeasures

It can be seen that the GA preferred using most of IQA measures in the fusion. However, not all of them are equally important, and taking into account the practical usage of the obtained multimeasure, it would be desirable to have a small number of IQA measures involved in the fusion. Therefore, for each fusion measure t-statistics and their corresponding p-values were determined in order to show which IQA measures significantly contribute to the fusion (i.e., multiple regression model). The hypothesis test on a given coefficient tests the null hypothesis that it is not significant as being equal to zero. It turned out that rSIM1 do not seem to differ significantly according to IW-SSIM, MSSIM, SR-SIM, or SFF at the 5 % significance level. Similar observations were made for rSIM2 (FSIM, FSIMc, GSM, MSSIM, SR-SIM, and SFF), rSIM3 (FSIM, FSIMc, IW-SSIM, PSNR, SSIM), and rSIM4 (VSI, FSIM, FSIMc, MSSIM, RFSIM, SR-SIM, SSIM, and SFF).

All obtained rSIMs consist of more than six IQA measures. A development of a multimeasure that uses even less IQA models while maintaining the state-of-the-art performance requires solving another optimisation problem. Therefore, the optimisation problem introduced in (7) was constrained and only fusion measures that are composed of two or three IQA measures were taken into account, i.e., the sum condition responsible for selection of IQA measures for regression in equation was changed into $\sum_{j=1}^{N} x_j = k$. This resulted in the following two sets of rSIMs, rSIM1-4^2 and rSIM1-4^3, where the number in superscript denotes the number of used IQA measures, $k = 2$ or $k = 3$. Such short rSIMs are shown below, as well, as a contribution of each fused measure, calculated as the percentage decrease of the sum of the squared residuals (i.e., observed values minus fitted values) of the fitted model without the considered IQA measure.

$$\text{rSIM1}^2 = -25.88 + 25.90VSI + 5.38IFS, \tag{12}$$

$$\text{rSIM1}^3 = -25.73 + 25.63VSI - 0.47PSNR + 5.53IFS, \tag{13}$$

$$\text{rSIM2}^2 = -14.91 + 13.08SR\text{-}SIM + 7.37IFS, \tag{14}$$

$$\text{rSIM2}^3 = 0.17 - 2.74MAD + 1.76VIF + 4.86IFS, \tag{15}$$

$$\text{rSIM3}^2 = 0.35 + 0.61MAD - 0.36VIF, \tag{16}$$

$$\text{rSIM3}^3 = 0.73 + 0.49MAD - 0.33VIF - 0.40IFS, \tag{17}$$

$$\text{rSIM4}^2 = 43.50 + 68.15MAD - 40.72VIF, \tag{18}$$

$$\begin{aligned}\text{rSIM4}^3 = &\ 33.55 + 7.52IW\text{-}SSIM + 75.83MAD \\ &- 38.50VIF.\end{aligned} \tag{19}$$

Table 4. Comparison of obtained multimeasures with state-of-the-art eight IQA measures on four benchmark datasets.

	VSI	FSIMc	MAD	MSSIM	PSNR	SR-SIM	IFS	SFF	rSIM1	rSIM2	rSIM3	rSIM4
TID2013												
SRCC	**0.8965**	0.8510	0.7807	0.7859	0.6395	0.7999	0.8697	0.8513	**0.9008**	0.8202	0.8225	0.8070
KRCC	**0.7183**	0.6665	0.6035	0.6047	0.4700	0.6314	0.6785	0.6581	**0.7264**	0.6482	0.6505	0.6287
PCC	0.9000	0.8769	0.8267	0.8329	0.0109	0.8590	0.8791	0.8706	**0.9171**	**0.9012**	0.8560	0.8669
RMSE	0.5404	0.5959	0.6975	0.6861	1.2396	0.6347	0.5909	0.6099	**0.4941**	**0.5372**	0.6409	0.6180
TID2008												
SRCC	0.8979	0.8840	0.8340	0.8542	0.5531	0.8913	0.8903	0.8767	**0.9124**	**0.9218**	0.9090	0.8881
KRCC	0.7123	0.6991	0.6445	0.6568	0.4027	0.7149	0.7009	0.6882	**0.7374**	**0.7572**	0.7366	0.7042
PCC	0.8762	0.8762	0.8306	0.8451	0.5734	0.8866	0.8810	0.8817	**0.9077**	**0.9259**	0.9043	0.8933
RMSE	0.6466	0.6468	0.7473	0.7173	1.0994	0.6206	0.6349	0.6333	**0.5631**	**0.5069**	0.5728	0.6033
CSIQ												
SRCC	0.9423	0.9310	0.9466	0.9133	0.8058	0.9319	0.9582	**0.9627**	0.9552	0.9609	**0.9738**	0.9601
KRCC	0.7857	0.7690	0.7970	0.7393	0.6084	0.7725	0.8165	**0.8288**	0.8129	0.8260	**0.8588**	0.8262
PCC	0.9279	0.9192	0.9500	0.8991	0.8000	0.9250	0.9576	0.9643	0.9620	0.9657	**0.9789**	**0.9680**
RMSE	0.0979	0.1034	0.0820	0.1149	0.1575	0.0997	0.0757	0.0695	0.0717	0.0682	**0.0537**	**0.0659**
LIVE												
SRCC	0.9524	0.9645	0.9669	0.9513	0.8756	0.9618	0.9599	0.9649	0.9643	0.9670	**0.9701**	**0.9754**
KRCC	0.8058	0.8363	0.8421	0.8045	0.6865	0.8299	0.8254	0.8365	0.8286	0.8369	**0.8444**	**0.8616**
PCC	0.9482	0.9613	**0.9675**	0.9489	0.8723	0.9553	0.9586	0.9632	0.8999	0.9653	0.9672	**0.9765**
RMSE	8.6816	7.5297	**6.9073**	8.6188	13.3597	8.0813	7.7765	7.3461	11.9133	7.1314	6.9430	**5.8913**
Overall direct												
SRCC	**0.9223**	0.9076	0.8821	0.8762	0.7185	0.8962	0.9195	0.9139	**0.9332**	0.9175	0.9189	0.9077
KRCC	0.7555	0.7427	0.7218	0.7013	0.5419	0.7372	0.7553	0.7529	**0.7763**	0.7671	**0.7726**	0.7552
PCC	0.9131	0.9084	0.8937	0.8815	0.5642	0.9065	0.9191	0.9200	0.9217	**0.9395**	**0.9266**	0.9262
RMSE	0.4283	0.4487	0.5089	0.5061	0.8322	0.4517	0.4338	0.4376	**0.3763**	**0.3708**	0.4225	0.4291
Overall weighted												
SRCC	**0.9102**	0.8851	0.8412	0.8425	0.6691	0.8628	0.8988	0.8877	**0.9194**	0.8851	0.8850	0.8709
KRCC	**0.7370**	0.7107	0.6711	0.6623	0.4984	0.6981	0.7220	0.7121	**0.7541**	0.7255	0.7266	0.7054
PCC	0.9036	0.8931	0.8625	0.8599	0.3780	0.8876	0.9004	0.8981	**0.9187**	**0.9247**	0.8998	0.9017
RMSE	0.5025	0.5333	0.6150	0.6050	1.0253	0.5456	0.5226	0.5313	**0.4481**	**0.4536**	0.5270	0.5260

Note: The best two measures for each performance index are written in bold.

Table 5. The summary of significance tests.

	VSI	**FSIMc**	**MAD**	**MSSIM**	**PSNR**	**SR-SIM**	**IFS**	**SFF**
TID2013, TID2008, CSIQ, LIVE								
rSIM1	1,1,1,-1	1,1,1,-1	1,1,1,-1	1,1,1,-1	1,1,1,1	1,1,1,-1	1,1,0,-1	1,1,0,-1
rSIM2	1,1,1,1	1,1,1,0	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,0,0
rSIM3	-1,1,1,1	-1,1,1,1	1,1,1,0	-1,1,1,1	1,1,1,1	0,1,1,1	0,1,1,1	0,1,1,1
rSIM4	-1,1,1,1	-1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	0,0,1,1	-1,1,1,1	0,1,0,1
rSIM1^3	1,0,1,0	1,0,1,-1	1,1,0,-1	1,1,1,0	1,1,1,1	1,1,1,1	1,0,-1,-1	1,0,-1,-1
rSIM2^3	-1,1,1,1	0,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	1,1,1,1	0,1,1,1	-1,1,1,1
rSIM3^3	-1,1,1,1	-1,1,1,1	1,1,1,0	1,1,1,1	1,1,1,1	0,1,1,1	-1,1,1,1	0,1,1,1
rSIM4^3	-1,1,1,1	-1,1,1,1	-1,1,1,1	-1,1,1,1	1,1,1,1	-1,0,1,1	-1,1,1,1	-1,1,1,1
rSIM1^2	0,1,-1,-1	1,1,-1,-1	1,1,-1,-1	1,1,1,-1	1,1,1,1	1,0,-1,-1	1,1,-1,-1	1,1,-1,-1
rSIM2^2	-1,1,1,-1	1,1,1,-1	1,1,-1,-1	1,1,1,-1	1,1,1,-1	1,1,1,-1	1,1,0,-1	1,1,-1,-1
rSIM3^2	-1,0,1,1	-1,0,1,0	-1,1,1,-1	-1,1,1,1	1,1,1,1	-1,1,1,1	-1,1,1,1	-1,1,1,-1
rSIM4^2	-1,0,1,1	-1,0,1,1	-1,1,1,1	-1,1,1,1	1,1,1,1	-1,1,1,1	-1,1,1,1	-1,1,1,1

Note: The fusion measure in the row is significantly better than the IQA measure in the column ('1'), worse ('-1'), or indistinguishable ('0'). Results for datasets are separated by commas.

Contributions to rSIM1^2 are as follows: VSI 43.60%, IFS 38.94%, to rSIM1^3: VSI 44.56%, PSNR 2.34%, IFS 40.47%, to rSIM2^2: SR-SIM 36.20%, IFS 39.28%, to rSIM2^3: MAD 27.49%, VIF 19.78%, IFS 17.37%, to rSIM3^2: MAD 63.00%, VIF 43.04%, to rSIM3^3: MAD 45.32%, VIF 37.69%, IFS 8.86%, to rSIM4^2: MAD 32.43%, VIF 38.67%, and to rSIM4^3: IW-SSIM 3.15%, MAD 52.73%, VIF 35.57%.

The contribution of used measures seems to be equally distributed, except for PSNR in rSIM1^3 and IWSSIM in rSIM4^3. Here, comparing rSIM1^2 to rSIM1^3, and rSIM4^2 to rSIM4^3, the second multimeasure is not influenced much by the presence of PSNR or IW-SSIM, respectively. It can be seen that VSI, IFS, SFF, MAD, and VIF are among the most frequently aggregated IQA measures. There are also measures which are often coupled together, i.e., MAD with VIF, and VSI with IFS, or with its earlier version, SFF. Measures in such pairs seem to complement each other.

3.3. *Measure selection choices*

Since some of the techniques used in experiments tend to be present in rSIMs more often than others, it is desirable to discuss their similarities, in terms of quality of produced objective scores. In order to show similarities between scores returned by IQA techniques, SRCCs between them were obtained on LIVE benchmark (see Table 2). Some IQA measures seem to overlap others, e.g., SRCC between FSIMc and SR-SIM is equal to 0.9938. Also, SFF is highly correlated with IFS (0.9928). The conjunction of MAD with VIF seems to be interesting, both measures are less correlated (0.9632) with each other than with other measures, what may produce, together with their good single performance on LIVE dataset, the well-performing fusion measures. VSI is mostly used in fusion measures developed on images from TID2013, i.e., rSIM1$^{2\text{-}3}$. Here, VSI is the best single performing metric (see section 4), and that was the main reason for using this measure in the fusion. MAD and VIF perform well on LIVE and CSIQ, and that together with their lower mutual correlation led to the emergence of most short rSIM3s, or rSIM4s. Furthermore, the contribution of MAD to the resulting rSIMs is larger than VIF's due to its better performance on considered benchmarks.

3.4. *Elastic net regularization*

In the proposed approach, the GA was forced to reduce the number of used measures. Such simple multimeasures should avoid overfitting that can characterise regression models using more IQA measures. However, in cases where many correlated predictors are involved in creating the regression model, the multicollinearity can arise in which the least-squares estimate becomes sensitive to random errors in the response. This, along with the lack of independence of used IQA measures for some distortion types or the need for suppressing errors caused by some used predictors, can be responsible for negative regression coefficients in the obtained models. A possible approach to the multicollinearity is to use the elastic net technique [45]. The elastic net is able to identify the important predictors and remove redundancies. This method creates zero-valued coefficients for unimportant predictors and solves the regularization problem [45]. Taking that into account, an experiment was carried out in which the elastic net was fitted using all 16 predictors on LIVE benchmark. In the experiment, in order to ensure better generalisation of the resulting model, 10-fold cross-validation was employed. The following fitted model was obtained: $\text{rSIM4}^{elastic\ net} = 2.26VSI - 29.67FSIM - 20.69FSIMc + 174.3GSM + 1.19IFC + 18.54IW\text{-}SIM + 58.07MAD + 7.98MSSIM + 8.97NQM - 1.72PSNR - 2.51RFSIM - 39.77VIF - 33.01IFS - 9.81SFF$. Here, SSIM and SR-SIM were excluded by the elastic net from the model, since there are many derivatives of SSIM in the used

set of IQA measures. Furthermore, the high correlation of SR-SIM with other measures can be seen in Table 2. The rSIM4$^{elastic\ net}$ aggregates 14 IQA measures and behaves similarly to rSIM4. This also can be said about related rSIMs which were developed using the elastic net on other image benchmarks. Therefore, in the rest of the paper the previously presented multimeasures are evaluated and compared.

4. Evaluation

In this section the proposed multimeasures are evaluated on standard datasets and compared with other techniques.

Table 3 contains results of such evaluation of all developed fusion measures on four benchmarks in terms of SRCC. The best two measures are written in bold. Since measures were developed using some images from benchmark datasets, obtained results on other datasets allow drawing conclusions on their generalisation abilities. For example, measures developed using LIVE or CSIQ benchmarks, which share most of distortion types, tend to perform weakly on images from TID benchmarks, where many new distortion types were added. Here, measures with longer equations, rSIM3-4, performed better than their shorter equivalents since they better evaluate images with known distortions. Similar observation can be made for rSIM1-2. Weighted results were obtained using the number of images in the benchmark as its weight. Overall results are TID2013 biased, i.e., they are better for techniques that outperform other approaches on TID2013 benchmark. They show that all reduced measures, i.e., rSIM1-4$^{2\text{-}3}$, are promising, as well as rSIM3. Matlab scripts which can be used to reproduce these results for rSIMs can be downloaded at `http://marosz.kia.prz.edu.pl/rSIM.html`. It is worth noting that multimeasures with a smaller number of fused measures performed close to their long equivalents, what is important in practice, where only scores for two or three single measures could be required in order to provide acceptable performance.

Table 4 presents evaluation results for the best seven measures, PSNR and rSIM1-4. However, non-rSIM techniques presented in the table can be also compared with reduced rSIMs, i.e.,rSIM1-4$^{2\text{-}3}$, using Table 3 as the reference. The top two models for each criterion are shown in boldface. Obtained results clearly indicate that the developed family of fusion measures outperformed the measures used in the optimisation. In these tests, rSIM1, rSIM2, and rSIM3 were better than other techniques. Furthermore, reduced rSIMs also outperformed these methods. Among other techniques, VSI, MAD, and IFS performed better than other non-rSIM measures.

The results obtained on the basis of four performance indices are promising, but it would be desirable to determine if obtained results are statistically better. In evaluation of the statistical significance, hypothesis tests based on the prediction residuals of each measure after non-linear mapping were conducted using left-tailed F-test [23]. Here, smaller residual variance denoted the better prediction. The summary of statistical significance tests is presented in Table 5. The test covers the best eight used measures and PSNR. The symbol "1", "0" or "-1" in the cell denotes that the measure in the row is statistically better with the confidence greater than 95%, indistinguishable, or worse than the measure in the column, respectively. Obtained results confirm some findings on the performance of measures that were developed using a subset of images from a benchmark that has a completely different set of distortions than the benchmark used for tests. For example, rSIM1 was statistically better than other measures on TID2013 and worse than most of them on LIVE. This can also indicate the overfitting. What is interesting, the behaviour of rSIM2 was better, i.e., this multimeasure was never significantly worse than other measures. Taking into account significance tests, rSIM3 and rSIM4 are also worthy of interest, since they were only worse on TID2013 benchmark. Among short rSIMs, rSIM2^3 was better than other measures. Overall, all obtained multimeasures performed statistically better than any compared measure. This provides an additional motivation for their use.

The superior performance of rSIM family was shown using typical performance indices. It can also be seen on scatter plots with objective and subjective scores. Such scatter plots for rSIM2 and the best three IQA measures for each benchmark dataset are presented in Fig.1. Here, different types of distortions are represented by differently coloured circles. The colours share the distortion type within a dataset. It can be seen that the compared techniques yielded less accurate quality predictions for large DMOS values and small MOS values (i.e., in presence of severe distortions) than rSIM2.

Obtained multimeasures statistically outperformed popular measures, even in cases where only several measures were selected by the GA. However, it would be desirable to compare them with other fusion IQA measures. Table 6 contains comparative evaluation of rSIMs with known fusion measures based on published SRCC values. SRCC was chosen for this purpose, since it is the most often reported performance index in the literature. Unavailability of the sourcecode, or objective scores, of the majority of compared fusion measures prevents more detailed tests which were shown in the previous paragraphs of this section. In the table, the best three results for a given benchmark are written in boldface, results not reported are denoted by "-". Some measures are not benchmark independent, i.e., their authors reported evaluation results on the benchmark that took part in the development of the measure without providing results for other benchmarks, or prepared a one IQA measure for each benchmark, also without cross-benchmark tests, e.g., [24], [26], [27], [29], [30], [31], [35], [36]. Results for approaches that are not dataset independent were excluded from comparison, they are written in italics in the table. Overall results take into account IQA measures for which independent results are known. TID2013 was excluded, since most measures were not evaluated on this benchmark. The comparison with other fusion measures revealed that the proposed rSIMs performed better, in terms of SRCC values, than their competitors. Here, rSIM2, rSIM3, and rSIM2^3 were better than other multimeasures, what is shown by overall values. Among other measures, the fusion

Table 6. Comparison of fusion IQA measures.

IQA multimeasure	TID2008	CSIQ	LIVE	Overall direct	Overall weighted
[26]	*0.8720*	-	-	-	-
[29]	-	-	*0.9500*	-	-
[47]	0.8617	0.9333	0.9460	0.9137	0.9001
[32]	*0.9471*	-	-	-	-
[30]	*0.8882*	*0.9573*	*0.9711*	-	-
[24]	*0.8569*	*0.9453*	*0.9633*	-	-
[48]	0.8902	0.9401	0.9580	0.9294	0.9190
[27]	*0.9098*	*0.9498*	*0.9622*	-	-
[31]	*0.9487*	*0.9755*	*0.9732*	-	-
[25]	0.8849	0.9549	0.9631	0.9343	0.9186
[49]	0.8100	0.9630	*0.9570*	0.9100	0.8843
[46]	*0.9259*	0.9204	0.9423	0.9295	0.9282
[36]	*0.8865*	*0.9141*	*0.9574*	-	-
[28]	**0.9107**	*0.9733*	0.9722	**0.9521**	**0.9395**
[34]	*0.9073*	**0.9688**	**0.9730**	0.9497	0.9365
rSIM1	**0.9124**	0.9552	0.9643	0.9440	0.9337
rSIM2	*0.9218*	0.9609	0.9670	**0.9499**	**0.9409**
rSIM3	**0.9090**	*0.9738*	0.9701	**0.9510**	**0.9384**
rSIM4	0.8881	0.9601	*0.9754*	0.9412	0.9239
$rSIM1^2$	0.9056	0.9542	0.9582	0.9393	0.9287
$rSIM1^3$	0.8962	0.9339	0.9520	0.9274	0.9165
$rSIM2^2$	*0.9038*	0.9577	0.9639	0.9418	0.9297
$rSIM2^3$	*0.9073*	**0.9695**	**0.9730**	**0.9499**	0.9367
$rSIM3^2$	0.8890	*0.9666*	**0.9731**	0.9429	0.9260
$rSIM3^3$	0.9014	*0.9685*	**0.9736**	0.9478	0.9333
$rSIM4^2$	0.8890	**0.9666**	*0.9731*	0.9429	0.9260
$rSIM4^3$	0.8875	0.9670	*0.9731*	0.9425	0.9253

Note: Tests are based on SRCC. The best three measures are shown in boldface, the results in italics are not dataset independent.

measure introduced in [28] also obtained good results, but was worse than rSIM2 (overall weighted). Measures trained on images from a given benchmark tend to perform worse on other benchmarks where some unknown distortion types are introduced, as for, e.g., [46] or [47]. For all examined IQA benchmark datasets, rSIM family of multimeasures showed superior performance: rSIM1 on TID2008; $rSIM2^3$, $rSIM4^3$, and $rSIM4^2$ on CSIQ; and $rSIM2^3$, $rSIM3^2$, and $rSIM3^3$ on LIVE.

5. Conclusions

In this paper, an approach to the fusion of full-reference IQA measures was presented. The fusion was obtained by the genetic algorithm that selected some IQA measures used as predictors in the multiple linear regression. The usage of these two techniques for the development of a hybrid full-reference IQA measure is among the contributions of this work. Furthermore, the genetic algorithm was able to find well-performing multimeasures, which are composed of a predefined number of IQA techniques. Tests using such constrained regression models were also presented. The resulting family of multimeasures, rSIMs, was extensively evaluated in terms of SRCC, KRCC, PCC, and RMSE on four largest IQA benchmark datasets. The results of comparison revealed that the proposed approach is significantly better than popular state-of-the-art IQA measures and better than fusion approaches.

References

[1] Chandler, D. M. (2013). Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, 2013, art. ID 905685.

[2] Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J.,Vozel, B., Chehdi, K., Carli, M., Battisti, F., Kuo, C.-C. J. (2015). Image database TID2013: Peculiarities results and perspectives. *Signal Processing: Image Communication*, 30, 57–77.

[3] Anbarjafari, G. (2015). An objective no-reference measure of illumination assessment. *Measurement Science Review*, 15(6), 319–322.

[4] Valenzise, G., Magni, S., Tagliasacchi, M., Tubaro, S. (2012). No-reference pixel video quality monitoring of channel-induced distortion. *IEEE Transactions on Circuits and Systems for Video Technology*, 22 (4), 605–618.

[5] Li, X., Guo, Q., Lu, X. (2016). Spatiotemporal statistics for video quality assessment. *IEEE Transactions on Image Processing*, 25 (7), 3329–3342.

[6] Damera-Venkata, N., Kite, T. D., Geisler, W. S., Evans, B. L., Bovik, A. C. (2000). Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9 (4), 636–650.

[7] Wang, Z., Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 9 (3), 81–84.

[8] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13 (4), 600–612.

[9] Wang, Z., Simoncelli, E. P., Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *37th Asilomar Conference on Signals, Systems & Computers*. IEEE, 1398–1402.

[10] Wang, Z., Li, Q. (2011). Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20 (5), 1185–1198.

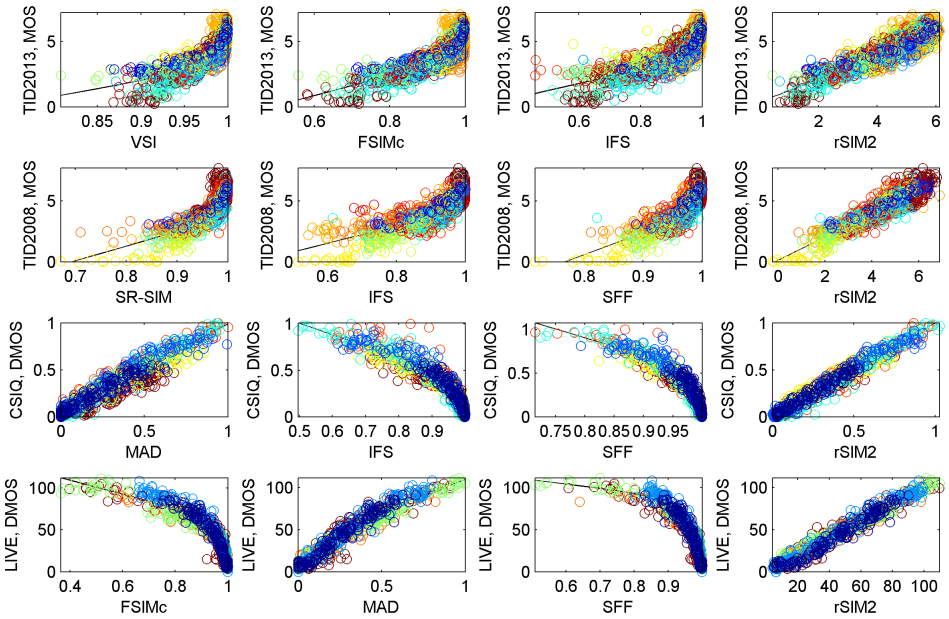

Fig.1. Scatter plots for rSIM2 and the best three state-of-the-art IQA measures for each dataset. Subjective opinion scores are plotted against scores obtained by the measures. Colours represent different distortions.

[11] Sheikh, H. R., Bovik, A. C. (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, 15 (2), 430–444.

[12] Sheikh, H., Bovik, A., de Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14 (12), 2117–2128.

[13] Zhang, L., Zhang, L., Mou, X., Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20 (8), 2378–2386.

[14] Zhang, L., Li, H. (2012). SR-SIM: A fast and high performance IQA index based on spectral residual. In: *19th IEEE International Conference on Image Processing*. IEEE, 1473–1476.

[15] Zhang, L., Shen, Y., Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23 (10), 4270–4281.

[16] Zhang, L., Zhang, L., Mou, X. (2010). RFSIM: A feature based image quality assessment metric using Riesz transforms. In *2010 IEEE International Conference on Image Processing*. IEEE, 321–324.

[17] Wang, F., Sun, X., Guo, Z., Huang, Y., Fu, K. (2015). An object-distortion based image quality similarity. *IEEE Signal Processing Letters*, 22 (10), 1534–1537.

[18] Wu, J., Lin, W., Shi, G. (2014). Image quality assessment with degradation on spatial structure. *IEEE Signal Processing Letters*, 21 (4), 437–440.

[19] Liu, A., Lin, W., Narwaria, M. (2012). Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21 (4), 1500–1512.

[20] Zhou, F., Lu, Z., Wang, C., Sun, W., Xia, S.-T., Liao, Q. (2015). Image quality assessment based on inter-patch and intra-patch similarity. *PLoS ONE*, 10 (3), e0116312.

[21] Guo, S., Xiang, T., Li, X. (2015). Image quality assessment based on multiscale fuzzy gradient similarity deviation. *Soft Computing*, doi:10.1007/s00500-015-1844-9.

[22] Liu, M., Yang, X. (2009). A new image quality approach based on decision fusion. In *Fifth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD '08)*. IEEE, 10–14.

[23] Larson, E. C., Chandler, D. M. (2010). Most apparent distortion: Full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19 (1), 011006.

[24] Peng, P., Li, Z.-N. (2012). Regularization of the structural similarity index based on preservation of edge direction. In: *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2127–2132.

[25] Saha, A., Wu, Q. M. J. (2014). Full-reference image quality assessment by combining global and local distortion measures. arXiv:1412.5488 [cs.CV].

[26] Okarma, K. (2010). Combined full-reference image quality metric linearly correlated with subjective assessment. In *Artificial Intelligence and Soft Computing*.

Springer, 539–546.

[27] Okarma, K. (2013). Extended Hybrid Image Similarity - combined full-reference image quality metric linearly correlated with subjective scores. *Elektronika ir Elektrotechnika*, 19 (10), 129–132.

[28] Oszust, M. (2016). Full-reference image quality assessment with linear combination of genetically selected quality measures. *PLoS ONE*, 11 (6), 1–17.

[29] Lahouhou, A., Viennet, E., Beghdadi, A. (2010). Selecting low-level features for image quality assessment by statistical methods. *Journal of Computing and Information Technology – CIT*, 18 (2), 183–189.

[30] Peng, P., Li, Z.-N. (2012). A mixture of experts approach to multi-strategy image quality assessment. In *Image Analysis and Recognition*, Springer, LNCS 7324, 123–130.

[31] Liu, T.-J., Lin, W., Kuo, C.-C. (2013). Image quality assessment using multi-method fusion. *IEEE Transactions on Image Processing*, 22 (5), 1793–1807.

[32] Jin, L., Egiazarian, K., Kuo, C.-C. (2012). Perceptual image quality assessment using block-based multimetric fusion (BMMF). In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 1145–1148.

[33] Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Carli, M. (2011). Modified image visual quality metrics for contrast change and mean shift accounting. In *11th International Conference – The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*. IEEE, 305–311.

[34] Oszust, M. (2016). Image quality assessment with lasso regression and pairwise score differences. *Multimedia Tools and Applications*, doi:10.1007/s11042-016-3755-x.

[35] Lukin, V. V., Ponomarenko, N. N., Ieremeiev, O. I., Egiazarian, K. O., Astola, J. (2015). Combining full reference image visual quality metrics by neural network. In *Human Vision and Electronic Imaging XX*, SPIE, Vol. 9394, 93940K.

[36] Yuan, Y., Guo, Q., Lu, X. (2015). Image quality assessment: A sparse learning way. *Neurocomputing*, 159, 227–241.

[37] Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (1996). *Applied Linear Statistical Models*, Vol. 4. McGraw-Hill/Irwin.

[38] Sheikh, H. R., Sabir, M. F., Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15 (11), 3440–3451.

[39] Video Quality Experts Group (VQEG). (2003). *Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr_tv2)*. `http://bit.ly/2g1TeXz`.

[40] Yang, Y., Huang, S., Gao, J., Qian, Z. (2014). Multi-focus image fusion using an effective discrete wavelet transform based algorithm. *Measurement Science Review*, 14 (2), 102–108.

[41] Chang, H.-W., Zhang, Q.-W., Wu, Q.-Q., Gan, Y. (2015). Perceptual image quality assessment by independent feature detector. *Neurocomputing*, 151, 1142 – 1152.

[42] Chang, H.-W., Yang, H., Gan, Y., Wang, M.-H. (2013). Sparse feature fidelity for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 22 (10), 4007–4018.

[43] Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F. (2009). TID2008 – A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10, 30–45.

[44] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Professional.

[45] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67 (2), 301–320.

[46] Pei, S.-C., Chen, L.-H. (2015). Image quality assessment using human visual DOG model fused with random forest. *IEEE Transactions on Image Processing*, 24 (11), 3282–3292.

[47] Li, S., Zhang, F., Ma, L., Ngan, K. N. (2011). Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13 (5), 935–949.

[48] Wu, J., Lin, W., Shi, G., Liu, A. (2013). Perceptual quality metric with internal generative mechanism. *IEEE Transactions on Image Processing*, 22 (1), 43–54.

[49] Barri, A., Dooms, A., Jansen, B., Schelkens, P. (2014). A locally adaptive system for the fusion of objective quality measures. *IEEE Transactions on Image Processing*, 23 (6), 2446–2458.

Received August 18, 2016.
Accepted November 28, 2016.