Reprinted from
Journal of Imaging Science and Technology® 65(6): 060409-1–060409-13, 2021.

Multi-gene Genetic Programming based Predictive Models for Full-reference Image Quality Assessment

Naima Merzougui
LESIA laboratory, University Mohamed Khider Biskra, Algeria
Department of Computer Science and Information Technologies, University of Kasdi Merbah Ouargla, Algéria
E-mail: merzougui.naima@univ-ouargla.dz

Leila Djerou
LESIA laboratory, University Mohamed Khider Biskra, Algeria

Abstract. *Many objective quality metrics for assessing the visual quality of images have been developed during the last decade. A simple way to fine tune the efficiency of assessment is through permutation and combination of these metrics. The goal of this fusion approach is to take advantage of the metrics utilized and minimize the influence of their drawbacks. In this paper, a symbolic regression technique using an evolutionary algorithm known as multi-gene genetic programming (MGGP) is applied for predicting subject scores of images in datasets using a combination of objective scores of a set of image quality metrics (IQM). By learning from image datasets, the MGGP algorithm can determine appropriate image quality metrics, from 21 metrics utilized, whose objective scores employed as predictors in the symbolic regression model, by optimizing simultaneously two competing objectives of model 'goodness of fit' to data and model 'complexity'. Six large image databases (namely LIVE, CSIQ, TID2008, TID2013, IVC and MDID) that are available in public domain are used for learning and testing the predictive models, according the k-fold-cross-validation and the cross dataset strategies. The proposed approach is compared against state-of-the-art objective image quality assessment approaches. Results of comparison reveal that the proposed approach outperforms other state-of-the-art recently developed fusion approaches.* © *2021 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2021.65.6.060409]

1. INTRODUCTION

Images have an important role in the media, but their use depends on their quality. They are affected by a wide variety of deformations across different stages of the distribution chain (acquisition, storage, transmission, and delivery to the user). Identifying the distortion factor and quantifying the level of image degradation due to low-resolution, extreme weather illumination, system noise, blurring, compression etc, have been investigated [1–4]. Authors in [1] have presented a study to quantify the level of image degradation due to optical turbulence in natural waters. Authors in [2] have shown a large list of nuisance factors, which would be taken into account when doing real-world image super-resolution, including blur (e.g. motion or defocus), compression artefacts, color and sensor noise. However, cloud cover and cloud shadow are two of the most common noise sources for the majority of Remote Sensing (RS) data in the range of the visible and infrared spectra [3]. A general overview of image degradation and recent progress in the field of image quality assessment is provided in [4].

A large amount of Image Quality Assessment (IQA) methods have been introduced in the last decade, which are categorized into subjective and objective ones. The subjective IQA methods are carried out by human observers, where the image quality assessment score is determined by averaging the scores assigned by a panel of human observers following specific protocols. These methods are precise in estimating the visual quality of an image but they take considerable time and requiring a large number of observers. Moreover, they cannot be automated.

Conversely, the objective IQA methods are computer-based methods that can automatically predict the perceived image quality. These methods are usually developed to take into account the human visual system (HVS), and thus have the goal of correlating with subjective assessment. They are relatively quicker and cheaper than subjective assessment.

Depending on the nature of the information required to perform the assessment, existing objective IQA methods can be classified into three categories [5, 6]: Methods that compare the version of the distorted image with a reference version are usually called Full-Reference (FR) [7]. Methods comparing a description of the image to be evaluated with just partial information about the original image are called Reduced Reference (RR) [8]. Finally, methods that require neither the reference image nor any of its are called No-Reference techniques (NR) [9].

In response to the universality and good performance of FR methods, a large number of various FR-IQA metrics have been proposed in the literature, including VSI [10], FSIM [11], FSIMc [11], GSM [12], IFC [13], IW-SSIM [14], MAD [15], MSSIM [16], NQM [17], PSNR [18], RFSIM [19], SR-SIM [20], VIF [21], IFS [22], SFF [23], SIM [24], COHERENSI [25], UNIQUE [26], MSUNIQUE [27], Per-SIM [28], RVSIM [29]. Each one has its own intended use or construction.

Received Aug. 7, 2021; accepted for publication Nov. 18, 2021; published online Dec. 7, 2021. Associate Editor: Danli Wang.
1062-3701/2021/65(6)/060409/13/$25.00

However, the quality assessment of images subjected to various types of distortions is still one of the most challenging problems in computer vision and image analysis [30, 31]. For this reason, many recent studies have adopted a new strategy that involves various fusion techniques for quantifying image quality. They used different types of information so that their combination provided new possibilities leading to better correlation with the subjective scores. Based on the usage information, the IQA fusion algorithms can be classified into two categories; in the first, the fusion concerns a few features extracted from the image [32–34]. However, in the second, the fusion is based on the combination of certain FR-IQA metrics [35–37].

Moreover, recent trends related to the use of learning techniques has a strong competitor in the IQA fusion algorithms since it is difficult to predict visual quality under various distortion and rich image contents using a single formula. Most of the machine learning techniques used in IQA fusion algorithms are conventional learning methods such as Artificial Neural Networks (ANN), and Support Vector Machines (SVMs). Examples of learning-based IQA methods can be found in [38–40], among many others.

Learning-based methods have proven to be successful in assessing image quality, and this paper aims to follow this direction of research by adopting another framework of machine learning called Multi-gene genetic programming (MGGP) [41].

MGGP is a biologically inspired machine learning method belonging to the class of Evolutionary Algorithms (EA) that evolves computer programs (represented by tree structures) to perform a task. It is one of the most advanced variants of Genetic Programming (GP) algorithm [42], that linearly combines low depth GP trees in order to improve fitness of the standard GP.

MGGP is considered as an efficient method for solving the symbolic regression problem [43], by searching the space of mathematical expressions to find the model that best fits a given dataset, both in terms of accuracy and simplicity. Each regression model (individual or solution of the MGGP population) is a weighted linear combination of low-order non-linear transformations of the input variables [44]. Even when large numbers of input variables are utilized, this technique can automatically select the most contributed variables in the model, formulate the structure of the model, and solve the coefficients in the regression equation, while simultaneously optimizing for both accuracy and complexity [45]. As there are two objective functions, that are to be considered simultaneously (complexity and accuracy of models), the MGGP technique gives rise to a number of optimal solutions, known as Pareto-optimal solutions; each one describes an accurate equation of regression model [44]. These properties enable MGGP to be a powerful technique for solving regression problems, compared to other conventional learning methods such as Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) [44, 46, 47], which are known as trajectory-based algorithms and black-boxes, because they select a single solution through the learning process and the mechanisms behind are difficult to understand and analyze [44].

In this study, we exploit the advantages of MGGP, regarding extraction of the most information from the data and building of the accurate regression models, to automatically generate predictive models for FR-IQA. The predicted score of an image is obtained by the weighted sum of objective scores of a mix of image quality metrics (IQM). Based on their intended use or construction and the fact that the code of all of them is publicly available, 21 image quality metrics are utilized: VSI [10], FSIM [11], FSIMc [11], GSM [12], IFC [13], IWSSIM [14], MAD [15], MSSIM [16], NQM [17], PSNR [18], RFSIM [19], SRSIM [20], VIF [48], IFS [22], SFF [23], SSIM [24], COHERENSI [25], UNIQUE [26], MSUNIQUE [27], PerSIM [28], RVSIM [29]. By learning from benchmark image datasets, the MGGP can determine the appropriate image quality metrics, from these 21 metrics, whose objective scores are utilized as predictors, in the symbolic regression, by optimizing simultaneously two competing objectives of model 'goodness of fit' to data and model 'complexity'. To evaluate the performance of the proposed method, which is called Full Reference metric, based Multi Gene Genetic Programming (FR-MGGP), and present its properties, several experiments are carried, with different aspects, on six distinct image databases (namely LIVE, CSIQ, TID2008, TID2013, IVC and MDID), such as the number, types and levels of distortions, in each distorted image. Due attention is paid to the demonstration the generalization powerof the proposed approach from the k-fold-cross-validation and the cross dataset experiments.

The rest of the paper is organized as follows. Section 2 briefly reviews the related works. Section 3 presents Multi Gene Genetic Programming (MGGP). Section 4 describes the proposed FR-MGGP method in detail. Experimental results are discussed in Section 5. The paper ends with the conclusion followed by the references.

2. RELATED WORK

Considering the necessity to have IQA metric, which can efficiently estimate image quality across all degradations, numerous IQA fusion algorithms have been proposed, in the literature [35, 46, 47], with varying degrees of success. The IQA fusion algorithms can be classified into two categories. In the first category, the IQA fusion algorithms aim at predicting image quality from features extracted from images after computing a mix of selected features, a combination of elements in this set is analyzed through permutations and combinations to produce the quality score of the images. For example, in [15], Chandler proposed Most Apparent Distortion algorithm (MAD), where the local luminance and contrast masking evaluate high-quality images, and local statistics of spatial-frequency components assess low-quality images. A nonlinear combination of features extracted is applied in [49]. In [50], the authors developed a non-distortion specific IQA technique which uses features derived from the conformity of the first digit distribution (FDD) of natural images in the transform domain with Benford's law

(BL); a non-linear mapping of these features is trained using Gaussian process regression with a rational quadratic kernel. In [51], a similarity-based FR-IQA model is introduced without learning procedure, which combines three feature information processing parts: visual saliency, structure and chrominance features.

In the second category, the IQA fusion algorithms are based on the combination of two or more elementary metrics. The predicted image quality is the result of the combined metrics values, according to the strategy applied. For example, a canonical correlation analysis was used to combined SNR, SSIM, VIF, and VSNR in [52]. The regularized regression was used to combine up to seven IQA models in [53]. A nonlinear combination of scores MSSIM, VIF and R-SVD was proposed in [54]. Thereafter the metric R-SVD was replaced with FSIMc [55] as well as a nonlinear combination of RFSIM and FSIMc metrics in [56]. Another combination with RFSIM and weighted FSIMc metric led to the Extended Hybrid Image Similarity (EHIS) metric [35]. The verification of this approach for multiple distorted images was discussed in [57]. Machine learning approach was used in [58] to combine the scores of multiple FRIQA determined by a selection algorithm. Another approach based on conditional Bayesian mixture of experts model was proposed in [59]. In that paper, a support vector machine classifier was used to predict the type of distortion, and then SSIM, VSNR, and VIF were fused with k-nearest-neighbor regression. Other mechanisms were used to combine objective quality measures; for example, internal generative in [60] and adaptive weighting in [61]. Also in [38], the neural network was used to combine six IQA measures. In [62, 63], preliminary work with a non-linear combination of several IQA measures selected by a genetic algorithm was shown, except that in [63] it is the weighted sum instead of the weighted product in [62]. The same author in [37], use the multiple linear regression of opinions provided by genetically selected IQA measures. A linear regression is also applied in [64] to combine three sub-measures which are based on maximizing contrast with minimum artifact. A Robust linearized combined metrics [65] have been designed for three configurations: median of three estimates of MOS resulting from elementary metrics, median of five estimates, alpha-trimmed mean of five estimates. Another metric is proposed in [39]; the authors used the particle swarm optimization schema to select a set of relevant IQA metrics, and then a support vector regression based fusion strategy is adopted to derive the overall index of image quality. An artificial neural network was used to combining several no-reference metrics in [40], where several types of such networks with different configurations and the influence of various factors on the final accuracy of such metrics have been studied.

3. MULTI GENES GENETIC PROGRAMMING

Genetic programming GP [42] is a symbolic optimization technique inspired by Darwin's theory of evolution. It allows discovering and optimizing complete mathematical and computational models that best describe some desired phenomena by mimicking the process of evolution in nature. GP generates both of model types and its coefficients automatically based on the given input data. It employs a population of individuals (or solutions), each of them denoted by a tree structure that codifies a mathematical equation, which describes the relationship between a set of input features and the output. The population is then, repeatedly updated under fitness-based selection, by applying genetic operators until the desired solution is found. A detailed review of GP literature has been authored by [66].

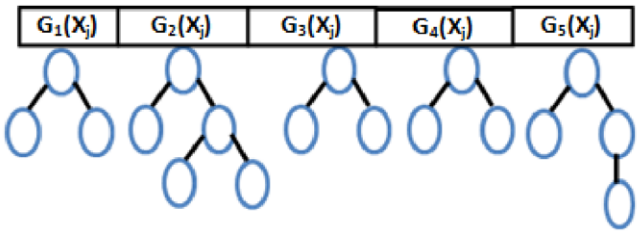

Figure 1. Example of multi-genes individual.

In recent years, several improvements of GP have been suggested. Multi-Gene Genetic Programming (MGGP) is one of the most recent advancements that combines the ability of the standard GP in constructing the model structure with the capability of traditional regression in parameter estimation [43].

In traditional GP the population is a set of individuals, each of them denoted by a tree structure that is composed of the terminal and function set. The elements of the terminal set X can input process variables and random constants. The function set codifies a mathematical equation, which describes the relationship between the output Y and a set of input variables X. Based on these ideas, MGGP generalizes GP as it denotes as a set of tree structures, commonly called multi-genes (Figure 1), that similarly receives X and tries to predict Y.

In MGGP, each prediction of the output variable Y is formed by a weighted output of each of the trees/genes in the multi-genes individual plus a bias term. The mathematical form of the multi-genes representation can be written as:

$$Y = \sum_{i=1}^{N} d_i G_i(X) + d_0, \tag{1}$$

where G_i is the value of the ith gene (it is a function of one or more of the input variables of the terminal set X, di is the ith weighting coefficient, N is the number of genes, d_0 is a bias term, and Y is the predicted output.

During the MGGP run, in addition to the standard GP subtree crossover (a low-level crossover), genes can be acquired or deleted using a tree crossover operator called high-level crossover. In the low-level crossover, a gene is

chosen at random from each parent individual. Then, the standard sub-tree crossover is applied and the created trees will replace the parent trees in the otherwise unaltered individual in the next generation. The high-level crossover allows the exchange of one or more genes with another selected individual subject to the Gmax constraint. If an exchange of genes results in any individual containing more genes than Gmax, the genes will be randomly selected and deleted until the individual contains Gmax genes [43]. With respect to genetic operators, a mutation in MGGP is similar to that in GP. As for the crossover, the level at which the operation is performed must be specified: it is possible to apply crossover at high and low levels.

In general, the evolutionary process in MGGP differs from that in GP due to the addition of two parameters [43]: maximum number of trees per individual and high level crossover rate. A high value is normally used for the first parameter to assure a smooth evolutionary process. On the other hand, the high-level crossover rate, similar to other genetic operator rates, needs to be adjusted.

In addition, the MGGP simultaneously optimizes two objectives (maximizing the goodness-of-fit and minimizing the model complexity), through a non-dominated concept. While the GP uses only one objective; maximizing the goodness-of-fit, in the process of developing the model quality of fit to training data. Using a single objective in the optimization process makes the models developed become too complex.

4. THE PROPOSED METHOD

In this paper, we exploit the flexibility of MGGP to propose a novel and explicit formulation of combination of image quality metrics (IQMs) which provide some new possibilities leading to better correlation with subjective scores of various kinds of distortions. The idea was inspired from the IQA problematic:

- Despite all available IQMs, there is no single metric that can predict or measure image quality containing various types of distortions. Using a fusion of IQA metrics into some combined ones should provide the opportunity to compute the objective score, which is linearly correlated with subjective perception of various kinds of distortions.
- However, finding the "optimal" number of the appropriate IQMs, used for fusion, from a large set of available IQMs, is usually a challenge since it requires a priori knowledge and large experiments. Using MGGP aims to find an explicit formulation of image quality metrics (IQM) combination, in a regression model form, which best fits a given dataset, both in terms of accuracy and simplicity.

In this context, we have chosen 21 IQMs: VSI [10], FSIM [11], FSIMc [11], GSM [12], IFC [13], IWSSIM [14], MAD [15], MSSIM [16], NQM [17], PSNR [18], RFSIM [19], SRSIM [20], VIF [48], IFS [22], SFF [23], SSIM [24], COHERENSI [25], UNIQUE [26], MSUNIQUE [27], PerSIM [28], RVSIM [29], based on their intended use or construction and the fact that the code of all of them is publicly available. The MGGP may exploit huge search space that consists of all possible combinations of objective scores of these IQMs and find the best regression model, considering its simplicity as well as its performance on the learning data, which describes the weighted sum of IQM objective scores for predicting the subject scores of images in datasets, without making any a-prior assumptions about the model structure.

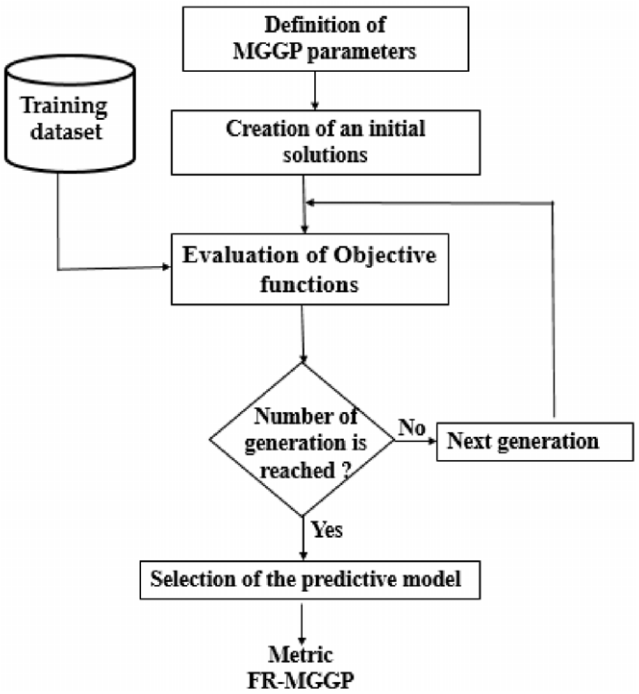

Figure 2. Flowchart of the proposed metric.

For facility the implementation of multi gene genetic programming, we have chosen the widely used technology platform for symbolic regression via MGGP "GPTIPS", which is a free open source MATLAB toolbox developed by Dominic Searson [43]. An improved version GPTIPS2 [67] is available since May 2015.

A flow chart of the proposed method, which is called Full Reference metric, based Multi Gene Genetic Programming (FR-MGGP), is shown in Figure 2. The description of its important features (training dataset, solution, objective functions and MGGP parameters related) is below.

4.1 *Training Dataset*

In order to build the MGGP models and test their efficiency and the generalization their ability, we used six popular images benchmarks; Laboratory for Image and Video Engineering (LIVE [24]), Categorical Subjective Image Quality (CSIQ [15]), Tampere Image Database 2008 (TID2008 [68]), Image and Video Communication (IVC [69]), Tampere Image Database 2013 (TID2013 [70]) and Multiply Distorted

Figure 3. (a) An example of image of LIVE database, and its distorted versions by: (b) jpeg2000, (c) jpeg, (d) White Noise, (e) Gaussian Blur and (f) Fast Fading Rayleigh.

Table I. IQA benchmark image datasets.

	LIVE	CSIQ	TID2008	TID2013	IVC	MDID
Number of ref. images	29	30	25	25	10	20
Number of dis. Images	779	866	1700	3000	185	1600
Distortion types	5	6	17	24	4	5
Distortion levels	5	4–5	4	5	5	4
Type of database	STD	STD	STD	MTD	STD	MTD
Number of dist. in an image	1	1	1	1–2	1	1–4
Format of the subjective scores	DMOS	DMOS	MOS	MOS	DMOS	MOS
Range of scores	0–100	0–1	0–9	0–9	1–5	0–8
Image format	BMP	PNG	BMP	BMP	BMP	BMP
Year	2006	2010	2008	2013	2005	2017

Table II. The utilized full reference IQA metrics measures.

x_1: VSI [10]	x_{12}: SR-IM [20]
x_2: FSIM [11]	x_{13}: VIF [48]
x_3: FSIMc [11]	x_{14}:IFS [22]
x_4: GSM [12]	x_{15}: SFF [23]
x_5: IFC [13]	x_{16}: SSIM [24]
x_6: IW-SSIM [14]	x_{17}:COHERENSI [25]
x_7: MAD [15]	x_{18}: UNIQUE [26]
x_8: MSSIM [16]	x_{19}: MSUNIQUE [27]
x_9:NQM [17]	x_{20}: PerSIM [28]
x_{10}: PSNR [18]	x_{21}: RVSIM [29]
x_{11}: RFSIM [19]	

Image Database (MDID [71]). Each one contains reference images, distorted images, the corresponding subjective score, e.g., mean opinion score (MOS) or differential mean opinion score (DMOS). However, they are distinct in term, the number of reference and distorted images, type of data; images contain simultaneously multiple types of distortions MTD, or single Type of Distortion (STD), number, types and levels of distortions in distorted image. For example in the case of LIVE, distorted images were made using five types of degradation at five different levels. Which are: jpeg2000, jpeg, White Noise, Gaussian Blur and Fast Fading Rayleigh. An example of a reference image and its 5 distorted versions is shown in Figure 3. The detailed information of the used IQA databases is presented in Table I.

To ensure a robustness of results, multiple training sets were constructed, according to k-fold-cross-validation strategy; each image dataset was divided into 5 sets of approximately equal size (i.e., the 5-fold method is used); 20% of total images in each database are selected as the test set, and the remaining 80% were used for training. This was done 5 times, where each set was used as the testing set once; each image get to be tested exactly once and is used in training k-1 times. The average accuracy, of the tests over 5 sets, is taken as the performance measure.

The training set consists of a set of images used only for learning and played a role in building the MGGP models. The testing set is not used to evolve the models and serves to give an indication of how well the models generalize to new data.

Each image in training set is described by subjective scores y (MOS or DMOS), and its evaluations by 21 utilized IQMs presented as objective scores $x_i : i = 1, \ldots 21$ (see Table II). The sample structure, in the training set, is illustrated in Figure 4.

The subjective score y is output/response variable, the objective scores, $x_i : i = 1, \ldots 21$, provided by the utilized IQMs, are the input variables used to predict the quality score. The predictive model of y is a weighted linear combination of low order non-linear transformations of the input variables.

Image index	y (MOS/DMOS)	(1) : x_1	(2) : x_2	...	(21) : x_{21}

Figure 4. Simple structure in training set.

4.2 *Solutions*

In MGGP, each solution (individual) is usually expressed as a structure of trees, also called genes. It contains randomly between 1 and Gmax genes (trees) with the tree depth chosen randomly between 1 and Dmax; Gmax and Dmax are parameters set by the user. Each tree is composed of the terminal and function set. The elements of the terminal set T are the index of the IQMs, which are randomly selected between 1 and 21. The function set $F = \{\text{“+”}, \text{“−”}\}$ is internal nodes of tree.

Each solution codifies a mathematical equation, which describes a regression model for the predicted score (prediction quality) $\hat{y}$ of each image in training set. The predicted scores $\hat{y}$ can be written as the combination of weighted objective scores xi of the IQMs whose index i, $i \in \{1, 2, \ldots, 21\}$, are presented in terminal set of the solution, plus a bias term;

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_G x_G; \quad G \in \{1, 2, \ldots, 21\}, \tag{2}$$

where b_0 is bias (noise) term and b_G are the regression coefficients (i.e. gene weights). Generally, the coefficients are determined by the ordinary least squares method for each MGGP individual [67].

4.3 *Objective Functions*

MGGP creates a sequence of populations, the generations, by applying genetic operators (selection, crossover and mutation) to the individuals. In the first generation, a population of random individuals is generated. Each individual contains randomly between 1 and Gmax genes. A tree representing each gene is formed (with the tree depth chosen randomly between 1 and Dmax) by randomly combining the elements from the functional set (+, −) and the some elements of terminal set (the index of the IQMs).

In each generation, individuals are evaluated simultaneously, using two objective functions, one expressing the complexity of individual structure codifying the model mathematical equation and the other, its accuracy [67].

The restriction on the maximum number of trees (genes), in individual structure, and depth of the gene exerts control over the complexity of the model and results in accurate and compact model. Minimizing the sum of the nodes, in structure of individual trees, is used as objective function to evaluate the complexity of model.

The accuracy of the model is presented by an objective function will typically be maximization or minimization of some aggregate function that combines results of applying every sample of the training set to the individual that codifies the mathematical equation of this model. Maximizing the correlation coefficient R^2 (or minimizing 1-R^2) can be used as an objective function to evaluate the model accuracy:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}, \tag{3}$$

where y_i and $\hat{y}_i$ are subjective and predicted scores for the ith image in training set, respectively, $\overline{y}$ is the average of the predicted scores and n is the number of images (samples) in training set.

As there are, two objective functions must be considered simultaneously, the problem is known as multi-objective optimization problem. Therefore, to satisfy these objective functions, a set of optimal solutions is required instead of one optimal solution. The reason for the optimality of many solutions is that no one objective function can be considered to be better than any other. These solutions are “Pareto-optimal” solutions or non-dominated solutions. The image of the Pareto optimal set in the objective space is called the Pareto front. The decision maker has to look at this set of Pareto-optimal solutions and make a choice of one solution.

In the MGGP algorithm, the determination of the Pareto-optimal solutions is done according to the principle of NSGA-II algorithm [72], which is incorporated into GPTIPS [67]. Two measures are used when comparing individuals (for selection and breeding): The first is the non-domination rank, which measures how close an individual is to the non-dominated front. An individual with a lower rank (closer to the front) is always preferred to an individual with a higher rank. If two individuals have the same non-domination rank, as a secondary criterion, a crowding distance is used to increase the diversity of the population giving more priority to the individual that have a large average crowding distance.

By learning from one training set, the proposed method gives rise to a number of optimal solutions, known as Pareto-optimal solutions; each one describes an accurate equation of regression model to predict the quality score. However, these models are of varying complexity and performance. Figure 5 display examples of the population of evolved models in terms of their complexity as well as their performance, for the five training sets (5-fold-cross-validation) in the case of LIVE benchmark. Where the blue circles show the results of all evolved models, the green circles comprises of the Pareto-optimal models in the population.

4.4 *Selection of the Predictive Model*

In order to select the predictive model, from one Pareto-optimal model, a tradeoff must be made between model complexity and performance. We noticed that the Pareto models in the lower left of the population (high accuracy and low complexity) are usually where a satisfactory solution may be found.

For each benchmark data set, we get five predictive models for full-reference image quality assessment (M_i, $i = 1..5$), as the result of applying the 5-fold cross-validation strategy. Therefore, the average of these five models is the accurate predictive model for predicting subject scores of images in

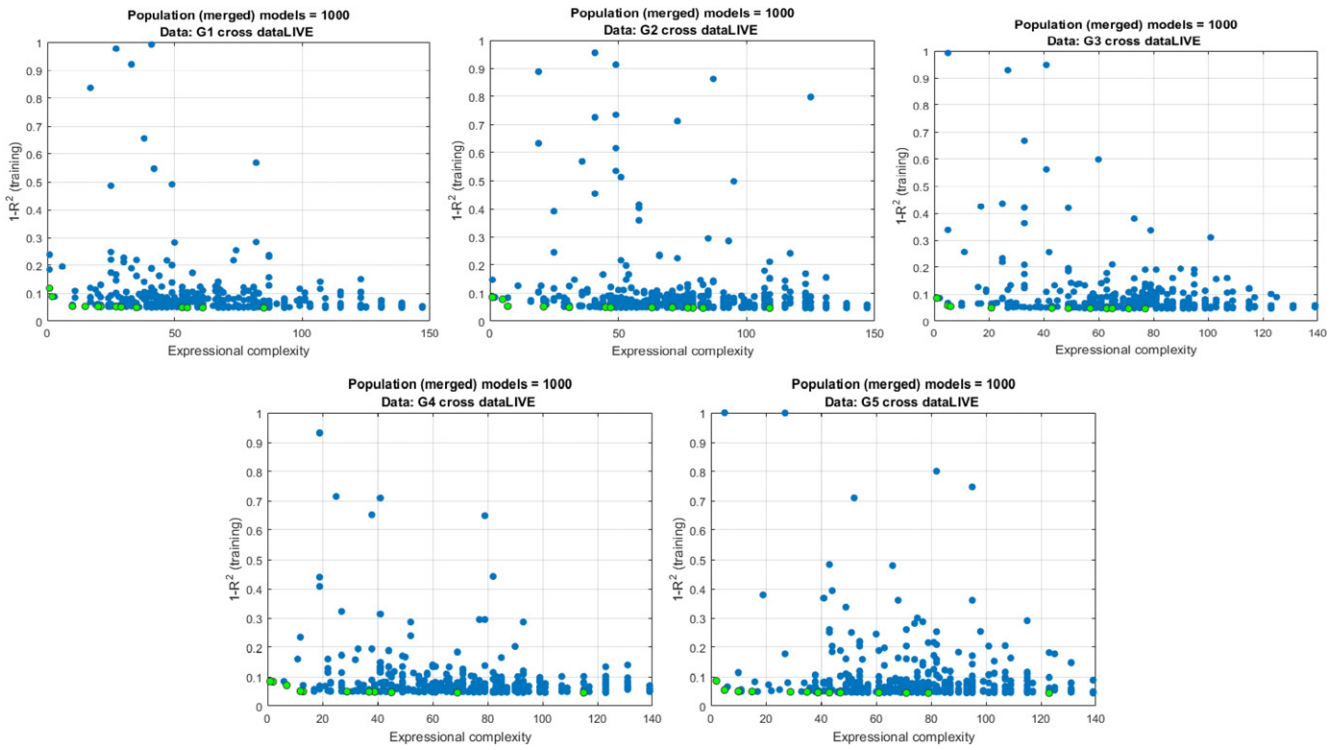

Figure 5. Evolved model population, in terms of model performance ($1 - R^2$) and model complexity, for the five training sets (5-fold-cross-validation) in the case of LIVE benchmark, models on the Pareto front are marked in green.

datasets considered, it is called Full Reference metric based Multi Gene Genetic Programming (FR-MGGP).

For the utilized benchmark datasets, the accurate predictive models are:

$$\begin{aligned}\text{FR-MGGP1(LIVE)} = {} & 9.8^{*}\text{IFC} + 69.0^{*}\text{MAD} + 13.0^{*}\\ & \text{NQM} - 44.0^{*}\text{IFS} - 31.0^{*}\text{SFF} - 40.0^{*}\\ & \text{UNIQUE} + 49.0^{*}\text{MSUNIQUE} + 53.0; \end{aligned} \tag{4}$$

$$\begin{aligned}\text{FR-MGGP2(CSIQ)} = {} & 2.4^{*}\text{MAD} + 3.2^{*}\text{MSSIM} - 1.1^{*}\text{RFSIM} - 1.4^{*}\text{SRSIM}\\ & - 0.51^{*}\text{IFS} - 0.46^{*}\text{SFF} - 0.51^{*}\text{MSUNIQUE} - 1.3\\ & ^{*}\text{RVSIM} + 1.9; \end{aligned} \tag{5}$$

$$\begin{aligned}\text{FR-MGGP3(TID2008)} = {} & 17.0^{*}\text{VSI} - 1.6^{*}\textit{MAD} + 2.1^{*}\text{PSNR} - 4.0^{*}\text{VIF} + 2.2\\ & ^{*}\text{IFS} + 3.6^{*}\textit{SFF} + 2.3^{*}\text{COHERENSI} - 15.0; \end{aligned} \tag{6}$$

$$\begin{aligned}\text{FR-MGGP4(TID2013)} = {} & 1.8^{*}\text{FSIMc} - 3.1^{*}\text{IW} - \text{SSIM} - 1.8^{*}\text{MAD} - 1.0\\ & ^{*}\text{NQM} + 2.4^{*}\text{SFF} + 3.4^{*}\text{COHERENSI} + 0.93^{*}\\ & \text{UNIQUE} + 1.6^{*}\text{PerSIM} + 0.045; \end{aligned} \tag{7}$$

$$\begin{aligned}\text{FR-MGGP5(IVC)} = {} & 0.2^{*}\text{VIF} - 0.023^{*}\text{MAD} - 5.7^{*}\text{MSSIM} - 0.06\\ & ^{*}\text{NQM} - 1.7^{*}\text{IWSSIM} + 4.8^{*}\text{IFS} - 2.7^{*}\text{SSIM} + 2.9^{*}\\ & \text{UNIQUE} + 0.46^{*}\text{PerSIM} + 4.9^{*}\text{RVSIM} + 4.2; \end{aligned} \tag{8}$$

$$\begin{aligned}\text{FR-MGGP6(MDID)} = {} & 68.0^{*}\text{IW} - \text{SSIM} - 9.7^{*}\text{MAD} - 69.0^{*}\textit{MSSIM} - 7.2^{*}\\ & \text{PSNR} + 21.0^{*}\text{VIF} + 40.0^{*}\text{IFS} - 8.2^{*}\text{COHERENSI}\\ & + 7.1^{*}\text{UNIQUE} - 7.1; \end{aligned} \tag{9}$$

4.5 *MGGP Parameters*

A set of parameters must be defined for MGGP's evolutionary process. The number of individuals (solutions or programs) in the population is determined by the population size (PS). The number of levels the algorithm, will use before the run terminates, is defined by the number of generations (NG). The size and various forms of the model to be searched for in the global solution space, are defined by the maximum number of genes (trees) allowed in an individual (Gmax) and the maximum tree depth (Dmax). The high level crossover rate to be employed to generate new genes for individuals as well as to reduce the overall number of genes for one model and increase the total number of genes for the other is determined by the crossover rate parameter (CR). Scattered crossover, Gaussian mutation and stochastic uniform selection rules were used. The values of these parameters are determined basing on previously suggested values that can be found in the literature [43, 67, 73] and by experiments. Table III shows the parameter settings used for the MGGP implementation in this study.

5. EXPERIMENTAL EVALUATION

To evaluate the performance of the proposed method, several carried experiments with different aspects, are analyzed in the following paragraphs.

Table III. Parameter settings for MGGP algorithm.

Parameter	Setting
Population size (PS)	100
Number of generations (NG)	100
Function set	+, −.
Maximum number of genes (Gmax)	3
Maximum tree depth (Dmax)	5
Tournament size	2
Elite_fraction	0.05
Crossover event (CR)	0,85
Mutation events	0,3

5.1 *Evaluation of Prediction Performance*

The relationship between image quality scores predicted by objective IQA metrics and subjective scores (typically expressed as MOS or DMOS) can be seen on scatter plots. Figure 6 presents the scatter plots for the proposed models (FR-MGGPs models) and the best three state-of-the-art IQA metrics for each benchmark dataset. Additionally, a fit with a logistic function as suggested in [18] and [74] is shown for easier comparison. Each point on the plot represents one image in the benchmark, the horizontal axis corresponds to the (scaled) metric score and the vertical axis corresponds to the subjective scores for that image. Compared with other scatter plots, the FR-MGGPs models show better linearity and correlation. We noticed that the resultant FR-MGGP is adequate for human perception on all benchmark dataset; where the percentage of outliers is decreased and a tendency to a monotonic behavior is increased.

To evaluate the six developed models FR-MGGPs, we used the following performance indices, concerning the prediction accuracy, monotonicity, and consistency [18, 74]; the Root Mean Square Error (RMSE), Pearson linear Correlation Coefficient (PCC), Spearman Rank order Correlation Coefficient (SRCC) and Kendall Rank order Correlation Coefficient (KRCC). Considering the nonlinear relationship between image distortions and their perceived quality, these performance indices are calculated after a nonlinear mapping function [5, 18, 75] between a vector of objective score S and subjective scores (MOS or DMOS), using the following mapping function for the non-linear regression (as recommended by the Video Quality Experts Group (VQEG) [74]):

$$S_m = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp\left(\beta_2 \left(S - \beta_3\right)\right)} \right) + \beta_4 S + \beta_5, \quad (10)$$

where $\{\beta_1, \beta_1, \beta_1, \beta_1, \beta_1\}$ are parameters to be fitted, and S_m is the non-linearly mapped S. Different values may lead to different PCC and RMSE values, but they do not affect SRCC and KRCC. This nonlinearity is applied to the FR-IQA algorithm score, which give a better fit for all data.

Table IV presents the evaluation results, for the best ten models (Among 21 utilized metrics) and FR-MGGPs. The top two models for each criterion are shown in boldface. According to [18, 74], Higher SRCC, KRCC and PCC values are considered better, in contrary to the values of RMSE. We observe for example, in LIVE database, the RMSE values of all metrics belong to the interval [5,913–12,022], where the smallest values are 5,913 and 6,816 for FRMGGP1 and FRMGGP3 respectively and the highest value is 12,022 for MS-UNIQUE. On the other hand, we note that the highest values of the three coefficients (SRCC, KRCC and PCC) are obtained by FRMGGP1 and FRMGGP3; which confirms that they are the two best methods for this database.

In summary, results of Table IV show that all FR-MGGP measures are clearly top performing models compared to measures of the other databases; the metrics FR-MGGP1 and FR-MGGP3 are significantly outperformed compared to measures on LIVE. For CSIQ the FR-MGGP2 is the best one. FR-MGGP3 and FR-MGGP4 yielded very best results on TID2008. And for TID2013, the FR-MGGP4 is better than all the compared metrics. FR-MGGP5 and FR-MGGP6 gives best performance on IVC and MDID respectively.

5.2 *Performance Comparisons with Fusion IQA Measures*

It would be agreeable to compare FR-MGGP measures with other related fusion IQA measures in terms of the commonly used performance measure of state-of-the-art IQA algorithms, SRCC values. We choose four comprehensive databases including TID2013, TID2008, CSIQ and LIVE, considering the most utilized to evaluate the fusion IQA measures.

Table V contains such comparative evaluation. Since values of PCC, KRCC and RMSE are often not available, SRCC was used as a basis for comparison. The two best results for a given benchmark dataset are written in boldface. Some results were not reported in referred work; therefore, they are denoted by "−". Results for No Reference method (are written in italics) were excluded from comparison, e.g. [39, 40], since our method is a full reference.

Evaluation results show that for the LIVE database, FR-MGGP1 outperformed other approaches. FR-MGGP2 is the best one on CSIQ database, and FR-MGGP3 and FR-MGGP4 performed well on TID2008 database. On TID2013 database, FR-MGGP4 provided superior result than the other approaches and is comparable with CNNM (this method is trained and tested on only one database).

It can be seen that measures trained on smaller benchmarks tend to perform poorly on the large one, e.g., FR-MGGP1, FR-MGGP2 on TID2013, and FR-MGGP2 on TID2008. The main factor that affects the performance of evaluated techniques is the number of distortions presented in benchmark databases. Approaches trained on large benchmarks, which contain a variety of distortions, usually perform better than techniques that are trained on smaller benchmarks, e.g., FR-MGGP4 on the LIVE, CSIQ and also in TID2008.

Moreover, overall results were calculated excluding TID2013 because some measures had not been evaluated on it, and take into account IQA measures for which

Table IV. Performance comparison of the proposed approach with the best ten models of IQA.

	VSI	SRSIM	IWSSIM	FSIMc	SFF	MAD	MSSIM	VIF	MS-UNIQUE	PerSIM	FR-MGGP1	FR-MGGP2	FR-MGGP3	FR-MGGP4	FR-MGGP5	FR-MGGP6
									LIVE							
SRCC	0,952	0,961	0,956	0,964	0,964	0,966	0,951	0,963	0,945	0,943	**0,975**	0,925	**0,970**	0,966	0,939	0,965
KRCC	0,805	0,829	0,817	0,836	0,836	0,842	0,804	0,828	0,795	0,795	**0,860**	0,767	**0,846**	0,834	0,790	0,831
PCC	0,948	0,955	0,952	0,961	0,963	0,967	0,948	0,941	0,898	0,920	**0,976**	0,916	**0,968**	0,965	0,923	0,962
RMSE	8,681	8,081	8,347	7,529	7,346	6,907	8,618	9,240	12,022	10,860	**5,913**	10,924	**6,816**	7,156	10,504	7,425
									CSIQ							
SRCC	0,942	0,931	0,921	0,931	0,962	0,946	0,913	0,919	0,929	0,929	**0,966**	**0,975**	0,965	0,952	0,920	0,948
KRCC	0,785	0,772	0,752	0,769	0,828	0,797	0,739	0,753	0,759	0,768	**0,839**	**0,862**	0,837	0,809	0,755	0,802
PCC	0,927	0,925	0,914	0,919	0,964	0,95	0,899	0,927	0,928	0,898	**0,972**	**0,979**	0,965	0,951	0,931	0,956
RMSE	0,097	0,099	0,106	0,103	0,069	0,082	0,114	0,098	0,097	0,115	**0,061**	**0,053**	0,068	0,080	0,095	0,076
									TID08							
SRCC	0,897	0,891	0,855	0,884	0,876	0,834	0,854	0,749	0,869	0,856	0,891	0,902	**0,914**	**0,914**	0,784	0,838
KRCC	0,712	0,714	0,663	0,699	0,688	0,644	0,656	0,586	0,681	0,679	0,707	0,727	**0,742**	**0,748**	0,611	0,653
PCC	0,876	0,886	0,857	0,876	0,881	0,830	0,845	0,808	0,845	0,837	0,896	0,904	**0,917**	**0,912**	0,829	0,862
RMSE	0,646	0,620	0,689	0,646	0,633	0,747	0,717	0,789	0,715	0,733	0,595	0,573	**0,533**	**0,548**	0,749	0,679
									TID13							
SRCC	**0,896**	0,799	0,777	0,851	0,851	0,780	0,785	0,676	0,870	0,853	0,829	0,813	0,846	**0,907**	0,740	0,798
KRCC	**0,718**	0,631	0,597	0,666	0,658	0,603	0,604	0,514	0,687	0,677	0,643	0,640	0,668	**0,742**	0,565	0,611
PCC	**0,9**	0,859	0,831	0,876	0,870	0,826	0,832	0,773	0,854	0,854	0,864	0,865	0,880	**0,920**	0,808	0,832
RMSE	**0,540**	0,634	0,688	0,595	0,609	0,697	0,686	0,785	0,646	0,759	0,622	0,620	0,587	**0,484**	0,729	0,686
									IVC							
SRCC	0,899	0,926	0,912	**0,929**	0,924	0,914	0,898	0,896	0,912	0,894	0,914	0,914	0,913	0,918	**0,954**	0,916
KRCC	0,721	0,756	0,733	**0,763**	0,755	0,740	0,720	0,715	0,745	0,713	0,739	0,741	0,736	0,747	**0,812**	0,742
PCC	0,912	0,936	0,923	**0,939**	0,932	0,921	0,910	0,902	0,924	0,900	0,915	0,921	0,919	0,918	**0,959**	0,915
RMSE	0,499	0,428	0,468	**0,418**	0,440	0,474	0,502	0,523	0,465	0,530	0,490	0,473	0,479	0,481	**0,341**	0,489
									MDID							
SRCC	0,856	0,852	0,891	0,890	0,839	0,724	0,829	**0,930**	0,871	0,819	0,013	0,804	0,663	0,833	0,870	**0,950**
KRCC	0,670	0,668	0,709	0,712	0,659	0,533	0,636	**0,771**	0,689	0,628	0,018	0,611	0,478	0,644	0,691	**0,805**
PCC	0,870	0,868	0,898	0,899	0,859	0,755	0,841	**0,936**	0,880	0,831	0,133	0,823	0,667	0,844	0,886	**0,952**
RMSE	1,085	1,092	0,968	0,961	1,128	1,444	1,189	**0,771**	1,043	1,225	2,183	1,251	1,640	1,180	1,020	**0,670**

independent results are available for the other databases. So they are compared with the five FR fusion IQA measures which their results are known. Overall results confirmed the outstanding performances of FR-MGGP3 and FR-MGGP4. The fusion measure introduced in [63] also obtained good results, but was worse than FR-MGGP3 (overall weighted).

With regards to the recently published database MDID, a comprehensive evaluation of 32 state-of-the-art FR-IQA metrics was presented in [76], where the authors demonstrated that there is still a lot of space for the improvement of FR-IQA algorithms because only the single metric HaarPSI [77] was able to produce SRCC values higher than 0.9. Nevertheless, we managed with our new approach to get a higher value than this method ($FR-MGGP = 0.9505$ and $HaarPSI = 0.9028$).

5.3 *Cross Database Test*

Furthermore, to evaluate the generalization of the proposed FR-MGGP method, we trained the system based on one database and tested it on the other five databases. The results are shown in Table VI. It can be seen that the SRCC values are over 0.901 for most cases except when the system is trained on IVC or tested on TID2013 and MDID. This exception can be explained as the tested set involves a few distortion types, which are not addressed in the training set, for example, TID13 and MDID are much larger than the other datasets in terms of the number of images, the number of quality distortion types, and the multiply distorted images. The results show that the proposed metric performs well in terms of SRCC correlation and the performance does not depend on the database. On the other hand, IVC is the smallest one, which can not completely cover all distortion types.

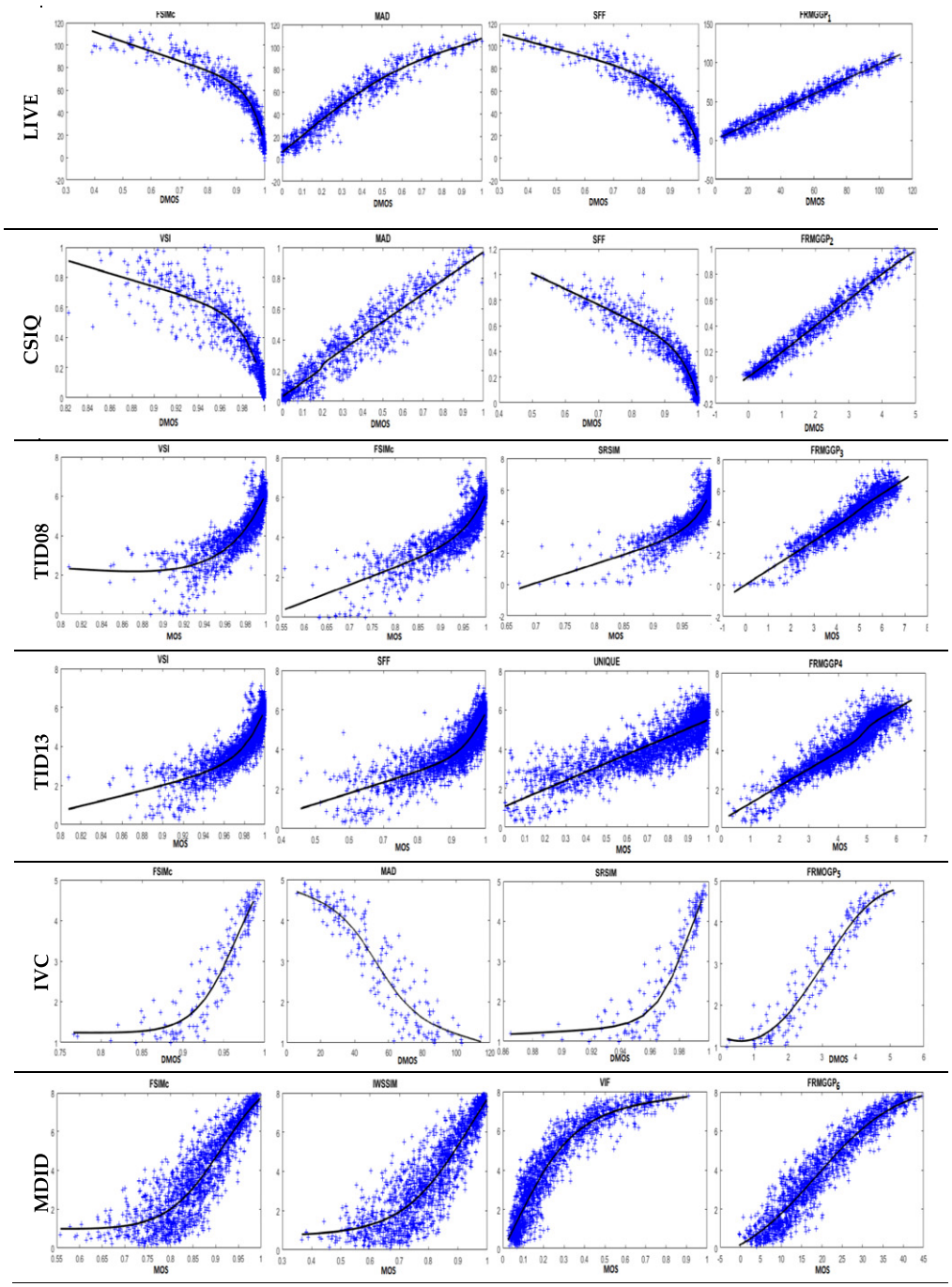

Figure 6. Correlations of subjective and objective assessment obtained and the best three state-of-the-art IQA measures for each dataset (Each data point represent one test image).

6. CONCLUSION

In this paper, we have developed an evolutionary learning methodology, to automatically generate predictive models for full-reference image quality assessment, using Multi Gene Genetic Programming method. The predicted score of image is obtained by the weighted sum of objective scores of a small number of image quality metrics (IQM). By learning from benchmark image datasets, the MGGP can determine the appropriate image quality metrics, from 21 utilized metrics, whose objective scores employed as predictors, in the symbolic regression, by optimizing simultaneously two competing objectives of model 'goodness of fit' to data and model 'complexity'.

To evaluate the performance of the proposed method, several experiments were carried, with different aspects, using six largest image benchmarks (LIVE, CSIQ, TID2008, TID2013, IVC and MDID) and four performance indices (SRCC, PCC, KROCC, RMSE). For each benchmark, we have obtained one predicted model namely Full Reference metric based Multi Gene Genetic Programming (FR-MGGPs).

Table V. Comparison of obtained measures with other fusion IQA measures based on SRCC values.

IQA measure	TID13	TID08	CSIQ	LIVE	Overall direct	Overall weighted weighted
CNNM [38]	**0,93**	–	–	–	–	–
NR-SVR [39]	*0,859*	–	*0,920*	*0,967*	–	–
NR-ANN [40]	*0.772*	–	–	*0.982*	–	–
VCGS [51]	0.892	0.897	0.944	**0.976**	0.939	0.928
IGN [60]	–	0,890	0,940	0,958	0,929	0,919
LAF [61]	–	0,810	0,963	0,957	0,910	0,885
ESIM [62]	0,880	0,902	0,962	0,942	0,935	0,927
LCSIM [63]	0,830	0,910	**0,973**	0,972	**0,952**	0,939
RLCM [65]	0,887	–	0,945	0,963	–	–
FR-MGGP1	0,829	0,891	0,966	**0,975**	0,944	0,930
FR-MGGP2	0,813	0,902	**0,975**	0,925	0,934	0,927
FR-MGGP3	0,834	**0,926**	0,959	0,960	**0,948**	**0,943**
FR-MGGP4	**0,907**	**0,914**	0,952	0,959	0,942	0,935

Table VI. Cross-database SRCC performance of FR-MGGP.

Test /Train	LIVE	CSIQ	TID2008	TID2013	IVC	MDID
LIVE	–	0.96	0.90	0.80	0.91	0.81
CSIQ	0.92	–	0,90	0.78	0.91	0.79
TID2008	0.96	0.95	–	0.82	0.91	0.77
TID2013	0.96	0.96	0.90	–	0.91	0.86
IVC	0.86	0.90	0.73	0.66	–	0.67
MDID	0.96	0.94	0.83	0.76	0.92	–

The presented results confirm the superior performance of obtained FR-MGGPs in comparison with state-of-the-art IQA measures, including other recently published fusion approaches. In future works, we would like to extend this approach to video and 3D quality assessment.

REFERENCES

[1] W. Hou, S. Woods, E. Jarosz, W. Goode, and A. Weidemann, "Optical turbulence on underwater image degradation in natural environments," Appl. Opt. **51**, 2678 (2012).

[2] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," *Computer Vision--ECCV 2018*, Lecture Notes in Computer Science (Springer International Publishing, Cham, 2018), Vol. 11210, pp. 187–202.

[3] Li Xinghua, Shen Huanfeng, Zhang Liangpei, Zhang Hongyan, and Yuan Qiangqiang, "Gang yang recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning," IEEE Trans. Geosci. Remote Sens. **52**, 7086–7098 (2014).

[4] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," Sci. China Inf. Sci. **63**, 211301 (2020).

[5] D. M. Chandler, "Seven challenges in image quality assessment: past, present, and future research," ISRN Signal Process. **2013**, 1–53 (2013).

[6] T.-J. Liu, Y.-C. Lin, W. Lin, and C.-C. J. Kuo, "Visual quality assessment: recent developments, coding applications and future trends," APSIPA Trans. Signal Inf. Process. **2** (2013).

[7] B. S. Phadikar, G. K. Maity, and A. Phadikar, "Full reference image quality assessment: A survey," *Industry Interactive Innovations in Science, Engineering and Technology* (Springer, Singapore, 2018), Vol. 11, pp. 197–208.

[8] A. Rehman, "Zhou Wang reduced-reference image quality assessment by structural similarity estimation," IEEE Trans. Image Process. **21**, 3378–3389 (2012).

[9] I. T. Ahmed, C. S. Der, and B. T. Hammad, "A survey of recent approaches on no-reference image quality assessment with multiscale geometric analysis transforms," Int. J. Sci. Eng. Res. **7**, 10 (2016).

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

[11] Z. Lin, Z. Lei, M. Xuanqin, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," IEEE Trans. Image Process. **20**, 2378–2386 (2011).

[12] L. Anmin, L. Weisi, and M. Narwaria, "Image quality assessment based on gradient similarity," IEEE Trans. Image Process. **21**, 1500–1512 (2012).

[13] 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. **14**, 2117–2128 (2005).

[14] W. Zhou, "Qiang Li information content weighting for perceptual image quality assessment," IEEE Trans. Image Process. **20**, 1185–1198 (2011).

[15] D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," J. Electron. Imaging **19**, 011006 (2010).

[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Proc. The Thrity-Seventh Asilomar Conf. on Signals, Systems & Computers* (IEEE, Piscataway, NJ, 2003), pp. 1398–1402.

[17] N. Damera-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. **9**, 636–650 (2000).

[18] 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. **15**, 3440–3451 (2006).

[19] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using riesz transforms," *Proc. 2010 IEEE Int'l. Conf. Image Processing* (IEEE, Piscataway, NJ, 2010), pp. 321–324.

[20] L. Zhang and H. Li, "SR-SIM: A fast and high performance iqa index based on spectral residual," *Proc. 2012 19th IEEE Int'l. Conf. on Image Processing* (IEEE, Piscataway, NJ, 2012), pp. 1473–1476.

[21] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Trans. Image Process. **15**, 430–444 (2006).

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

[23] H.-W. Chang, H. Yang, Y. Gan, and M.-H. Wang, "Sparse feature fidelity for perceptual image quality assessment," IEEE Trans. Image Process. **22**, 4007–4018 (2013).

[24] 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. **13**, 600–612 (2004).

[25] T. Hegazy and G. Al Regib, "COHERENSi: A new full-reference IQA index using error spectrum chaos," *Proc. 2014 IEEE Global Conf. on Signal and Information Processing (GlobalSIP)* (IEEE, Piscataway, NJ, 2014), pp. 965–969.

[26] D. Temel, M. Prabhushankar, and G. AlRegib, "UNIQUE: Unsupervised image quality estimation," IEEE Signal Process. Lett. **23**, 1414–1418 (2016).

[27] M. Prabhushankar, D. Temel, and G. AlRegib, "MS-UNIQUE: Multi-model and sharpness-weighted unsupervised image quality estimation," Electron. Imaging **2017**, 30–35 (2017).

[28] D. Temel and G. AlRegib, "PerSIM: Multi-resolution image quality assessment in the perceptually uniform color domain," *Proc. 2015 IEEE Int'l. Conf. on Image Processing (ICIP)* (IEEE, Piscataway, NJ, 2015), pp. 1682–1686.

[29] G. Yang, D. Li, F. Lu, Y. Liao, and W. Yang, "RVSIM: A feature similarity method for full-reference image quality assessment," EURASIP J. Image and Video Process. **2018**, 1–15 (2018).

[30] K. Okarma, "Current trends and advances in image quality assessment," ElAEE **25**, 77–84 (2019).

[31] S. Athar and Z. Wang, "A comprehensiv e performance evaluation of image quality assessment algorithms," IEEE Access **7**, 140030–140070 (2019).

[32] A. Saha and Q. M. J. Wu, "Full-reference image quality assessment by combining global and local distortion measures," Signal Process. **128**, 186–197 (2016).

[33] S. Li, F. Zhang, L. Ma, and K.N. Ngan, "Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments," IEEE Trans. Multimedia **13**, 935–949 (2011).

[34] W. Cai, C. Fan, L. Zou, Y. Liu, Y. Ma, and M. Wu, "Blind image quality assessment based on classification guidance and feature aggregation," Electronics **9**, 1811 (2020).

[35] K. Okarma, "Extended hybrid image similarity – combined full-reference image quality metric linearly correlated with subjective scores," Electron. Electr. Eng. **19**, 539–546 (2013).

[36] P. Lech and K. Okarma, "Prediction of the optical character recognition accuracy based on the combined asessment of image binarization results," ElAEE **21**, 62–65 (2015).

[37] M. Oszust, "A regression-based family of measures for full-reference image quality assessment," Meas. Sci. Rev. **16**, 316–325 (2016).

[38] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," Proc. SPIE **9394**, 93940K (2015).

[39] D. Borhen Eddine, H. Fella, and B. Azeddine, "Blind image quality assessment method based on a particle swarm optimization support vector regression fusion scheme," J. Electron. Imaging **25**, 061623.

[40] O. Ieremeiev, V. Lukin, N. Ponomarenko, and K. Egiazarian, "Combined no-reference iqa metric and its performance analysis," *Proc. IS&T Electronic Imaging: Image Processing, Algorithms and Systems XVII* (IS&T, Springfield, VA, 2019), pp. 260-1–260-6.

[41] A. H. Gandomi and A.H. Alavi, "A new multi-gene genetic programming approach to nonlinear system modeling. Part I: Materials and structural engineering problems," Neural Comput. Appl. **21**, 171–187 (2012).

[42] J. Koza, "Genetic programming: on the programming of computers by means of natural selection," *Complex Adaptive Systems* (MIT Press, Cambridge, Mass, 1992), ISBN 978-0-262-11170-6.

[43] C. Hii, D. P. Searson, and M. J. Willis, "Evolving toxicity models using multigenesymbolic regression and multiple objectives," Int. J. Mach. Learn. Comput. **1**, 30–35 (2011).

[44] A. H. Gandomi, S. Sajedi, B. Kiani, and Q. Huang, "Genetic programming for experimental big data mining: A case study on concrete creep formulation," Autom. Constr. **70**, 89–97 (2016).

[45] A. Danandeh Mehr and V. Nourani, "A pareto-optimal moving average-multigene genetic programming model for rainfall-runoff modelling," Environ. Modelling Softw. **92**, 239–251 (2017).

[46] J. M. Valencia-Ramirez, J. A. Raya, J. R. Cedeno, R. R. Suarez, H. J. Escalante, and M. Graff, "Comparison between genetic programming and full model selection on classification problems," *Proc. 2014 IEEE Int. Autumn Meeting on Power, Electronics and Computing (ROPEC)* (IEEE, Piscataway, NJ, 2014), pp. 1–6.

[47] M. Shafaei and O. Kisi, "Predicting river daily flow using wavelet-artificial neural networks based on regression analyses in comparison with artificial neural networks and support vector machine models," Neural Comput. Appl. **28**, 15–28 (2017).

[48] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Trans. Image Process. **15**, 430–444 (2006).

[49] P. Soo-Chang, "Li-heng chen image quality assessment using human visual DOG model fused with random forest," IEEE Trans. Image Process. **24**, 3282–3292 (2015).

[50] H. Al-Bandawi and G. Deng, "Blind image quality assessment based on Benford's law," IET Image Process. **12**, 1983–1993 (2018).

[51] C. Shi and Y. Lin, "Full reference image quality assessment based on visual salience with color appearance and gradient similarity," IEEE Access **8**, 97310–97320 (2020).

[52] M. Liu and X. Yang, "A new image quality approach based on decision fusion," *Fourth Int'l. Conf. on Fuzzy Systems and Knowledge Discovery* (IEEE, Piscataway, NJ, 2008), Vol. 4, pp. 10–14.

[53] A. Lahouhou, E. Viennet, and A. Beghdadi, "Selecting low-level features for image quality assessment by statistical methods," J. Comput. Inf. Technol. **18**, 183 (2010).

[54] K. Okarma, "Combined full-reference image quality metric linearly correlated with subjective assessment," *10th Int'l. Conf. on Artificial Intelligence and Soft Computing, ICAISC 2010*, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2010), Vol. 6113, pp. 539–546.

[55] K. Okarma, "Combined image similarity index," Opt. Rev. **19**, 349–354 (2012).

[56] K. Okarma, "Hybrid feature similarity approach to full-reference image quality assessment," *Proc. 2012 Int'l. Conf. on Computer Vision and Graphics*, Lecture Notes in Computer Science (Springer, Berlin, Heidelberg, 2012), Vol. 7594, pp. 212–219.

[57] K. Okarma, "Quality assessment of images with multiple distortions using combined metrics," ElAEE **20**, 128–131 (2014).

[58] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," IEEE Trans. Image Process. **22**, 1793–1807 (2013).

[59] P. Peng and Z.-N. Li, "A mixture of experts approach to multi-strategy image quality assessment," *Image Analysis and Recognition* (Springer, Berlin, Heidelberg, 2012), Vol. 7324, pp. 123–130.

[60] Wu Jinjian, Lin Weisi, and Shi Guangming, "Anmin liu perceptual quality metric with internal generative mechanism," IEEE Trans. Image Process. **22**, 43–54 (2013).

[61] A. Barri, A. Dooms, B. Jansen, and P. Schelkens, "A locally adaptive system for the fusion of objective quality measures," IEEE Trans. Image Process. **23**, 2446–2458 (2014).

[62] M. Oszust, "Decision fusion for image quality assessment using an optimization approach," IEEE Signal Process. Lett. **23**, 65–69 (2016).

[63] M. Oszust, "Full-reference image quality assessment with linear combination of genetically selected quality measures," PLoS One **11**, e0158333 (2016).

[64] M. Abdoli, F. Nasiri, P. Brault, and M. Ghanbari, "Quality assessment tool for performance measurement of image contrast enhancement methods," IET Image Process. **13**, 833–842 (2019).

[65] O. Ieremeiev, V. Lukin, N. Ponomarenko, and K. Egiazarian, "Robust linearized combined metrics of image visual quality," *Proc. IS&T Electronic Imaging: Image Processing, Algorithms and Systems XVI* (IS&T, Springfield, VA, 2018), pp. 260-1–260-6.

[66] R. Poli, W. B. Langdon, N. F. McPhee, and J. R. Koza, *A Field Guide to Genetic Programming* (Lulu Press, Morrisville, NC, 2008), ISBN 978-1-4092-0073-4.

[67] D. P. Searson, "GPTIPS 2: An open-source software platform for symbolic data mining," *Handbook of Genetic Programming Applications* (Springer, SpringerNature Switzerland AG, 2015), pp. 551–573.

[68] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "TID2008 – A database for evaluation of full-reference visual quality assessment metrics," Adv. Mod. Radioelectronics **10**, 30–45 (2009).

[69] A. Ninassi, P. Le Callet, and F. Autrusseau, "Pseudo no reference image quality metric using perceptual data hiding," Proc. SPIE **6057**, 60570G (2006).

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

[71] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," Pattern Recognit. **61**, 153–168 (2017).

[72] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Trans. Evolutionary Comput. **6**, 182–197 (2002).

[73] D. Searson, M. Willis, and G. Montague, "Co-evolution of non-linear PLS model components," J. Chemometr. **21**, 592–603 (2007).

[74] *VQEG: Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*; FR-TV Phase II http://www.vqeg.org/.

[75] K. Okarma, "On the optimisation of nonlinear mapping functions towards high correlation of full-reference image quality metrics and their combinations with subjective evaluations," *Computer Applications in Electrical Engineering*, 15.

[76] D. Varga, Empirical evaluation of full-reference image quality metrics on MDID database. arXiv: 1910.01050 [cs, eess] (2019).

[77] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A haar wavelet-based perceptual similarity index for image quality assessment," Signal Process. Image Commun. **61**, 33–43 (2018).