A Machine Learning Regression scheme to design a FR-Image Quality Assessment Algorithm

Christophe Charrier, Olivier Lezoray, Gilles Lebrun

A Machine Learning Regression scheme to design a FR-Image Quality Assessment Algorithm

Christophe Charrier Olivier Lézoray Gilles Lebrun
Université de Caen-Basse Normandie, UMR 6072 GREYC, F-14032 Caen, France

Abstract

A crucial step in image compression is the evaluation of its performance, and more precisely available ways to measure the quality of compressed images. In this paper, a machine learning expert, providing a quality score is proposed. This quality measure is based on a learned classification process in order to respect that of human observers. The proposed method namely Machine Learning-based Image Quality Measurment (MLIQM) first classifies the quality using multi Support Vector Machine (SVM) classification according to the quality scale recommended by the ITU. This quality scale contains 5 ranks ordered from 1 (the worst quality) to 5 (the best quality). To evaluate the quality of images, a feature vector containing visual attributes describing images content is constructed. Then, a classification process is performed to provide the final quality class of the considered image. Finally, once a quality class is associated to the considered image, a specific SVM regression is performed to score its quality. Obtained results are compared to the one obtained applying classical Full-Reference Image Quality Assessment (FR-IQA) algorithms to judge the efficiency of the proposed method.

Introduction

The way to evaluate the performance of any compression scheme is a crucial step, and more precisely available ways to measure the quality of compressed images. There is a very rich literature on image quality criteria, generally dedicated to specific applications (optics, detector, compression, restoration, ...).

When the presence of the original image is available, the usually applied scheme to design an FR-IQA (Full Reference Image Quality Assessment) algorithm consists in performing 1) a color space transformation to obtain decorrelated color coordinates and 2) a decomposition of these new coordinates towards perceptual channels. An error is then estimated for each one of these channels. A final quality score is obtained by pooling these errors in both spatial and frequency domain. The most common way to perform this pooling is to use the Minkowski error metric. Some studies [1] have shown that this summation does not perform well. The same final value can be computed for two different degraded images even if the visual quality of the two images is drastically different [2]. This is due to the fact that the implicit assumption of this metric is based on the independence of all signal samples. It is yet commonly assumed that this is not true when one uses perceptual channels. This explains why the Minkowski metric might fail to generate a good final score. The use of such a metric is not necessarily the best way to score the quality of a test image. Actually, in the recommendations given by the ITU [3], the human observers have to choose a quality class from an integer scale from 0 to 100. Those notes characterize the quality of the reconstructed images in semantic terms {`excellent,very good, good, bad, very bad`}. That way, the human observers make then neither more nor less one classification, and the given score could be interpreted as a confidence of the observer in its judgment. Since it is not natural for human beings to score the quality of an image, they prefer to give a semantic description of what they are watching. This semantic description is usually feeling description: "it is beautiful","it is bad", and so on.

Previous works tried to apply a machine learning-based approach, mainly based on standard back propagation neural network to predict the quality score of a test image [4, 5, 6]. e.g., in [4], Bouzerdoum *et al.* propose a FR-IQA algorithm based on a neural network approach. The chosen neural network is a standard back propagation neural network. Its input layer consists of as many neurons are parameters in the input vector. The network has two hidden layers of six neurons each, and one output neuron. The characteristic vector to be input into the neural network is chosen to be composed of several elements based on the Wang *et al.*'s features [7]. These include the image mean and the image standard-deviation of both the reference and the test image, the covariance and the MSE between the reference and the test image. More recently, NARWARIA *et al.* [8] propose an IQA algorithm based on support vector regression. The input features are the singular vectors out of singular value decomposition. Yet, the proposed approaches do not account for the intrinsic classification process of the quality judgment of human beings.

All IQA algorithms perform well (in sens of high correlation with human ratings) for very poor or very good quality images but in between there are big differences between algorithms. Firstly, one can assume that for medium quality images, predicted scores do not reflect very well human ratings and predicted scores are not as good as they would be. In a second interpretation, on can assume that an IQA algorithm using the same sensitivity across the quality continuum would not be able to refine its prediction for medium quality images. It should be better to develop a quality metric that can modulate its sensitivity with respect to image quality. One way to do so, is to classifiy image quality with respect to quality classes and from the obtained classification, to modelize the distribution of each class in order to design a quality function whose its sensitivity will differ from others.

In this paper, the modelization of the judgment of human beings by a machine learning expert to design a FR-IQA algorithm is proposed. Fig 1 displays the general scheme of the Machine Learning-based Image Quality Measure (MLIQM) used to predict the quality of a test image. After computing a feature vector including several local quality features, a SVM multiclass

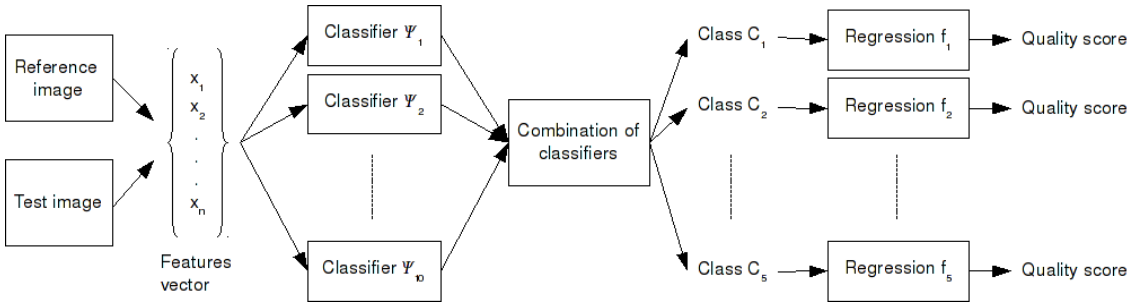

Figure 1. *General scheme of the proposed method to obtain the final quality score of a test image.*

classification process is performed to provide the final quality class $C_i, \forall i \in [1,\ldots,5]$ of the test image. Those five correspond to the quality classes as advocated by the recommendation ITU-R BT.500-11 [3]. Finally, from this classification a SVM regression process is applied to score the quality of the test image. This way, the proposed IQA method yiels a sensitivity adaptation to quality image in order to counterbalance medium prediction of usually used IQA techniques.

The Selected Full-Reference Features

To design the input features vector of the classification process, only derived full-reference characteristics are employed. A scalar is then generated for each trial feature. The whole set of computed scalars forms the feature vector associated to an image. This vector will be classified to designate the associated class of quality.

In [9], SHEIK *et al.* compared 10 recent IQA algorithms and determined which had particularly high levels of performance. They concluded that no IQA algorithm has been shown to definitively outperform all others for all possible degradations, although owing to the inclusion of both scene models and perceptual models, the MS-SSIM index outperform many with statistical significance. Thus, factors embedded in the MS-SSIM index will serve a spatial criteria as described in section .

Wang *et al.* [10] have shown that natural images are highly structured, in the sense that their pixels exhibit strong dependencies, and these dependencies carry important information about the visual scene. Structural information is located on visible edges of the image. These edges correspond to spatial frequency that infers in a positive or negative way with the other frequencies to produce spatial structures of the image. Thus, Spatial-frequency factors are computed to take into account structural information.

Spatial criteria (13 features)

The first selected criteria in our study concern the factors integrated in the MS-SSIM metric proposed by WANG and BOVIK [11]. These criteria allow us to measure 1) the luminance distortion, 2) the contrast distortion and 3) the structure comparison. Those criteria are computed considering only the Achromatic information. The authors proposed to represent an image as a vector in an image space. In that case, any image distortion can be interpreted as adding a distortion vector to the reference image vector. In this space, the two vectors that represent luminance and contrast changes span a plane that is adapted to the reference image vector. The image distortion corresponding to a rotation of such a plane by an angle can be interpreted as the structural change.

To obtain a multi-scale index, a low-pass filter is applied to the reference (I) and the distorted images (J). Next a downsampling of the filtered images by a factor of 2 is performed. Considering the initial design of the MS-SSIM indice that consists in computing the factors $c(\cdot)$ and $s(\cdot)$ at five different scales, and the luminance $l(\cdot)$ at the coarser level, 11 distortion maps are generated. Each of them is then pooled in a single scalar distortion score, providing 11 factors that are integrated in the feature vector.

Since previous criteria only concern the achromatic axis, two local descriptors dedicated to chromatic information are computed [12]. Those descriptors are not punctually defined in the image but with respect to the mean value of the local neighborhood of the pixel. The two used features are 1) a local chrominance distortion feature measuring the sensitivity of an observer to color degradation within a uniform area and 2) a local colorimetric dispersion feature that measures the spatio-colorimetric dispersion in each one of the two color images. The calculation of these two desriptors is performed in an antagonist Luminance-Chrominance color space, namely the CIE Lab colorspace [13]. These two criteria are also included in the feature vector.

Spatial-frequency criteria (12 features)

The aim of such features is to model, as well as possible, HVS-characteristics such as contrast masking effects, the luminance variation sensitivity, and so on. Many models exist to estimate the visibility of errors by simulating the relevant functional properties of the HVS. All these models perform decomposition of the input signal into a set of channels, each of them being selectively sensitive to a restricted range of spatial frequencies and orientations, in order to account for the spatial-frequency sensitivity of the HVS. Decompositions mainly differ from number radial bands, orientations and bandwith [14, 15, 16].

Among all existing decompositions, the steerable pyramid transform [17] is used in this paper to quantify contrast masking effects. The decomposition consists in many spatial frequency levels, which are further divided into a set of orientation bands. The basis function are directional derivative operators. In this paper, three levels with four orientation bands with bandwidths of 45 degrees $0, 45, 90, 135$ plus one isotropic lowpass filter are used. The coefficients induced by the decomposition are next squared to obtain local energy measures. As mentioned in [18], those coefficients are normalized to take into account the dynamic limited range of the mechanisms in the Human Visual Sys-

tem.

Let $a(x,y,f,\theta)$ be an original coefficient issued from the decomposition process located at the position (x,y) in frequency band y and orientation band θ. The associated squared and normalized sensor output $r(x,y,f,\theta)$ is defined as

$$r(x,y,f,\theta) = k\frac{(a(x,y,f,\theta))^2}{\sum_{\phi\in 0,45,90,135}(a(x,y,f,\phi))^2+\sigma^2}, \quad (1)$$

This procedure leads to normalized sensors having a limited dynamic range. Each sensor is able to discriminate contrast differences over narrow range of contrasts. This is why the use of multiple contrast bands (with different k's and σ's) is required to discriminate contrast changes over the full rang of contrast.

The final stage computes the simple squared error norm between the sensor outputs from the reference image $r_0(x,y,f,\theta)$ and the degraded images $r_1(x,y,f,\theta)$ for each frequency band t and orientation band θ:

$$\Delta r(f,\theta) = \|\sum_{x,y} r_0(x,y,f,\theta) - r_1(x,y,f,\theta)\|^2 \quad (2)$$

From this step, 12 scores are available and integrated within the feature vector.

SVM classification and regression

From all existing classification schemes, a Support Vector Machine (SVM)-based technique has been selected due to high classification rates obtained in previous works [19], and to their high generalization abilities. The SVMs were developed by VAPNIK *et al.* [20] and are based on the structural risk minimization principle from statistical learning theory. SVMs express predictions in terms of a linear combination of kernel functions centered on a subset of the training data, known as support vectors (SV).

Given the training data $\mathscr{S} = \{(x_i,y_i)\}_{i=\{1,\ldots,m\}}, x_i \in \mathbb{R}^n$, $y_i \in \{-1,+1\}$, SVM maps the input vector x into a high-dimensional feature space $\mathbf{H}$ through some non linear mapping functions $\phi : \mathbb{R}^n \to \mathbf{H}$, and builds an optimal separating hyperplane in that space. The mapping operation $\phi(\cdot)$ is performed by a kernel function $K(\cdot,\cdot)$ which defines an inner product in $\mathbf{H}$. The separating hyperplane given by a SVM is: $w\cdot\phi(x)+b=0$. The optimal hyperplane is characterized by the maximal distance to the closest training data. The margin is inversely proportional to the norm of w. Thus computing this hyperplane is equivalent to minimize the following optimization problem:

$$\mathscr{V}(w,b,\xi) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{m}\xi_i\right) \quad (3)$$

where the constraint $\forall_{i=1}^m : y_i[w\cdot\phi(x_i)+b] \geq 1-\xi_i, \xi_i \geq 0$ requires that all training examples are correctly classified up to some slack ξ and C is a parameter allowing trading-off between training errors and model complexity. This optimization is a convex quadratic programming problem. Its whole dual [20] is to maximize the following optimization problem:

$$\mathscr{W}(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \quad (4)$$

subject to $\forall_{i=1}^m : 0 \leq \alpha_i \leq C$, $\sum_{i=1}^m y_i\alpha_i = 0$. The optimal solution α^* specifies the coefficients for the optimal hyperplane $w^* = \sum_{i=1}^m \alpha_i^* y_i \phi(x_i)$ and defines the subset SV of all support vector (SV). An example x_i of the training set is a SV if $\alpha_i^* \geq 0$ in the optimal solution. The support vectors subset gives the binary decision function h:

$$h(x) = \text{sign}(f(x)) \text{ with } f(x) = \sum_{i\in SV}\alpha_i^* y_i K(x_i,x) + b^* \quad (5)$$

where the threshold b^* is computed via the unbounded support vectors [20] (*i.e.*, $0 < \alpha_i^* < C$). An efficient algorithm SMO (Sequential Minimal Optimization) [21] and many refinements [22, 23] were proposed to solve dual problem.

SVM model selection

Kernel function choice is critical for the design of a machine learning expert. Radial Basic Function (RBF) kernel function is commonly used with SVM. The main reason is that RBF functions work like a similarity measure between two examples.

In this paper, the common One-Versus-One (OO) decomposition scheme is used to create 10 binary classifiers. Let $t_{i,j}, \forall i \in [1,5], j \in [2,5], j > i$ be a binary problem with $t_{i,j} \in \{+1,-1\}$. Number 5 represents the final quality classes according to the ones recommended by the ITU. Let $h_i(\cdot)$ (Eq. 5) be the SVM decision function obtained by training it on the i^{th} binary problem. The binary problem transformation is the first part of a combination scheme. A final decision must be taken from all binary decision functions. Since the SVMs are binary classifiers, the resolution of a multi-class problem is achieved through a combination of binary problems in order to define a multi-class decision function D. One interesting way to achieve this combination is the use of the theory of evidence since the confidence one has in classifiers can be take into account for the final assignment decision.

The combination of binary classifiers

Once the multi-class classifier has been decomposed in ten binary classifiers, one needs to take a decision about the final quality class assignment of the input vector. This assignment is done using the theory of evidence framework (also known as the Dempster-Shafer theory or the belief functions theory) [24, 25]. Indeed, each of the binary classifier can be considered as an information source that can be imprecise and uncertain. Combining these different sources using the theory of evidence yields to process uncertain information to take the final assignment decision. Conceptually, the final decision is taken with respect to the confidence we have on the results of each binary classifier. The confidence index can be provided in many different ways: a recognition rate, a likelihood probability, an *a posteriori* probability, and so on. Yet, SVMs do not directly provide such a measure.

In this paper, an *a posteriori* probability is computed from the output of the SVM and will serve as confidence index. Instead of estimating the class-conditional densities $p(f|y)$, a parametric model is used to fit the posteriori $p(y=1|f)$ where f represents the uncalibrated output value of SVMs. PLATT [21] has proposed a method to compute the *a posteriori* probabilities from the obtained SVM parameters. The suggested formula is based on a parametric form of a sigmoid as:

$$p(y=1|f) = \frac{1}{1+\exp(Ef+F)}, \quad (6)$$

where the parameters E and F are fit using maximum likelihood estimation from a training set (f_i, y_i).

Elements of theory of evidence.

Let $\Omega = \{\omega_1, \ldots, \omega_N\}$ be the set of N final classes possible for the quality of an image, called the frame of discernment. In our study, $N = 5$ and Ω corresponds to the five final classes $(\omega_l)_{1 \leq l \leq 5}$ respectively representing the five quality classes {`excellent`,`very good`, `good`, `bad`, `very bad`} Instead of narrowing its measures to the set Ω (as performed by the theory of probability constrained by its additivity axiom), the theory of evidence extends on the power set Ω, labeled as 2^Ω, the set of the 2^N subsets of Ω. Then a mass function m is defined and represents the belief allowed to the different states of the system, at a given moment. This function is also known as the initial mass function $m(\cdot)$ defined from 2^Ω in $[0,1]$ and corroborating:

$$\sum_{A \subseteq \Omega} m(A) = 1 \quad \text{and} \quad m(\emptyset) = 0 \tag{7}$$

where $m(A)$ quantifies the belief that the search class belongs to the subset $A \subseteq \Omega$ (and to none other subset of A). Subsets A such as $m(A) > 0$ are referred to as *focal elements*. A represents either a singleton ω_j or a disjunction of hypothesis. In the case where the set of hypothesis is exhaustive and exclusive, the mass of the empty set is equal to 0. Such assumption means that the solution belongs to the frame of discernment.

In case of imperfect data (e.g., incomplete or uncertain data), fusion is an interesting solution to obtain more relevant information. In that case, the combination can be performed from the mass function in order to provide combined masses synthesizing the knowledge of the different sources.

Two initial mass functions m_1 and m_2 representing respectively the information providing from two independent sources, can be combined according to Dempster's rule [24]:

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - K}, \forall A \in 2^\Omega,\ A \neq \emptyset. \tag{8}$$

K is known as the *conflict factor* and represents the discrepancy between the two sources. It corresponds to the mass of the empty set if the masses are not normalized

$$K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C). \tag{9}$$

One notes that Dempster's combination, also known as orthogonal sum and written as $m = m_1 \oplus m_2$, is commutative and associative.

When performing the Dempster's combination, it is crucial to take into account the value of K, which is the normalization term of the combination: the higher the value, the more incoherent the combination is. When $k = 1$ one reaches a complete opposition and the data fusion is impossible. Several solutions have been developed to deal with this conflict term. For example SMETS [26] proposed to avoid the normalization step, since he considered the conflict can only come from a bad definition of Ω. In that case, K represents the mass associated to one or more new hypothesis that have not been initially taken into account.

After performing the combination, the decision associated to the most "probable" element Ω has to be quantified. Among the existing rules of decision, the most commonly used is the maximum of the pignistic probability. This decision rule, introduced by Smets [27] uses the pignistic transformation that allows to distribute the mass associated to a subset of Ω over each one of its elements:

$$\text{BetP}(\omega_l, m) = \sum_{\omega_l \in A \subseteq \Omega} \frac{m(A)}{|A|}, \forall \omega_l \in \Omega, \forall 1 \leq l \leq 5 \tag{10}$$

$|A|$ is the cardinal of A. The decision is executed from the elements of Ω the highest value of which

Mass function design

One of the main drawbacks of the theory of evidence is the design of mass functions: the quality of the fusion process depends on the quality of the mass function. The design of this mass function is deeply linked to the application.

Among all existing models, the one proposed by DENŒUX [28] has been retained in our study on account of its integration of both the distance to the neighbors and different criteria of neighborhood in its definition. Thus the mass $m(\{\omega_j\})$ is defined as a decreasing function of the distance d between the vector to classify and the barycenter of the class:

$$\begin{cases} m(\omega_l) = \alpha \exp(-\gamma_l d^2) \\ m(\Omega) = 1 - m(\omega_l) \end{cases} \tag{11}$$

where $0 < \alpha < 1$ is the *a posteriori* probability computed from the binary SVM dedicated to the class ω_l. γ_l depends on the class ω_l and is computed by minimization of an error criterion using the SEM (Stochastic Expectation Maximization) algorithm.

The mass functions yield to take into account the associated uncertainty to each one of the classifier. Thus, close classes are brought together in the same focal element, and the final decision is taken only after combining the obtained results from other projections.

To construct such a focal element, the input vector it not associated to only one class from$\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$, but to a subset of classes corresponding at most to Ω. To generate such a subset, the affectation constraint has to be loosened. One way to perform that is to generate an interval computed from the maximum value of the *a posteriori* probabilities to generate the subset A such as:

$$A = \{\omega_l \in \Omega / \max(p_l) - \delta_l \leq p_l \leq \max(p_l)\} \tag{12}$$

where $l \in \{1, \ldots, 5\}$ and δ_l is an ad-hoc constant depending on the used classifier.

In that case, all the classes for which their probabilities are included within this new interval are considered as candidates for classification during the fusion process.

SVM regression scheme

Even if scoring the quality of an image is not natural for human beings, it is quite necessary to obtain scalar quality score. The main reason is due to the fact that total order only exists in the real set $\mathbb{R}$.

SVMs can be applied not only to classification problems but also to the case of regression. Our SVM-based classifier does not directly provide any quality score. In order to provide such

a quality score, we use the support vector regression technique referred to as ν-SVR [29] which a commonly used to solve regression problems. In particular ν-SVR has the advantage of being able to automatically adjust the width of the ε-tube [29].

We first present the ε-SVR and then present ν-SVR as an improvement [29, 20]. Given the training data $\mathscr{S} = \{(x_i, y_i)\}_{i=\{1,\ldots,m\}}, x_i \in \mathbb{R}^n$, $y_i \in \{-1,+1\}$. In ε-SVR, x is first mapped to $z = \Phi(x)$ in feature space, then a linear function $f(x,w) = w^T z + b$ is constructed such that it deviates least from the training set according to a ε-insensitive loss function

$$|y - f(x)|_\varepsilon = \begin{cases} 0 & \text{if } |y - f(x)| < \varepsilon \\ |y - f(x)| - \varepsilon & \text{otherwise} \end{cases}$$

while $\|w\|$ is as small as possible. This is equivalent to minimize

$$\min \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{m}(\xi_i + \xi_i^*)\right)$$

subject to $\forall_{i=1}^m, y_i - f_i \leq \varepsilon + \xi_i^*, f_i - y_i \leq \varepsilon + \xi_i, \xi_i, \xi_i^* \geq 0$ where $f_i = f(x_i, w)$ and C is a user-defined constant. After training, those nonzero ξ_i's and ξ_i^*'s will be exactly equal to the difference between the corresponding y_i and f_i.

A drawback of ε-SVR is that ε can be difficult to tune. ν-SVR alleviated this problem trading off ε against model complexity and training error using parameter $\nu > 0$. Mathematically, the problem becomes

$$\min_{w,\varepsilon,\xi_i,xi_i^*} \frac{1}{2}\|w\|^2 + C\left(\nu\varepsilon + \frac{1}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*)\right) \tag{13}$$

subject to $\forall_{i=1}^m, y_i - f_i \leq \varepsilon + \xi_i^*, f_i - y_i \leq \varepsilon + \xi_i, \xi_i, \xi_i^* \geq 0$ and $\varepsilon \geq 0$. In [30], Schölkopf have shown that ν is an upper bound of the fraction of margin errors and a lower bound of the fraction of SV. Furtermore, he shown that, with probability 1, ν equal the both fractions. Thus, in situations where prior knowledge on these fractions is available, ν much be easier to adjust than ε.

In this paper, the RBF is chosen as kernel for ν-SVR. For each quality class, a ν-SVM is trained in order to estimate function f as defined in Eq. 5 using the quality scores of the training sets. In order to be coherent with the ITU scale, a numerical scale is assigned to each quality class. The range of the five quality scales is [0;5] and each quality scale has a numerical scale of length 1. Thus the quality class "very bad quality" is associated to the scale [0,1], the following one "bad quality" is associated to the scale]1;2], and so on until the final quality class "excellent" that is associated to the scale]4;5]. Thus, no overlap between scores obtained from different classes is possible.

Finally, one obtains five regression functions associated to each quality class applying the One-Versus-All approach. When a distorted image is first classified within a quality class, the associated regression function yields to score the quality of that image using a scalar number.

Experimental setup and performance measure

Experimental setup

The used image databases

To judge the performance of the proposed approach, two different image databases are used: 1) the LIVE database release 2 [31] and 2) the TID2008 database [32]. The LIVE database consists of 5 subsets of 5 types of distortions; 1) JPEG2000 distortions (227 images), 2) JPEG distortions (233 images), 3) White noise distortions (174 images), 4) Gaussian blur distortions (174 images), and 5) Fast-fading Rayleigh channel distortions (which are simulated with JPEG2000 compression followed by channel bit-errors) (174 images). The subjective ratings (that will serve as groundtruth) in its Differential Mean Opinion Score (DMOS) form are also available.

The TID2008 database contains 25 reference images and 1600 distorted images using 16 distortion types, as described in Table 1. The MOS value of each image is provided too.

The training and test sets design To apply the MLIQM classification process, two distinct sets have been generated from the trail databases: the training sets and the test sets. Since five quality classes are used, ten OO-SVM classifiers are designed.

One training set (TrainC1) is generated from LIVE database. This is composed of the degraded versions of 12 images of the LIVE image database, for all kind of degradation. The LIVE test set (TestC1) is composed of the degraded versions of the 13 remaining images.

To complete ν-SVM regression, five training sets (TrainR1, TrainR2, $\cdots$, TrainR5) are generated for each quality class, following the same previous design process. This will result in five regresion functions design, *i.e.*, one per quality class.

The parameters of both the SVM classification scheme and the ν-SVM regression scheme are determined using a 10-fold cross-validation technique on the training sets. In addition, a bootstrap process with 999 replicates is used to quantify the performance of MLIQM.

As training is only applied on LIVE subsets (TrainC1, TrainR1, TrainR2, $\cdots$, TrainR5), the entire TID2008 image database will serve as test set as well as the subset TestC1.

Performance measures To measure the performance of the proposed approach, a comparison with usual state-of-the-art FR IQA algorithms is performed. These FR-IQA techniques are MS-SSIM [7], VSNR [33], VIF [34] and PSNR. All these methods are computed using the luminance component of the images.

To provide quantitative performance evaluation, three measures of correlation have been used: 1) Pearson, 2) Kendall and 3) Spearman measures. To perform the Pearson correlation measures (CC), a logistic function (as adopted in the video quality experts group (VQEG) Phase I FR-TV test [35]) was used to provide a non-linear mapping between the predicted values and subjective scores. This function is a three-parameter logistic function

$$r(x) = \frac{b_1}{1 + \exp\left(-b_2(x - b_3)\right)} \tag{14}$$

This nonlinearity is applied to the FR-IQA algorithm score, which give a better fit for all data. Kendall (KROCC) and Spearman (SROCC) rank order correlation measures were computed between the DMOS values and the predicted scores obtained using any trial FR-IQA algorithms. Those measures can be interpreted as prediction accuracy measures (Pearson and Kendall coefficients) and prediction monotonicity measure (Spearman coefficient).

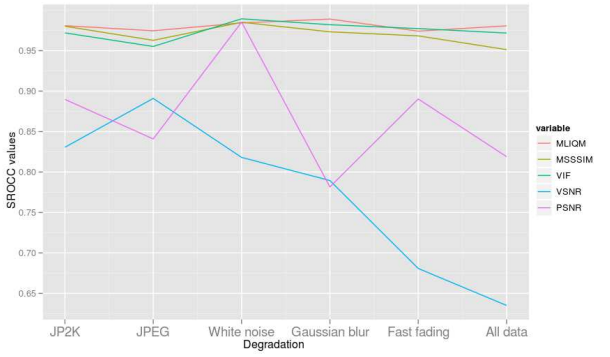

Figure 2. *Obtained correlation coefficients between the predicted DMOS values and the subjective DMOS scores considering LIVE database test set.*

Results

All three correlation coefficients (LCC, KROCC, SROCC) have been computed between the predicted values and the subjective DMOS scores considering the test set TestC1, the entire LIVE database and the entire TID2008 database. Since similar results have been obtained for the three correlation coefficients, only SROCC is reported.

Figure 2 presents SROCC values obtained between the predicted values and the subjective DMOS scores considering both the test set TestC1 and the entire LIVE database for all the five trial FR-IQA methods. Concerning the MLQIM algorithm, the displayed results are median values of SROCC. From the correlation evaluation results, we see that the performance of the MLIQM is significantly better than for the four tested FR-IQA algorithms when whole LIVE database is considered. For most subsets of LIVE, the use of MLIQM provides consistent improvement in the performance of IQA algorithms for different correlation coefficients. Even if improvements are not all significant (which is not really surprising since several trial IQA measures achieve high performance on LIVE), this consistency of improvement can be interpreted as an indicator of the validity of the proposed approach. A second interpretation concerns the selected features. As they are of prime importance to reach high quality results for machine learning classification and regression, this improvement tends to demonstrate that the used features are relevant to design SVM classification and regression-based NR-IQA algorithm. Even if MLIMQ seems to be less performant for fast fading degradation (that uses JP2K), the difference of correlation coefficients with the best IQA method is not significantly different.

These high obtained correlation coefficient values were expected since the training sets used to train the SVM classifier and the SVM regression scheme where generated from LIVE database.

Figure 3 illustrates some obtained results when the trial FR-IQA algorithms are performed on both an original image extracted from LIVE and some of its degraded versions.

Figure 4 displays the performance of the trail IQA algorithms with the TID2008 image database. No new training phase has been performed. This means that shown results are obtained from the MLIQM technique trained on TrainC1 and (TrainR1,$\cdots$, TrainR5) sets for, respectively, the SVM classification step and the SVM regression step. The proposed approach yields to obtain high SROCC values for most subsets of TID database. Except for degradation #5, #7, #12, #15, #16 and #17, MLIQM provides improvement of performance. In addition, when all subsets are considered, the proposed scheme significantly outperforms the trial NR-IQA algorithms, namely MS-SSIM, VSNR, VIF and PSNR. Degradation #5 and #7 respectively deals with high frequency noise and quantization noise. Considering the first kind of artefact, the difference of correlation between the best IQA algorithm (MS-SSIM) and the MLIQM approach is not statistically significant. This is not true if the second degradation is highlighted. This degradation can be interpreted as a loss of color, which induces artificial structural information (edges) for strong quantization. In that case, structural dissimilarities are high and are perfectly captured using MS-SSIM index. The used entry features for MLIQM contain many other features that could blur the information provided by dedicated structural features. Yet, the correlation difference between the two approach (MS-SSIM and MLIQM) is small.

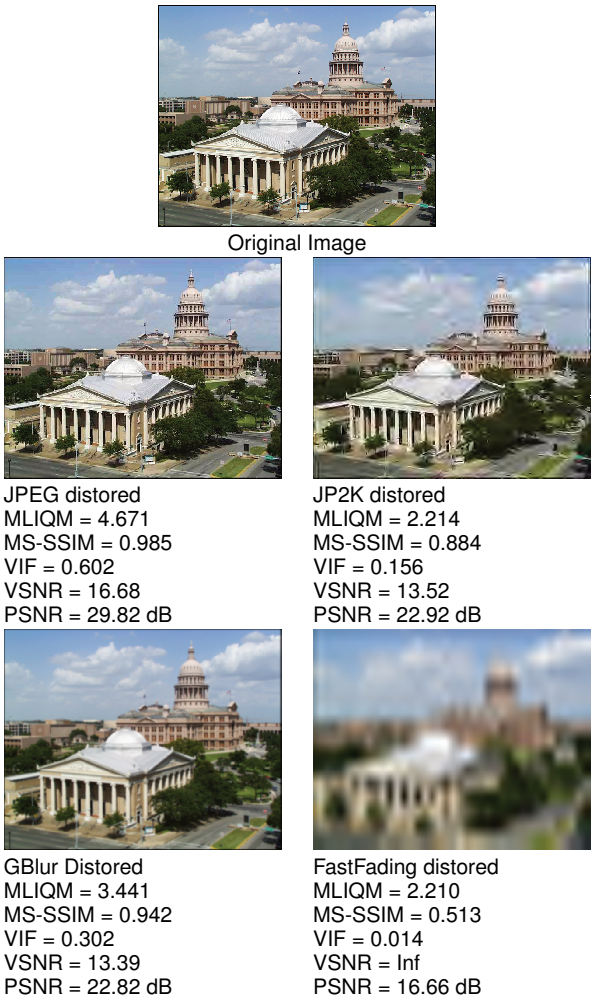

Figure 3. *Example of results obtained computing the trial FR-IQA algorithms on an original image (`churchandcapitol` extracted from LIVE and its degraded versions by applying JPEG (0.83865 bpp), JPEG2000 (0.194 bpp), Gaussian Blur ($\sigma = 1.565074$) and a fast fading process (receiver SNR=18.9).*

Considering compression oriented degradations, except for degradation #12 (JPEG transmission errors), MLIQM yields an increase of SROCC values for compression-degraded images. In

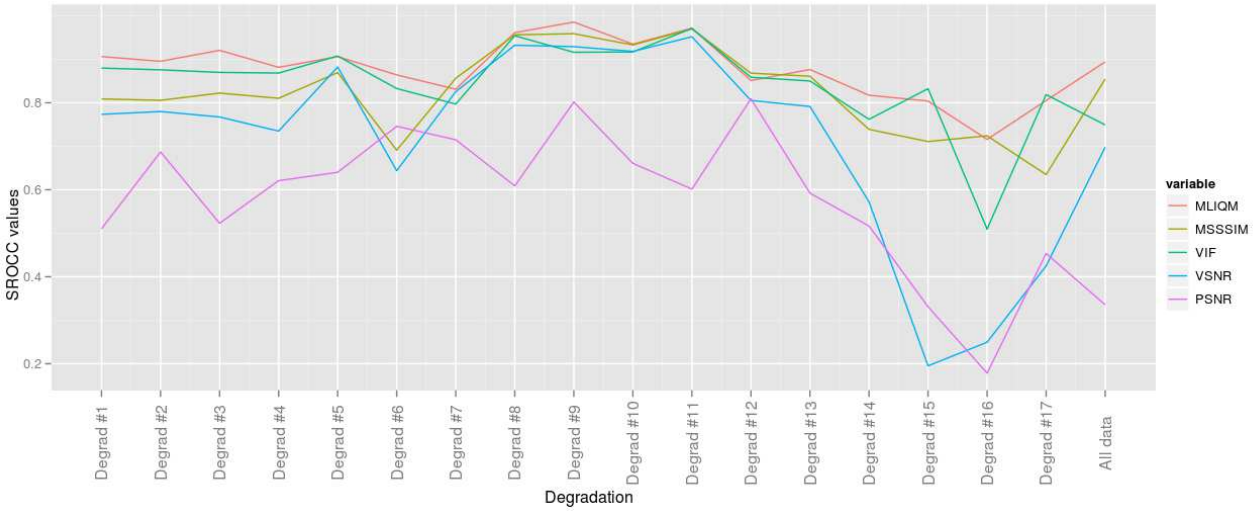

Figure 4. *Obtained Spearman rank order correlation coefficient (SROCC) between the predicted DMOS values and the subjective DMOS scores considering TID2008 database as test set. The type of degradations are described in table 1.*

addition, degradation #15 (local block-wise distortions of different intensity) can be considered as transmission errors since local blocks of the image are color degraded. As for degradation #12, a small correlation difference is noticeable between MS-SSIM and MLIQM. Degradations #16 and #15, respectively, concern a change of intensity and of contrast. They cannot be considered only as a degradation process, but also as a change of the naturalness of images. When analysing the images corresponding to the considered degradation, visible differences between the reference image and the degraded versions are not necessarily great. Nevertheless, for these degradation, a small difference of correlation is between the best IQA algorithm and the MLQIM.

Finally, considering the entire TID database, MLIQM yields 1) a higher correlation rate and 2) a difference with the other trial IQA schemes statistically significant. In addition, adding more elements associated to degradation for which MLIQM is less performant, the proposed approach should perform better (since 100 images for those degradations do not seems to reach a relevant training process). The same final remark formulated for obtained results on LIVE can be applied to TID : this consistency of improvement for subsets as for the entire TID database can be considered as an indicator of the validity of the proposed approach.

The complexity of the proposed approach relies on the training phase in order to design both the classification process and the regression scheme. This phase can (and should) be done offline, as a preprocessing stage. Actually, both SVMs and ν-SVRs training are of high complexity. Once MLIQM is trained, during the online stage, its complexity depends on the complexity of feature extraction process, since the complexity associated to both classification and regression stage can be neglicted. Even if this complexity is higher than simple IQA algorithms, it is acceptable since MLIQIM provides very high correlations obtained with respect to human judgments (and it outperforms IQA algorithms for some degradation).

Conclusion

In this paper a new approach to design a FR-IQA algorithm is proposed. This approach is based on a classification process

Degrad #	Type of distortion
1	Additive Gaussian noise
2	Additive noise in color components is more intensive than additive noise in the luminance component
3	Spatially correlated noise
4	Masked noise
5	High frequency noise
6	Impulse noise
7	Quantization noise
8	Gaussian blur
9	Image denoising
10	JPEG compression
11	JPEG 2000compression
12	JPEG transmission errors
13	JPEG2000 transmission errors
14	Non eccentricity pattern noise
15	Local block-wise distortions of different intensity
16	Mean shift (intensity shift)
17	Contrast change

Description of the 17 degradation types within the TID2008 database

such as the human being is supposed to proceed to judge the quality of an object. To apply the classification process, a vector of features has been generated. The selected features are chosen from full-reference image HVS-based features and full-reference image features, for both of them a reference image is needed.

The compared techniques with the proposed LMIQM method, are four state-of-the-art FR-IQA methods. The obtained results shown that LMIQM gives better results and yields a significant improvement of the correlation coefficients with the human judgments.

References

[1] Z. Wang, A. C. Bovik, and E. P. Simoncelli, "Structural approaches to image quality assessment," in *Handbook of Image and Video Processing*, pp. 961–974, Academic Press, 2nd ed., 2005.

[2] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? - a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117,

2009.

[3] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," tech. rep., International Telecommunication Union, Geneva, Switzerland, 2002.

[4] A. Bouzerdoum, A. Havstad, and A. Beghdadi, "Image quality assessment using a neural network approach," in *Fourth IEEE Inter. Symp. on Signal Proc. and Information Tech., 2004.*, pp. 330–333, 2004.

[5] P. Gastaldo, R. Zunino, I. Heynderickx, and E. Vicario, "Objective quality assessment of displayed images by using neural networks," *Signal Processing: Image Communication*, vol. 20, pp. 643–661, 2005.

[6] R. V. Babu, S. Suresh, and A. Perkis, "No-reference JPEG image quality assessment using GAP-RBF," *Signal Processing*, vol. 87, no. 6, pp. 1493–1503, 2007.

[7] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *IEEE Asilomar Conference on Signals, Systems, and Computers*, pp. 1398–1402, 2003.

[8] M. Narwaria and W. Lin, "Objective image quality assessment based on support vector regression," *IEEE Transactions on Neural Networks*, vol. 21, no. 3, pp. 515–519, 2010.

[9] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 5, no. 11, pp. 3441–3452, 2006.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, 2004.

[11] Z. Wang and A. C. Bovik, "A universal quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[12] A. Trémeau, C. Charrier, and E. Favier, "Quantitative description of image distorsions linked to compression schemes," in *Proceedings of The Int. Conf. on the Quantitative Description of Materials Microstructure*, (Warsaw), Apr. 1997. QMAT'97.

[13] M. W. Schwartz, W. B. Cowan, and J. C. Beatty, "An experimental comparaison of RGB, YIQ, L*a*b*, HSV, and opponent color models," in *ACM Transactions on Graphics*, vol. 6, pp. 123–158, Apr. 1987.

[14] A. B. Watson, "The cortex transform: Rapid computation of simulated neural images," *Computer Vis. Graphics and image proces.*, vol. 39, pp. 311–327, 1987.

[15] J. Lubin, *Digital Images and Human Vision*, ch. The use of psychophysical data and models in the analysis of display system performance, pp. 163–178. MIT Press, 1993.

[16] S. Daly, "A visual model for optimizing the design of image processing algorithm," in *ICIP*, vol. 2, pp. 16–20, 1994.

[17] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: a flexible architecture for multi-scale derivative computation," in *ICIP*, (Washington, DC), pp. 444–447, 1995.

[18] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *ICIP*, vol. 2, pp. 982–986, 1994.

[19] G. Lebrun, C. Charrier, O. Lezoray, C. Meurie, and H. Cardot, "Fast pixel classification by SVM using vector quantization, tabu search and hybrid color space," in *the 11th International Conference on CAIP*, (Rocquencourt, France), pp. 685–692, 2005.

[20] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[21] J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization, Advances in Kernel Methods-Support Vector Learning*. MIT Press, 1999.

[22] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.

[23] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines." Sofware Available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm, 2001.

[24] A. Dempster, "Upper and Lower Probablilities Induced by Multivalued Mapping," *Ann. Math. Statist.*, vol. 38, pp. 325–339, 1967.

[25] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[26] P. Smets and R. Kruse, "The transferable belief model fr belief representation," in *Uncertainty management in Information Systems: from Needs to Solutions* (P. S. A. Motro, ed.), Boston: Kluwer, 1997.

[27] P. Smets, "Constructing the pignistic probability function in a context of uncertainty," *Uncertainty in Artificial Intelligence*, vol. 5, pp. 29–39, 1990. Elsevier Science Publishers.

[28] T. Denoeux, "A k-nearest neighbor classification rule based on dempster-shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[29] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," Tech. Rep. NeuroCOLT Technical Report(NC2-TR-1998-030), Royal Holloway College, University of London, UK, 1998.

[30] B. Scholkopf and A. J. Smola, "New support vector algorithms," Tech. Rep. NeuroCOLT Technical Report(NC2-TR-1998-031), Royal Holloway College, University of London, UK, 1998.

[31] Laboratory for Image & Video Engineering, University of Texas (Austin), "LIVE Image Quality Assessment Database," *http://live.ece.utexas.edu/research/Quality*, 2002.

[32] N. Ponomarenko, M. Carli, V. Lukin, K. E. ans J. Astola, and F. Battisti, "Color image database for evaluation of image quality metrics," in *International Workshop on Multimedia Signal Processing*, (Australia), pp. 403–408, Oct. 2008.

[33] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, 2007.

[34] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, pp. 430–444, Feb. 2006.

[35] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," tech. rep., 2000.