Signal Processing: *Image Communication* 27 (2012) 209–219

Contents lists available at SciVerse ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Machine learning to design full-reference image quality assessment algorithm

Christophe Charrier*, Olivier Lézoray, Gilles Lebrun

Université de Caen Basse-Normandie, GREYC UMR CNRS 6072, Equipe Image, ENSICAEN, 6 Bd. Maréchal Juin, F-14050 Caen, France

ARTICLE INFO

Article history:
Received 13 April 2011
Accepted 3 January 2012
Available online 3 February 2012

Keywords:
FR-IQA algorithm
Classification
Theory of evidence
SVM classification
SVM regression

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 human observers. The proposed method namely Machine Learning-based Image Quality Measure (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.

1. Introduction

Nowadays, it has become an ordinary thing for anybody to take photos with digital cameras, to upload images on computers and to use softwares to apply many image processing algorithms on these images (compression, deblurring, denoising, etc.). This is a simple and representative example of the growing of digital media that is everywhere in our world. Many Tera-bytes thus transit on the Internet. In order to reduce the amount of transmitted data, one typical applied processing on an image is compression, so that little data is to be further transmitted. There are many compression schemes, such as the so-called JPEG that is the most famous and commonly used. Very recently, Google has launched their own compression scheme: WebP [1]. It is supposed to offer lesser bit rates than JPEG for the same quality. This implies that efficient performance measures are available. The way to evaluate this performance is a crucial step regarding image compression, 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, etc.). Quality evaluation can be divided into two main topics: (1) objective and (2) subjective evaluation.

The first topic gives place to two families of criteria: (1) basic and (2) Human Visual System (HVS)-based criteria. The first family corresponds to the traditional criteria known as mathematical measures, because they result from geometry (concept of distance) or from signal processing (signal to noise ratio). The second family of criteria takes into account the characteristics of the human visual system by a weighting of images' errors. This second topic relates to psychophysical experiments allowing to add a subjective dimension to the quality evaluation process. Due to the time-expensive aspect of

* Corresponding author.
E-mail address: christophe.charrier@unicaen.fr (C. Charrier).

0923-5965/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2012.01.002

this last topic, Image Quality Assessment (IQA) algorithms have been intensively investigated to quantify the quality of a compressed image.

IQA algorithms can be divided into three main topics: (1) Full-reference (FR) IQA methods, (2) Reduced-references (RR) IQA techniques and (3) No-reference (NR) IQA algorithms.

FR-IQA algorithms refer to algorithms that require the presence of a reference signal for the prediction of the quality of a test signal while RR-IQA techniques refer to algorithms that only require partial information about the reference signal in order to predict the quality of a test signal. NR-IQA solutions refer to algorithm for which the reference signal is not available. The two first classes of IQA algorithms can be considered as similarity measures since the main goal of those methods is to judge how two images are visually close. RR-IQA algorithms provide a solution that lies between full-reference and no-reference models.

The usually applied scheme to design an IQA 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 of these channels. A final quality score is obtained by pooling these errors in both spatial and frequency domains. The most common way to perform this pooling is to use the Minkowski error metric. Some studies [2] 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 [3]. 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 [4], the human observers have to choose a quality class from an integer scale from 0 to 100. Those scores 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 than one classification, and the given score could be interpreted as a confidence of the observer in its judgment. In addition, when a human being judges the quality of an image, many internal psychophysical scales come into play [5,6]. 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 [7–9]. *E.g.*, Bouzerdoum et al. [7] 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 as 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 [10] features. 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 and Lin [11] 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 terms 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 should be. In a second interpretation, one 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 classify 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 the sensitivity of which 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

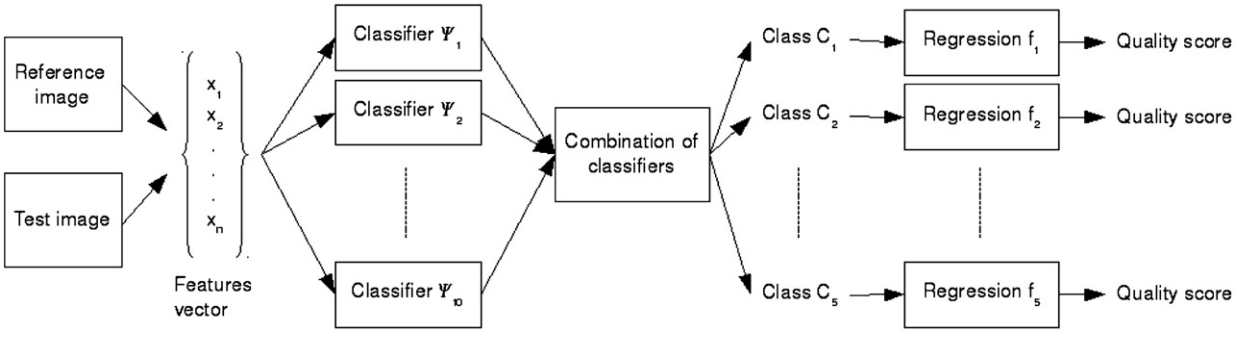

Fig. 1. General scheme of the proposed method to obtain the final quality score of a test image.

image. After computing a feature vector including several local quality features, a SVM multi-class classification process is performed to provide the final quality class C_i, $\forall i \in [1,...,5]$ of the test image. Those five correspond to the quality classes as advocated by recommendation ITU-R BT.500-11 [4]. Finally, from this classification, a SVM regression process is applied to score the quality of the test image as follows. Each quality class is associated with a score range of length 1: the first class is associated to the range [0,1[, the second one to the range [1,2[and so on until the fifth quality class that is associated to the range [4,5]. For each class C_i, $\forall i \in [1,...,5]$, a regression function is designed in order to score the quality of a degraded image within the associated range. Thus, the quality of a degraded image is scored between 0 and 5.

This way, the proposed IQA method yields a sensitivity adaptation to quality image in order to counterbalance medium prediction of usually used IQA techniques.

The paper is structured as follows. In Section 2, we briefly present the set of features used to describe the quality of images. Section 3 details how classification and regression are performed with Support Vector Machines. Section 4 presents obtained results. The last section concludes.

2. The selected full-reference features

When a human being judges the quality of an image, the employed internal scales are supposed to be scales of comparison for which a "conscious" reference image is required. This is to put in relation with the inherent conscience of each human being. The conscience is related to what Freud calls "the perception–consciousness system" [12]. It concerns a peripheral function of the psychic apparatus which receives information from the external world and those coming from the memories and the internal feelings of pleasure or displeasure. The immediate character of this perceptive function involves an inability for the conscience of keeping a lasting trace of this information. It sends them to the preconscious, a place of a first setting in memory. The consciousness perceives and transmits significant qualities. Freud employs a formula like "index of perception, of quality, of reality" to describe the content of the operations of the perception–conscience system.

Thus, 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.

This section describes the used set of features to measure degradations of the distorted image. This category of criteria is obtained measuring the dissimilarity between an original image and its degraded version. Many criteria have been developed considering either the definition domain of the image (*e.g.* color space), the frequency domain (*e.g.* Fourier transform, DCT) or the spatio-frequency space (*e.g.* wavelet transform).

Sheik et al. [13] compared 10 recent IQA algorithms and determined which had particularly high levels of performance. They concluded that more can be done to reduce the gap between machine and human evaluation of image quality. Seshadrinathan and Bovik [14] studied the relationship between the structural similarity (SSIM) [10] and VIF [15] frameworks and older metrics, *i.e.* the MSE and HVS-based quality metrics. They concluded that SSIM and VIF are closely related to the older IQA metrics under certain natural scene modeling assumptions. This also was recently studied by Horé and Ziou who defined a bijective relation between SSIM and PSNR yielding predictions of SSIM values from PSNR (and inversely) [16]. The global conclusion of all those comparison studies is 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 outperforms many with statistical significance. Thus, factors embedded in the MS-SSIM index will serve a spatial criterion as described in Section 2.1.

Wang et al. [17] 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 the structural information.

2.1. 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 [18]. 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.

The luminance comparison between an original image I and its degraded version J is defined as

$$l(I,J) = \frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1} \quad (1)$$

where μ_I and μ_J, respectively, represent the mean intensity of the images I and J, and C_1 is a constant for avoiding instability when $\mu_I^2 + \mu_J^2 \approx 0$. A common choice for the stabilizing constant is $C_1 = (K_1 L)^2$, where L is the theoretical dynamic range of the image's pixels and $K_1 = 0.01$.

The contrast distortion measure is defined to have a similar form:

$$c(I,J) = \frac{2\sigma_I\sigma_J + C_2}{\sigma_I^2 + \sigma_J^2 + C_2} \quad (2)$$

where C_2 is a non-negative constant commonly defined as $C_2 = (K_2 L)^2$ ($K_2 = 0.03$), and σ_I (resp. σ_J) represents the standard deviation.

The structure comparison is performed after luminance subtraction and contrast normalization. The structure comparison function is defined as

$$s(I,J)=\frac{\sigma_{I,J}+C_3}{\sigma_I\sigma_J+C_3} \tag{3}$$

where C_3 is a non-negative constant defined as $C_3=C_2/2$, and $\sigma_{IJ}=1/(N-1)\sum_{i=1}^{N}(I_i-\mu_i)(J_i-\mu_j)$. Substituting C_3 by $C_2/2$ in (3)

$$s(I,J)=\frac{2\sigma_{I,J}+C_2}{2\sigma_I\sigma_J+C_2} \tag{4}$$

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 index that consists of 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 [19]. 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 descriptors is performed in an antagonist Luminance–Chrominance color space, namely the CIE Lab color space [20]. These two criteria are also included in the feature vector.

2.2. 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. There are many models 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 bandwidth [21–23].

Among all existing decompositions, the steerable pyramid transform [24] is used in this paper to quantify contrast masking effects. The decomposition consists of many spatial frequency levels, which are further divided into a set of orientation bands. The basis function is 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. Fig. 2 presents an example decomposition of a synthetic image and the associated Fourier transform magnitude of the four used filters.

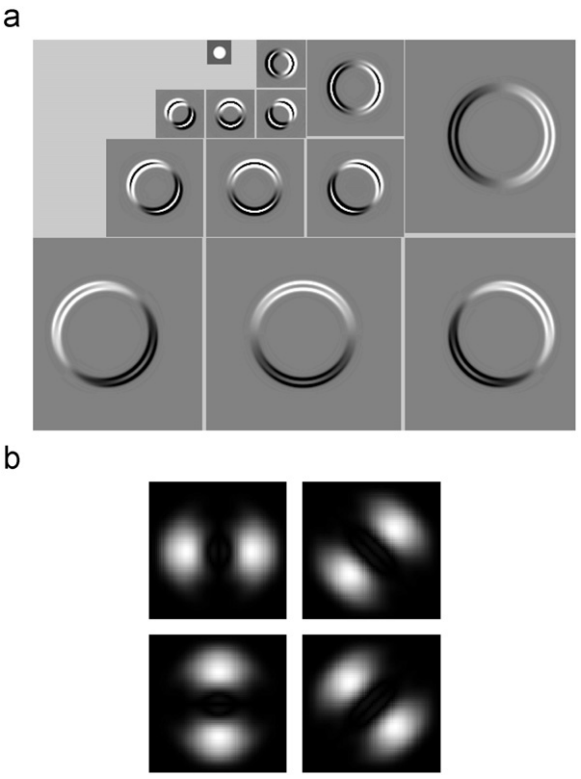

Fig. 2. Illustration of the (a) three levels steerable pyramidal decomposition of a synthetic image containing a white disk centered on a dark background and (b) the associated Fourier transform magnitude of the four used filters.

The coefficients induced by the decomposition are next squared to obtain local energy measures. As mentioned in [25], those coefficients are normalized to take into account the dynamic limited range of the mechanisms in the Human Visual System.

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

$$r(x,y,f,\theta)=k\frac{(a(x,y,f,\theta))^2}{\sum_{\phi\in 0,45,90,135}(a(x,y,f,\phi))^2+\sigma^2} \tag{5}$$

This procedure leads to normalized sensors having a limited dynamic range. Each sensor is able to discriminate contrast differences over a 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 range 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:

$$\Delta r(f,\theta)=\left\|\sum_{x,y} r_0(x,y,f,\theta)-r_1(x,y,f,\theta)\right\|^2 \tag{6}$$

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

3. 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 [26], and to their high generalization abilities. The SVMs were developed by Vapnik [27] and are based on the structural risk minimization principle from the 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 $\mathcal{S}=\{(x_i,y_i)\}_{i=\{1,\ldots,m\}}$, $x_i\in\mathbb{R}^n$, $y_i\in\{-1,+1\}$, SVM maps the input vector x into a high-dimensional feature space $\mathbf{H}$ through some non-linear mapping functions $\phi:\mathbb{R}^n\rightarrow\mathbf{H}$, 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:

$$\mathcal{V}(w,b,\xi)=\frac{1}{2}\|w\|^2+C\left(\sum_{i=1}^{m}\xi_i\right) \tag{7}$$

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 [27] is to maximize the following optimization problem:

$$\mathcal{W}(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \tag{8}$$

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 vectors (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:

$$h(x)=\mathrm{sign}(f(x)) \quad \text{with} \quad f(x)=\sum_{i\in SV}\alpha_i^* y_i K(x_i,x)+b^* \tag{9}$$

where the threshold b^* is computed via the unbounded support vectors [27] (*i.e.*, $0<\alpha_i^*<C$). An efficient algorithm SMO (Sequential Minimal Optimization) [28] and many refinements [29,30] were proposed to solve dual problem.

3.1. 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 RBFs works 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 $t_{i,j}$, $\forall i\in[1,5]$, $j\in[2,5]$ 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. (9)) be the SVM decision function obtained by training it on the ith binary problem. Table 1 gives binary problems transformation used in the OO scheme.

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. Several combination schemes of binary classifiers exist [31,32]. One interesting way to achieve this combination is the use of the theory of evidence [33,34] since the confidence one has in classifier can be taken into account for the final assignment decision.

3.2. The combination of binary classifiers

Once the multi-class classifier has been decomposed in 10 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) [35,33]. Indeed, each of the binary classifiers 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 *a posteriori* $p(y=1|f)$ where f represents the uncalibrated output value of SVMs. Platt [28] has proposed a method to compute

Table 1
Binary problems transformation used in a One-Versus-One combination scheme.

Class	$t_{5,4}$	$t_{5,3}$	$t_{5,2}$	$t_{5,1}$	$t_{4,3}$	$t_{4,2}$	$t_{4,1}$	$t_{3,2}$	$t_{3,1}$	$t_{2,1}$
5	+1	+1	+1	+1	–	–	–	–	–	–
4	−1	–	–	–	+1	+1	+1	–	–	–
3	–	−1	–	–	−1	–	–	+1	+1	–
2	–	–	−1	–	–	−1	–	−1	–	+1
1	–	–	–	−1	–	–	−1	–	−1	−1

the *a posteriori* probabilities from the obtained SVM parameters. The suggested formula is based on a parametric form of a sigmoid as

$$p(y=1|f)=\frac{1}{1+\exp(Ef+F)} \tag{10}$$

where parameters E and F are fit using maximum likelihood estimation from a training set (f_i,y_i). Those parameters are found by minimizing the negative log likelihood of the training data, which is a cross-entropy error function defined as

$$\min -\sum_i t_i\log(p_i)+(1-t_i)\log(1-p_i) \tag{11}$$

where $t_i=(y_i+1)/2$ represents the target probabilities from a new training set (f_i,t_i), and $p_i=1/(1+\exp(Ef_i+F))$. This sigmoid model is equivalent to assume that the SVM outputs are proportional to the log odds of a positive example.

3.2.1. 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 $(\omega_l)_{1\leq l\leq 5}$ representing the five quality classes {`excellent, very good, good, bad, very bad`} [4]. 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

$$\sum_{A\subseteq\Omega} m(A)=1 \quad \text{and} \quad m(\emptyset)=0 \tag{12}$$

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 an 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 the information providing from two independent sources, can be combined according to Dempster's [35] rule:

$$m(A)=\frac{\sum_{B\cap C=A} m_1(B)m_2(C)}{1-K}, \quad \forall A\in 2^\Omega,\ A\neq\emptyset \tag{13}$$

where 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:

$$K=\sum_{B\cap C=\emptyset} m_1(B)m_2(C) \tag{14}$$

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 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. 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 and Kruse [36] proposed to avoid the normalization step, since they 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 hypotheses 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 [37], uses the pignistic transformation that allows one to distribute the mass associated to a subset of Ω over each one of its elements:

$$\text{BetP}(\omega_l,m)=\sum_{\omega_l\in A\subseteq\Omega}\frac{m(A)}{|A|}, \quad \forall\omega_l\in\Omega,\ \forall 1\leq l\leq 5 \tag{15}$$

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

3.2.2. 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. Yet, there are three commonly used models: (1) the distance-based model introduces by Denœux [38] and Denœux and Zoushal [39], (2) Shafer's model [33] based on a likelihood function where the conditional *a priori* probability function is supposed to be known and (3) Appriou's models [40] also based on likelihood functions. In [40], the author proposed two models to manage the uncertain learning in the framework of evidence theory. Those models are consistent with the Bayesian approach when the mass is only allocated to singletons.

Among the three previous models, the one proposed by Denœux [38] 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:

$$\begin{cases} m(\omega_l)=\alpha\exp(-\gamma_l d^2) \\ m(\Omega)=1-m(\omega_l) \end{cases} \tag{16}$$

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 [41].

The mass functions yield to take into account the associated uncertainty to each one of the classifiers. 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 is 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

$$A=\{\omega_l\in\Omega/\max(p_l)-\delta_l\le p_l\le\max(p_l)\} \tag{17}$$

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.

3.3. SVM regression scheme

Even if scoring the quality of an image is not natural for human beings, it is quite necessary to obtain a 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 v-SVR [42] which is commonly used to solve regression problems. In particular v-SVR has the advantage of being able to automatically adjust the width of the ϵ-tube [42].

We first present the ϵ-SVR and then present v-SVR as an improvement [42,27]. Given the training data $\mathcal{S}=\{(x_i,y_i)\}_{i=\{1,\ldots,m\}}$, $x_i\in\mathbb{R}^n$, $y_i\in\{-1,+1\}$. In ϵ-SVR, x is first mapped to $z=\Phi(x)$ in feature space, then a linear function $f(x,w)=w^Tz+b$ is constructed in such a way that it deviates as less as possible from the training set according to a ϵ-insensitive loss function:

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

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 $\forall_{i=1}^{m}, y_i-f_i\le\epsilon+\xi_i^*, f_i-y_i\le\epsilon+\xi_i,\xi_i,\xi_i^*\ge 0$ 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. v-SVR alleviated this problem trading off ϵ against model complexity and training error using parameter $v>0$. Mathematically, the problem becomes

$$\min_{w,\epsilon,\xi_i,xi_i^*}\frac{1}{2}\|w\|^2+C\left(v\epsilon+\frac{1}{m}\sum_{i=1}^{m}(\xi_i+\xi_i^*)\right) \tag{18}$$

subject to $\forall_{i=1}^{m}, y_i-f_i\le\epsilon+\xi_i^*, f_i-y_i\le\epsilon+\xi_i,\xi_i,\xi_i^*\ge 0$ and $\epsilon\ge 0$. Schölkopf and Smola [43] have shown that v is an upper bound of the fraction of margin errors and a lower bound of the fraction of SV. Furthermore, they have shown that, with probability 1, v equals to both fractions. Thus, in situations where prior knowledge on these fractions is available, v is much easier to adjust than ϵ.

In this paper, the RBF is chosen as kernel for v-SVR. For each quality class, a v-SVM is trained in order to estimate function f as defined in Eq. (9) 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 depending on the associated quality class. When all the score ranges for all five regression functions are considered, a continuous score scale from 0 to 5 is available to predict the quality of a candidate image.

4. Experimental setup and performance measure

4.1. Experimental setup

4.1.1. The used image databases

To judge the performance of the proposed approach, two different image databases are used: (1) the LIVE database release 2 [44] and (2) the TID2008 database [45]. The LIVE database consists of five subsets of five 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 2. 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.

Table 2
Description of the 17 degradation types within the TID2008 database.

Degradation #	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	JPEG2000 compression
12	JPEG transmission errors
13	JPEG2000 transmission errors
14	Non-eccentricity pattern noise
15	Local block-wise distortions of different intensities
16	Mean shift (intensity shift)
17	Contrast change

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 kinds of degradation. The LIVE test set (TestC1) is composed of the degraded versions of the 13 remaining images.

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

The parameters of both the SVM classification scheme and the v-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, …, 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 [10], VSNR [46], VIF [15] 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 [47]) was used to provide a non-linear mapping between the predicted values and subjective scores. This function is a three-parameter logistic function:

$$r(x) = \frac{b_1}{1+\exp(-b_2(x-b_3))} \tag{19}$$

This nonlinearity is applied to the FR-IQA algorithm score, which gives 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).

4.2. Results

All three correlation coefficients (LCC, KROCC, SROCC) have been computed between the predicted values and the subjective DMOSs 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.

Fig. 3 presents SROCC values obtained between the predicted values and the subjective DMOSs 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 all improvements are not 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. Fig. 4 presents an example of predicted quality scores by MLIQM on degraded images.

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

Fig. 5 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, …, 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 degradations #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.

Degradations #5 and #7, respectively, deal with high frequency noise and quantization noise. Considering the first kind of artifact, 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,

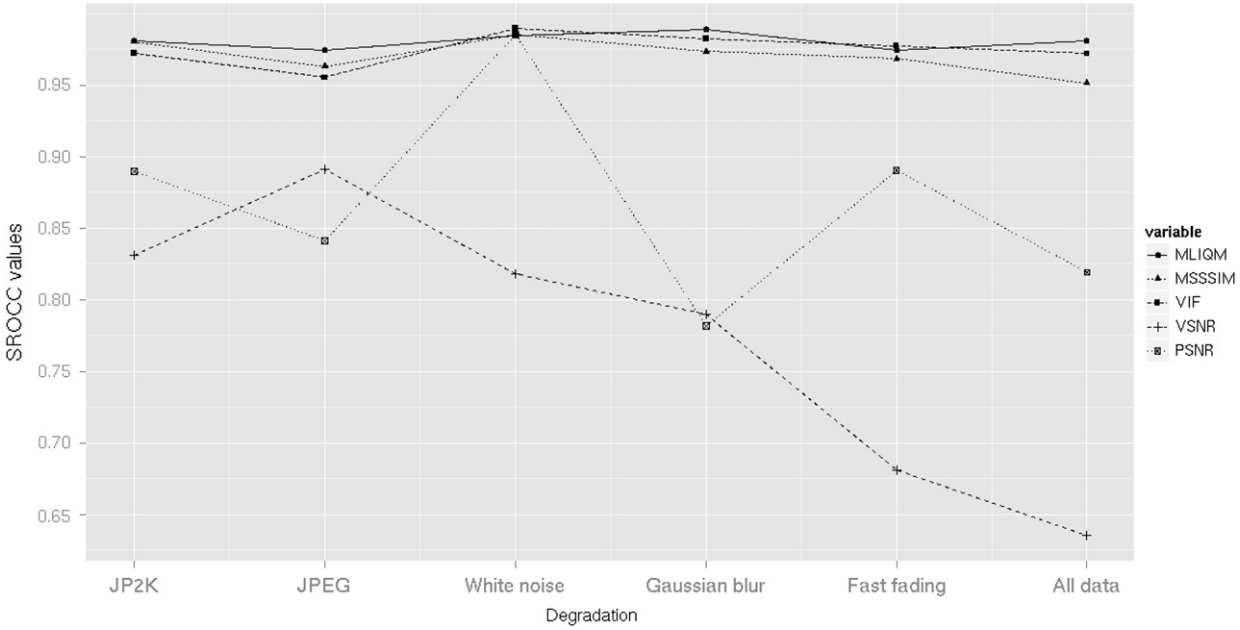

Fig. 3. Obtained correlation coefficients between the predicted DMOS values and the subjective DMOSs considering LIVE database test set.

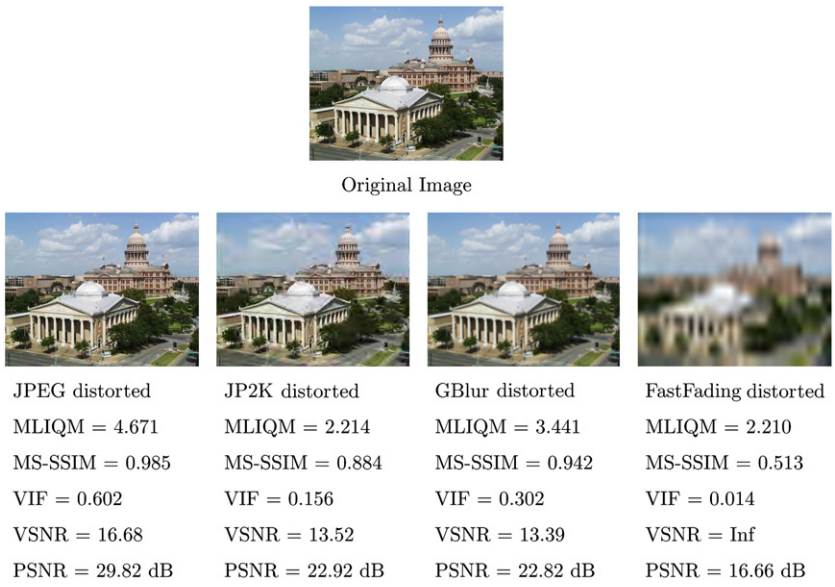

Fig. 4. Example of results obtained computing the trial FR-IQA algorithms on an original image (`churchandcapitol` extracted from LIVE and its degraded versions by applying JPEG (0.83865 bpp), JPEG2000 (0.194 bpp), Gaussian blur ($\sigma = 1.565074$) and a fast fading process (receiver SNR = 18.9).

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 approaches (MS-SSIM and MLIQM) is small.

Considering compression oriented degradations, except for degradation #12 (JPEG transmission errors), MLIQM yields an increase of SROCC values for compression-degraded images. In addition, degradation #15 (local block-wise distortions of different intensities) 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 analyzing the images corresponding to the considered degradation, visible differences between the reference image and the

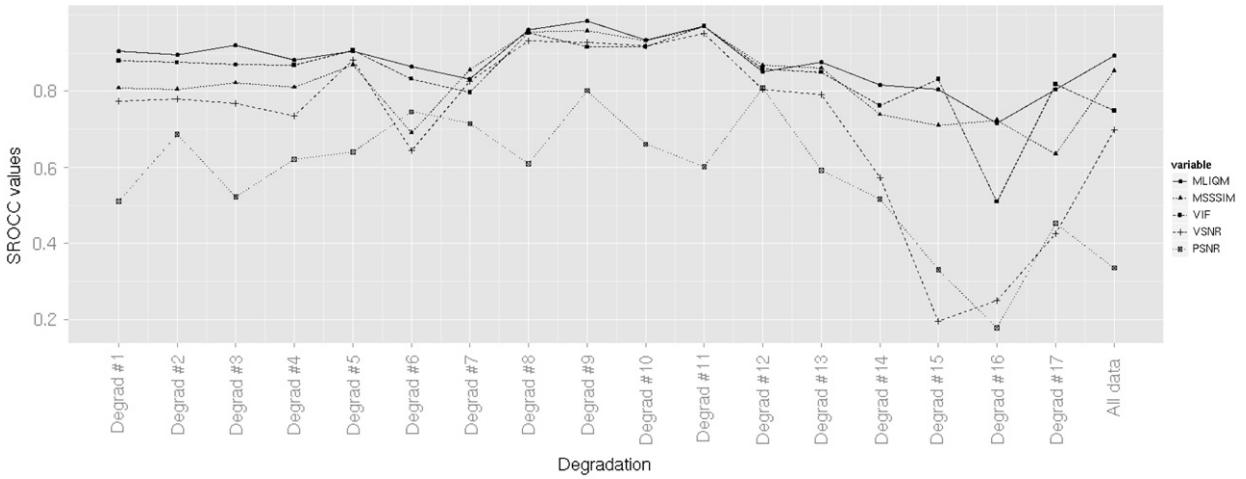

Fig. 5. Obtained Spearman rank order correlation coefficient (SROCC) between the predicted DMOS values and the subjective DMOSs considering TID2008 database as test set. The type of degradations is described in Table 2.

degraded versions are not necessarily great. Nevertheless, for these degradations, 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 statistically significant difference with the other trial IQA schemes. 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 seem 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 v-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 neglected. 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).

5. Discussion

The proposed FR-IQA algorithm based on SVM classification and regression to compute the quality score of an image seems to be a promising new way of design, since whatever the used database (LIVE and TID), the consistency of the correlation improvement is observed. This means that the formulated hypothesis concerning the classification process used by human beings when watching scenes is valid. Furthermore, this classification strategy can be modeled by a high dimensional classifier since many details can modify the final human judgment. The fact that humans are able to rank order the quality of images can be modeled by a decision function. This function can be formulated as a non-linear regression function. The belief any human being can formulate about his decision can be summarized by soft margin used to define the non-linear decision function in the regression process. This is a new framework to design FR-IQA algorithms.

Yet, even if the proposed scheme seems to be validated, the obtained results are deeply linked to the extracted features: in this case, a mere adjustment of the used vector of features and of the used kernel functions is required.

The fact that correlation results obtained with MLIQM (which integrate MS-SSIM factors) are most of the time better than those obtained with MS-SSIM yields to perhaps hypothesize that the original combination of the MS-SSIM factors is not necessarily optimal. Maybe some artifacts might not be well taken into account too.

6. Conclusion

In this paper a new approach to design a FR-IQA algorithm is proposed. This approach is based on a classification process 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 show that LMIQM gives better results and yields a significant improvement of the correlation coefficients with human judgments.

Acknowledgment

This work is supported by the ANR project #ANR-08-SECU-007-04.

References

[1] Google, WebP compression scheme, 〈http://code.google.com/intl/fr/speed/webp/〉.

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

[3] Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? A new look at signal fidelity measures, IEEE Signal Processing Magazine 26 (1) (2009) 98–117.

[4] ITU-R Recommendation BT.500-11, Methodology for the Subjective Assessment of the Quality of Television Pictures, Technical Report, International Telecommunication Union, Geneva, Switzerland, 2002.

[5] H. Hemminger, P. Mahler, Psychophysical saturation scales and the spectral sensitivity in human vision, Psychological Research 42 (1980) 207–212.

[6] R.N. Shepard, Psychological relations and psychophysical scales: on the status of direct psychophysical measurement, Journal of Mathematical Psychology 24 (1) (1981) 21–57.

[7] A. Bouzerdoum, A. Havstad, A. Beghdadi, Image quality assessment using a neural network approach, in: Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004, pp. 330–333.

[8] P. Gastaldo, R. Zunino, I. Heynderickx, E. Vicario, Objective quality assessment of displayed images by using neural networks, Signal Processing: Image Communication 20 (2005) 643–661.

[9] R.V. Babu, S. Suresh, A. Perkis, No-reference JPEG image quality assessment using GAP-RBF, Signal Processing 87 (6) (2007) 1493–1503.

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

[11] M. Narwaria, W. Lin, Objective image quality assessment based on support vector regression, IEEE Transactions on Neural Networks 21 (3) (2010) 515–519.

[12] D. Rapaport, M.M. Gill, The points of view and assumptions of metapsychology, The International Journal of Psychoanalysis 40 (1959) 153–162.

[13] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transactions on Image Processing 5 (11) (2006) 3441–3452.

[14] K. Seshadrinathan, A.C. Bovik, Unifying analysis of full reference image quality assessment, in: IEEE International Conference on Image Processing (ICIP), 2008, pp. 1200–1203.

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

[16] A. Horé, D. Ziou, Image quality metrics: Psnr vs. ssim, in: IEEE International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010, pp. 2366–2369.

[17] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error measurement to structural similarity, IEEE Transactions on Image Processing 13 (1) (2004) 1–14.

[18] Z. Wang, A.C. Bovik, A universal quality index, IEEE Transactions on Image Processing 9 (3) (2002) 81–84.

[19] A. Trémeau, C. Charrier, E. Favier, Quantitative description of image distortions linked to compression schemes, in: Proceedings of the International Conference on the Quantitative Description of Materials Microstructure, Warsaw, 1997, qMAT'97.

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

[21] A.B. Watson, The cortex transform: rapid computation of simulated neural images, Computer Vision, Graphics and Image Processing 39 (1987) 311–327.

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

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

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

[25] P.C. Teo, D.J. Heeger, Perceptual image distortion, in: ICIP, vol. 2, 1994, pp. 982–986.

[26] G. Lebrun, C. Charrier, O. Lezoray, C. Meurie, 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, 2005, pp. 685–692.

[27] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.

[28] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: Advances in Kernel Methods—Support Vector Learning, MIT Press, 1999.

[29] R. Collobert, S. Bengio, SVMTorch: support vector machines for large-scale regression problems, Journal of Machine Learning Research 1 (2001) 143–160.

[30] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, Software, 2001. Available at: 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉.

[31] C.-W. Hsu, C.-J. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks 13 (3) (2002) 415–425.

[32] O. Lezoray, H. Cardot, Comparing combination rules of pairwise neural networks classifiers, Neural Processing Letters 27 (1) (2008) 43–56.

[33] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.

[34] B. Quost, T. Denœux, M.-H. Masson, Pairwise classifier combination using belief functions, Pattern Recognition Letters 28 (5) (2007) 644–653.

[35] A. Dempster, Upper and lower probabilities induced by multivalued mapping, Annals of Mathematical Statistics 38 (1967) 325–339.

[36] P. Smets, R. Kruse, The transferable belief model for belief representation, in: P.S.A. Motro (Ed.), Uncertainty Management in Information Systems: From Needs to Solutions, Kluwer, Boston, 1997.

[37] P. Smets, Constructing the pignistic probability function in a context of uncertainty, in: Uncertainty in Artificial Intelligence, vol. 52, Elsevier Science Publishers, 1990, pp. 29–39.

[38] T. Denœux, A *k*-nearest neighbor classification rule based on Dempster–Shafer theory, IEEE Transactions on Systems, Man and Cybernetics 25 (5) (1995) 804–813.

[39] T. Denœux, L.M. Zouhal, Handling possibilistic labels in pattern classification using evidential reasoning, Fuzzy Sets and Systems 122 (2) (2001) 47–62.

[40] A. Appriou, Probabilités et incertitude en fusion de données multisenseurs, Revue Scientifique et Technique de la Défense 11 (1991) 27–40.

[41] G. Celeux, J. Diebolt, The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem, Computational Statistics Quarterly (2) (1985) 73–82.

[42] A.J. Smola, B. Scholkopf, A Tutorial on Support Vector Regression, Technical Report, NeuroCOLT Technical Report (NC2-TR-1998-030), Royal Holloway College, University of London, UK, 1998.

[43] B. Schölkopf, A.J. Smola, New Support Vector Algorithms, Technical Report, NeuroCOLT Technical Report (NC2-TR-1998-031), Royal Holloway College, University of London, UK, 1998.

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

[45] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, F. Battisti, Color image database for evaluation of image quality metrics, in: International Workshop on Multimedia Signal Processing, Australia, 2008, pp. 403–408.

[46] D.M. Chandler, S.S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Transactions on Image Processing 16 (9) (2007) 2284–2298.

[47] VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Technical Report, 2000.