Sparse representation-based image quality assessment

Tanaya Guha*, Ehsan Nezhadarya, Rabab K. Ward

Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

Abstract

A highly promising approach to assess the quality of an image involves comparing the perceptually important structural information in this image with that in its reference image. The extraction of the perceptually important structural information is however a challenging task. This paper employs a sparse representation-based approach to extract such structural information. It proposes a new metric called the sparse representation-based quality (SPARQ) index that measures the visual quality of an image. The proposed approach learns the inherent structures of the reference image as a set of basis vectors. These vectors are obtained such that any structure in the image can be efficiently represented by a linear combination of only a few of these basis vectors. Such a sparse strategy is known to generate basis vectors that are qualitatively similar to the receptive field of the simple cells present in the mammalian primary visual cortex. To estimate the visual quality of the distorted image, structures in the visually important areas in this image are compared with those in the reference image, in terms of the learnt basis vectors. Our approach is evaluated on six publicly available subject-rated image quality assessment datasets. The proposed SPARQ index consistently exhibits high correlation with the subjective ratings of all datasets and overall, performs better than a number of popular image quality metrics.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Digital images incur a variety of distortions during their acquisition, compression, transmission, storage or reconstruction. Such processes often degrade the visual quality of images. In order to monitor, control and improve the quality of images produced at the various stages, it is important to automatically quantify the image quality. As the end-users of the majority of image-based applications are humans, the automatic quantification of image quality requires the understanding of human perception of image quality, so as to mimic it as closely as possible.

The mean squared error (MSE) and the peak signal-to-noise ratio (PSNR) have been traditionally used to measure the image quality degradations. These metrics are mathematically convenient to use but they do not correlate well with human perception of image quality [1]. A considerable amount of research effort has been put towards quantifying the quality of images as perceived by humans, and a number of objective image quality assessment algorithms that agree with the subjective judgment of human beings have been developed. The objective quality assessment methods, depending on how much information about the original undistorted image they use, are broadly classified into three categories: no-reference, reduced-reference and full-reference. This paper concentrates on the full-reference quality estimation approach.

Earlier focus of the full-reference image quality assessment research has been on building a comprehensive and accurate model of the human visual system (HVS) and its...
quality of images. The structural approaches [7] have evolved to extract information from natural scenes and therefore, use the natural scene statistics to estimate the perceptual quality of images. The statistical approaches [5,6] hypothesize that the HVS has over the years evolved to extract information from natural scenes and therefore, use the natural scene statistics to estimate the perceptual quality of images. The structural approaches [7–13] on the other hand operate on the basis of a rather important aspect of the HVS – its sensitivity towards the image structures for developing cognitive understanding. In this approach, image quality is estimated in terms of the fidelity of the structures in the distorted image relative to that in the reference image. An image quality metric that is representative of the class of structural information-based metrics is the structural similarity index (SSIM) [7]. SSIM treats the structural distortions separately from the non-structural distortions (such as luminance and contrast change). The visual quality of a patch in the distorted image is measured by comparing it with the corresponding patch in the reference image in terms of three components: luminance, contrast and structure. A global quality score is computed by combining the effects of the three components over all image patches. SSIM achieved much success because of its simplicity, and its ability to tackle a wide variety of distortions. Due to its pixel-domain implementation, SSIM is highly sensitive to geometric distortions like scaling, translation, rotation and other misalignments [2]. To improve the performance of SSIM, multiscale extension [9], wavelet transform-based modification [12], gradient-domain implementation [10] and various pooling strategies [11,14] have been proposed.

The underlying assumption behind utilizing the structural information is that the HVS uses the structures it extracts from the viewing field for its cognitive understanding. Therefore, for an image to be considered of high-quality, all the structural information present in its reference image should be well preserved. From this viewpoint, the efficient capture of the structural information of images is the key to developing a successful image quality assessment algorithm. But extracting or analyzing the structural information in a perceptually meaningful way is a non-trivial task. Although the use of predefined, data-independent basis functions is prevalent in analyzing image structures, their success is often limited by the degree as to how suitable the basis functions are in capturing the structural information of the signals under consideration.

We propose the use of a more generalized approach to analyzing image structures in the context of image quality assessment. This involves learning from the training data a set of basis elements that could be adapted to represent the inherent structures of the signal in question. These learnt basis elements are collectively known as a dictionary. As each basis vector could be tailored to represent a significant part of the structures present in the given data, a learnt dictionary is more efficient in capturing the structural information compared to a predefined set of bases. In the last few years, several practical dictionary learning algorithms have been developed [15,16]. It has been shown that the data-dependent, learnt dictionaries, due to their superior ability to efficiently model the inherent structures in the data, can outperform predefined dictionaries like wavelets in several image processing tasks [15,17,18]. Dictionary learning ideas have also been explored in the context of measuring generic image similarity, applicable to problems like clustering or retrieval [19,20]. The problem of perceptual image quality assessment, although related to image similarity measures, is more specific and needs to be addressed as a different problem.

This dictionary learning approach is connected to an important result obtained by Olshausen and Field in 1996 [21]. They proposed to learn a set of basis elements from the input itself instead of designing a new basis function that could sparsely represent the input signal. They enforced (i) a sparsity prior – an assumption that it is possible to describe the input using a small number of basis elements, (ii) overcompleteness – the number of basis elements is greater than the vector space spanned by the input. They showed that this strategy results into a set of basis elements that are localized, oriented, and bandpass in nature, which resemble the properties of the receptive field of simple cells in the primary visual cortex [21].

In this paper, we develop a full-reference image quality assessment metric which we call the sparse representation-based quality (SPARQ) index. This metric relies on capturing the inherent structures in the reference image as a set of basis vectors which collectively form an overcomplete dictionary. These vectors are obtained such that any structure (patch) in the image can have a sparse representation w.r.t. the dictionary so as to resemble the visual cortex. To estimate the visual quality of the distorted image the visually important regions in this image are compared with those in the reference image. These regions are detected by using a saliency detection algorithm. Each patch in the visually important region in the distorted image and its corresponding patch in the reference image are decomposed in terms of the learnt dictionary. Their sparse coefficients are then compared to yield a measure of quality of the distorted image. Since our method analyzes image structures by building a cortex-like model of the stimuli, we expect the extracted structural information to be important to the HVS, and perceptually more meaningful compared to the structural information used in existing methods.

To evaluate the efficacy of the proposed metric, we perform various experiments on six publicly available, subject-rated image quality assessment datasets. The proposed SPARQ index shows great promise as it consistently exhibits high correlation with the subjective scores and often outperforms its competitors.

The rest of the paper is organized as follows. Section 2 describes the proposed quality assessment approach in detail; this is followed by experimental results and discussion in Section 3. Section 4 concludes the paper and suggests possible directions to future work.
2. The proposed approach

Our image quality assessment approach is divided into two phases:

- **Training** phase – learns an overcomplete dictionary for a given reference image in order to capture its inherent structures.
- **Quality estimation** phase – generates a quality score for a given distorted image by comparing the structures in this image with the corresponding ones in its reference image, in terms of the learnt dictionary.

Fig. 1 presents an overview of the proposed approach and each step is described below in detail.

2.1. Training phase

Natural viewing fields are highly structured and spatially correlated. The light rays that reflect off various structures in the viewing field get focused onto an array of photoreceptors present in the retina. The visual information is then encoded as complex statistical dependencies among the photoreceptor activities [22]. The goal of the primary visual cortex, as indicated in several seminal studies [21,22], is to reduce these statistical dependencies in order to discover the intrinsic structures that gave rise to the image. This is achieved by enforcing sparsity and overcompleteness as discussed in the previous section.

A reasonable strategy towards mimicking this phenomenon is to describe the image (or an image patch) in terms of a linear superposition of a small number of basis vectors. These basis vectors form a subset of a larger, overcomplete set of basis vectors (dictionary) that are adapted to the given image so as to best represent all structures in that image [21,22]. It has been shown that on employment of this strategy, the resulting basis elements of the dictionary are qualitatively similar to the receptive field of the cortical simple cells [21]. The importance of sparsity as an important prior, as shown in [21], is based on the observation that natural images contain sparse structures and can be described by a small number of structural primitives like lines, edges and corners [22,23].

**Dictionary learning**: To design an overcomplete dictionary for the reference image \( I \in \mathbb{R}^{n} \), the image structures are first extracted as image patches. A large number of distinct, possibly overlapping patches of dimension \( \sqrt{n} \times \sqrt{n} \) are extracted randomly from \( I \). Ideally, one patch centered at every pixel should be extracted; but in practice, extracting any large number of patches is sufficient for learning a good dictionary. After extracting a large number of random patches, the patches with low or no structural information are discarded (by removing the patches whose variance is zero or close to zero). The remaining (say, \( k \)) patches are selected and each of these \( k \) patches is converted to a vector of length \( n \). These patch vectors are concatenated to form a matrix \( P \in \mathbb{R}^{n \times k} \).

Using the \( k \) patches as input, we intend to learn an overcomplete dictionary \( \Phi = \{ \phi_i \}_{i=1}^{m} \), \( \phi_i \in \mathbb{R}^{n} \). For overcompleteness, it is required that \( n < m \) i.e. \( \Phi \) should have more basis vectors than the dimensionality of the input. An overcomplete dictionary offers greater flexibility in representing the essential structures in a signal. It is also robust to additive noise, occlusion and small translation [24].

With overcompleteness however, greater difficulties arise; because a full-rank, overcomplete \( \Phi \) creates an underdetermined system of linear equations having an infinite number of solutions. To narrow down the choice to one well-defined solution, constraints (e.g. minimum norm, sparsity) are required. We enforce a constraint of sparsity in order to mimic the cortical model in [21]. Let the sparse representation of \( P \) over the dictionary \( \Phi \) be denoted by \( X = \{ x_i \}_{i=1}^{m} \), \( x_i \in \mathbb{R}^{n} \) where any patch vector in \( P \) can be represented by a linear superposition of no more than \( \tau_1 \) dictionary columns where \( \tau_1 \ll m \). This is formally written as the following sparse optimization problem:

\[
\min_{\Phi} \| P - \Phi X \|_F^2 \quad \text{s.t. } \forall i \| x_i \|_0 \leq \tau_1
\]

where \( \| . \|_F \) is the Frobenius norm (square root of the sum of the squared values of all elements in a matrix) and \( \| . \|_0 \) is the \( C_0 \) semi-norm that counts the number of non-zero elements in a vector.

![Fig. 1. Overview of the proposed sparse representation-based image quality assessment approach.](image)
The $\ell_0$ semi-norm provides a straightforward notion of sparsity, but it renders the problem non-convex. Thus obtaining an accurate solution of (1) is NP hard. Nevertheless, in the last few years researchers have found practical and stable ways to solve such underdetermined systems via convex optimization [25] and greedy pursuit algorithms [26]. To solve (1), a popular learning algorithm known as the K-SVD [15] is employed. K-SVD iteratively solves (1) by performing two steps at each iteration: (i) sparse coding and (ii) matching pursuit. In the sparse coding step, $\Phi$ is kept fixed and the coefficients in $X$ are computed by a greedy algorithm called the orthogonal matching pursuit (OMP) [26]:

$$\min_{x} ||P - \Phi X||_2^2 \quad \text{subject to} \quad ||x||_0 \leq \tau$$

(2)

In the dictionary update step, each basis vector $\phi_i \in \Phi$ is updated sequentially, allowing the corresponding coefficients in $X$ to change as well. Dictionary learning can be understood as a generalization of the vector quantization method. In vector quantization, every input vector is approximated to one and only one basis vector that is closest to the input. This can be considered as an extreme sparsity constraint. In dictionary learning, the input vector is allowed to be approximated using a small number of basis vectors. For the details of this learning algorithm, refer to the original work [15].

2.2. The quality estimation phase

To estimate the quality of an image, we compare the local patches of this image with the corresponding patches in its reference image. Instead of using all possible image patches (as in SSIM and its variants), we intend to compare only a set of carefully selected image patches. These selected patches are considered to be visually more important than others. Later (Section 3, Fig. 4), we also show that higher quality scores are obtained using the visually important patches as opposed to the scores obtained using all image patches. A global measure of quality is then computed by aggregating the scores obtained at the local level.

2.2.1. Detection of the visually important patches

It is well-known that not every pixel (or region) in an image receives the same level of visual importance. Several studies have shown that a significant improvement in the performance of quality metrics can be achieved by detecting the perceptually important regions [27–29].

In order to detect the visually important regions in an image, any visual saliency detection method can be used. In this work, we experiment with the following salient patch detection methods:

(a) Itti–Koch saliency model [30].
(b) Graph-based visual saliency [31].
(c) Spectral residual [32].
(d) Entropy-based method.

The first three methods mentioned above are well known saliency detection methods and the last one is a rather simple approach. Our experiments show that the (c) spectral residual and the (d) entropy-based methods yield the best results for our purpose (details in Section 3.3). A brief description of each of the four salient patch detection methods is provided in Appendix A.

Let us denote the saliency maps pertaining to $I_r$ and $I_d$ by $M_r$ and $M_d$, computed by one of the four saliency detection methods mentioned above. A combined saliency map of the same size as $M_r$ and $M_d$ is then created as max($M_r, M_d$) i.e. by taking the maximum of the values of $M_r$ and $M_d$ at each point. The locations of the pixels in the combined map that have high saliency scores are found. These points are used to select the corresponding $q$ patches from $I_r$ and $I_d$ as the visually important patches (see Fig. 2). These patches upon extraction from $I_r$ and $I_d$ are vectorized and arranged in columns of the matrices $P_r \in \mathbb{R}^{n \times q}$ and $P_d \in \mathbb{R}^{n \times q}$ respectively.

2.2.2. Computation of the quality score

At this point, we have two sets of visually important patches: $P_r$ and $P_d$, extracted from the same locations in the reference and the distorted images. The next step is to compare these patches w.r.t. the dictionary $\Phi$.

Let us consider any patch vector $p_r \in P_r$ from $I_r$ and its corresponding patch vector $p_d \in P_d$ from $I_d$. The patches $p_r$ and $p_d$ are decomposed using $\Phi$ to obtain their respective sparse coefficient vectors $x_r$ and $x_d$:

$$\min_{x_r} ||p_r - \Phi x_r||_2^2 \quad \text{subject to} \quad ||x_r||_0 \leq \tau_2$$

(3)

$$\min_{x_d} ||p_d - \Phi x_d||_2^2 \quad \text{subject to} \quad ||x_d||_0 \leq \tau_2$$

(4)

Note that, each of $x_r$ and $x_d$ contains only $\tau_2$ non-zero elements. The locations (indices) of these non-zero coefficients indicate those specific basis vectors in $\Phi$ which actually contribute to the approximation of the input patch. These active basis vectors are called the support of the input. The amplitudes of these non-zero coefficients are the weights by which these support vectors are combined. The support vectors and their weights together are indicative of the structural and non-structural distortions (e.g. luminance or contrast change) between the two input patches. Ideally, $p_r$ and $p_d$ would have different sets of support vectors whenever there exist any structural distortions between them. Otherwise, if the two patches undergo purely non-structural distortions, the supports would remain the same but their weights may change.

In order to quantify the perceptual quality of $p_d$ w.r.t. $p_r$, we compare their sparse representations $x_r$ and $x_d$. A simple but effective way to compare two vectors is to compute their normalized correlation coefficient. A parameter $\alpha$ is computed based on the correlation coefficient between $x_r$ and $x_d$ as follows:

$$\alpha(p_r, p_d) = \frac{|x_r^T x_d| + c_1}{||x_r||_2 ||x_d||_2 + c_1}$$

(5)

where $c_1$ is a small positive constant added to avoid instability when the denominator is close to zero. Clearly, $0 < \alpha \leq 1$. When $x_r$ and $x_d$ are orthogonal, $|x_r^T x_d| = 0$; but due to the presence of $c_1$, the parameter $\alpha$ is slightly greater than zero. Due to normalization, $\alpha$ is unaffected by the lengths of $x_r$ and $x_d$. Thus $\alpha$ is unable to measure distortions that cause the length of $x_d$ to change.
To account for these types of distortions as well, we introduce another parameter. An important measure of similarity (or difference) between two vectors is their pointwise difference. Hence, we compute another quantity $\beta$ which uses the length of the difference vector $\|x_r - x_d\|$: $\beta(p_r; p_d) = 1 - \frac{\|x_r - x_d\|}{\|x_r\| + \|x_d\| + c_2}$ (6)

where $c_2$ is a small positive constant. It is easy to see that $0 < \beta < 1$, for non-empty $x_r$ and $x_d$.

We propose a function $S(p_r, p_d)$ that measures the perceptual quality of $p_d$ w.r.t $p_r$ as follows:

$S(p_r, p_d) = \alpha(p_r, p_d) \beta(p_r, p_d)$ (7)

Let $S(p^i_r, p^i_d)$ be the quality measure of $p^i_d$—the $i$th salient patch in $I_d$, w.r.t. $p^i_r$ — the corresponding patch in $I_r$. The proposed global image quality $\text{SPARQ}(I_r, I_d)$ is computed by averaging over all $q$ visually important patches:

$\text{SPARQ}(I_r, I_d) = \frac{1}{q} \sum_{i=1}^{q} S(p^i_r, p^i_d)$ (8)

Remarks.

- The SPARQ index (8) is bounded: $0 < \text{SPARQ} < 1$; it is always non-negative since each of its components is non-negative.
- The highest value of SPARQ is attained when $I_r = I_d$.
- The index is not symmetric i.e. $\text{SPARQ}(I_r, I_d) \neq \text{SPARQ}(I_d, I_r)$. This is because the dictionary $\Phi$ is trained on the reference image only. Symmetry can be achieved by repeating the quality estimation stage with a dictionary trained on the distorted image and averaging the resulting quality scores obtained using the two dictionaries. Our experiments show that achieving symmetry has little or no significance on the performance of the SPARQ index.

3. Experimental validation

This section presents a critical evaluation of the proposed image quality metric, the SPARQ index, on the six publicly available image databases whose subjective
quality ratings are available. These databases are A57 [33], CSIQ [34], LIVE [35], MICT [36], TID [37] and WIQ [38]. The images in these databases contain a variety of distortions such as compression artifacts, blurring, flicker noise, and wireless transmission artifacts. First, we discuss how to set the parameter values required to compute the SPARQ index. Experiments are carried out to select the salient patch detection method for which SPARQ performs the best. The performance of a quality metric is evaluated by computing the correlation between its objective scores and the available subjective ratings. To compare the performance of SPARQ with state-of-the-art, correlation scores of SPARQ are compared with those of the seven well-known image quality metrics: PSNR, SSIM [7], PHVSM [39], VIF [6], VSNR [3], MAD [40] and IWSSIM [11].

3.1. The databases

A brief description of each of the six datasets used in this work is provided below.

The Cornell-A57 dataset [33,33] consists of 54 distorted images created from 3 original grayscale images. The images are subject to the following 6 types of distortions: JPEG compression (JPEG), JPEG 2000 compression (JP2K), additive white Gaussian noise (AWGN), Gaussian blur, JPEG2000 compression with dynamic contrast-based quantization algorithm, and uniform quantization of LH sub-bands of a 5-level discrete wavelet transform at all scales.

The CSIQ database [34] has 30 original images which were used to create 866 distorted images. The 6 distortion types (at four to five distortion levels) include JPEG, JP2K, global contrast decrements, AWGN, and Gaussian blur.

The LIVE database [7,35] contains 779 distorted images created from 29 original color images. Each distorted image exhibits one of the five types of distortions: JP2K, JPEG, AWGN, Gaussian blur and fastfading channel distortion of JPEG2000 compressed bitstreams.


The TID database [37] is so far the largest subject-rated image dataset for quality evaluation. It has 1700 images generated from 25 reference images with 17 distortion types at four distortion levels. The distortion types are AWGN, additive noise in color components, spatially correlated noise, masked noise, high frequency noise, impulse noise, quantization noise, Gaussian blur, image denoising, JPEG, JP2K, JPEG transmission errors, JP2K transmission errors, non-eccentricity pattern noise, local block-wise distortions of different intensity, mean shift, and contrast change.

The WIQ database [38,41] consists of 80 distorted images generated from 7 reference images. The images exhibit wireless imaging artifacts which are not considered in other datasets. Due to the complex nature of a wireless communication channel, the images contain more than one artifact.

3.2. Evaluation methodology

To perform a comparison between objective and subjective scores pertaining to image quality, we use a set of evaluation measures suggested by the video quality expert group (VQEG) [42] and other experts [6,11]. These evaluation measures are Spearman’s rank order correlation coefficient (SROCC), Kendall’s rank order correlation coefficient (KROCC), the Pearson linear correlation coefficient (PLCC), mean absolute error (MAE) and root mean squared error (RMSE).

The SROCC and KROCC are used to measure the prediction monotonicity of the objective scores, while PLCC, MAE and RMSE measure the prediction accuracy of the same. The SROCC and KROCC are computed from the raw objective scores. In order to compute PLCC, MAE and RMSE, a five-parameter logistic function (refer to (9) and (10)) is fitted to the raw objective scores separately for each dataset. A particular objective score, s, is mapped to a new score, Q(s), using a non-linear mapping function Q(·) which is defined as follows:

\[
Q(s) = \gamma_1 \log(\gamma_2 + (s - \gamma_3)) + \gamma_4 + \gamma_5
\]

(9)

\[
\logistic(\sigma, s) = \frac{1}{2} - \frac{1}{1 + \exp(\sigma, s)}
\]

(10)

A MATLAB function called fminunc is used for fitting to obtain the parameters \(\gamma_1, \gamma_2, \ldots, \gamma_5\). The values of PLCC, MAE and RMSE are computed after the above monotonic non-linear mapping between the subjective and objective scores is performed. Note that, SROCC and KROCC are non-parametric rank correlation metrics and are independent of any nonlinear monotonic mapping between the subjective and the objective scores. For details of the evaluation methodology see [6,11,42].

A good image quality assessment metric is expected to have high SROCC, KROCC and PLCC scores, and low MAE and RMSE values.

3.3. Implementation details

3.3.1. Preprocessing

Before training and quality assessment, two preprocessing steps are executed: (1) every color image in each dataset is converted to grayscale image, and (2) all images (reference and distorted) are downsampled by a factor F (bicubic rescaling) so as to account for different viewing conditions [7] (such as, image size, viewing distance). The value of F is obtained by using the following empirical formula [7]:

\[
F = \max(1, \text{round}(g/256))
\]

(11)

where \(g = \min(\#\text{rows in } I_r, \#\text{columns in } I_r)\).

3.3.2. Training

In the training phase, there are 4 parameters to be set by the user:

- \(\sqrt{n_i}\): patch size,
- \(k\): number of random patches to be extracted from a reference image in order to train the dictionary,
- \(m\): number of basis vectors in the dictionary,
- \(r_1\): sparsity constraint.

Unfortunately, there are no theoretical guidelines to determine the values of these parameters, so we rely on previous work and empirical methods. A patch size of
\[ \sqrt{m} \times \sqrt{n} = 11 \times 11 \] is used following the patch-size specification of SSIM [7]. A collection of as large as \( k = 3000 \) patches are extracted randomly from every reference image to train its corresponding dictionary. We set the overcompleteness factor \( (m/n) \) to 2 which yields \( m = 242 \). Although larger overcompleteness factor may yield higher accuracy, it increases the computational load significantly. It has also been shown that for a low overcompleteness factor, sparse representations are stable in the presence of noise [43]. The value of \( \tau \) is set to 12.

### 3.3.3. Selecting the visually important patch detection method

As mentioned before, 3 well-known saliency detection methods (Itti–Koch saliency model [30], graph-based visual saliency [31], spectral residual [32]) and a simple entropy-based approach are used to detect the visually important regions in images. In order to investigate the effect of these methods on the quality assessment task, we employed each method in our framework to observe and compare their performances on the six datasets.

Fig. 3 compares the 4 patch selection methods in terms of computation time. Table 1 compares them in terms of SROCC indicating their impact on quality assessment. Each method uses the same number of \( q \) salient patches (see Section 3.3.4 for how to determine the value of \( q \)). The result of selecting random patches is also presented in Table 1. Random patch selection result serves as a baseline.

It is clear from Table 1 that carefully selecting and using the visually important patches for quality assessment improve the performance of the proposed quality metric. Another important observation is that the simple entropy-based method performs better or at par with the well-known saliency methods in the context of quality assessment.

Considering the performance (Table 1) and speed (Fig. 3) of the competing patch detection methods, we observe that the spectral residual method and the entropy-based approach are the two better methods. From this point, we will use the following notations of SPARQ depending on which patch selection method it uses:

- **SPARQ\(_{sr}\)** – uses entropy for patch selection.
- **SPARQ\(_{sp}\)** – uses spectral residual method for patch selection.

### 3.3.4. Quality estimation

In the quality estimation phase, we need to set the following parameters:

- \( c_1, c_2 \): stabilizing constants used in (5) and (6).
- \( q \): number of salient patches.
- \( \tau_2 \): sparsity constraint.

The constants are chosen to have small positive values, \( c_1 = 256 \times 0.01, c_2 = 0.01 \) so as to have minimal influence on the quality score.

The value of \( q \) is determined empirically. For each database, the number of visually important patches, \( q \), is varied, and the performance of SPARQ\(_{sr}\) is measured in terms of SROCC. This is presented in Fig. 4. The value of \( q \) is varied from 2% to 100% of \( N \) where \( N \) is the total number of pixels in \( I \) or \( I_b \). In five out of the six datasets, the best performance of the SPARQ index is observed when \( q = 0.15N \) i.e. 15% of \( N \). We also observe that, when all patches in \( I \) are used, the performance of SPARQ degrades. This confirms our assumption that only the visually important areas are useful for quality assessment.

In order to determine the value of \( \tau_2 \), it is varied from 2 to 12, and the changes in SROCC scores are plotted in Fig. 5 for SPARQ\(_{sr}\) and SPARQ\(_{sp}\). The results are shown for the larger datasets available: the TID and CSIQ datasets. From Fig. 5, we see that \( \tau_2 = 6 \) provides the best overall trade-off.

### 3.4. Performance comparison

Table 2 compares the performances of SPARQ\(_{sr}\) and SPARQ\(_{sp}\) with the state-of-the-art quality metrics in terms of SROCC, PLCC and RMSE (KROCC and MAE are left out for easier comprehension of the results and also because they reflect similar performance trends as SROCC and RMSE, respectively). PSNR is used as a baseline method. For the implementation of SSIM, PHVS-M, VIF, VSNR, MAD and IWSSIM, we have used the original MATLAB codes provided by the respective authors. The parameters of each of these methods are set to their default values as suggested in the original references.

The best two results in Table 2 are written in bold for each dataset. As can be seen in the comparison, no single metric performs the best on all datasets. Nevertheless, the performances of SPARQ\(_{sr}\) and SPARQ\(_{sp}\) are consistently high over all

![Fig. 3. Comparison of the 4 visually important patch detection methods in terms of computation time for the same pair of images (size 256 × 256).](image-url)
datasets. Results in Table 2 show that SPARQ performs better than PSNR, SSIM, PHVS-M and VSNR on all datasets. Also note that the WIQ dataset is the only dataset which contains more than one artifacts due to the nature of wireless imaging. SPARQ handles such complex artifacts better than other metrics. This indicates the potential of SPARQ index to be used in complex practical systems where degradation of images is likely to be caused by more than one factors. Overall, SPARQ achieves highest correlation score in 3 (A57, CSIQ and WIQ) out of the 6 datasets. For the remaining three datasets (LIVE, MICT and TID), VIF and IWSSIM perform better than SPARQ. This is because, SPARQ does not handle non-structural distortion (e.g. contrast and illumination) very well. Performance of SPARQ index for separate distortion types is presented in Appendix B.

To obtain further insight into each metric’s performance, statistical significance tests are performed on the

<table>
<thead>
<tr>
<th>Patch selection method</th>
<th>A57</th>
<th>CSIQ</th>
<th>LIVE</th>
<th>MICT</th>
<th>TID</th>
<th>WIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.875</td>
<td>0.904</td>
<td>0.870</td>
<td>0.766</td>
<td>0.674</td>
<td>0.778</td>
</tr>
<tr>
<td>Itti-Koch saliency [30]</td>
<td>0.926</td>
<td>0.941</td>
<td>0.915</td>
<td>0.848</td>
<td>0.805</td>
<td>0.800</td>
</tr>
<tr>
<td>Graph-based [31]</td>
<td>0.909</td>
<td>0.939</td>
<td>0.914</td>
<td>0.865</td>
<td><strong>0.806</strong></td>
<td>0.807</td>
</tr>
<tr>
<td>Spectral residual [32]</td>
<td>0.920</td>
<td>0.946</td>
<td>0.930</td>
<td><strong>0.872</strong></td>
<td>0.792</td>
<td>0.816</td>
</tr>
<tr>
<td>Entropy-based</td>
<td><strong>0.943</strong></td>
<td><strong>0.950</strong></td>
<td><strong>0.933</strong></td>
<td>0.870</td>
<td>0.774</td>
<td><strong>0.851</strong></td>
</tr>
</tbody>
</table>

Fig. 4. Performance of the SPARQ index (correlation with subjective scores measured in terms of SROCC) with the varying number of high-entropy patches used in the quality estimation phase.

Fig. 5. Effect of sparsity on the performances of SPARQṛ and SPARQsr on TID and CSIQ datasets.
prediction errors or residuals of each metric. We adopt a similar procedure as described in [3]. Assuming that the residuals follow a Gaussian distribution, F-test is used to assess if two sets of residuals produced by two different metrics, $M_1$ and $M_2$, correspond to the same population, and if one has significantly larger residuals than the other. Let $\sigma_1$ and $\sigma_2$ be the variances of two different distribution of residuals. The F-statistic is given by $F = \sigma_1^2 / \sigma_2^2$. Values of $F > F_{\text{critical}}$ (or $F < 1 / F_{\text{critical}}$) indicate that, at a given confidence level, $M_1$ has significantly larger (or, smaller) residuals compared to $M_2$. The performance of each metric is evaluated w.r.t. SPARQ, using F-test and corresponding F-values are listed in Table 3.

Table 3 shows that SPARQ performs significantly better (in terms of prediction error) in the majority of cases. Even for datasets like LIVE or TID where SPARQ was not among the top performing metrics, it produces significantly lower error compared to other quality metrics. We also observe that although IWSSIM produces higher SROCC, it yields significantly larger errors compared to SPARQ, for five databases. VIF shows more consistent performance producing lower errors in three out of six databases. Performances of SPARQ and SPARQsr seem to be comparable.

### 3.4.1. Computational complexity

In order to compute SPARQ, the two steps that require the bulk of computation are (i) the dictionary learning step in the training phase and (ii) the sparse coding step in the quality estimation phase. The computational load of the dictionary learning step in turn is dominated by the sparse coding step (2) performed as part of the learning process. Hence, it is the sparse coding step that we should be concerned with.

Our implementation uses an efficient sparse coding algorithm called the Batch-OMP [44]. Its computational complexity is $O(mnr)$ per training signal, where $n$ is the dimension of a patch vector, $m$ is the number of dictionary elements and $r$ is the sparsity constraint ($r < m$).

To give an idea of the computation time, a basic MATLAB implementation (on a computer with Intel Q9400 processor at 2.66 GHz) takes on average 3.4 s for the dictionary learning step using the parameter values specified in this paper. The quality estimation step for SPARQ takes 1.7 s and for SPARQsr it takes 1.0 s.

### 3.4.2. Limitations of SPARQ

Due to its dependence on sparse coding (Eqs. (2)–(4)), SPARQ is computationally demanding (still less expensive compared to the HVS-based models like MAD). Nevertheless, considering the rapid growth of the area, we are hopeful that faster sparse coding algorithms will be available soon.

---

**Table 2**

Performance comparison of various quality assessment metrics over six datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SROCC-based comparison</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS7</td>
<td>0.575</td>
<td>0.806</td>
<td>0.896</td>
<td>0.622</td>
<td><strong>0.935</strong></td>
<td>0.864</td>
<td>0.775</td>
<td><strong>0.943</strong></td>
<td>0.920</td>
</tr>
<tr>
<td>CSIQ</td>
<td>0.800</td>
<td>0.858</td>
<td>0.822</td>
<td>0.919</td>
<td>0.809</td>
<td>0.899</td>
<td>0.921</td>
<td><strong>0.950</strong></td>
<td><strong>0.946</strong></td>
</tr>
<tr>
<td>LIVE</td>
<td>0.852</td>
<td>0.947</td>
<td>0.922</td>
<td><strong>0.963</strong></td>
<td>0.912</td>
<td>0.943</td>
<td><strong>0.956</strong></td>
<td>0.933</td>
<td>0.931</td>
</tr>
<tr>
<td>MICT</td>
<td>0.613</td>
<td>0.875</td>
<td>0.848</td>
<td>0.907</td>
<td>0.860</td>
<td><strong>0.908</strong></td>
<td>0.920</td>
<td>0.870</td>
<td>0.872</td>
</tr>
<tr>
<td>TID</td>
<td>0.552</td>
<td>0.773</td>
<td>0.561</td>
<td>0.749</td>
<td>0.704</td>
<td>0.770</td>
<td><strong>0.853</strong></td>
<td>0.774</td>
<td><strong>0.792</strong></td>
</tr>
<tr>
<td>WIQ</td>
<td>0.485</td>
<td>0.758</td>
<td>0.757</td>
<td>0.692</td>
<td>0.656</td>
<td>0.790</td>
<td>0.786</td>
<td><strong>0.851</strong></td>
<td><strong>0.816</strong></td>
</tr>
<tr>
<td>PLCC-based comparison</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS7</td>
<td>0.561</td>
<td>0.802</td>
<td>0.875</td>
<td>0.614</td>
<td>0.914</td>
<td>0.881</td>
<td>0.765</td>
<td><strong>0.945</strong></td>
<td><strong>0.925</strong></td>
</tr>
<tr>
<td>CSIQ</td>
<td>0.746</td>
<td>0.758</td>
<td>0.772</td>
<td>0.927</td>
<td>0.735</td>
<td>0.820</td>
<td>0.914</td>
<td><strong>0.945</strong></td>
<td><strong>0.939</strong></td>
</tr>
<tr>
<td>LIVE</td>
<td>0.808</td>
<td>0.941</td>
<td>0.917</td>
<td><strong>0.944</strong></td>
<td>0.917</td>
<td>0.939</td>
<td><strong>0.951</strong></td>
<td>0.930</td>
<td>0.928</td>
</tr>
<tr>
<td>MICT</td>
<td>0.632</td>
<td>0.705</td>
<td>0.839</td>
<td><strong>0.902</strong></td>
<td>0.855</td>
<td><strong>0.911</strong></td>
<td>0.802</td>
<td>0.873</td>
<td>0.872</td>
</tr>
<tr>
<td>TID</td>
<td>0.519</td>
<td>0.727</td>
<td>0.552</td>
<td>0.808</td>
<td>0.682</td>
<td>0.748</td>
<td><strong>0.851</strong></td>
<td>0.805</td>
<td><strong>0.820</strong></td>
</tr>
<tr>
<td>WIQ</td>
<td>0.568</td>
<td>0.641</td>
<td>0.749</td>
<td>0.730</td>
<td>0.763</td>
<td><strong>0.830</strong></td>
<td>0.660</td>
<td><strong>0.836</strong></td>
<td>0.801</td>
</tr>
<tr>
<td>RMSE-based comparison</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AS7</td>
<td>0.203</td>
<td>0.147</td>
<td>0.119</td>
<td>0.194</td>
<td>0.099</td>
<td>0.116</td>
<td>0.105</td>
<td><strong>0.080</strong></td>
<td><strong>0.093</strong></td>
</tr>
<tr>
<td>CSIQ</td>
<td>0.375</td>
<td>0.171</td>
<td>0.167</td>
<td>0.058</td>
<td>0.178</td>
<td>0.150</td>
<td>0.106</td>
<td><strong>0.086</strong></td>
<td><strong>0.090</strong></td>
</tr>
<tr>
<td>MICT</td>
<td>0.969</td>
<td>0.887</td>
<td>0.680</td>
<td><strong>0.540</strong></td>
<td>0.648</td>
<td><strong>0.515</strong></td>
<td>0.748</td>
<td>0.611</td>
<td>0.612</td>
</tr>
<tr>
<td>TID</td>
<td>1.147</td>
<td>0.921</td>
<td>1.119</td>
<td>0.790</td>
<td>0.981</td>
<td>0.890</td>
<td><strong>0.689</strong></td>
<td>0.796</td>
<td><strong>0.768</strong></td>
</tr>
</tbody>
</table>

**Table 3**

Statistical test results on RMSE between subjective and objective scores. Orange cells indicate where SPARQ produces significantly lower prediction error, blue cells indicate where the competing method produces significantly lower prediction error than SPARQ, and no color indicates that the difference is not statistically significant.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AS7</th>
<th>CSIQ</th>
<th>LIVE</th>
<th>MICT</th>
<th>TID</th>
<th>WIQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_{\text{critical}}$</td>
<td>1.421</td>
<td>1.091</td>
<td>1.096</td>
<td>1.213</td>
<td>1.064</td>
<td>1.334</td>
</tr>
<tr>
<td>$1 / F_{\text{critical}}$</td>
<td>0.704</td>
<td>0.917</td>
<td>0.912</td>
<td>0.775</td>
<td><strong>0.939</strong></td>
<td>0.750</td>
</tr>
<tr>
<td>PSNR</td>
<td>4.428</td>
<td>3.818</td>
<td>1.931</td>
<td>2.161</td>
<td>1.784</td>
<td>2.818</td>
</tr>
<tr>
<td>SSIM [7]</td>
<td>3.624</td>
<td>3.003</td>
<td>1.272</td>
<td>1.441</td>
<td>0.980</td>
<td>1.891</td>
</tr>
<tr>
<td>PHVS [39]</td>
<td>2.314</td>
<td>3.742</td>
<td>1.278</td>
<td>0.999</td>
<td>2.005</td>
<td>1.237</td>
</tr>
<tr>
<td>VIF [6]</td>
<td>7.339</td>
<td>0.686</td>
<td>0.698</td>
<td>0.641</td>
<td>0.951</td>
<td>1.431</td>
</tr>
<tr>
<td>MAD [40]</td>
<td>2.241</td>
<td>3.399</td>
<td>0.773</td>
<td>1.101</td>
<td><strong>1.722</strong></td>
<td>1.501</td>
</tr>
<tr>
<td>SPARQ</td>
<td>1.401</td>
<td>2.628</td>
<td>1.003</td>
<td>1.024</td>
<td>0.843</td>
<td>1.232</td>
</tr>
<tr>
<td>SPARQsr</td>
<td>1.401</td>
<td>2.628</td>
<td>1.003</td>
<td>1.024</td>
<td>0.843</td>
<td>1.232</td>
</tr>
</tbody>
</table>
The present version of SPARQ index works on grayscale images and thus is blind to degradations in the color components. Like most of the existing image quality assessment metrics, SPARQ relies on signal fidelity to quantify perceptual quality where fidelity is one of the several factors in determining the perceptual quality [45].

4. Conclusion

The main contribution of this work is the development of a new image quality assessment metric that quantifies perceptual quality of a distorted image with respect to a reference image. This metric measures the structural fidelity between the reference and the distorted images. It learns an overcomplete dictionary from the reference image to capture its inherent structures. The structural information in the visually important regions in the distorted image is then compared with those in the reference image in terms of the sparse coefficients obtained using learnt dictionary.

The proposed metric, called the SPARQ index, consistently produces high correlation score over all six databases, achieving the highest correlation score in three databases. Statistical tests demonstrate that SPARQ produces significantly lower prediction error compared to most of the competing metrics in five databases. The proposed sparse representation-based quality assessment method is simple and has shown much promise in analyzing structural information in the context of image quality assessment.

The proposed index can be improved in several ways. Possible directions include combining SPARQ with various pooling strategies, learning multiscale dictionaries, using more efficient sparse solvers, and extending it to work for color images and videos.

Acknowledgments

The authors are thankful to the editor and the anonymous reviewers whose comments have helped improve the value of this work. This work was supported by NSERC Canada and by Qatar National Research Fund (QNRF no. NPRP 09-310-1-058).

Appendix A. Saliency detection methods

Itti-Koch saliency model: This classical method decomposes the input image into a set of feature maps by extracting multiple low level features (such as intensity, color and orientation) at different scales. These feature maps are normalized and combined across scales to form conspicuity maps, one for each feature. These conspicuity maps are then combined to create one saliency map of the image. For details, refer to the work of Itti et al. [30].

The graph-based visual saliency (GBVS): This successful method uses the computational power and the parallel nature of the graph algorithms to compute saliency map of images. Like [30], GBVS also computes multiple feature maps in order to find out the points that are unusual in its neighborhood using a graph-based algorithm. The maps that identify unusual points are called the activation maps. These activation maps are normalized and combined to create the saliency map. Details can be found in the original reference [31].

The spectral residual approach: This is a state-of-the-art saliency detection method that can compute robust saliency map of natural images very fast. This method analyzes the frequency spectrum of an input image obtained by the Fourier transform. The method extracts the points of statistical singularities in the spectrum which corresponds to the salient regions in the spatial domain. For details, refer to the original work [32].

Entropy-based visually important patch detection: A common hypothesis is that the HVS is an efficient extractor of information, and therefore the image regions that contain high information attract more visual attention [14,11]. Based on this hypothesis, we take an information theoretic approach towards detecting the visually important patches. One way to quantify the local information content of an image is by computing Shannon’s entropy of each patch. The information content or entropy of a discrete random variable \( z \) with probability distribution \( P_z = \{p_1, p_2, \ldots, p_J\} \) is defined as

\[
H(z) = H(P_z) = - \sum_{j=1}^{J} p_j \log_2 p_j
\]  

Similarly, an image patch can also be analyzed as a random variable. Let us consider an image patch \( z \) of dimension \( \sqrt{n} \times \sqrt{n} \) where each pixel in \( z \) is assumed to be independent and identically distributed. If \( z \) contains \( J \) distinct intensity values, its probability distribution, \( P_z \), is given by \( P_z = \{p_1, p_2, \ldots, p_J\} \), where \( J \leq 2^8 \) for an 8-bit grayscale image; \( p_j \) is the probability of the pixel intensity value \( j \). The probability \( p_j \) is defined as \( p_j = f_j / n \), where \( f_j \) is the number of pixels (frequency) with intensity value \( j \) occurring in the image patch \( z \) and \( n \) is the total number of pixels in \( z \). The entropy of every \( \sqrt{n} \times \sqrt{n} \) patch (a patch around every pixel) in the reference image \( I_r \in \mathbb{R}^N \) is computed.
puted as

$$H(z) = -\sum_{j=1}^{n} p_j \log_2 p_j = -\sum_{j=1}^{n} \frac{f_j}{n} \log_2 \left( \frac{f_j}{n} \right) \tag{A.2}$$

The larger the value of $H$, the higher is the information content of a patch.

Appendix B. Distortion specific performance

Since SPARQ and SPARQ$_*$ yield comparable results, we provide the performance of only one of them for different types of distortions. The distortion-specific performance of SPARQ$_*$ is presented in Table A1 for the most common distortion types: JPEG, JPEG2000, AWGN and Gaussian Blur.

References