Signal Processing: *Image Communication* 29 (2014) 293–302

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Image quality assessment using a SVD-based structural projection

Anzhou Hu, Rong Zhang *, Dong Yin, Yibing Zhan

University of Science and Technology of China, Hefei, Anhui 230027, China

ARTICLE INFO

Article history:
Received 15 October 2012
Received in revised form
17 January 2014
Accepted 17 January 2014
Available online 4 February 2014

Keywords:
Image quality assessment (IQA)
Human visual system (HVS)
Singular value decomposition (SVD)
Structural projection
Neural networks (NN)
Spatial pooling

ABSTRACT

The development of objective image quality assessment (IQA) metrics aligned with human perception is of fundamental importance to numerous image-processing applications. Recently, human visual system (HVS)-based engineering algorithms have received widespread attention for their low computational complexity and good performance. In this paper, we propose a new IQA model by incorporating these available engineering principles. A local singular value decomposition (SVD) is first utilised as a structural projection tool to select local image distortion features, and then, both perceptual spatial pooling and neural networks (NN) are employed to combine feature vectors to predict a single perceptual quality score. Extensive experiments and cross-validations conducted with three publicly available IQA databases demonstrate the accuracy, consistency, robustness, and stability of the proposed approach compared to state-of-the-art IQA methods, such as Visual Information Fidelity (VIF), Visual Signal to Noise Ratio (VSNR), and Structural Similarity Index (SSIM).

1. Introduction

Image quality assessment (IQA) has become a fundamentally important and challenging task in numerous digital image applications with the rapid development of visual communication and information technologies [1]. A reliable IQA scheme can be used to choose the parameters in a coding system, dynamically monitor image/video quality, and compare the performance of different image processing algorithms. Because human beings represent terminals for the majority of processed digital images, subjective evaluation is the most straightforward and reliable IQA method [2]. However, subjective assessment is as cumbersome in practice as it is laborious, expensive, complex, and time-consuming. Thus, the aim of objective IQA is to automatically evaluate, using computers, the quality of images as near to human perception as possible.

Based on the availability of a reference image, objective image quality prediction schemes can be classified into three categories: full reference (FR), reduced reference (RF), and no reference (NF) approaches [3]. In this paper, we focus on the FR IQA, which means that the original "perfect-quality" image is fully accessible, and IQA refers to the measurement of visual fidelity between the reference and degraded images.

The simplest and most popular IQA methods adopt a mathematical statistic to represent pixel distortions, e.g., Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR) [4]. It is well known that these methods do not match well with the perceived visual quality (with a correlation coefficient of 0.87 on the *Laboratory for Image & Video Engineering* (LIVE) database) [5]. Over the course of the last 30 years, a number of researchers have devoted time to improving the assessment accuracy of objective IQA methods by taking advantage of known human visual

* Corresponding author. Tel.: +86 551 3603262.
E-mail addresses: haz@mail.ustc.edu.cn (A. Hu), zrong@ustc.edu.cn (R. Zhang), yindong@ustc.edu.cn (D. Yin), zybjy@mail.ustc.edu.cn (Y. Zhan).

http://dx.doi.org/10.1016/j.image.2014.01.007
0923-5965

system (HVS) characteristics. These HVS-based IQA techniques can be put into two basic categories based on the types of methodology used: vision bionics metrics and HVS-based engineering metrics [6]. The first category attempts to simulate the well modelled functionalities of the HVS, for example, the Daly and Lubin models [7]. Although the perceptual criteria used in these bionics approaches are generally accepted, a lot of drawbacks, such as unreasonable assumptions and suprathreshold problems, restrict their further development [8]. The second category only considers the relationship of the input and output of the HVS and models it as a "black box" [1]. In recent years, these engineering algorithms have received widespread attention for their low complexity and good performance. The most advanced state-of-the-art engineering schemes include Information Fidelity Criterion (IFC) [9], Visual Information Fidelity (VIF) [10], M Block Singular Value Decomposition (MSVD) [11], Visual Signal to Noise Ratio (VSNR) [12], and different versions of Structural Similarity Index (SSIM) [13–17] (with an average correlation coefficient of 0.92 on the LIVE database).

With the growth of these engineering metrics, many plausible HVS models have been proposed in the existing literature, including the following examples:

- *Structure information extraction* refers to the hypothesis that the HVS is highly adapted to extracting structural or edge information from the visual scene [13,18].
- *Multi-channel decomposition* refers to the large number of neurons of the primary visual cortex, which are tuned to visual stimuli with specific spatial locations, frequencies, and orientations. Many signal decomposition methods have been introduced to mimic this effect [19,20].
- *Visual importance-based spatial pooling strategy* is based on the idea that different image regions enjoy different image perceptual significances. The IQA performance is improved by assigning visual importance weights to local distortions [16,21].
- *Distortion feature fusion using machine learning* is an efficient pooling scheme that has been demonstrated to have the ability to determine the complex mapping between a number of visual features and the cognitive quality with sufficient training data [22,23].

These models separately represent different visual stages from low-level to high-level HVS processing mechanisms. It is intuitive to combine such programs, because the visual-cognitive processes are not isolated. However, satisfactory IQA research results are non-existent. To further improve the performance of IQA, in this paper, we propose a novel method that incorporates the ideas of multi-channel structural distortion measurement, perceptual spatial pooling, and neural networks (NN). First, a comprehensive theoretical analysis of full-parameter SVD is explored from the perspective of signal representation, by which the SVD of an image can be regarded as a set of structural projections on a group of under-complete base functions. Second, based on the aforementioned analysis results of SVD, the reference and distorted images are divided into several local blocks and the SVD is carried out for each block. The local multi-dimensional feature vectors are extracted by comparing the similarity of all of the layers of structural projections between the original and degraded images. Then, a spectral residual-based visual saliency weighting algorithm is utilised to converge the local feature vectors into a global distortion vector by taking into account the region of interest (ROI) property of HVS. Finally, the global features are integrated into a single quality score with a back-propagation feed-forward neural network (BPNN) program.

A set of elaborate experiments was carried out to examine the accuracy and robustness of our proposal. Three benchmark IQA databases were used to establish different training and test group with sufficient ground truth for NN learning. Performance comparison results with subjective evaluations show that the proposed metric achieves a higher consistency than the relevant state-of-the-art methods, such as Multi-scale SSIM (MS-SSIM) [14], VSNR, Perceptual Image Quality Index (PIQ) [17], and VIF (with correlation coefficients higher than 0.92 on the LIVE, IVC, and MICT databases).

The rest of this paper is organised as follows. Section 2 illustrates the analysis of SVD-based structural projection. Section 3 describes the details of the proposed IQA metric. Experimental results compared with other schemes are provided in Section 4. Finally, the conclusions are presented and future work is detailed in Section 5.

2. SVD and the structural projection

SVD is one of the most useful tools of matrix analysis, which has been applied in many digital image-processing tasks, including image compression [24], sparse representation [25], and IQA [11,26]. Traditional IQA approaches using SVD only measure the errors of SVD results without considering the perception of image structure. In this section, we investigate the relationship between SVD and image structure from the perspective of signal representation. This analysis demonstrates that the SVD-based signal projection can adaptively capture the efficient structural information of an image, which is suitable for reliable IQA schemes.

The SVD of a $m \times n$ real matrix $\boldsymbol{\Lambda}$ can be expressed as

$$\boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T \tag{1}$$

and:

$$\boldsymbol{U} = \left[\boldsymbol{u}_1, \quad \boldsymbol{u}_2 \quad \dots \quad \boldsymbol{u}_i \quad \dots \quad \boldsymbol{u}_m\right]_{m\times n}$$

$$\boldsymbol{V} = \left[\boldsymbol{v}_1, \quad \boldsymbol{v}_2 \quad \dots \quad \boldsymbol{v}_j \quad \dots \quad \boldsymbol{v}_m\right]_{m\times n}$$

$$\boldsymbol{S} = \mathrm{diag}(\sigma_1, \quad \sigma_2 \quad \dots \quad \sigma_k \quad \dots \quad \sigma_r, \quad 0 \quad \dots \quad 0)_{m\times n}$$

Here, $\boldsymbol{U}$ is the left singular vector matrix, $\boldsymbol{V}$ is the right singular vector matrix, and $\boldsymbol{S}$ is the diagonal matrix of singular values, σ_k, with descending order. $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are left singular vectors and right singular vectors, respectively. r is the rank of $\boldsymbol{\Lambda}$. Moreover, $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrixes.

Eq. (1) can be rewritten as

$$\boldsymbol{\Lambda} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{2}$$

In Eq. (2), the original matrix $\boldsymbol{\Lambda}$ is represented as a linear combination of a set of images of atom. Three inferences can be deduced from this formula. More specific inference procedures are detailed in Appendix A. The inferences are

(1) $\|\boldsymbol{u}_i\boldsymbol{v}_i^T\|_E = 1$ for $i = 1, 2 \ldots r$

(2) $\left\langle \boldsymbol{u}_i\boldsymbol{v}_i^T, \boldsymbol{u}_j\boldsymbol{v}_j^T \right\rangle = 0$ for $i = 1, 2 \ldots r$; $j = 1, 2 \ldots r$ and $i \neq j$

(3) $\|\boldsymbol{\Lambda}\|_E^2 = \sum_{i=1}^{r} \sigma_i^2$

where the operator $\| \ \|_E$ denotes the Euclidean norm and $\langle \rangle$ refers to the inner product of two vectorised matrixes.

From inference (1) and (2), we can see that the vectorised matrixes $\boldsymbol{u}_i\boldsymbol{v}_i^T$ are orthonormal. Therefore, they can be regarded as a set of orthonormal basis functions with the number r for signal representation in the $m \times n$-dimensional real vector space. Because the number of basic functions is much less than the dimension of the total space, these bases are under-complete and can only span a subspace of the total space. The subspace property is illustrated through the example shown in Fig. 1. The original and distorted images are taken from the Laboratory for Image & Video Engineering (LIVE) Database [27]. In this figure, (c) is reconstructed using the singular vectors of the distorted image and singular values of the original images. In other words, the reconstructed and distorted images belong to the same subspace. It can be observed that the reconstructed image enjoys the same structure information as the distorted image, particularly the blur and ringing around the edges, which implies that the under-complete basis functions (i.e., "$\boldsymbol{u}_i\boldsymbol{v}_i^T$") can capture the structure information of the decomposed image. Otherwise, compared to the complete orthogonal transforms, such as DFT and DCT with fixed bases, SVD bases are unique for each image. Thus, a SVD-based projection can represent the image structure better than the general image transforms. Actually, this SVD property has been utilised to establish the famous K-SVD algorithm in the field of dictionary learning [25]. From inference (3) and Fig. 1 we can also see that the singular values denote the energy factor of the image, which is similar to the Parseval's theorem [28]. Therefore, we refer to Eq. (2) as a specific structural projection. The projection bases ($\boldsymbol{u}\boldsymbol{v}_i^T$) and projection coefficients (σ_i) adaptively preserve the structural and energy information of the decomposed image, respectively. This projection provides an effective tool for objective algorithm design because the detection of structural change is the prerequisite for IQA.

Different projection bases of SVD represent different spatial frequency characteristics, as shown in Fig. 2(a–c) show the 1-st, 30-th, and 70-th projection bases of the image "coinsinfountain". The lower-rank basis appears coarser and the higher-rank basis appears finer, which means more high-frequency components will appear as the rank of basis is increased. This is also similar to the general signal transforms (e.g., DFT and DCT) and coincides with the multi-channel HVS model implicitly. Thus we can relate the projection coefficients (singular values) to the weight of different structural components with a specific spatial frequency. Fig. 2(d) shows the singular value variation tendency of the original and the fastfading distorted image for all ranks. One can note that the distorted images contain much less high-frequency components than the original image, which corresponds with the fact that the dominant artefacts in fastfading distorted images are blur and ringing, which always lead to the loss of image details.

In short, SVD can be regarded as an under-complete orthonormal projection transform from the perspective of signal representation. Projection bases represent the structural components with different frequency and projection coefficients indicating the separate energy of these components. Due to the separation of structure and energy and the adaptive capture of structural components subject to a given image, the SVD-based projection is more advantageous for detecting structural information than the general image transforms. Hence, an effective IQA framework can be established through an elaborate comparison of such projection bases and coefficients.

3. Proposed metric

Inspired by the study in the previous section, we propose a novel IQA method using the SVD-based structural projection named the NN-pooled Structural Projection Similarity Index (NNSPM). The proposal works only with the luminance of images by separating luminance and chrominance in the YCbCr colour space. The flowchart depicting the proposed scheme is shown in Fig. 3, which consists of three major sections: (1) extraction of local feature vectors based on block structural projection, (2) application of the spatial pooling strategy for the global feature vector, and (3) integration of NNSPM using BPNN. The following subsections provide the details of each stage.

Fig. 1. Illustration of the subspace property: (a) original image "coinsinfountain", with $\|\boldsymbol{\Lambda}\|_E^2 = 4.7665 \times 10^9$; (b) fastfading distorted image, with $\|\boldsymbol{\Lambda}\|_E^2 = 4.6460 \times 10^9$; (c) image reconstructed from $\boldsymbol{U}$, $\boldsymbol{V}$ of (b) and singular values of (a), with $\|\boldsymbol{\Lambda}\|_E^2 = 4.7665 \times 10^9$.

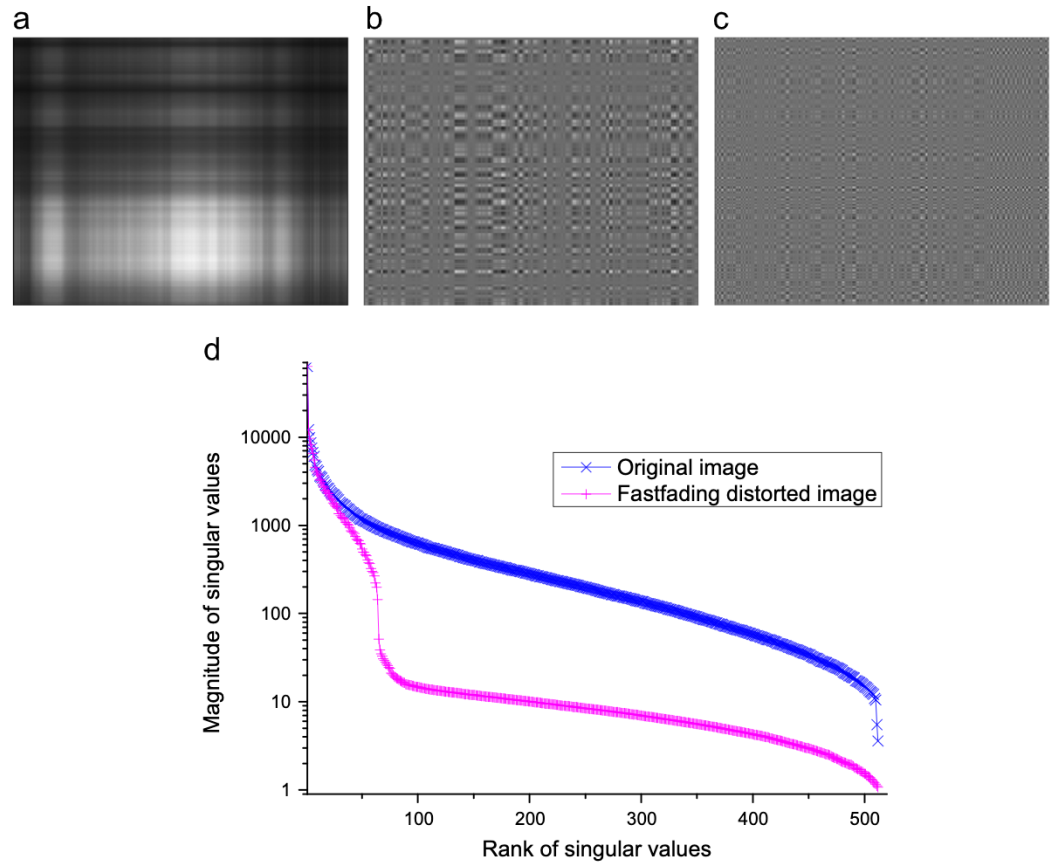

Fig. 2. Spectral characteristics of a SVD-based structural projection: (a) $\boldsymbol{u}_1\boldsymbol{v}_1^T$ of the original image "coinsinfountain"; (b) $\boldsymbol{u}_{30}\boldsymbol{v}_{30}^T$ of the original image; (c) $\boldsymbol{u}_{70}\boldsymbol{v}_{70}^T$ of the original image; (d) singular value variation tendency of the original image and the corresponding fastfading distorted image.

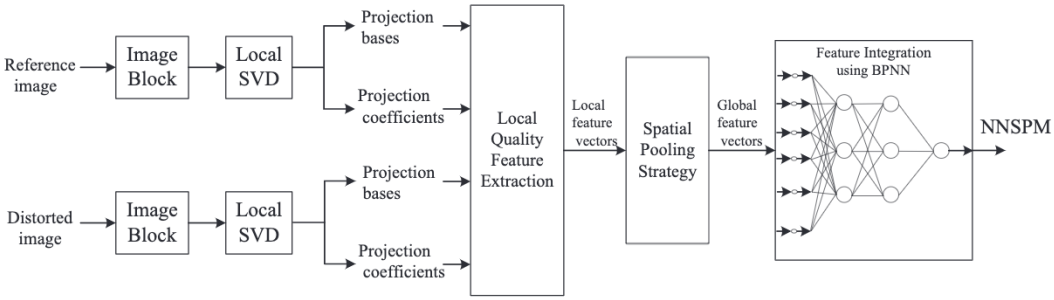

Fig. 3. The flowchart of the proposal.

3.1. Local features extraction

As illustrated in Fig. 3, the local distortion features are detected by comparing the projection bases and the projection coefficients between the local SVD of the original and distorted image blocks. The local blocks are obtained by applying an $N\times N$ half-overlapped sliding window. Denoting the corresponding local blocks of the reference image and distorted image by x and y, respectively, the SVD-based structural projections are given by

$$x=\boldsymbol{U}^x\boldsymbol{S}^x\boldsymbol{V}^{x^T}=\sum_{i=1}^{r_x}\sigma_i^x\boldsymbol{u}_i^x\boldsymbol{v}_i^{x^T}$$

$$y=\boldsymbol{U}^y\boldsymbol{S}^y\boldsymbol{V}^{y^T}=\sum_{i=1}^{r_y}\sigma_i^y\boldsymbol{u}_i^y\boldsymbol{v}_i^{y^T} \tag{3}$$

where r_x and r_y are the ranks of x and y.

Before extracting the distortion features, we first construct two vectors, $\boldsymbol{s}_x$ and $\boldsymbol{s}_y$, using the singular values of x and y, which are shown in Eq. (4).

$$\begin{aligned}\boldsymbol{s}_x&=[\sigma_1^x,\sigma_2^x\ldots\sigma_N^x]\\ \boldsymbol{s}_y&=[\sigma_1^y,\sigma_2^y\ldots\sigma_N^y]\end{aligned} \tag{4}$$

The distortion feature vector for each block consists of $N+1$ features:

$$\boldsymbol{f}=[f_0,f_1\ldots f_N] \tag{5}$$

The j-th feature, f_j, is calculated according to the following rules:

For j equals 0:

$$f_0 = \frac{|\langle \boldsymbol{s}_x, \boldsymbol{s}_y \rangle|}{\|\boldsymbol{s}_x\|_E \cdot \|\boldsymbol{s}_y\|_E} \tag{6}$$

For j from 1 to $max\{r_x, r_y\}$:

$$f_j = |\langle \boldsymbol{u}_j^x \boldsymbol{v}_j^{x^T}, \boldsymbol{u}_j^y \boldsymbol{v}_j^{y^T} \rangle| \tag{7}$$

For j from $max\{r_x, r_y\}+1$ to N:

$$f_j = 0 \tag{8}$$

As seen from Eqs. (6)–(8), all of the distortion features inherit the properties of symmetry, boundedness, normalisation within the range [0, 1], and a unique maximum, which is obtained from a perfect match between the distorted and reference images. The feature values decrease as the distortions increase.

3.2. Spatial pooling strategy

The HVS processes local regions of the image with different visual significances and represents an increased sensitivity with respect to the distortions that are present in certain regions of visual interest. Using this idea, the performance of IQA is improved by assigning different visual importance weights to local distortions, e.g., entropy weighting [29], local-contrast weighting [30], information content weighting, [16] and visual attention data-based weighting [21].

In this study, we use a visual saliency weighting approach. The adopted saliency-detection algorithm is a natural image statistics-driven method, named the spectral residual model [31]. In this theory, the authors believe that the statistical singularities in the spectrum may be responsible for anomalous regions in the image, where proto-objects are propped up. The algorithm process is implemented through the analysis of a log-spectrum representation, which can be summarised as

$$\begin{aligned} R &= \log(A) - h * \log(A) \\ S &= |\mathfrak{F}^{-1}(\exp(R + jP))| \end{aligned} \tag{9}$$

Here, A and P denote the amplitude spectrum and phase spectrum of the original image, respectively; h is a local average filter; $\mathfrak{F}^{-1}$ is the inverse DFT; and S is the obtained saliency map. A spatial pooling strategy is used subsequently to generate the $N+1$ dimensional global feature vector $\boldsymbol{F}$, which is shown in Eq. (10).

$$\boldsymbol{F} = [F_0, F_1 \ldots F_N] = \frac{\sum_{k=1}^{N_{\text{block}}} \boldsymbol{f}_k \cdot s_k}{\sum_{k=1}^{N_{\text{block}}} s_k} \tag{10}$$

where s_k denotes the mean saliency value of the k-th local image block.

To interpret the characteristics of the global feature vector, we take N equals 32 as an example without the loss of generality. The reference image "lighthouse" and the corresponding white noise (WN) and JPEG distortion images from the LIVE database are employed in the investigation. These two types of distortion are representative of additive impairments and detail losses. For each distortion, three degradation levels from severe to mild are taken into account. All of the images and the distribution diagrams of the 33-dimensional global features are presented in Fig. 4.

All of the amplitudes of the 33-dimensional features are attenuated when the original image undergoes distortions. The attenuations become more serious as the distortions increase. On the other hand, more attenuation occurs for high-dimension features than low-dimension features, because image distortions always suppress the perception of image details.

3.3. BPNN for feature integration

The last step of our proposal is the integration of the global feature vector $\boldsymbol{F}$ for a single quality score. The simplest and most widely used pooling technique is averaging. By averaging all of the features, we identify a concise IQA measure that can be referred to as the Mean Structural Projection Similarity Index (MSPM), which is given in Eq. (11). This index will be used as a control method in the experiments of Section 4.

$$MSPM = \frac{\sum_{j=0}^{N} F_j}{N+1} \tag{11}$$

However, the complicated non-linear relationships between multiple features and perception of visual quality of the HVS contribute to the non-reliability of this averaging approach. The pooling of these features can be described as a multiple nonlinear regression problem. The goal is to find a mapping function $Q = g(F_0, F_1 \ldots F_N)$ that mimics the actual target function (the subjective quality score). In the field of machine learning, neural networks (NN) have been demonstrated to have the ability to learn complex, high-dimensional, nonlinear data structures, and approximate any continuous mapping. Recently, NNs have been widely used in the processing chain, from the preprocessing/filtering level to the image understanding level [32]. Therefore, we introduce a NN-based scheme for feature pooling in this paper.

The utilised NN is a back-propagation feed-forward network (BPNN) that, consists of $N+1$ input features, two hidden layers with three and six neurons, and an output layer with one neuron. The network structure is shown in Fig. 5. Empirically, this number of artificial neurons can provide a sufficient capacity of the system and does not require a training set that is too large. Each layer performs a weighted sum of its inputs followed by a transfer function. The chosen transfer functions for all of the layers are sigmoid functions. Thus the output of the network is normalised, with range [0, 1]. Because different IQA databases use different subjective score ranges, all of the training data were normalised linearly with the theoretic maximum and minimum in the experiment. This strategy is rational due to the use of another nonlinear regression (given in Eq. (12)) that will be used to compensate for different database rating scales. For the sake of improving the generalisation capability of the network, the Bayesian regularisation (BR) training metric with early termination (ET) was adopted. It has been demonstrated that this learning metric has the advantages of fewer

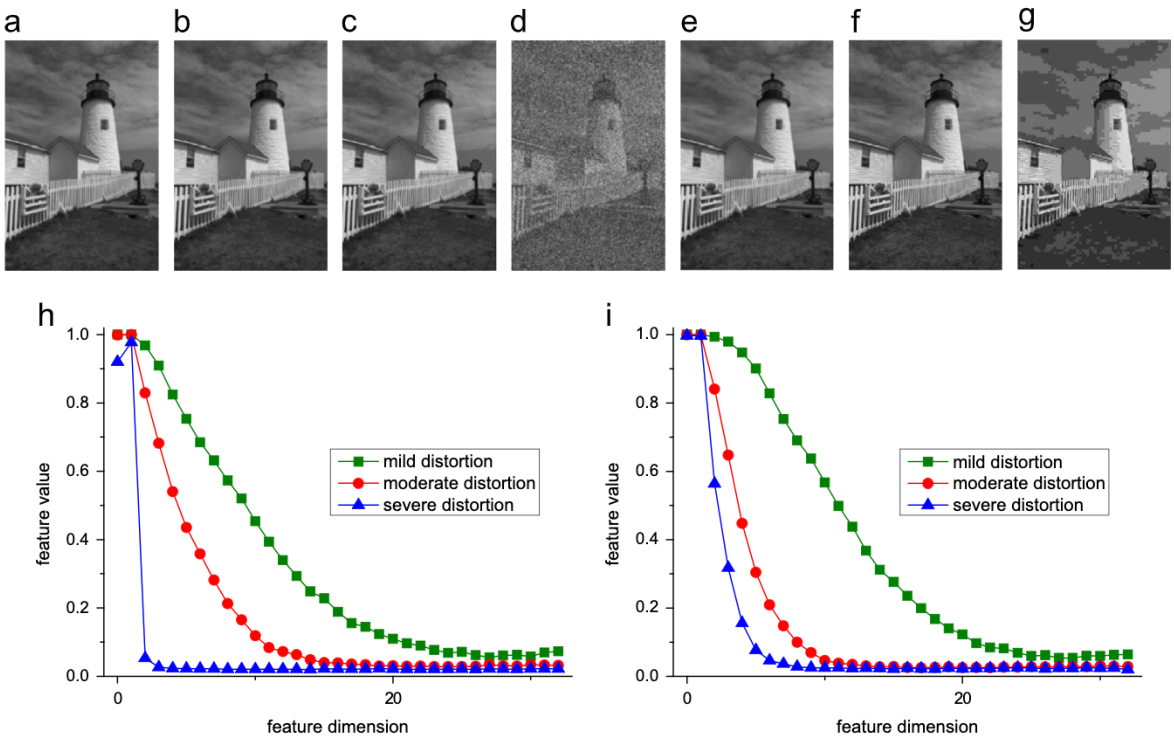

Fig. 4. Characteristics of the global feature vector: (a) original image "lighthouse"; (b) mild WN distortion; (c) moderate WN distortion; (d) severe WN distortion; (e) mild JPEG distortion; (f) moderate JPEG distortion; (g) severe JPEG distortion; (h) distribution diagrams of global features for WN distortion; (i) distribution diagrams of global features for JPEG distortion.

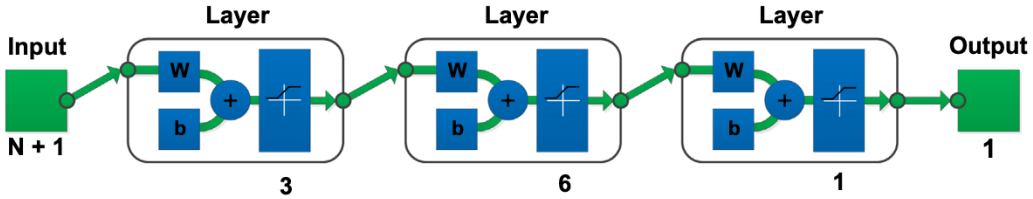

Fig. 5. Network structure of the BPNN that was used.

iterations and high training accuracy while avoiding overfitting [33].

4. Experimental results and discussion

4.1. Impact of a single feature of the global feature vector

Before comparing the proposed NNSPM approach with the state-of-the-art IQA methods, we first discuss the impact of each single feature, F_j, of the global feature vector $\boldsymbol{F}$. We calculated the Spearman Rank Correlation Coefficient (SROCC) [34] between each individual feature of all of the images and their corresponding subjective scores on the LIVE, IVC [35], and MICT [36] databases. SROCC is one of the most common parameters used in evaluating IQA performance. A higher SROCC indicates that an objective IQA achieves better performance. To determine the blocks size, we tested the average SROCC of MSPM on these databases, with the block sizes of 16, 32, 48, 64, and 80. The best performance was obtained for a 32×32 block size (average SROCC=0.919). Therefore, we set the block size at 32×32 in the following experiments. This result is reasonable to some extent, for blocks that are too small may not retain sufficient structural information and blocks that are too large cannot satisfy the local property of spatial pooling.

Fig. 6 shows the plots of SROCC versus the dimension of single features for three subjective IQA databases. In this figure, we observe that the scatter plots share similar bell shaped curves and better performances concentrate approximately between F_5 and F_{10}, suggesting that human eyes are more sensitive to the intermediate-frequency image content than high- and low-frequency image content. This finding correlates to the band-pass nature of the Contrast Sensitivity Function (CSF) [37] and demonstrates the accuracy of the previous analysis of the spatial frequency characteristics of the structural projection basis in Section 2.

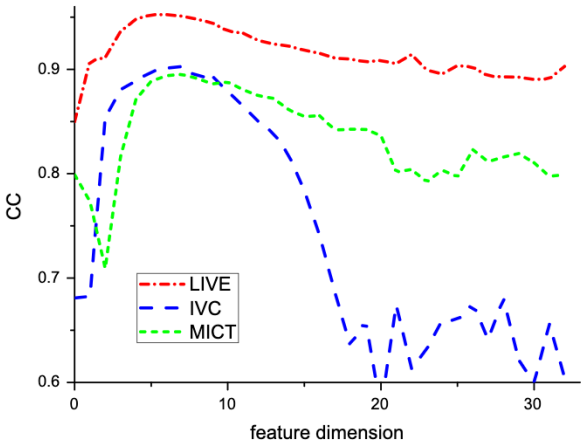

Fig. 6. Plots of the SROCC as functions of the individual feature of the global feature vector for the LIVE, IVC, and MICT databases.

Based on the inference that human visual sensitivity peaks at middle frequencies, researchers have employed weighed linear averaging to fuse IQA features [38], in which spectral varying weights are assigned over different frequency channels. However, in our study, we discovered that the learned weights of the utilised NN are not consistent with a simple linear function. This result partly reveals that relationships between multi-channel features and HVS's perceptions are more complicated than the conventional models. In the following sub-sections, we provide comprehensive performance validation results for the proposed IQA scheme, including elaborate discussions of the utilised BPNN-based feature fusion.

4.2. Overall performance and statistical significance

In this subsection, the performance of the proposed metric NNSPM is compared with the state-of-the-art IQA schemes, including PSNR, Mean SSIM (MSSIM) [13], Multi-scale SSIM (MSSSIM) [14], NNet [39], MSVD [11], VIF [10], VSNR [12], full-parameter SVD (FPSVD) [26], and the aforementioned MSPM.

This jointly subjective and objective study was undertaken using the LIVE, IVC, and MICT databases. In addition to SROCC, we also report two other performance criteria: the Root Mean Squared Error (RMSE) and the Pearson Linear Correlation Coefficient (CC) between objective results and subjective scores (difference of the mean opinion, DMOS or the mean opinion, MOS) after a five-parameter logistic regression, which is given in Eq. (12) [34], according to the video quality experts group (VQEG) phaseIITR-TV [40]. The CC and RMSE evaluate the prediction accuracy and the SROCC measures the prediction monotonicity. A better IQA measure is expected to have a higher CC and SROCC, while having a lower RMSE. Moreover, an F Statistic was performed to assess the statistical significance of the proposed metric's performance relative to the traditional metrics. The F Statistic is given by the ratio of the variance of the residuals from one IQA method to another. Values of $F > F_{critical}$ (or $F < 1/F_{critical}$) represent that, at a given confidence level, the compared method is statistically better (or worse) than the reference method (NNSPM was utilised as the reference method in this test). The reference images in the databases were excluded in our experiment to carry out an accurate nonlinear regression.

$$\text{Quality}(x) = \beta_1 \text{logistic}(\beta_2, (x-\beta_3)) + \beta_4 x + \beta_5$$
$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1+\exp(\tau x)} \quad (12)$$

The 10-fold cross-validation method was employed to train and test the traditional NN-based metric (NNet) and our proposed BPNN. The images and corresponding subjective scores of each database were randomly divided into 10 equal groups. One group was used for testing, and the remaining nine groups were used for training. The test was repeated by applying each of the 10 groups for testing. The performance evaluation of NNSPM was expressed as the average consistency of the tests over the 10 groups.

The results of the CC, RMSE, SROCC, and F-statistic are listed in Table 1, where values of best performance and $F > F_{critical}$ are shown in boldface. The confidence level used here is 99%. As shown in Table 1, the proposed scheme achieves a good consistency with subjective scores. It is the best-performer with all of three databases, except for the performance of SROCC on the MICT database. Although the performance of NNSPM is not always the best, it is only slightly worse than the best results, and the difference between NNSPM and the best performer is not statistically significant. We note that $F > F_{critical}$ occurs for more than three quarters of cases, and no case of $F < 1$ occurs, which means that the proposal is not statistically worse than any existing method with the databases under comparison.

The rational of the proposed NNSPM metric can be demonstrated by comparing the results between MSVD, FPSVD, MSPM, NNSPM, and NNet. Among the approaches based on SVD, MSVD is not a very reliable perceptual quality index, as it only introduces the change of singular values. FPSVD outperforms MSVD by measuring the degradations of both singular values and singular vectors, but it does not account for the signal representation nature of SVD and spatial sensitivities of HVS. However, MSPM employs all of these properties and performs better than MSVD and FPSVD. The best performance is yielded by NNSPM, which applies the machine learning method for feature integration instead of averaging, as used in MSPM. On the other hand, NNSPM also outperforms NNet, which uses the features of simple statistics, such as image means, derivations, and MSE, as well as a machine-learning paradigm. Although the utilised NN of NNet is similar to our metric, the extracted features are based on grey-scale statistics in the pixel domain and do not consider the perception of image structure (which is the same problem as PSNR). As mentioned above, a SVD-based image projection can represent image structural components with specific spatial frequencies. Therefore, the prediction accuracy of BPNN is better than NNet. The success of NNSPM provides a strong indicator of the effectiveness and reliability of combining the available ideas of structural distortion measurement, perceptual spatial pooling, and neural networks.

Table 1
Performance comparison and statistical significance test for 10 IQA methods on the LIVE, IVC, and MICT databases. CC: Pearson linear correlation coefficient after logistic regression; SROCC: Spearman rank correlation coefficient; RMSE: root mean squared error after logistic regression. The values of the best performance and $F > F_{critical}$ are shown in boldface.

Database	Criteria	PSNR	MSSIM	MSSSIM	NNet	VIF	VSNR	MSVD	FPSVD	MSPM	NNSPM
LIVE	CC	0.8723	0.9059	0.9496	0.9390	0.9604	0.9229	0.8933	0.9191	0.9490	**0.9664**
Fcritical = 1.1817	SROCC	0.8756	0.9122	0.9521	0.9391	0.9636	0.9271	0.8950	0.9238	0.9487	**0.9644**
1/*Fcritical* = 0.8462	RMSE	13.4021	11.6104	8.5914	9.4288	7.6387	10.5557	12.3227	10.8030	8.6420	**7.0461**
	F-statistic	**3.6179**	**2.7152**	**1.4867**	**1.7907**	1.1753	**2.2443**	**3.0585**	**2.3507**	**1.5043**	1
IVC	CC	0.7032	0.7572	0.9109	0.9324	0.8962	0.8034	0.8146	0.8169	0.9128	**0.9444**
Fcritical = 1.4110	SROCC	0.6908	0.7499	0.8989	0.9245	0.8979	0.8012	0.8075	0.8112	0.9068	**0.9339**
1/*Fcritical* = 0.7087	RMSE	0.8782	0.8068	0.5097	0.4465	0.5479	0.7354	0.7164	0.7124	0.5044	**0.4062**
	F-statistic	**4.6742**	**3.9450**	**1.5745**	1.2083	**1.8194**	**3.2777**	**3.1105**	**3.0759**	**1.5420**	1
MICT	CC	0.6487	0.8125	0.8955	0.8947	0.9092	0.8705	0.7158	0.7843	0.9012	**0.9145**
Fcritical = 1.4356	SROCC	0.6133	0.8046	0.8915	0.8829	**0.9083**	0.8601	0.6908	0.7825	0.9001	0.9044
1/*Fcritical* = 0.6966	RMSE	0.9670	0.7385	0.5656	0.5674	0.5290	0.6253	0.8871	0.7883	0.5491	**0.5125**
	F-statistic	**3.5601**	**2.0764**	1.2180	1.2257	1.0654	**1.4886**	**2.9961**	**2.3659**	1.1479	1

Table 2
Cross-distortion validation with the CC, SROCC, and RMSE or individual distortion types of the LIVE database. The values of the best or second best performance are shown in boldface.

Criteria	Distortion	PSNR	MSSIM	MSSSIM	NNet	VIF	VSNR	MSVD	FPSVD	MSPM	NNSPM
	FF	0.8926	0.9458	0.9411	0.9354	0.9696	0.9031	0.9060	0.9373	**0.9715**	**0.9737**
	Gblur	0.7841	0.8809	0.9576	0.8918	**0.9745**	0.9337	0.7223	0.8987	0.9665	**0.9677**
CC	JPEG	0.8896	0.9517	**0.9835**	0.9274	**0.9873**	0.9744	0.9472	0.9685	0.9672	0.9697
	JP2K	0.8997	0.9421	**0.9689**	0.9277	**0.9781**	0.9627	0.9383	0.9604	0.9631	0.9644
	WN	**0.9879**	0.9768	0.9862	0.9761	**0.9904**	0.9814	0.9755	0.9724	0.9797	0.9810
	FF	0.8907	0.9421	0.9408	0.9304	**0.9650**	0.9024	0.9066	0.9377	0.9647	**0.9659**
	Gblur	0.7823	0.9036	0.9414	0.8908	**0.9728**	0.9410	0.6990	0.9001	0.9577	**0.9590**
SROCC	JPEG	0.8809	0.9458	**0.9818**	0.9179	**0.9846**	0.9648	0.9364	0.9578	0.9670	0.9701
	JP2K	0.8955	0.9359	**0.9630**	0.9197	**0.9696**	0.9551	0.9374	0.9564	0.9534	0.9549
	WN	**0.9854**	0.9614	0.9774	0.9704	**0.9858**	0.9785	0.9840	0.9778	0.9727	0.9714
	FF	13.0721	9.4117	9.8060	10.2497	7.0966	12.4474	12.2747	10.1071	**6.8683**	**6.6102**
	Gblur	11.6657	8.89712	5.4164	8.5059	**4.2187**	6.7294	13.2130	8.2447	4.8227	**4.7363**
RMSE	JPEG	14.7586	9.9182	**5.8444**	12.0924	**5.1396**	7.2614	10.3362	8.0269	8.2155	7.9007
	JP2K	11.1771	8.5872	**6.3390**	9.5646	**5.3301**	6.9345	8.8587	7.1390	6.8976	6.7717
	WN	**4.4106**	6.0985	4.7111	6.1908	**3.9388**	5.4726	6.2695	6.6484	5.7028	5.5108

4.3. Robustness evaluation for untrained distortions

The robustness of the proposed metric was determined by developing a cross-distortion validation experiment. The LIVE database was utilised in this evaluation because it contains the most common distortion types, including JPEG2000 compression (JP2K), JPEG compression, White Noise (WN), Gaussian Blur (GBlur), and Fastfading transmission errors (FF). For each distortion type, we tested the prediction accuracy of NNSPM and NNet given that the training was performed only with the remaining distortion types. The experiment was repeated for the five distortion types. The results of the CC, SROCC, and RMSE between our proposed and subjective scores compared with the performance of MSSIM are shown in Table 2, which also includes the performance of the other considered objective quality measures (PSNR, MSSIM, MSSSIM, VIF, VSNR, MSVD, FPSVD, MSPM).

As observed, the proposed method predicts the DMOS effectively, even with the untrained distortions. The performance is more consistent and stable than PSNR, MSSIM, VSNR, MSVD, FPSVD, especially MSPM and NNet, which share a similar feature extraction or feature pooling strategy to our scheme, and are comparable with the leading performers (MSSSIM and VIF). Such results are expected due to the global distortion features of different distortion types share a similar distribution characteristic, which is shown in Fig. 4. In addition, the utilised BPNN establishes an optimal and generalised mapping of the features to the perceptual quality assessment without any prior knowledge of the distortion type. This is most likely a stronger and more straightforward demonstration of the feature extraction and pooling strategy of our scheme.

4.4. Cross-database validation

To validate the generalisation capabilities of the proposed metric, an extended experiment was explored using cross-database validation. The LIVE database was used for training, and the other two databases (IVC and MICT) were used for testing because the LIVE database is much larger than IVC and MICT. Table 3 lists the CC, SROCC, and RMSE on the test group. It can be observed that our approach achieves good prediction for untrained databases, with CC and SROCC both higher than 0.9. This result outperforms most of the state-of-the-art metrics in Table 1. As image content and distortion types vary across databases, the trained mapping $Q = g(F_0, F_1 \ldots F_N)$ is dependent only on

Table 3
Performance validation for cross-databases with the CC, SROCC, and RMSE.

Criteria	IVC	MICT
CC	0.9184	0.9057
SROCC	0.9088	0.9034
RMSE	0.4888	0.5387

the loss of visual quality. Thus, it again confirms the robustness and the rational of using BPNN in modelling complex input–output relationships of the HVS. Moreover, this is crucial from a practical perspective because most of the computational power is spent in the off-line training phase. A real-time IQA can, therefore, be implemented using this well-trained NN-based system.

5. Conclusion

In this paper, we propose a novel IQA model that incorporates the available ideas of structural distortion measurement, perceptual spatial pooling, and machine learning. From the perspective of signal representation, the SVD-based structural projection is introduced to select the local structural distortions of different spatial frequency channels. The local feature vectors are pooled into a global vector by using a spectral residual-based visual saliency weighting algorithm. We also use BPNN to mimic the high-level HVS processing mechanism that maps the global feature vector to a single perceptual quality score. By conducting experiments on three subjective IQA databases that span a wide variety of visual and distortion content, the proposed metric achieves a better performance that is consistent with human judgment when compared to the state-of-the-art IQA methods. In addition, we demonstrate the robustness, stability, and generalisation of the proposed BPNN scheme to untrained distortions and databases. Further work will focus on taking into account the chrominance information and extending this method to video quality assessment.

Acknowledgements

This work is supported by National Grand Fundamental Research 973 Program of China under Grant No. 2010 CB731904, and National Nature Science Foundation of China under Grant No. 61172154. We also would like to thank the anonymous referees for their thoughtful comments that helped us to improve this paper.

Appendix A

In this appendix, we present the specific derivation procedures of inference (1), (2), and (3) in Section 2. These inferences prove the orthonormal nature of the projection bases and the Parseval's theorem of the projection coefficients. This helps us compare the SVD with general orthogonal transforms, such as DFT and DCT, and illustrate the advantage of SVD for extracting structural information from the images.

The derivation procedure of inference (1):

$$\begin{aligned}\|\boldsymbol{u}_i\boldsymbol{v}_i^T\|_E &= \sqrt{\textstyle\sum_{s=1}^{m}\sum_{t=1}^{n}(\boldsymbol{u}_i\boldsymbol{v}_i^T)_{s,t}^2}\\ &= \sqrt{\textstyle\sum_{s=1}^{m}\sum_{t=1}^{n}u_{is}^2\cdot v_{it}^2}\\ &= \sqrt{\textstyle\sum_{s=1}^{m}(u_{is}^2\cdot\sum_{t=1}^{n}v_{it}^2)}\end{aligned}$$

where u_{is} (v_{it}) indicates the element of the i-th column and s-th (t-th) row of matrix $\boldsymbol{U}$ ($\boldsymbol{V}$). Because $\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrixes, it is obvious that $\sqrt{\sum_{s=1}^{m}u_{is}^2}=1$ and $\sqrt{\sum_{t=1}^{n}v_{it}^2}=1$ for $i=1, 2\ldots r$. Hence, $\|\boldsymbol{u}_i\boldsymbol{v}_i^T\|_E=1$.

The derivation procedure of inference (2):

$$\begin{aligned}\langle\boldsymbol{u}_i\boldsymbol{v}_i^T,\boldsymbol{u}_j\boldsymbol{v}_j^T\rangle &= \textstyle\sum_{s=1}^{m}\sum_{t=1}^{n}[(\boldsymbol{u}_i\boldsymbol{v}_i^T)_{s,t}\cdot(\boldsymbol{u}_j\boldsymbol{v}_j^T)_{s,t}]\\ &= \textstyle\sum_{s=1}^{m}\sum_{t=1}^{n}(u_{is}\cdot v_{it}\cdot u_{js}\cdot v_{jt})\\ &= \textstyle\sum_{s=1}^{m}u_{is}\cdot u_{js}\cdot(\sum_{t=1}^{n}v_{it}\cdot v_{jt})\end{aligned}$$

$\boldsymbol{U}$ and $\boldsymbol{V}$ are both orthogonal matrixes. Therefore, $\sum_{s=1}^{m}u_{is}\cdot u_{js}=0$ and $\sum_{t=1}^{n}v_{it}\cdot v_{jt}=0$ when $i\neq j$. Then, $\langle\boldsymbol{u}_i\boldsymbol{v}_i^T,\boldsymbol{u}_j\boldsymbol{v}_j^T\rangle=0$.

The derivation procedure of inference (3):

$$\begin{aligned}\|\boldsymbol{\Lambda}\|_E^2 &= tr(\boldsymbol{\Lambda}^T\boldsymbol{\Lambda})\\ &= \textstyle\sum_{i=1}^{r}\lambda_i\\ &= \textstyle\sum_{i=1}^{r}\sigma_i^2\end{aligned}$$

where λ_i are eigenvalues of matrix $\boldsymbol{\Lambda}^T\boldsymbol{\Lambda}$.

References

[1] X. Gao, W. Lu, D. Tao, X. Li, Image quality assessment and human visual system, Proc. SPIE 7744 (2010). (77440Z-1-7740Z-10).
[2] W. Lin, C.C. Jay Kuo, Perceptual visual quality metrics: a survey, J. Vis. Commun. Image Represent. 22 (4) (2011) 297–312.
[3] K. Thung, P. Raveendran, A survey of image quality measures, International Conference for Technical Postgraduates (December 2009) 1–4.
[4] A.M. Eskicioglu, Quality measurement for monochrome compressed images in the past 25 years, IEEE Int. Conf. Acoust. Speech Signal Process. 4 (2000) 1907–1910.
[5] Z. Wang, A.C. Bovik, L. Lu, Why is image quality assessment so difficult, IEEE Int. Conf. Acoust. Speech Signal Process. 4 (2002) 3313–3316.
[6] W. Lin, M. Narwaria, Perceptual image quality assessment: recent progress and trends, Proc. SPIE 7744 (2010). (774403-1-77403-9).
[7] T.N. Pappas, R.J. Safranek, J. Chen, Perceptual criteria for image quality evaluation, in: A.C. Bovik (Ed.), Handbook of Image and Video Processing, Academic Press, San Diego, 2000, pp. 669–684.
[8] Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? A new look at signal fidelity measures, IEEE Signal Process. Mag. 26 (1) (2009) 98–117.
[9] H.R. Sheikh, A.C. Bovik, G.D.E. Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128.
[10] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[11] A. Shnayderman, A. Gusev, A.M. Eskicioglu, An SVD-based grayscale image quality measure for local and global assessment, IEEE Trans. Image Process. 15 (2) (2006) 422–429.
[12] D.M. Chandler, S.S. Hemami, SNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298.
[13] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[14] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multi-scale structural similarity for image quality assessment, Asilomar Conference on Signals, Systems, and Computers (2003) 1398–1402.
[15] M.P. Sampat, Z. Wang, S. Gupta, A.C. Bovik, M.K. Markey, Complex wavelet structural similarity: a new image similarity index, IEEE Trans. Image Process. 18 (11) (2009) 2385–2401.
[16] Z. Wang, Q. Li, Information content weighting for perceptual image quality assessment, IEEE Trans. Image Process. 20 (5) (2011) 1185–1198.
[17] A. Hu, R. Zhang, X. Zhan D. Yin, Image Quality assessment incorporating the interaction of spatial and spectral sensitivities of HVS, in: IASTED International Conference on Signal and Image Processing (2011) 1–7.
[18] L. Capodiferro, G. Jacovitti, E.D. Di Claudio, Two-dimensional approach to full-reference image quality assessment based on positional structural information, IEEE Trans. Image Process. 21 (2) (2012) 505–516.
[19] M. Liu, X. Yang, Image quality assessment using contourlet transform, Opt. Eng. 48 (10) (2009). (107201-1-107201-10).
[20] Z. Haddad, A. Beghdadi, A. Serir, A. Mokraoui, Image quality assessment based on wave atoms transform, in: IEEE International Conference on Image Processing (2010) 305–308.
[21] L. Hantao, I. Heynderickx, Studying the added value of visual attention in objective image quality metrics based on eye movement data, in: IEEE International Conference on Image Processing (2009) 3097–3100.
[22] P. Le Callet, C. Viard-Gaudin, D. Barba, A convolutional neural network approach for objective video quality assessment, IEEE Trans. Neural Netw. 17 (5) (2006) 1316–1327.
[23] M. Narwaria, W. Lin, Objective image quality assessment based on support vector regression, IEEE Trans. Neural Netw. 21 (3) (2010) 515–519.
[24] H.S. Prasantha, H.L. Shashidhara, K.N. Balasubramanya Murthy, Image compression Using SVD, Int. Conf. Comput. Intell. Multimed. Appl. 3 (2007) 143–145.
[25] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process. 54 (11) (2006) 4311–4322.
[26] R. Wang, Y. Cui, Y. Yuan, Image quality assessment using full-parameter singular value decomposition, Opt. Eng. 50 (5) (2011). (057005-1-057005-9).
[27] H.R. Sheikh, Z. Wang, L. Cormack, A.C. Bovik, LIVE Image Quality Assessment Database Release 2, 2003, Available: ⟨http://live.ece.utexas.edu/research/quality⟩.
[28] A.V. Oppenheim, A.S. Willsky, Signals and Systems, NJ: Prentice-Hall Inc., Englewood Cliffs, 1983.
[29] D.V. Rao, I.R. Bahu, L.P. Reddy, Image quality assessment complemented with visual regions of interest, in: International Conference on Computing: Theory and Applications (2007) 681–687.
[30] S. Rezazadeh, S. Coulombe, A novel approach for computing and pooling Structural Similarity index in the discrete wavelet domain, in: IEEE International Conference on Image Processing (2009) 2209–2212.
[31] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, in: IEEE Conference on Computer Vision and Pattern Recognition (2007) 1–8.
[32] M. Egmont-Petersen, D. de Ridder, H. Handels, Image processing with neural networks – a review, Pattern Recognit. 35 (10) (2002) 2279–2301.
[33] F. Dan Foresee, M.T. Hagan, Gauss-Newton approximation to Bayesian learning, Int. Conf. Neural Netw. 3 (1997) 1930–1935.
[34] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3441–3451.
[35] A. Ninassi, P. Le Callet, F. Autrusseau, Subjective Quality Assessment-IVC Database, 2005, Available: ⟨http://www2.irccyn.ecnantes.fr/ivcdb⟩.
[36] Y. Horita, K. Shibata, Y. Kawayoke, Z.M.P. Sazzad, MICT Image Quality Evaluation Database, 2000, Available: ⟨http://mict.eng.u-toyama.ac.jp/mict/index2.html⟩.
[37] J. Mannos, D. Sakrison, The effects of a visual fidelity criterion of the encoding of images, IEEE Trans. Inf. Theory 20 (4) (1974) 525–536.
[38] C. Yang, W. Gao, L. Po, Discrete wavelet transform-based structural similarity for image quality assessment, in: IEEE International Conference on Image Processing (2008), 377–380.
[39] A. Bouzerdoum, A. Havstad, A. Beghdadi. Image quality assessment using a neural network approach, in: Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology (2004) 330–333.
[40] VQEG: Final Report from the video quality experts group on the validation of objective models of video quality assessment, FR-TV Phase II, August 2003, Available: ⟨http://www.vqeg.org/⟩.