SVD-Based Quality Metric for Image and Video Using Machine Learning

Manish Narwaria and Weisi Lin, *Senior Member, IEEE*

***Abstract*—We study the use of machine learning for visual quality evaluation with comprehensive singular value decomposition (SVD)-based visual features. In this paper, the two-stage process and the relevant work in the existing visual quality metrics are first introduced followed by an in-depth analysis of SVD for visual quality assessment. Singular values and vectors form the selected features for visual quality assessment. Machine learning is then used for the feature pooling process and demonstrated to be effective. This is to address the limitations of the existing pooling techniques, like simple summation, averaging, Minkowski summation, etc., which tend to be ad hoc. We advocate machine learning for feature pooling because it is more systematic and data driven. The experiments show that the proposed method outperforms the eight existing relevant schemes. Extensive analysis and cross validation are performed with ten publicly available databases (eight for images with a total of 4042 test images and two for video with a total of 228 videos). We use all publicly accessible software and databases in this study, as well as making our own software public, to facilitate comparison in future research.**

***Index Terms*—Image structure, singular value decomposition (SVD), support vector regression (SVR), visual quality assessment.**

I. Introduction

THE RAPID growth of digital imaging and visual communication technologies has led to a large number of applications that produce digital images and videos. Visual signals can be affected by a wide variety of distortions during the process of acquisition, compression, processing, transmission, and reproduction, which generally results in loss of visual quality. In most cases, since humans are the end receivers, they are the best judges of perceptual visual quality. Therefore, subjective assessment is the accurate and reliable way of assessing visual quality if the number of subjects is sufficiently large. However, subjective assessment is cumbersome, expensive, and unsuitable for in-service and real-time applications. Therefore, objective visual quality assessment has attracted significant attention in the recent years because it is a useful tool in many image and video processing systems. Today, picture quality assessment plays a central role [3], [12], [49], [76] in shaping most (if not all) visual processing algorithms and systems, as well as their implementation, optimization, and testing. For instance, measuring image quality enables the parameter adjustment of image processing techniques to maximize image quality or to reach a given quality in applications like image coding. Another practical use of image quality assessment (IQA) can be found in the area of information hiding [69], where secret messages are embedded into images so that an unauthorized user cannot detect the hidden messages. Since such an embedding process will degrade the image quality, an IQA metric can help in guiding the optimization process between the desired quality and the strength of message to be embedded. It is also widely used to evaluate/compare the performance of processing systems and/or optimize processing algorithms. For example, the well-known IQA metric structural similarity index measure (SSIM) [11] has recently been used as the optimization criterion in H.264 video coding [70], [71]. Other examples of technological dependence upon IQA include signal acquisition, synthesis, enhancement, transmission, storage, retrieval, reconstruction, authentication, display, and printing. Furthermore, a picture quality metric can be used as the preprocessor in many other applications, such as object recognition, video summarization, and so on.

The simplest and most widely used quality metrics are the mean square error (MSE) and the peak signal-to-noise ratio (PSNR); however, they can be poor predictors of visual quality [1], particularly when noise is not additive. The major reason for the overall poor performance of MSE (or PSNR) is its assignment of equal importance to all the changes in a visual signal (image or video) regardless of their perceptual significance. Objective evaluation of picture quality in line with the human perception is a difficult task [2], [3], [32] due to the complex multidisciplinary nature of the problem (related to physiology, psychology, vision research, and computer science) and the limited understanding of the human visual system (HVS) mechanism. There has not been a clear-cut and general scheme so far that can account for all the related characteristics of the HVS (please refer to [3] and [76] for recent reviews). There exist two types of approaches [49] for developing visual quality metrics: 1) the vision based models [4], [5], and 2) those based on extraction and analysis of features in images [3], [11], [23], [50], [51]. More recent research effort has been directed to the second type because the first type involves expensive computation and difficulties due to the gap between vision research [usually with simplistic, single stimulus (or a small number of stimuli)] and engineering implementation (for real-world complex stimuli).

Manuscript received April 27, 2010; revised April 6, 2011 and June 12, 2011; accepted July 7, 2011. Date of publication September 29, 2011; date of current version March 16, 2012. This work was supported by Singapore Ministry of Education Academic Research Fund Tiers 1 and 2 under Grant T208B1218. This paper was recommended by Associate Editor S. Singh.

The authors are with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: mani0018@ntu.edu.sg; wslin@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2011.2163391

II. Key Issues in Objective Visual Quality Assessment

Assessment of perceptual visual quality can be considered as a two-stage process consisting of feature extraction followed by feature pooling into a single number to represent the quality score. As for the first stage, features selected have to be an effective representative of visual quality variations, while the second stage determines the relationship among different features and the perceived visual quality.

Pertaining to the issue of feature extraction, in [6], a study was conducted to investigate and analyze the distortion criteria of the human viewing. It was found that for pictures where distortion is greater at edges, the MSE (or one of its relatives) is less satisfactory. Similar results have also been demonstrated in [7] and [17], where distortions at the edges have been differentiated. As a result, image structure needs to be accounted for effective visual quality assessment. Therefore, during the recent years, there has been a growing interest to take image structure into account for picture quality evaluation, because structural properties play a big role in the human perception [2], [8], [12] as well as image content recognition [13]. Evaluating the loss of structure is therefore expected to give good assessment of visual quality. A well-known metric is the SSIM [10], [11], which is based on the idea of equating the perceived image distortion to the measurement of structural distortion. In SSIM, the mean of quality scores of individual image blocks gives the overall image quality score.

Some other image quality assessment metrics [14], [16], [23], [50], [65] have also been proposed. The metric known as M block singular value decomposition (MSVD) [23] evaluates quality of each image block based on the error in singular values of the block as a result of singular value decomposition (SVD). The overall image quality score is computed as the average absolute difference between each block's error and the median error over all blocks. Another improved scheme has been proposed in [50] to measure the change in singular vectors of the reference and distorted images to compute the structural changes, and the overall quality is determined by a Minkowski summation. A no-reference metric using the SVD of local image gradients has been proposed in [79] and used for proper selection of the parameters of image denoising algorithms. The authors in [72] proposed an SVD-based method in which the difference between the reflection coefficients (obtained by projecting the two images onto the right singular vectors) of the original and distorted images is used for quality assessment. The method proposed in [74] also projects the distorted image on the singular vectors of the original image and uses a referee matrix of the distorted image to assess quality. In [16], the Harris response has been used to describe the geometric structure on a pixel-by-pixel basis with the overall quality score being computed by a simple averaging over all the image pixels. A reduced-reference metric has been proposed in [51] that extracts structural features from the images. The overall quality score is then computed as a weighted sum of the features where the weights have been determined by subjective experiments. The authors in [77] and [78] have explored the combination of multiscale SSIM, visual information fidelity (VIF) [34], and reflection SVD (R-SVD) [72] algorithms to assess image quality.

As for the issue of feature pooling (also known as error pooling), literature survey shows that scant research effort has been directed to develop effective cognitive models to map the features into a quality score, with the major reason being the complexity and limited knowledge about the HVS. Researchers have employed techniques like simple summation-based fusion, Minkowski combination, linear (i.e., weighted) combination, etc., to fuse the visual features into a quality score, and some examples have already been given. These pooling techniques, however, impose constraints on the relationship between the features and the quality score. A simple summation or averaging of features implicitly constraints the relationship to be linear. A weighted summation requires the determination of appropriate weighting coefficients, and there is no general method available for this. Subjective experiments may be used to compute the weights [51], but such a method is less consistent and unsuitable for real-time applications. The use of Minkowski summation for the spatial pooling of the features/errors implicitly assumes that errors at different locations are statistically independent. In addition, there is no systematic method to determine the proper/optimal value of the Minkowski summation exponent and is generally determined experimentally. Another method has been developed [21], which involves weighting quality scores as a monotonic function of quality. The weights are determined by local image content, assuming the image source to be a local Gaussian model and the visual channel to be an additive Gaussian model. However, there is a lack of convincing ground for these assumptions. Recently, two pooling strategies have been proposed [52] for the SSIM metric. Instead of using a simple mean as the overall quality score, these approaches attempt to weigh the quality scores of different blocks based on visual importance. The first strategy is based on the idea that lower quality regions in images attract more attention than those with higher quality; the second strategy uses visual attention (VA) to provide weighting [53]–[57], which is based on the idea that certain regions attract more human attention than others. The strategy of feature pooling using VA while intuitive may suffer from drawbacks due to the fact that it is not always easy to find regions that attract VA. Furthermore, improvement in quality prediction by using VA is not yet clearly established and still open to further investigations [53], [57]. One reason for this is that the perception in still images varies with allowed observation time. That is, if an observer has time long enough to perceive an image, then every point of the image can eventually become the attention center. This may render direct VA-based pooling less effective. In summary, the existing feature pooling techniques tend to suffer from one or more drawbacks, and there is a need for a more systematic and effective feature pooling strategy. We believe that multiple features jointly affect the HVS's perception of visual quality, and their relationship with the overall quality is possibly nonlinear and difficult to be determined *a priori*.

The rest of this paper is organized as follows: Section III presents an overview of the proposed algorithm based upon SVD for feature detection and machine learning technique for feature pooling, as a more comprehensive approach in comparison to the existing relevant work [23] and our initial work [50], [63], with the justification for each stage. Section IV discusses the theoretical and experimental aspects of characterization of images by SVD. In Section V, we discuss the details of the proposed visual quality metric using support vector regression (SVR). We describe the databases, the training, and the test methodology in Section VI. Substantial experimental results and the related analysis are presented in Section VII. Finally, Section VIII gives the concluding remarks.

III. Overview of the Proposed Algorithm

In this paper, we tackle the two stages of the problem earlier stated. First, an effective and general representation of the structural changes is devised for feature detection in visual signal since as aforementioned the HVS is sensitive to such changes. In addition, luminance changes and the related changes in texture (roughness/smoothness variation) also affect the picture quality and need to be considered along with the structural changes. Although structural and textural changes all need to be accounted for quality evaluation, they have to be distinguished (or decomposed) first since they deserve different treatments (otherwise, the mistake in MSE or PSNR is repeated). The proposed scheme is based on SVD for feature detection. We propose to quantify the visual distortions by using singular vectors and values together. Both singular vectors and values are the features to be detected in the proposed scheme [as a more reasonable and comprehensive approach than using singular values alone (like in [23]) or even singular vectors alone (like in [50] and [63])] and are differentiated for their effects on perceptual quality.

In addition, to overcome the limitations of the existing pooling schemes, we propose the use of a machine learning technique for feature fusion; this is because such a technique is general, more systematic, and reasonable, and the related model parameters (weights) are estimated via training from the sufficient available data (i.e., the substantial ground truth). Given the strong theoretical foundations and proven success of machine learning techniques in numerous applications (such as face detection [18], handwriting/signature verification [19], video surveillance [20], robot tutoring [40], speech quality assessment [66], and so on), we believe that it can be exploited for IQA. In contrast to the existing pooling methods, a machine learning technique in visual quality evaluation helps in avoiding assumptions on the relative significance and relationship of different distortion statistics (i.e., feature changes). In our opinion, it will be useful in determining the underlying complex relationship between a set of visual features and the perceptual quality. Since the required weights/parameters for pooling the features will be determined by training with sufficient data, it can help to overcome the limitations of the existing pooling schemes. There has been some early work in applying machine learning techniques for visual quality evaluation. In [58] and [59], objective quality assessment of video using neural networks (NNs) has been reported, whereas the use of NNs has been demonstrated in [60] and [61] for image cases. However, in these approaches, effort has not been directed to justify the feature selection, which is very important in machine learning [62], and overall, machine learning in visual quality evaluation remains a largely uninvestigated area. In our previous work [63], we had briefly introduced the idea of using a machine learning technique in image quality assessment, and encouraging initial results were obtained. In this paper, we provide a more comprehensive and general visual quality metric using machine learning and full SVD.

The contributions of the current work in comparison with our previous work [50], [63], as well as the relevant existing work [23], are as follows:

1) A comprehensive analysis of SVD has been presented for visual quality assessment.A new visual quality metric is proposed and demonstrated with both singular values and vectors as visual features and machine learning technique for feature pooling.
2) A thorough set of experimental results, related analysis, and cross validation (CV) are done to provide a strong ground for the proposed approach. This involves the use of ten publicly available independent databases (eight for images and two for videos).
3) To alleviate the difficulty of repeating and comparing different metrics in future research, we have used all publicly accessible software and databases so that the results reported in this paper can be reproduced easily.

In addition, the Matlab implementation of the proposed metric and other useful results from this study are made publicly available[1] to facilitate future comparisons and study in the said area.

IV. Feature Extraction With SVD

Visual features must effectively be extracted for objective perceptual quality assessment. Various transforms like discrete Fourier transform (DFT), discrete cosine transform (DCT), discrete wavelet transform (DWT), contourlet transforms, etc., can be used. In general, any 2-D transform (SVD, DFT, DCT, etc.) decomposes the image into several basis images weighted by transformation coefficients. Visual quality can be assessed by measuring the changes in transformation coefficients [14]–[16], [23], [65]. For example, in [65], image quality was predicted by computing the difference between frequency-domain coefficients of the original and distorted images. For frequency-domain transforms like DFT and DCT, the basis images (accounting for image structure) are same for all the images, so the changes in visual signal can be captured only by the transformation coefficients. On the contrary, the basis images for SVD are unique for each image and are expected to be able to represent the structure of an individual image better; hence, any change caused in an image is reflected in the individualized basis images with SVD. Due to this, SVD is more advantageous for capturing structural components in the visual signal. As stated before, effective differentiation of structural change is the prerequisite for its deserved treatments in visual quality evaluation to remedy the mistake in MSE/PSNR and other existing metrics.

The SVD [25] of an image matrix $\boldsymbol{X}$ (size $r \times c$) yields the left singular vector matrix $\boldsymbol{U}$, the right singular vector matrix

[1] [Online]. Available: http://www.ntu.edu.sg/home/wslin/codes_smc.rar

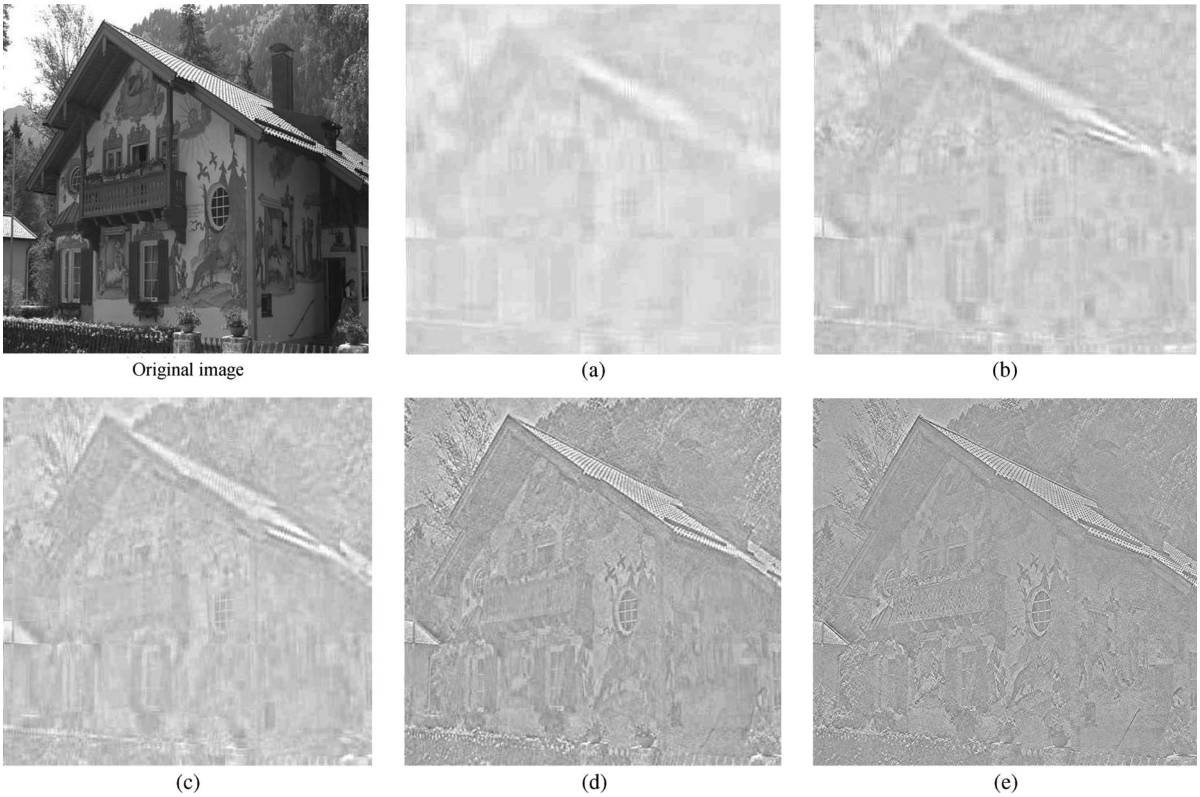


Fig. 1. $\mathbf{X_z}$ as defined by (3) for different z values. (a) $z = 10$. (b) $z = 20$. (c) $z = 30$. (d) $z = 100$. (e) $z = 512$.

$\boldsymbol{V}$, and the diagonal matrix of singular values $\boldsymbol{\sigma}$, i.e.,

$$\begin{aligned}\boldsymbol{U} &= [\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \ldots \boldsymbol{u}_r]\\ \boldsymbol{V} &= [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \ldots \boldsymbol{v}_c]\\ \boldsymbol{\sigma} &= \mathrm{diag}(\sigma_1, \sigma_2, \ldots \sigma_t)\end{aligned}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are column vectors, whereas σ_k is a singular value $(i = 1, 2, \ldots, r,\ j = 1, 2, \ldots c,\ k = 1, 2, \ldots t,$ and $t = min(r, c)$.). The singular values appear in descending order, i.e., $\sigma_1 > \sigma_2 > \ldots \sigma_t$.

A. Analysis of Singular Vectors

Any row of $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{p}_i = \sum_k u_{ik}\sigma_k \boldsymbol{v}_k^T. \tag{1}$$

Similarly, any column of $\boldsymbol{X}$ can be expressed as

$$\boldsymbol{q}_j = \sum_k \boldsymbol{u}_k \sigma_k v_{jk}. \tag{2}$$

Therefore, $\boldsymbol{p}_i$ is a linear combination of the right singular vectors $\boldsymbol{v}_j$, and $\boldsymbol{q}_j$ is a linear combination of the left singular vector $\boldsymbol{u}_i$.

The matrix $\boldsymbol{U}\boldsymbol{V}^T$ can be interpreted as the ensemble of the basis images, whereas the singular values $\boldsymbol{\sigma}$ are the weights assigned to these basis images. The image structure can therefore be represented as

$$\boldsymbol{X}_z = \sum_{i=1}^{z} \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{3}$$

where $z (z \leq t)$ is the number of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ pairs used. Each basis image (i.e., $\boldsymbol{u}_i\boldsymbol{v}_i^T$) specifies a layer of the image geometry, and the sum of these layers denotes the complete image structure. The first few singular vector pairs account for the major image structure, whereas the subsequent $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ account for the finer details in the image. We illustrate this point through an example shown in Fig. 1, where the image size is 512 × 512, and thus, $t = 512$. We can see that the first 20 pairs $(z = 20)$ of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ [i.e., $i = 1$ to 20 in (3)] capture the major image structure, and the subsequent pairs of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ signify the finer details in image structure. As an increasing number of $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ pairs are used, the finer image structural details appear. $\boldsymbol{U}$ and $\boldsymbol{V}$ can therefore be used to represent the structural elements in images.

Because $\boldsymbol{V}$ is square, it is also row orthogonal. We can write the SVD of $\boldsymbol{X}$ as

$$\boldsymbol{X}_{i,j} = \sum_k u_{ik}\sigma_k v_{jk}^T = \sum_k c_{ik} v_{jk}^T. \tag{4}$$

We can compare this with the DFT, which decomposes the original data into an orthogonal basis that can be expressed as follows:

$$\boldsymbol{X}_{i,j} = \sum_k b_{ik} e^{\mathbf{i}2\pi jk/r}. \tag{5}$$

We can see from (4) and (5) that SVD is similar to DFT in the sense that the cyclical term $e^{\mathbf{i}2\pi jk/r}$ is replaced by the normalized vector term $\boldsymbol{v}_{jk}^T$. Although the coefficient matrix $\boldsymbol{C} = \{c_{ik}\}$ of SVD is orthogonal (since $\boldsymbol{U}$ is orthogonal), the coefficient matrix $\boldsymbol{G} = \{g_{ik}\}$ of the DFT is not orthogonal

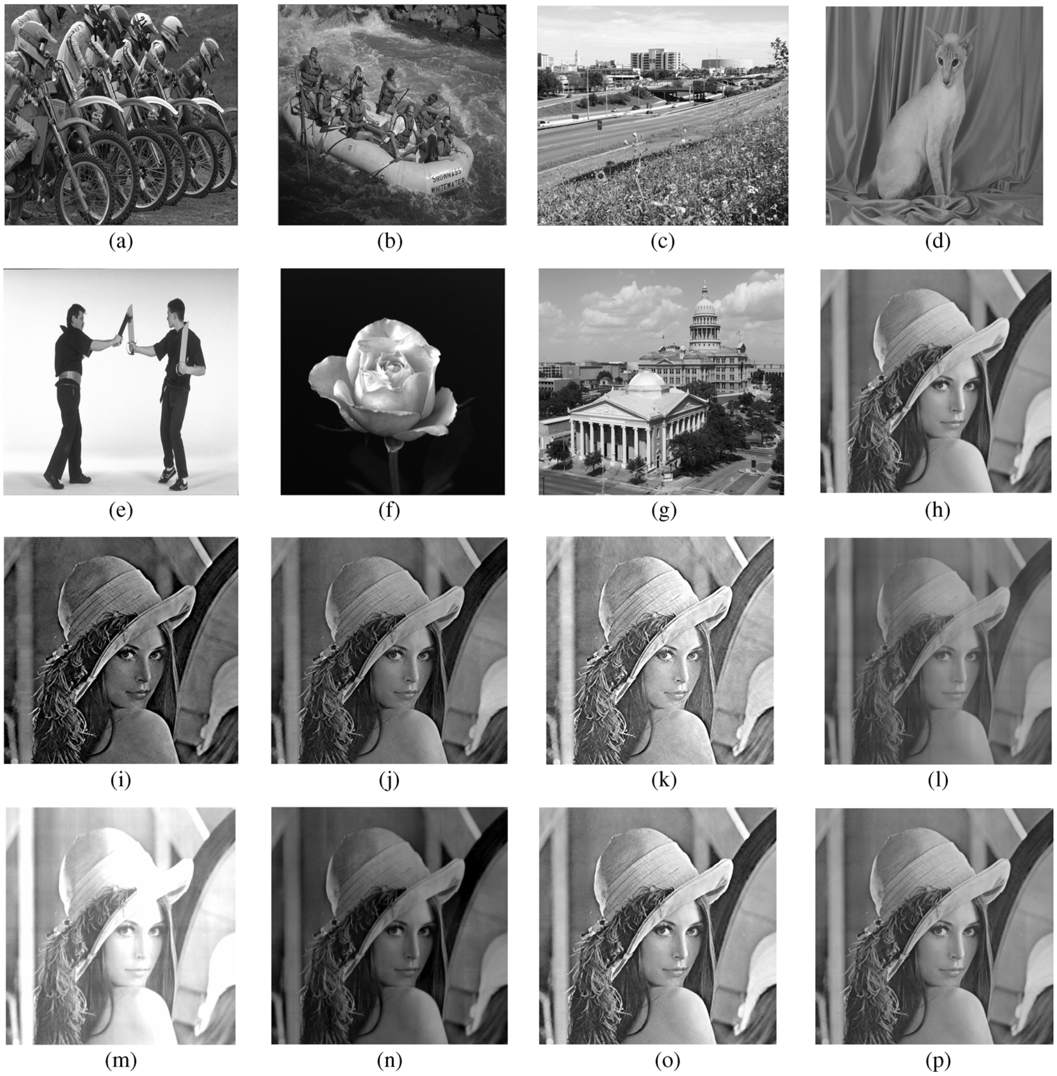


Fig. 2. Effect of changing $\boldsymbol{\sigma}$ in images. Images (i)–(o) are constructed from $\mathbf{U}$, $\mathbf{V}$ of original "Lena" image (h), and the $\boldsymbol{\sigma}$ matrices of images from (a)–(g), respectively. Image (p) is constructed from $\mathbf{U}$, $\mathbf{V}$ of original "Lena" image (h), and the average of $\boldsymbol{\sigma}$ matrices of images from (a)–(g).

in general. Nevertheless, this demonstrates that the SVD is similar to the DFT, where the basis images are determined in a very specific way from image data rather than being given at the outset as for the DFT. In view of the analogy between SVD and DFT, the first few singular vectors denote the low frequency components of the image, whereas the subsequent vectors account for the higher frequency, as can be seen in Fig. 1. We can see that using the first 10 or 20 vectors, mainly the low frequency components are visible. The high-frequency components appear as the number of vectors is increased. The major advantage of using SVD in comparison with DFT is that the basis images adaptively defined in (3) leads to the possibility of representing the image structure better.

From the perturbation analysis theory [44], [45], $\boldsymbol{U}$ and $\boldsymbol{V}$ are found to be sensitive to perturbation. Therefore, any changes introduced in the image (due to distortion) affect the singular vectors significantly. The sensitivity of singular vectors can be exploited to assess the visual quality since the changes in visual quality are characterized by structural changes. For example, blurring artifact affects the structure in an image by damaging edges and high-frequency regions. The commonly used JPEG image compression scheme damages the structure by introducing blockiness; JPEG-2000, which is a more recent compression standard based on the wavelet transform, makes images blurry along the edges and in high-frequency areas. As shown in [63], different types of distortions (added noise, blurring, and JPEG/JPEG-2000 compression) affect the structure of the visual signal represented by $\boldsymbol{U}$ and $\boldsymbol{V}$.

Since the changes in adaptively determined $\boldsymbol{U}$ and $\boldsymbol{V}$ account for such structural changes, they provide an effective basis for assessing visual quality.

B. Further Analysis With Singular Values

The $\boldsymbol{\sigma}$ values are mainly related to the luminance changes in images, as shown in Fig. 2: (a)–(h) show eight test images; (i)–(o) show the "Lena" image (h) constructed with its own $\boldsymbol{U}$ and $\boldsymbol{V}$, but using the $\boldsymbol{\sigma}$ matrices of the other images (a)–(g), respectively; in (p), we also show the "Lena" image constructed with its own $\boldsymbol{U}$ and $\boldsymbol{V}$ and the average $\boldsymbol{\sigma}$ of images (a)–(g). We can observe the luminance changes in the reconstructed "Lena" images (i)–(p). A closer examination of Fig. 2 reveals that the images (e) and (f) are with much brighter and much darker luminance, respectively, compared with the other images. The corresponding luminance changes can be seen in (m) and (n), which are formed from the $\boldsymbol{\sigma}$ matrices of images (e) and (f), respectively. In the MSVD metric [23], $\boldsymbol{\sigma}$ was used on the basis that it denotes the activity level in an image block. The activity level is defined as the luminance variation in pixels of an image block. A high activity level represents roughness or strong texture. Similarly, a low activity level corresponds to smoothness or weak texture. Due to its ability to characterize luminance changes, $\boldsymbol{\sigma}$ has also been used for image texture classification [26]. To illustrate this point further, we consider two 8 × 8 blocks taken from "bikes" image [shown in Fig. 2(a)] of the LIVE image database [35], one with a larger pixel intensity variation (denoted by $\mathbf{B}_H$) and the other with a smaller variation in pixel intensities (denoted by $\mathbf{B}_L$), i.e.,

$$\mathbf{B}_H = \begin{bmatrix} 88 & 85 & 56 & 19 & 61 & 114 & 111 & 105 \\ 95 & 87 & 87 & 51 & 19 & 70 & 120 & 109 \\ 62 & 103 & 92 & 92 & 51 & 22 & 81 & 122 \\ 28 & 72 & 107 & 89 & 71 & 22 & 29 & 92 \\ 31 & 35 & 66 & 52 & 41 & 30 & 21 & 38 \\ 34 & 34 & 32 & 31 & 31 & 33 & 31 & 24 \\ 27 & 31 & 34 & 34 & 33 & 34 & 33 & 29 \\ 56 & 27 & 34 & 34 & 32 & 31 & 31 & 30 \end{bmatrix}$$

$$\mathbf{B_L} = \begin{bmatrix} 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 98 & 98 & 98 \\ 99 & 99 & 99 & 99 & 98 & 95 & 96 & 93 \\ 99 & 99 & 99 & 99 & 99 & 97 & 98 & 98 \\ 99 & 99 & 99 & 99 & 99 & 98 & 96 & 99 \\ 99 & 99 & 99 & 99 & 98 & 97 & 99 & 103 \\ 99 & 99 & 99 & 99 & 102 & 105 & 104 & 106 \end{bmatrix}.$$

The singular values of $\mathbf{B}_H$ and $\mathbf{B}_L$ are as follows:

$$\begin{aligned} &\operatorname{diag}(\boldsymbol{\sigma}_{\mathrm{H}}) \\ &\quad = [478.75, 129.22, 64.71, 40.68, 26.4, 15.42, 4.84, 1.05] \\ &\operatorname{diag}(\boldsymbol{\sigma}_{\mathrm{L}}) \\ &\quad = [791.68, 10.42, 4.25, 2.17, 0.69, 0, 0, 0]. \end{aligned}$$

The ratio of the largest to the second largest singular value can be used to indicate the activity level [23]. In this example, this ratio is 3.70 for $\mathbf{B}_H$ (with high pixel variation), and it is 75.97 for $\mathbf{B}_L$ (with low pixel variation). The extreme case of $\mathbf{B}_L$ is a block in which all the pixel values are equal to, for example, q (i.e., no variation in pixel intensity); for such a block, the first singular value will be $8 \times q$, and the rest will be all zero. In this case, the said ratio is infinite, indicating no variation in pixel luminance. Different types of distortions bring about different changes in image luminance (with the related textural changes), which are captured reasonably well by the changes in singular values.

As mentioned before, singular values are the weights for the basis images, which can also be related to the changes in the frequency components of the image. Consider an 8 × 8 block of the "rapids" image [shown in Fig. 2(b)] of the LIVE image database. This block is then distorted by noise and blur. We show the singular values of the original, noisy, and blurred block in Fig. 3. One can notice that $\boldsymbol{\sigma}$ of the noisy block has higher values than that of the original block, and they decay slower. We can interpret $\boldsymbol{\sigma}$ to denote the effect of change in frequency because the noise increases the frequency, and this is captured in the increased $\boldsymbol{\sigma}$ values.

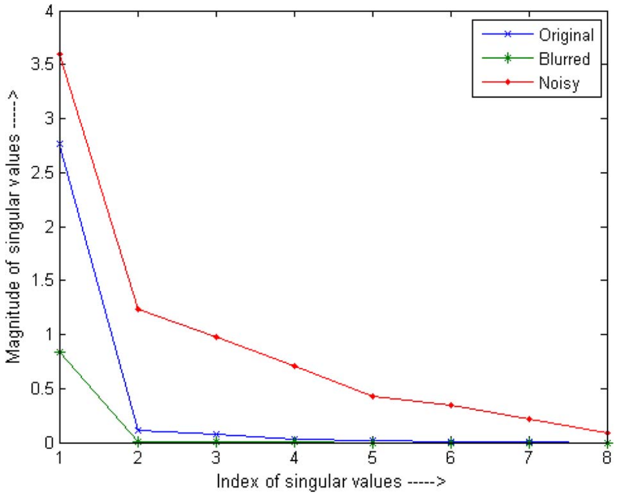


Fig. 3. Behavior of singular values for noise and blur distortion.

On the other hand, blur reduces the frequency, and the reader will notice that $\boldsymbol{\sigma}$ of the blurred block has lower values as compared with the original block, and it decays very fast, implying loss of frequency. In view of these, $\boldsymbol{\sigma}$ can account for the frequency changes induced in images due the distortion and thus provide useful information to characterize the quality.

In summary, the SVD transform analyzed in these two subsections has two major advantages over the other transforms (DFT, DCT, DWT, etc.) for visual quality evaluation: 1) the adaptively derived singular vectors allow better representation of image structure, and 2) the separation of structure and luminance components enables more effective differentiation of their effects on perceptual quality (while in other transforms, all the changes are reflected in the transform coefficients).

V. Proposed Visual Quality Metric

Based on the discussion and analysis in the previous section, we use SVD for feature extraction. $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\sigma}$ contain different information about the image, and quality degradation is reflected by their changes. This is also the basis of image noise filtering methods with SVD [46], [47]. As aforementioned, different types of distortions (like JPEG artifacts, blur, etc.) affect the visual quality in a largely similar fashion: by introducing structural and luminance changes that lead to different extents of perceived quality degradations. The proposed feature extraction with SVD aims at extracting the commonality behind seemly diverse degradations. Due to the existence of the underlying common patterns associated with quality degradation, a machine learning technique can be exploited to develop a general model by learning through examples.

A. SVD-Based Feature Detection

Features can be detected globally or locally in small blocks. We found that global SVD gives better prediction performance than local SVD. This can possibly be due to two reasons. First, use of local SVD means much larger number of features. For instance, for a block size of 16 × 16 (image size 512 × 512), one would need 32 768 dimensional vectors (16 384 features each for singular vectors and values). A large feature vector may contain redundant information, which leads to performance degradation. The second reason is that when small blocks are employed in SVD-based feature detection, they are assumed to be completely independent, which may not always be true. A global SVD, on the other hand, can tackle the interaction/ dependencies between the blocks better. However, with the global SVD approach, the feature vector size will depend on the image size, and this will result in feature vectors of unequal dimensions for the images in different databases. This will lead to mismatch in the dimension of the training and test feature vectors in case of cross database evaluation (i.e., training with images from one database and test set comes from the other databases, as detailed later in Section VI-B). There are two ways to tackle this. The first way to make the feature vector dimension equal for all images is to resize them to a common size. This is a straightforward solution, but such interpolation approach may introduce or remove some distortions, and so the original subjective scores may or may not be valid. To tackle the drawback associated with image resizing, we use an approach in between of the global (to maintain prediction accuracy) and local (to make the feature vector length the same for different images) approaches. We divide the given image into blocks of size $B \times B$ and compute the SVD for each block. For $B = 128$, an image (size 512 × 512) will have 16 blocks, whereas there will be 12 such blocks for a 512 × 384 image. Then, we use the average of the feature values of these blocks to define the final feature vector that will be $2B$ dimensional (B features for singular vectors and B for the singular values). The only requirement is that the image size should be $\geq B \times B$. For images with smaller size, we must use a smaller block size and proceed in a similar way, keeping in mind that the smaller the block size, the larger the number of features, and this generally will lower the prediction accuracy. We now outline the feature detection procedure.

First, the original and distorted images are divided into nonoverlapping blocks of size $B \times B$. Let us denote the kth (the total number of blocks is denoted as N_{block}) block in the original image as $\boldsymbol{A}_k$ and that in the distorted image as $\boldsymbol{A}_k^{(d)}$. We then obtain the respective singular values and singular vectors by applying SVD. We then measure the change in singular vectors as

$$\alpha_{jk} = \boldsymbol{u}_{jk} . \boldsymbol{u}_{\boldsymbol{jk}}^{(\boldsymbol{d})} \tag{6}$$

$$\beta_{jk} = \boldsymbol{v}_{jk} . \boldsymbol{v}_{\boldsymbol{jk}}^{(\boldsymbol{d})} \tag{7}$$

where α_{jk} ($j = 1$ to B and $k = 1$ to N_{block}) represents the dot product between the unperturbed and perturbed jth left singular vectors ($\boldsymbol{u}_j$ and $\boldsymbol{u}_{\boldsymbol{j}}^{(\boldsymbol{d})}$), and β_{jk} denotes that for the right singular vectors ($\boldsymbol{v}_j$ and $\boldsymbol{v}_{\boldsymbol{j}}^{(\boldsymbol{d})}$) of the kth block.

To illustrate the meaning of (6) [and also for (7)], we take a further look at the dot product between two vectors $\boldsymbol{u}_j$ and $\boldsymbol{u}_{\boldsymbol{j}}^{(\boldsymbol{d})}$ (angle between them is θ_u), which is defined as

$$\boldsymbol{u}_{jk} . \boldsymbol{u}_{\boldsymbol{jk}}^{(\boldsymbol{d})} = |\boldsymbol{u}_j| \left| \boldsymbol{u}_j^{(d)} \right| \cos(\theta_u). \tag{8}$$

In the case of singular vectors, the magnitude of each vector is unity, i.e., $|\boldsymbol{u}_j| = |\boldsymbol{u}_j^{(d)}| = 1$. Thus, the dot product between the unperturbed and perturbed singular vectors [as given by (6) and (7)] directly measures the cosine of the angle between the two singular vectors, i.e., $-1 \leq \alpha_{jk}, \beta_{jk} \leq 1$. We then define the feature vector $\boldsymbol{\Gamma}_k$ for the kth block for representing the change in $\boldsymbol{U}$ and $\boldsymbol{V}$ as follows, after the absolute valuation (for the reason explained at the end of this section) and normalization (for the values to range between 0 and 1):

$$\boldsymbol{\Gamma}_k = \{(|\alpha_{jk}| + |\beta_{jk}|)/2\} \qquad (j = 1 \text{ to } B). \tag{9}$$

To measure the change in singular values (let $\boldsymbol{\sigma}_k$ and $\boldsymbol{\sigma}_k^{(d)}$ denote the original and distorted singular value matrices), we let $\boldsymbol{s} = \text{diag}(\boldsymbol{\sigma})$ and $\boldsymbol{s}_{\boldsymbol{k}}^{(\boldsymbol{d})} = \text{diag}(\boldsymbol{\sigma}_{\boldsymbol{k}}^{(\boldsymbol{d})})$. We then define the feature vector for representing the change in singular values as

$$\boldsymbol{\tau}_k = \left(\boldsymbol{s}_k - \boldsymbol{s}_k^{(d)}\right)^2. \tag{10}$$

The lengths of $\boldsymbol{\Gamma}_k$ and $\boldsymbol{\tau}_k$ will be B. From (10), it is easy to see that all the elements of $\boldsymbol{\tau}_k$ are greater than or equal to 0. It is found that for natural images, the dynamic range of $\boldsymbol{\tau}_k$ is very large. Therefore, we divide each element in $\boldsymbol{\tau}_k$ by the maximum value in $\boldsymbol{\tau}_k$ for normalization to the range [0, 1], and define the resultant vector $\boldsymbol{\lambda}_k$ as

$$\boldsymbol{\lambda}_k = \boldsymbol{\tau}_k / max(\boldsymbol{\tau}_k). \tag{11}$$

The feature vector for the kth block is then defined as

$$\boldsymbol{x}_k = \{\boldsymbol{\Gamma}_k, \boldsymbol{\lambda}_k\}. \tag{12}$$

It follows that vector $\boldsymbol{x}_k$ will be of length $2B$. The final feature vector for the image is then obtained by averaging out the features over all the blocks, i.e.,

$$\boldsymbol{x} = \frac{1}{N_{\text{block}}} \sum_{k=1}^{N_{\text{block}}} \boldsymbol{x}_k. \tag{13}$$

We have found that the prediction errors were reduced significantly when we used the absolute magnitudes of α_{jk} and β_{jk} in (9), with the explanation as follows. By definition, $-1 \leq \alpha_{jk}$, $\beta_{jk} \leq 1$, and so $(\alpha_{jk} + \beta_{jk})$ can be positive or negative. Thus, two coefficients next to each other can be of similar magnitude but opposite sign, causing a large swing in the input data. This may affect the generalization performance of a machine learning algorithm. Therefore, we have used the absolute values as the feature input for the machine learning stage. A similar conclusion can be found in [28], which discusses the application of SVR for image coding when the absolute magnitudes of DCT coefficients were used as the input to the SVR.

In this paper, we used a block size of 128 (i.e., $B = 128$). We also experimented with smaller block sizes of 64 × 64, 32 × 32, etc., but the prediction performance, particularly for cross database evaluation, is better at bigger block size. There are two reasons for this observation:

1) As already mentioned, smaller blocks may not take into account the dependencies or interactions between them because features are extracted for each block independent of the other blocks.
2) With smaller block size, for example, 8 × 8, there will be 16 features for each block, but there will be a total of 4096 blocks. It is quite possible that in such a case the useful information about change in quality may be suppressed due to averaging over a large number of blocks. In fact, we use a machine learning technique in the first place to avoid such direct averaging/pooling methods. Nevertheless, with a larger block size, such as 128 × 128, the average of features is computed over fewer blocks and therefore more reasonable.

The reader will note that with the chosen block size of 128 × 128, we can handle almost all the existing image and video resolutions. For example, the typical resolution for DVD, miniDV, and Digital8 is 720 × 480, whereas newer technologies use higher resolutions (for instance, Blue ray uses 1280 × 720, 2K digital cinema uses 2048 × 1080, and so on), while the commonly used video resolutions are CIF (352 × 288), QCIF (176 × 144), 4 CIF (704 × 576), QVGA (320 × 240), VGA (640 × 480), XVGA (1024 × 768), DVD NTSC (720 × 480), DVD Pal (720 × 576), HDTV 720p (1280 × 720), etc. Note that for image sizes that are not multiples of the block size, one can use overlapping blocks (or zero padding) to compute the averaged feature vector, as outlined. We found that overlapped blocks (or zero padding) do not affect the prediction accuracy much, apart from the additional computational time required.

B. Combining Features Into a Perceptual Quality Score

Our aim is to represent the quality score Q as a function of the proposed feature vector $\boldsymbol{x}$, i.e.,

$$Q = f(\boldsymbol{x}) \tag{14}$$

where f is a function relating the elements of $\boldsymbol{x}$ to the final quality score and is difficult to be determined *a priori* in practice due to the limited knowledge and complexity of the HVS's perception of picture quality. To estimate f, we use a machine learning approach that is expected to give a more reasonable estimate compared with the existing pooling approaches, particularly when the number of features to be pooled is large. In this paper, we use SVR to map the high dimensional feature vector into a perceptual quality score by estimating the underlying complex relationship among the changes in $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{\sigma}$, and perceptual quality score. This exploits the advantage of machine learning with the ability to learn complex data patterns for an effective and generalized feature mapping. Although other choices of machine learning techniques are possible, we have used SVR because it is popular and well established. Furthermore, with SVR, one can obtain the support vectors, which are critical data points for SVR learning. The analysis of the support vectors can provide additional insights about the learning problem in hand, as will be shown in Section VII-F.

C. SVR

The goal of SVR is to find f based on training samples. Suppose that $\boldsymbol{x}_{\mathrm{i}}$ is the feature vector of the ith image in the training image set ($i = 1, 2 \ldots n_{\mathrm{tr}}$; n_{tr} is the number of training images). In the $\in$-SV regression [29], [30], the goal is to find a function $f(\boldsymbol{x}_{\mathrm{i}})$ that has the deviation of at most $\in$ from the targets s_i (being the corresponding subjective quality score) for all the training data and at the same time is as flat as possible [29]. The function to be learned is $f(\boldsymbol{x}) = \mathbf{w}^T\varphi(\boldsymbol{x}) + \mathrm{b}$, where $\varphi(\boldsymbol{x})$ is a nonlinear function of $\boldsymbol{x}$, $\mathbf{w}$ is the weight vector, and b is the bias term. We find the unknowns $\mathbf{w}$ and b from the training data such that the error

$$|\mathrm{s}_i - f(\boldsymbol{x}_i)| \leq \in \tag{15}$$

for the ith training sample $\{\boldsymbol{x}_i, \mathrm{s}_i\}$. It has been shown [29] that

$$\mathbf{w} = \sum_{i=1}^{n_{\mathrm{sv}}} (\eta i^* - \eta i)\varphi(\boldsymbol{x}_i) \tag{16}$$

where ηi^* and ηi $(0 \leq \eta i^*, \eta i \leq C)$ are the Lagrange multipliers used in the Lagrange function optimization, C is the tradeoff error parameter, and n_{sv} is the number of support vectors. For data points for which inequality (15) is satisfied, i.e., the points that lie within the $\in$ tube, the corresponding ηi^* and ηi will be zero so that the Karush–Kuhn–Tucker conditions are satisfied [29]. We have a sparse expansion of $\mathbf{w}$ in terms of $\boldsymbol{x}_i$ (i.e., we do not need all $\boldsymbol{x}_i$ to describe $\mathbf{w}$). The samples that come with nonvanishing coefficients (i.e., non zero ηi^* and ηi) are support vectors, and the weight vector $\mathbf{w}$ is defined only by the support vectors (not all training data). The function to be learned then becomes

$$\begin{aligned} f(\boldsymbol{x}) = \mathbf{w}^T\varphi(\boldsymbol{x}) + \mathrm{b} &= \sum_{i=1}^{n_{\mathrm{sv}}} (\eta i^* - \eta i)\varphi(\boldsymbol{x}_i)^T\varphi(\boldsymbol{x}) + \mathrm{b} \\ &= \sum_{i=1}^{n_{\mathrm{sv}}} (\eta i^* - \eta i)K(\boldsymbol{x}_i, \boldsymbol{x}) + \mathrm{b} \end{aligned} \tag{17}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}) = \varphi(\boldsymbol{x}_i)^T\varphi(\boldsymbol{x})$, being the kernel function. In SVR, the actual learning is based only on the critical points (i.e., the support vectors). In the training phase, the SVR system is presented with the training set $\{\boldsymbol{x}_i, \mathrm{s}_i\}$, and the unknowns $\mathbf{w}$ and b are estimated to obtain the desired function (17). During the test phase, the trained system is presented with the test feature vector $\boldsymbol{x}_j$ of the jth test image and predicts the estimated objective score s_j ($j = 1$ to n_{te}; n_{te} is the number of test images). We have used the radial basis function (RBF or Gaussian kernel) as the kernel function, which is of the form $K(\boldsymbol{x}_i, \boldsymbol{x}) = \exp(-\rho\|\boldsymbol{x}_i - \boldsymbol{x}\|^2)$, where ρ is a positive parameter controlling the radius. The RBF kernel is widely used and has been shown to achieve good performance in many applications [30]. In our case also, it was found to give better performance than other kernels like linear, polynomial, sigmoid, etc. We used a validation set to determine the SVR parameters, namely ρ, C, and $\in$. The reader will notice that

TABLE I
MAJOR CHARACTERISTICS OF THE TEN SUBJECTIVELY RATED DATABASES USED IN THIS PAPER

	No. of reference images/videos	No. of distorted images/videos	No. of distortion types	Typical image/frame size	Subjective score format (Range)
LIVE	29	779	5	768 × 512	DMOS (0-100)
CSIQ	30	866	6	512 × 512	DMOS (0-1)
IVC	10	185	4	512 × 512	MOS (1-5)
Toyama	14	168	2	512 × 768	MOS (1-5)
A57	3	54	6	512 × 512	DMOS (0-1)
TID	25	1700	17	512 × 384	MOS (0-9)
WIQ	7	80	1	512 × 512	DMOS (0-100)
Watermarked image database	5	210	1	512 × 512	MOS (1-5)
LIVE video database	10	150	4	768 × 432	DMOS (0-100)
EPFL video database	6	78	1	352 × 288	MOS (1-5)

the function $f(\boldsymbol{x})$ in (17) is a linear combination of Gaussian functions scaled by a factor of $(\eta i^{*} - \eta i)$. Hence, by using SVR, we attempt to approximate the desired mapping function from features to a quality score via a combination of Gaussian functions. The kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$ can be interpreted as the distance (or measure of similarity) between the ith SV $\boldsymbol{x}_i$ and the test vector $\boldsymbol{x}$ in the transformed space. We can also interpret $K(\boldsymbol{x}_i, \boldsymbol{x})$ as the cosine of the angle between the two Gaussian functions centered on $\boldsymbol{x}_i$ and $\boldsymbol{x}$. It is also easy to observe from (17) that the predicted value is a weighted sum of the distances (or "similarities") between all the SVs and test vector $\boldsymbol{x}$. Due to this, SVs are the critical points with regard to the learning phase in SVR.

VI. DATABASES AND TRAINING

Visual quality metrics must be tested on a wide variety of visual contents and distortion types to make meaningful conclusions about their performance. Evaluating a metric with one single subjective database might not be sufficient and general [31]. We have therefore conducted extensive experiments on ten open databases in total. As will be shown in Section VII, a metric performing well in one database may not necessarily do well on all the other databases. In this section, we describe the databases used for the experiments and provide the details of the training and test procedures adopted to verify the proposed approach.

A. Database Description

In this paper, we use total of ten subjectively rated image and video databases, namely, LIVE image database [35], CSIQ database [64], IVC database [37], Toyama database [36], A57 database [41], TID database [38], WIQ database [51], [67], and the watermarked image database [75]. The two video databases used are LIVE video database [24] and EPFL video database [68]. We provide a brief summary of these databases in Table I and refer the reader to the respective references for more details.

B. Test Procedure

We evaluate the performance of the proposed scheme in two different ways. First, we have employed the k-fold CV strategy [48] for each database separately: the data were split into k chunks, one chunk was used for test, and the remaining $(k-1)$ chunks were used for training. The experiment was repeated with each of the k chunks used for testing. The average accuracy of the tests over the k chunks was taken as the performance measure. The splitting of the data into k chunks was done carefully so that the image contents present in one chunk did not appear in any of the remaining chunks (and this chunk is used as the test set). One image content is defined as all the distorted versions of an original image. As an example, consider the CSIQ database that consists of 30 original images. In this case, the first chunk included all the distorted versions of the first three original images. The second chunk consisted of distorted versions of the next three original images and so on. Thus, for the CSIQ database, there were a total of ten chunks, each of which comprised different image contents. In the same way, the Toyama database (with 14 original images) was split into seven chunks with each chunk comprising of two image contents. The LIVE database with 29 original images was split into ten chunks, with the first nine chunks consisting of three image contents each, whereas the last chunk included two image contents. A similar splitting procedure was followed for the other databases as well. This way, it was ensured that the images appearing in the test set are not present in the training set.

Since the proposed metric involves training, we further need to examine the feasibility and robustness of such machine learning based system to untrained image and distortion types. To that end, we use the cross database validation: one database is used for training, and others are used for validation. In this paper, we use the notation Q_{database} to denote training with a particular database. Therefore, Q_{CSIQ}, Q_{LIVE}, and Q_{TID} denote that training is done with the CSIQ, LIVE, and TID databases, respectively, and a similar notation has been followed for the other databases as well. However, some databases have a few images in common, e.g., LIVE, TID, and Toyama. Therefore, we have reported the cross-database evaluation results for the cases when none of the images in the training set has appeared in the test set. This is again to ensure that the system is trained and tested on entirely different sets of images. For Q_{vector}, we have reported only the best results among those obtained on training with different databases.

A five-parameter logistic mapping between the objective outputs and the subjective quality ratings was also employed, following the Video Quality Experts Group (VQEG) Phase-I/II test and validation method [32], to remove any nonlinearity due to the subjective rating process and to facilitate the comparison of the metrics in a common analysis space. The experimental

TABLE II
IMPLICATIONS FOR DIFFERENT RANGES OF F VALUES COMPUTED WITH RESPECT TO THE PROPOSED Q FOR ANY METRIC X UNDER COMPARISON

$\mathrm{F} > \mathrm{F}_{\mathrm{critical}}$	$1 < \mathrm{F} < \mathrm{F}_{\mathrm{critical}}$	$1/\mathrm{F}_{\mathrm{critical}} < \mathrm{F} < 1$	$\mathrm{F} < 1/\mathrm{F}_{\mathrm{critical}}$
X has significantly larger residuals than Q, so Q is statistically better than X.	Although Q performs better than X since F>1, both Q and X are statistically indistinguishable.	Although X performs better than Q since F<1, both Q and X are statistically indistinguishable.	X has significantly smaller residuals than Q, so Q is statistically worse than X.

results are reported in terms of the three criteria commonly used for performance comparison, namely, Pearson linear correlation coefficient C_P (for prediction accuracy), Spearman rank order correlation coefficient C_S (for monotonicity), and root MSE (RMSE), between the subjective score and the objective prediction. For a perfect match between the objective and subjective scores, $C_P = C_S = 1$, and RMSE $= 0$. A better quality metric has higher C_P and C_S and lower RMSE.

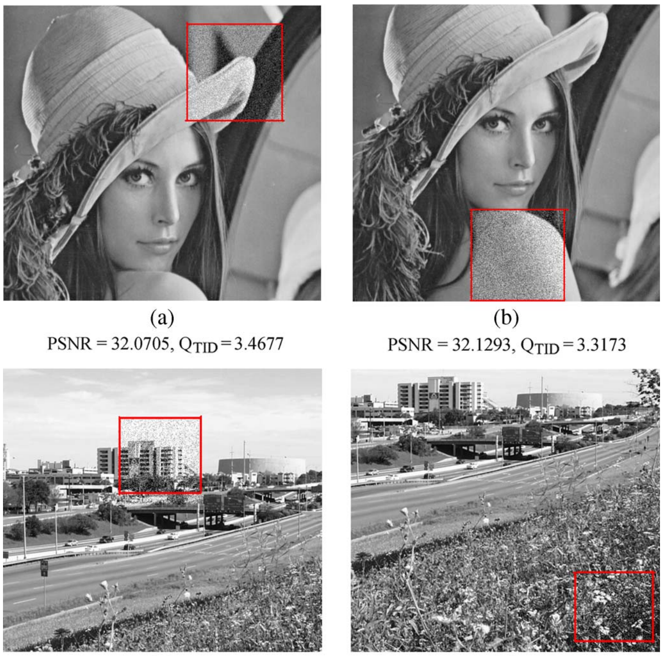
We have also compared the performance of the proposed Q (with k-fold CV) with the following existing visual quality estimators: PSNR, SSIM [11], MSVD [23], VSNR [33], IFC [22], VIF [34], and the method proposed in [73]. For VSNR, VIF, IFC, and SSIM implementation, we have used the publicly accessible Matlab package that implements a variety of visual quality assessment algorithms [39]; they are the original codes provided by the image quality assessment algorithm designers. The MSVD method was implemented in Matlab. In addition, we also report the results for the metric developed in our previous work denoted as Q_{vector} (also with k-fold CV) [63] for comparison. The publicly available LibSVM software package [30] was used to implement the SVR algorithm. Since we have used all publicly accessible software and databases in this paper, the results reported in this paper can be reproduced for any future research.

C. Statistical Significance Test

To assess the statistical significance of the proposed metric's performance relative to the other metrics, an F-test [32], [42], [43] was performed on the prediction residuals between the objective predictions (after applying the logistic mapping) and the subjective scores. Obviously, the smaller the residuals, the better the metric. The test is based on an assumption of Gaussianity of the residual differences. Letting σ_Q^2 and σ_X^2 denote the variances of the residuals from the proposed metrics Q and X, respectively, then the F-statistic with respect to Q is given by $F = \sigma_X^2/\sigma_Q^2$; obviously, Q is better than X when $F > 1$. The F value is compared with the critical F-statistic (denoted as F_{critical}) which is computed based on the number of residuals and the desired confidence level, to judge if Q and X are statistically indistinguishable. Table II lists the implications of different ranges of F values. In this paper, we have used a 99% confidence level to compute the F_{critical} values.

VII. EXPERIMENTAL RESULTS AND RELATED ANALYSIS

Most of the existing visual quality metrics work only with the luminance component of the image/video. Therefore, all the experimental results reported in this paper are for the luminance component only (because the luminance component plays a more significant role in human visual perception than color components).



(a)
PSNR = 32.0705, Q_{TID} = 3.4677

(b)
PSNR = 32.1293, Q_{TID} = 3.3173

(c)
PSNR = 33.8977, Q_{TID} = 4.7885

(d)
PSNR = 33.1732, Q_{TID} = 5.4665

Fig. 4. (a) White noise distorted hat part. (b) White noise distorted shoulder part. (c) White noise distorted building part. (d) White noise distorted plants part. The objective predictions from PSNR and Q_{TID} have been indicated below each image. For reference, $Q_{\text{TID}} = 5.7966$ for the image with no distortions. The images have been cropped for visibility. (a) PSNR $= 32.0705$, $Q_{\text{TID}} = 3.4677$. (b) PSNR $= 32.1293$, $Q_{\text{TID}} = 3.3173$. (c) PSNR $= 33.8977$, $Q_{\text{TID}} = 4.7885$. (d) PSNR $= 33.1732$, $Q_{\text{TID}} = 5.4665$.

A. Visual Quality Prediction Test

To demonstrate that the proposed method properly accounts for the distortion in different image areas, we show four images in Fig. 4. The source or the original images are shown in Fig. 2. First, we consider the "hat" part and the "shoulder" part of the "Lena" image, as indicated by the red boxes in Fig. 4(a) and (b). We distorted these two portions in the image by adding white Gaussian noise.

Note that the amount of noise in the two blocks in Fig. 4(a) and (b) is the same. Because the effect of white noise is uniformly distributed, it can be observed that it does not cause too much damage to the edge in the "hat," and the structure is largely preserved. As a result, noise in the "hat" part is less annoying. Of course, there will be loss of visual quality. On the other hand, the reader will notice that the shoulder in the "Lena" image is smooth due to which the added noise is clearly visible and therefore more annoying to the human eye. This leads to a higher level of annoyance in the shoulder as compared with the hat in spite of the same amount of distortion introduced in the two portions. We have indicated the objective quality scores from PSNR and proposed Q_{TID} (which means the TID

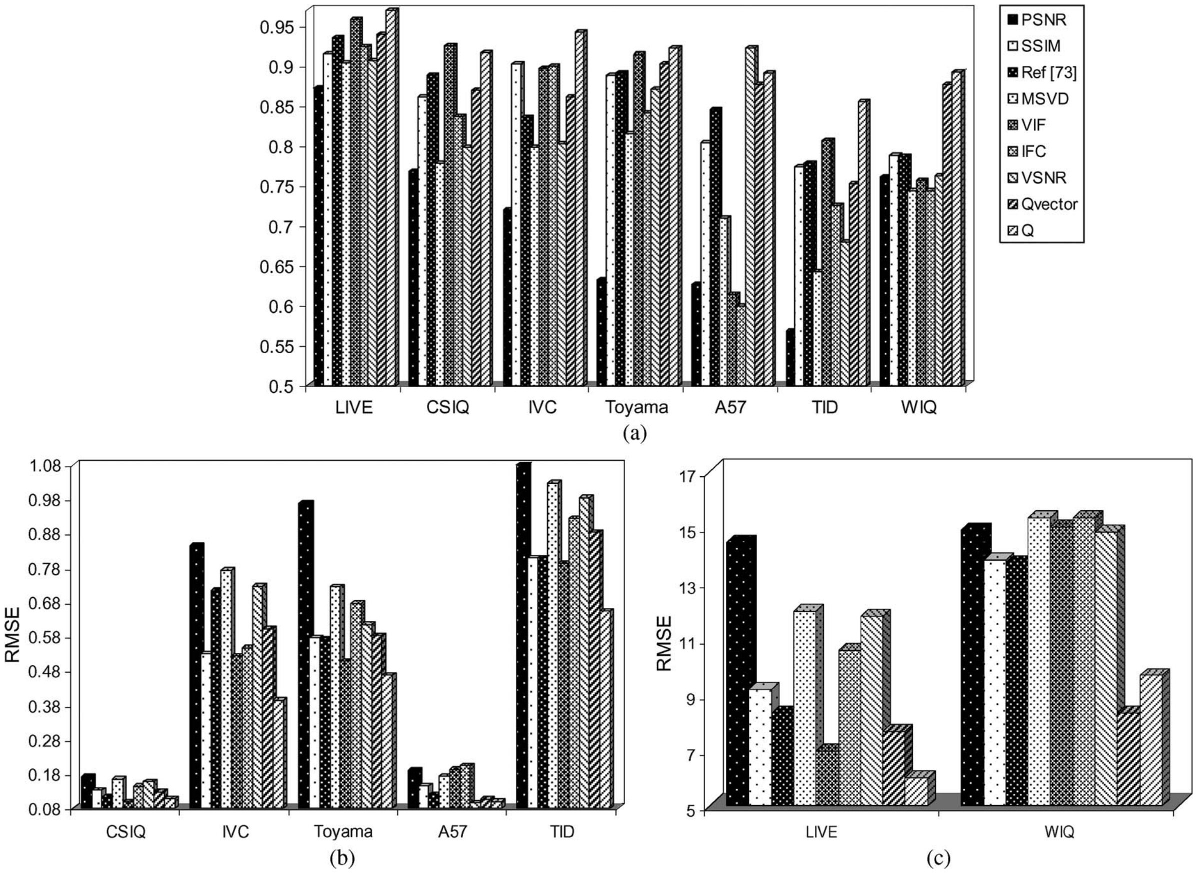


Fig. 5. (a) C_P comparison on different image databases. (b) RMSE for CSIQ, IVC, A57, and TID databases. (c) RMSE for LIVE and WIQ databases.

database is the training set). Note that Q_{TID} will predict scores in the form of mean opinion score (MOS) because the training database (TID) comprises of MOS. Therefore, a higher Q_{TID} means better quality. One can see that PSNR predicts a higher score for the image that has noise in the smooth part (more annoying) as compared with the other image, which is not consistent with HVS. In contrast, Q_{TID} predicts a lower score for the image with distortion in the smooth part (shoulder) and a higher score for the other image. Next, we consider the images shown in Fig. 4(c) and (d). We have indicated two portions in this image by red boxes. We added the same amount of white Gaussian noise to these two portions in the image. As can be seen, the distortion in the "building" part is more clearly visible and thus more annoying to the human eye. On the other hand, the area with "plants" is textured and can tolerate such distortions [3]. In fact, the white noise in that part cannot easily be noticed by the human eye. It can be noted that PSNR gives a higher score for the image in Fig. 4(c) and a lower score for the image in Fig. 4(d), where most of the noise is not visible due to masking. We have already mentioned that this happens because PSNR assigns equal importance to all the errors independent of their perceptual impact. On the other hand, the proposed approach is able to capture the effect of noise masked due to texture and assigns a higher score to Fig. 4(d) and a lower score to the image in Fig. 4(c). It may be mentioned that in the four images shown in Fig. 4, the distortion (in this case white noise) has been added in different parts of the image. In Fig. 4(a), mainly the edge part is distorted; in image (b), a smooth portion has been corrupted; in image (c), a visually more salient region has been distorted; and in image (d), the textured portion is distorted. It may further be mentioned that according to the Q_{TID} scores given in Fig. 4, noise in smooth portion causes the largest perceptual annoyance (Q_{TID} is smallest) followed by noise in edge regions, whereas the perceived loss of visual quality is the least in the textured region (Q_{TID} is the largest). This confirms that the perceptual impact of distortion in different portions is reasonably well handled by the proposed scheme. This demonstrates the effectiveness of the proposed SVD-based features and their proper pooling via SVR.

The foregoing discussion and analysis were meant to provide a visual illustration of the effectiveness of the proposed method and how it can handle distortions according to their perceptual significance. In the following sections, we provide the test results for the proposed method and eight existing objective image quality metrics using ten publicly available databases. The large number of images and distortion types spanned by these databases helps in a more thorough and comprehensive metric validation.

B. Performance Evaluation on Image Databases

In Fig. 5, the results for the proposed Q and other existing metrics are presented. The C_P values of different metrics are shown in Fig. 5(a), where we can see that the proposed scheme Q performs well in general. We can also see that the existing metrics do not perform well for all the test databases. For

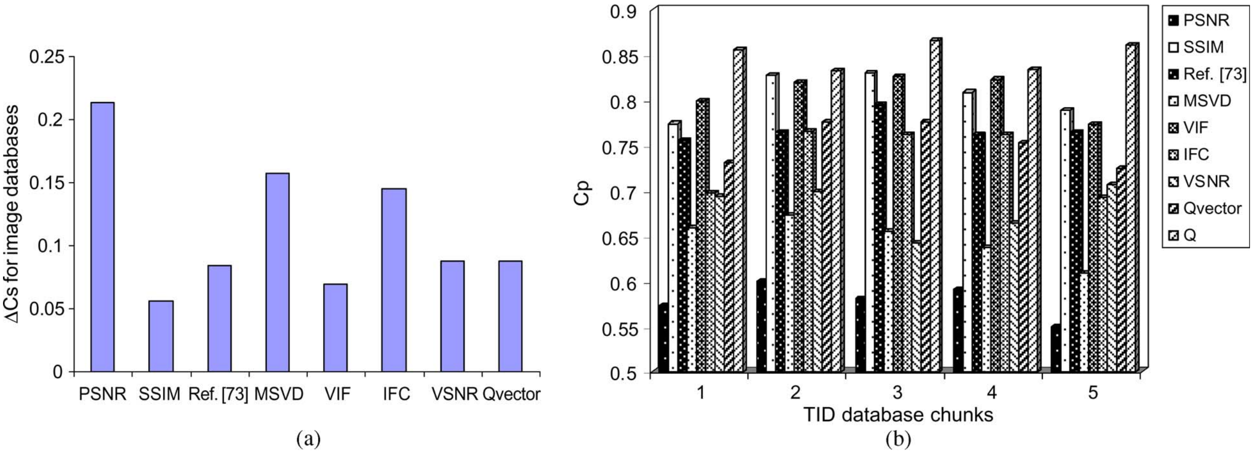


Fig. 6. (a) Average ΔCS values with respect to Q over the seven image databases for different metrics. (b) C_P values for five chunks of TID database.

example, we note that the performance of PSNR, VIF, VSNR, SSIM, MSVD, and IFC is worse on the WIQ database since these metrics generally perform better for images containing a single artifact in the image [51]. As aforementioned, the images in the WIQ database can contain more than one artifact (like blocking and ringing together in a same image) due to the complex nature of a wireless communication link. Similarly, for the A57 database, the performance of PSNR, VIF, SSIM, MSVD, and IFC is relatively poor. As can also be seen, VSNR, which performs well for A57 database, does not perform as well on the other databases. The IFC metric performs well for LIVE and IVC databases, but its prediction performance is worse on the remaining databases. By contrast to the existing metrics, the performance of Q is more consistent across all the databases and generally better than all the other metrics being compared.

Recall that for Q, none of the images in the training set appear in the test set. Therefore, the proposed metric exhibits robustness, and training with specific image contents is not necessary. It has also been found that, in general, the prediction performance of the proposed scheme is consistent over all the test chunks. We illustrate this through the performance on the TID database, which consists of 25 original images. Following the splitting procedure detailed in Section VI-B, we obtained five chunks each with 340 images. The C_P values of the different metrics for each TID test chunk are shown in Fig. 6(b). The proposed system performs well consistently for all the test chunks and is better than the other metrics. The consistency in prediction performance was similarly observed for all the other databases. This indicates that the proposed system performs well across varied images and distortions and does not show any dependency on any specific image/distortion content.

We further present the results of the F-test in Fig. 9. According to Table II, the points that lie above the F_{critical} boundary denote the cases for which the proposed scheme is better and also statistically distinguishable than the existing metric under comparison. We can see from the figure that a large number of points (about 70% of them) are above the F_{critical} boundary, indicating that the proposed Q is statistically better in comparison with the other metrics. The points that lie between the F_{critical} curve and the line $F = 1$ (i.e., $1 < F < F_{\mathrm{critical}}$) denote that the cases for the proposed Q is still better than the corresponding metric since $F > 1$ but statistically indistinguishable. We note that only two points (2.8% of the cases) fall below the $F = 1$ boundary. In these two cases, the proposed Q performs worse than the corresponding metric since $F < 1$ and is statistically indistinguishable from those two metrics (VIF and VSNR for CSIQ and A57, respectively). There is no single case for which $F < 1/F_{\mathrm{critical}}$, i.e., the proposed Q has not been statistically worse than any existing metric with any database under comparison.

Since C_P and C_S exhibit similar trends, we only show the average difference in C_S values (with respect to Q) over the seen image databases for the eight existing metrics in Fig. 6(a). As can be seen, all the ΔC_S are positive, indicating the better performance of Q. Fig. 5(b) and (c) also indicates that Q outperforms the other metrics in terms of RMSE.

C. Cross Database Validation

For the cross database evaluation, we selected the three biggest image databases available, namely, TID (1700 images), CSIQ (866 images), and LIVE (779 images), for training Q. As can be seen in Table III with different databases as training and test sets, Q again performs well across all the databases, similar to the k-fold CV tests. We have reported only the C_P values in Table III since C_S and RMSE show similar results as C_P. We can also see in Table III that Q_{CSIQ} gives a C_P value of 0.7550 for the TID database, which is comparable with other metrics like SSIM and VIF and better than PSNR, VSNR, IFC, and MSVD. This is significant since in this case, the training set (866 images) is only about half of the size of the test set (1700 images of different visual contents). The proposed metric also performs better than all the other metrics, as indicated by the higher C_P values achieved by Q_{CSIQ}, Q_{LIVE}, and Q_{TID} for the WIQ database. Overall, we can see from Table III that the proposed metric is consistent and gives good prediction performance for the cross database evaluation. We have also shown the scatter plot for the LIVE image database with Q_{CSIQ} as the objective metric in Fig. 7. The data points corresponding to the five types of distortions present in this database are highlighted using different notations/colors. As can be seen, the plot is compact around the logistic fitting curve and shows low

TABLE III
C_P VALUES FOR THE CROSS DATABASE VALIDATION (CROSS DATABASE EVALUATION HAS BEEN AVOIDED FOR SAME IMAGES APPEARING IN THE TRAINING AND TEST SETS; SUCH CASES ARE DENOTED BY "–" IN THE TABLE)

Test database/ Model	LIVE	CSIQ	IVC	Toyama	A57	TID	WIQ
Q_{CSIQ}	0.9086	--	0.8828	0.8327	0.8843	0.7550	0.7764
Q_{LIVE}	--	0.8581	0.8877	--	0.8807	--	0.7314
Q_{TID}	--	0.8831	0.8755	--	0.8854	--	0.7580
$Q_{\text{watermark}}$	0.9004	0.8267	--	0.8782	0.8064	0.7219	0.7202
Q_{vector}	--	0.8525	0.7884	--	0.8223	--	0.7573

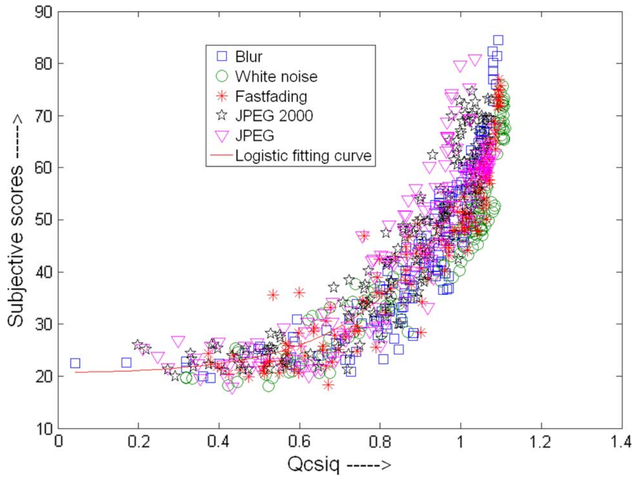


Fig. 7. Scatter plot for the LIVE image database with Q_{CSIQ} as the objective metric.

scattering around it. Therefore, the prediction performance of the proposed metric is good for all the distortions as none of the data points scatter too much around the logistic fitting curve. Note that a large scatter would imply poorer performance.

As mentioned before, we also use the image database with watermarked images. This type of distortion is different from other commonly occurring distortions (like JPEG, Blur, white noise distortion, etc.) due to the specific processing that images undergo. We used this database only as a training set to further confirm the robustness of the proposed scheme to new and untrained distortions. Similar to the previous notations, $Q_{\text{watermark}}$ denotes the training with a watermarked image database. However, out of five, we only used three original images and their distorted versions as the training set. This was done again to ensure that the images used for training are excluded from the test sets. Note that we excluded two images, namely, "monarch" and "rapids," which are present in many other databases. As can be seen in Table III, $Q_{\text{watermark}}$ performs quite well. This further confirms that quality degradation due to different distortion types can be assessed by exploiting the underlying common patterns characterized by the structure loss. Note that for $Q_{\text{watermark}}$, the training set that consists of 126 images and their corresponding subjective scores is relatively small as compared with test databases like TID, LIVE, and CSIQ. Thus, the results obtained for $Q_{\text{watermark}}$ are significant because the system is trained on a completely different distortion to those in the test databases.

It is also worth pointing out that the subjective quality score range is different for all the databases. For instance, LIVE includes subjective scores as difference MOS (DMOS) in the range of 0–100, whereas TID gives subjective results in the form of MOS in the range of 0–9. The IVC database consists of MOS in the range of 0–5, whereas the CSIQ database reports DMOS in the range of 0–1. The A57 database includes subjective scores as DMOS in the range of 0–1. Thus, Q_{TID}, before the logistic fitting, gave a C_P value of -0.8755 for the CSIQ database, -0.7656 for the WIQ database, and -0.8752 for the A57 database. All the resulting correlations here are negative due to the fact that the system was trained with MOS while it was tested with DMOS, which has an opposite range of valuation in quality specification.

Another aspect of note is the robustness to untrained distortions. For the cross database tests, since the training and test sets come from different databases, many of the distortion types appearing in the test set are not represented in the training set. The good performance of Q_{CSIQ}, Q_{LIVE}, Q_{TID}, and $Q_{\text{watermark}}$ for the WIQ database shows the robustness of Q to complex distortions, which are not present in the training set. Similarly, many of the distortion types present in the TID database do not occur in the CSIQ database, and hence, the C_P value of 0.7550 given by Q_{CSIQ} is noteworthy. Similar observations hold for Q_{LIVE}, Q_{TID}, and $Q_{\text{watermark}}$.

To further test the robustness to untrained distortions, we tested the images from the TID database, which were distorted by five types (from the total of 17 types) of distortions: image denoising, noneccentricity pattern noise, local block-wise distortions of different intensity, mean shift (intensity shift), and contrast change. These five distortions were chosen since they do not appear in any other database and also form a challenging set of distorted images to be assessed for visual quality. For example, consider the case of denoised images. The PSNR for a denoised image is generally higher than that of the original noisy image, but at the same time the denoised image may visually look worse than the corresponding original noisy image [38]. Hence, quality assessment of such images is not straightforward. The next distortion type considered is the local block-wise distortions of different intensity. For the first level of distortion, 16 image blocks (block size is 32 × 32) were distorted in each image; for the second level of distortion, eight blocks were distorted; for the third level of distortion, four blocks were distorted; and for the fourth level, two blocks were distorted. Recall that for the TID database, the first distortion level corresponds to the highest PSNR, whereas the fourth level of distortion corresponds to the lowest PSNR. It has been found that [38] an image in which two blocks have been corrupted (i.e., the fourth distortion level) is perceived as having a better visual quality (although it has smaller PSNR) than the image in which 16 blocks have been corrupted (i.e., the first distortion

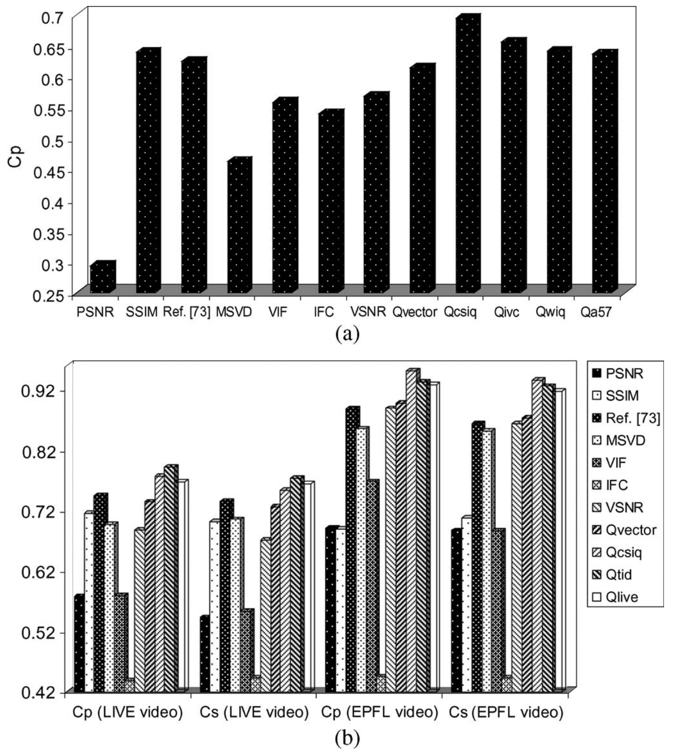


Fig. 8. (a) C_P values for the 500 images from TID database with five distortion types (see text for further explanation). (b) C_P and C_S values for LIVE and EPFL video databases.

level). This suggests that a lower amount of distortion spread over a larger area is likely to cause more quality degradation than a higher amount of distortion spread over a smaller area. Therefore, quality assessment of such images can be tricky for the metrics. Likewise, contrast and intensity changes (up to a certain level) generally do not affect the visual quality substantially (in spite of the presence of pixel errors), although the PSNR may change considerably. Hence, images with these five distortion types are indeed challenging for metrics. We have tested these 500 images (100 images for each of the five distortion types) with the proposed Q trained on the CSIQ database (Q_{CSIQ}), IVC database (Q_{IVC}), WIQ database (Q_{WIQ}), and A57 database (Q_{A57}). By training with these databases, it is ensured that the training and test images are different. We have also computed the results for the other metrics for comparison. We can see from Fig. 8(a) that metrics like VIF, VSNR, PSNR, and MSVD do not perform well ($C_P < 0.6$ for these metrics), whereas Q performs better than the other metrics. It may be noted that Q_{CSIQ}, Q_{IVC}, Q_{WIQ}, and Q_{A57} all perform quite well. This result is significant since the IVC, WIQ, and A57 databases contain significantly less number of images than the number of test images.

D. Performance Evaluation on Video Database

The performance of Q has been evaluated on the video databases using the cross database evaluation. It can be recalled that Q_{CSIQ}, Q_{LIVE}, and Q_{TID} denote training with the CSIQ, LIVE, and TID databases, respectively. The trained system is used to predict the quality score of each individual frame. The same procedure was also adopted for evaluating the other metrics. In this paper, the overall quality score of the video is determined as the average of the scores of all the frames in the video. We present the C_P and C_S values of different metrics for the two video databases in Fig. 8(b). As can be seen, Q_{CSIQ}, Q_{LIVE}, and Q_{TID} all perform better than the existing metrics under comparison. One can also note that the C_P and C_S values lead to a similar conclusion regarding metric performance. The RMSE values (not shown here to save space) were also found to be consistent with C_P and C_S. Since the training is done with image databases only, the good performance of Q is indicative of its generalization ability to new visual/distortion content. The F-test results for video have also been indicated in Fig. 9. For the video databases, the F values were calculated against residuals of Q_{TID} (i.e., training is done with the TID database). The two video databases used in this study (LIVE and EPFL) represent different visual contents since they use different original video sequences and thus provide diverse visual contents for testing the robustness of the proposed Q.

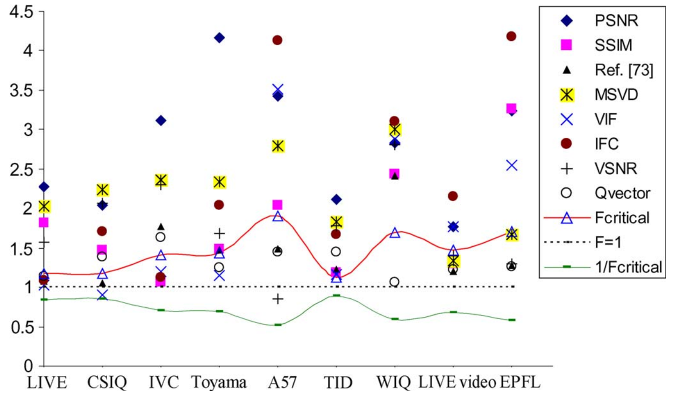


Fig. 9. F-test plot for different image and video databases (the points above the F_{critical} boundary denote the cases for the proposed scheme to be statistically better than the corresponding metric).

Interestingly, we can see from Fig. 8(b) that the metrics give relatively better performance for EPFL video database than for the LIVE video database. One reason for this is that the LIVE video database includes four distortion types as compared with the EPFL video database, in which the sequences are impaired only by packet loss. Another reason is that in the LIVE video database, the distortion strength has been adjusted perceptually [24]. As an example of the perceptual adjustment, consider four labels for visual quality ("Excellent," "Good," "Fair," and "Poor") and one reference video sequence "Tractor" from the LIVE video database. Four MPEG-2 compressed versions of "Tractor" are chosen to approximately match the four labels for visual quality. A similar procedure is applied to select H.264 compressed, wireless, and IP distorted versions. The "Excellent" MPEG-2 video and the "Excellent" H.264 video are designed to have approximately same visual quality, and a similar perceptual adjustment has been made for the other distortion categories and quality labels. On the other hand, for the EPFL database, the packet loss rates have been fixed *a priori*. It has been argued [24] that adjusting the distortion strength perceptually, as done for LIVE video database, is far more effective toward challenging and distinguishing the performance of visual quality metrics than, for instance, fixing the compression rates/packet loss rates across sequences. Due to

TABLE IV
AVERAGE EXECUTION TIME (IN SECONDS PER IMAGE) FOR DIFFERENT METRICS

Metrics	SSIM	MSVD	VIF	VSNR	PSNR	IFC	Ref. [73]	Ref. [64]	Proposed
Time	0.0454	0.6036	3.4829	0.4452	0.0037	4.4490	5.0333	5.1723	1.03

these two reasons, the LIVE video database is more challenging for visual quality metrics.

The adopted procedure of assessing video quality by using the average quality scores of frames takes into account the spatial information in the video, but the temporal information is disregarded in this case. Nonetheless, in this work, our aim is to demonstrate the performance of the proposed system to untrained visual/distortion contents. Incorporating temporal information for video quality assessment is an area of further research.

E. Computational Complexity

In this section, we provide an indication of the execution time of different metrics, i.e., time required for predicting the quality of an image. We measured the average execution time required per image in the A57 database (image resolution is 512 × 512) on a PC with 2.40-GHz Intel Core2 CPU and 2 GB of RAM. Table IV shows the average time required per image (in seconds), with all the codes implemented in Matlab. We note that the proposed method is computationally more expensive than metrics like PSNR and SSIM due to the fact that SVD is computationally intensive. An exact SVD of a $r \times c$ matrix has a time complexity $O(\min\{rc^2, r^2c\})$. However, the computational cost and time are reduced due to the fact that we use block-based SVD (although the block size is large but still smaller than the full image). Furthermore, many fast and efficient implementations of SVD are available, which can lead to a decrease in SVD computation. Training the SVR is of higher computational requirement, but the model training can be done offline.

To give more precise estimates of the time required for training and testing, we present an example below with TID as the training database and A57 being the test database. First, we extract the features for the images in the TID database for training the system, and the time taken is about 1306 s (totally, there are 1700 images in the TID database), which means about 0.7687 s/image (note that the image size is 512 × 384 in the TID database). Next, we train to obtain the model $Q_{\rm TID}$ by training with the features extracted. It took about 2.5776 s to obtain the trained model $Q_{\rm TID}$. Therefore, the total time for developing $Q_{\rm TID}$ is approximately 1309 s. This of course can be developed offline. Note that the training time is directly proportional to the number of training samples used. For testing, the time required for feature extraction per image is about 1 s/image (note that the image size is 512 × 512 in the A57 database) as measured from the 54 images of the A57 database (it took 53.7765 s for extracting the feature vectors of the 54 images in the database). The time required for the prediction of quality (after extracting the features) using $Q_{\rm TID}$ is negligible (only about 0.03 s/image). Because the prediction model (in this example $Q_{\rm TID}$) is developed offline, it takes approximately $1\ (\text{feature extraction}) + 0.03 (\text{for prediction}) = 1.03$ seconds to predict the quality of a 512 × 512 image. The proposed method is, however, less complex than more sophisticated metrics like VIF and IFC, which employ wavelet decomposition.

F. Further Observations

As aforementioned, SVs are samples for which the inequality (15) is not satisfied, i.e., they lie outside the ε tube. They are the critical data points that can be considered as the representative of the whole training set. The solution (i.e., weights) can be expressed as a linear combination of the SVs as given by (16), and thus, their analysis provides more insights about the problem. In our experiments, we observed that the SVR algorithm tends to select images that have either near-threshold distortions (i.e., low distortion level) or images with much higher distortion levels as the SVs. For example, consider the CSIQ database for which DMOS is in the range of [0, 1]: a DMOS close to 0 implies low distortion while that close to 1 means high distortion, as perceived by the subjects. We have found that samples that were chosen as the SVs for the CSIQ database corresponded to either DMOS less than 0.056 or DMOS greater than 0.846. Similarly for the other databases, the selected SVs corresponded to either relatively low or high distortion levels. This appears to be a reasonable and intuitive selection of SVs for visual quality assessment because images with very low and very high distortions are representative of the overall visual quality range variations. The significance of this can be explained using (17), where one can see that the term $K(\boldsymbol{x}_i, \boldsymbol{x})$ represents the similarity between the SVs $\boldsymbol{x}_i$ and the test image $\boldsymbol{x}$. Obviously, if the test image is of higher quality, it will yield a greater kernel similarity value [i.e., $K(\boldsymbol{x}_i, \boldsymbol{x})$ will be bigger] with the SVs, which represent a higher quality signal. On the other hand, it will have a lower similarity [(i.e., $K(\boldsymbol{x}_i, \boldsymbol{x})$ will be smaller] with the SVs representing low quality signals. To further illustrate this point, we considered two distorted images: 1) image with white Gaussian noise, and 2) blurred image. The noisy image was of higher visual quality than the blurred image. We denote the feature vector of the noisy image as $\boldsymbol{x}_n$, whereas $\boldsymbol{x}_b$ denotes that for the blurred image. We then computed the kernel similarity scores $K(\boldsymbol{x}_i, \boldsymbol{x}_n)$ and $K(\boldsymbol{x}_i, \boldsymbol{x}_b)$ by measuring their distances from the SVs $\boldsymbol{x}_i$. Note that $K(\boldsymbol{x}_i, \boldsymbol{x}_n)$ and $K(\boldsymbol{x}_i, \boldsymbol{x}_b)$ will be n_{SV} (the number of support vectors) dimensional vectors, and their elements denote the similarity scores of the respective image feature vectors with the SVs (0 indicates no similarity and 1 means completely similar). We show the kernel similarity of the feature vectors for noisy and blurred images in Fig. 10, where the plot in (a) are the similarity scores with the SVs corresponding to lower quality images ($\text{MOS} < 2$), while the plot in (b) shows the similarity with the SVs corresponding to higher quality images ($\text{MOS} > 6.5$). We chose $\text{MOS} < 2$ and $\text{MOS} > 6.5$ because in the TID database $0 < \text{MOS} < 9$ with 0 denoting worse quality and 9 indicating best quality. One can observe from Fig. 10 that the noisy image tends to have A higher similarity with SVs corresponding to higher quality images and lower similarity

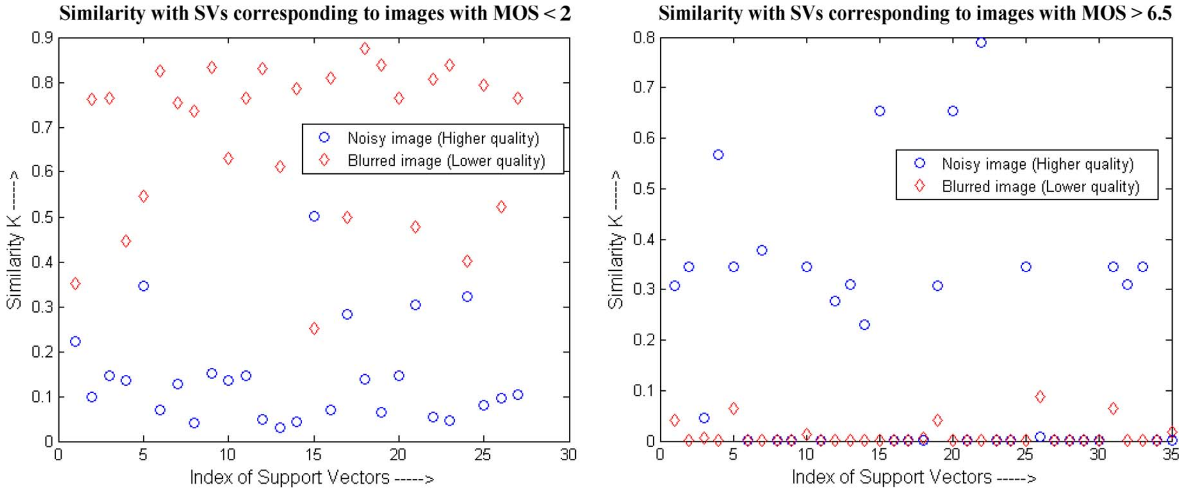


Fig. 10. (a) Kernel similarity scores of the noisy and blurred images with the SVs corresponding to lower quality images $(\mathrm{MOS} < 2)$. (b) Kernel similarity scores of the noisy and blurred images with the SVs corresponding to lower higher images $(\mathrm{MOS} > 6.5)$.

with SVs corresponding to lower quality images. On the other hand, the blurred image shows the opposite trend. Examination of the corresponding scaling factors $(\eta i^* - \eta i)$ reveals that they are generally large and positive for the SVs corresponding to higher quality images. In contrast, they are either small or negative for the SVs corresponding to lower quality images. Because the final quality score is a summation [see (17)] of the similarity scores scaled by the corresponding factor (the bias is same), this results in a higher quality score for noisy image and lower quality score for the blurred image. Therefore, the selection of SVs provides an insightful explanation of how the system predicts quality. This also highlights the effectiveness of the proposed SVD features since they enable proper selection of the SVs by allowing adequate distinction between images of different perceived qualities. As mentioned before, the linear kernel resulted in lower prediction accuracy than the RBF kernel. This is because with the use of the nonlinear kernel (RBF), the aforesaid kernel similarity is measured in a new transformed space instead of the original feature space. This enables the SVR algorithm to better distinguish/differentiate between images of different qualities, which may otherwise not be easily distinguished in the original space.

We also observed that the number of SVs was much smaller compared with the number of training samples. This is advantageous from the point of view of computation requirement since we can see from (16) that the weight vector calculation depends on the number of SVs. Naturally, a smaller number of SVs reduces the computational requirement. The number of SVs was found to decrease with increasing $\in$ value, which is expected since more samples fall within the $\in$-tube, and the associated performance changes were graceful. For example, the experiments with the LIVE image database show that the number of SVs decreases from 295 for which $C_P = 0.9677$ to as low as 54 (i.e., only 7% of data points) for which $C_P = 0.9579$.

We have a few additional remarks for feature selection. First of all, the smaller number of SVs as a result of SVR training indicates the efficiency of the proposed feature selection and SVR formulation. Second, as we know, the metrics MSVD, Q_{vector}, and Q use singular values, singular vectors, and their combination, respectively; as demonstrated consistently in Figs. 5, 6, 8, and 9, the performance of these three increases in the aforementioned order with each performance assessment criteria, namely, C_P, C_{s}, RMSE, and F-test. It can be concluded that as analyzed and expected, singular vectors and singular values together provide a more comprehensive basis for visual quality assessment.

Finally, to enable other researchers to use and compare the proposed method with new metrics and/or on new image/video databases, we have made the Matlab implementation of our codes and other useful parameters (such as the selected SVs) publicly available at http://www.ntu.edu.sg/home/wslin/codes_smc.rar. We have also included the four models, namely, Q_{CSIQ}, Q_{LIVE}, Q_{TID}, and $Q_{\mathrm{watermark}}$. As pointed out earlier, we used publicly available image/video databases, and these can be obtained from their respective references. In view of these, the results reported in the paper can easily be reproduced and can be helpful for further studies.

VIII. Conclusion

Visual quality assessment is an important research problem with wide ranging applications in visual processing systems. In this paper, to tackle effective feature detection and fusion in visual quality evaluation, we have proposed an SVR-based metric that operates with SVD-based features as the input. The feature selection based on comprehensive SVD analysis is novel, since adaptively determined singular vectors allow the capturing of structural information for each image (or a frame in video), and the separation of luminance and structural information enables their differentiation toward the assessment of perceptual quality.

We have used SVR to result in a model for combining the SVD features to predict the perceptual quality score. With the proposed model, we have avoided *a priori* assumptions on the distortion statistics (as an important advantage over the existing pooling methods) and exploited the underlying common patterns associated with visual quality degradation characterized by structural and luminance/textural changes (that is, training with specific visual and/or distortion content is not necessary). Each high dimensional feature vector was mapped into a perceptual quality score that is better aligned with the subjective viewing ground truth. We have devoted a significant portion of this paper for the experimental results and the related analysis to provide thorough and convincing ground for the proposed scheme. The proposed scheme is found to be consistently better in its prediction accuracy than the eight existing metrics across all the ten public databases, which span a wide variety of visual and distortion content. It performs well for visual and distortion content that do not appear in the training set (within a same database and also across different databases).The robustness to untrained images and distortions is crucial since in practice the visual and distortion contents are generally unknown. Finally, this paper also provides more insights regarding the support vectors and their role in visual quality prediction.

REFERENCES

[1] B. Girod, "What's wrong with mean-squared error?" in *Digital Images and Human Vision*, A. B.Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 207–220.

[2] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2002, pp. IV-3313–IV-3316.

[3] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[4] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 179–206.

[5] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 163–178.

[6] J. O. Limb, "Distortion criteria of the human viewer," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-9, no. 12, pp. 778–793, Dec. 1979.

[7] W. Lin, L. Dong, and P. Xue, "Visual distortion gauge based on discrimination of noticeable contrast changes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 900–909, Jul. 2005.

[8] P. G. J. Barten, "Contrast sensitivity of the human eye and its effects on image quality," in *Proc. SPIE*, Bellingham, WA, 1999, pp. 1–212.

[9] B. A. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer Assoc., 1995.

[10] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[11] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 1–14, Apr. 2004.

[12] Z. Wang and A. Bovik, *Modern Image Quality Assessment*. San Rafael, CA: Morgan & Claypool Publ., 2006.

[13] D. Rouse and S. Hemami, "Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM," in *Proc. Western New York Image Process. Workshop*, Rochester, NY, Oct. 2007.

[14] D. Tao, X. Li, W. Lu, and X. Gao, "Reduced-reference IQA in contourlet domain," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 6, pp. 1623–1627, Dec. 2009.

[15] H. Han, D. Kim, and R. Park, "Structural information-based image quality assessment using LU factorization," *IEEE Trans. Consum. Electron.*, vol. 55, no. 1, pp. 165–171, Feb. 2009.

[16] D. Kim and R. Park, "New image quality metric using the Harris response," *IEEE Signal Process. Lett.*, vol. 16, no. 7, pp. 616–619, Jul. 2009.

[17] S. Karunasekera and N. Kingsbury, "A distortion measure for blocking artifacts in images based on human visual sensitivity," *IEEE Trans. Image Process.*, vol. 4, no. 6, pp. 713–724, Jun. 1995.

[18] C. Warring and X. Liu, "Face detection using spectral histograms and SVMs," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 3, pp. 467–476, Jun. 2005.

[19] C. Gruber, T. Gruber, S. Krinninger, and B. Sick, "Online signature verification with support vector machines based on LCSS kernel functions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 4, pp. 1088–1100, Aug. 2010.

[20] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "Biologically inspired features for scene classification in video surveillance," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 307–313, Feb. 2011.

[21] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE ICIP*, 2006, pp. 2945–2948.

[22] H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[23] A. Eskicioglu, A. Gusev, and A. Shnayderman, "An SVD-based grayscale image quality measure for local and global assessment," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 422–429, Feb. 2006.

[24] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.

[25] D. Kalman, "A singularly valuable decomposition: The SVD of a matrix," *College Math. J.*, vol. 27, no. 1, pp. 2–23, 1996.

[26] A. Targhi and A. Shademan, "Clustering of singular value decomposition of image data with applications to texture classification," in *Proc. SPIE Vis. Commun. Image Process.*, Lugano, Switzerland, Jul. 2003, vol. 5150, pp. 972–979.

[27] Y. Tian, T. Tan, Y. Wang, and Y. Fang "Do singular values contain adequate information for face recognition?" *Pattern Recognit.*, vol. 36, pp. 649–655, 2003.

[28] J. Robinson and V. Kecman, "Combining support vector machine learning with the discrete cosine transform in image compression," *IEEE Trans. Neural Netw.*, vol. 14, no. 4, pp. 950–958, Jul. 2003.

[29] B. Scholkopf and A. Smola, *Learning With Kernels*. Cambridge, MA: MIT Press, 2002.

[30] C. Chang and C. Lin, LIBSVM: A Library for Support Vector Machines, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[31] S. Tourancheau, F. Autrusseau, P. Z. M. Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *Proc. IEEE Int. Conf. Image Process.*, 2008, pp. 365–368.

[32] VQEG, Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II, Aug. 2003. [Online]. Available: http://www.vqeg.org

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

[34] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[35] H. Sheikh, Z. Wang, A. Bovik, and L. Cormack, *Image and Video Quality Assessment Research at LIVE*. [Online]. Available: http://live.ece.utexas.edu/research/quality/

[36] Y. Horita, Y. Kawayoke, and Z. Sazzad, *Image Quality Evaluation Database*. [Online]. Available: http://160.26.142.130/toyama_database.zip

[37] P. Le Callet and F. Autrusseau, *Subjective Quality Assessment IRCCyN/IVC Database*. [Online]. Available: http://www2.irccyn.ec-nantes.fr/ivcdb/

[38] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti, "Color image database for evaluation of image quality metrics," in *Proc. Intern. Workshop Multimedia Signal Process.*, Australia, Oct. 2008, pp. 403–408.

[39] M. Gaubatz, *Metrix MUX Visual Quality Assessment Package*. [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/

[40] J. Tani, R. Nishimoto, and M. Ito, "Codevelopmental learning between human and humanoid robot using a dynamic neural network model," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 1, pp. 43–59, Feb. 2008.

[41] A57 Dataset. [Online]. Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html

[42] D. Montgomery and G. Runger, *Applied Statistics and Probability for Engineers*. New York: Wiley-Interscience, 1999.

[43] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[44] G. Stewart, "Stochastic perturbation theory," *SIAM Review*, vol. 32, no. 4, pp. 579–610, Dec. 1990.

[45] J. Liu, X. Liu, and X. Ma, "First order perturbation analysis of singular vectors in singular value decomposition," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3044–3049, Jul. 2008.

[46] Z. Devèi and S. Lonèari, "SVD block processing for non-linear image noise filtering," *J. Comput. Inf. Technol.*, vol. 7, no. 3, pp. 255–259, 1999.

[47] W. Qi, A. Morimoto, R. Ashino, and R. Vaillancourt, "Image denoising using spline and block singular value decomposition," in *Sci. Proc. Riga Tech. Univ.*, 2004, vol. 21, pp. 36–46.

[48] P. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Mach. Learn.*, vol. 48, no. 1–3, pp. 85–113, Jul. 2002.

[49] S. Winkler, "Perceptual video quality metrics—A review," in *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. Boca Raton, FL: CRC Press, 2005, ch. 5.

[50] M. Narwaria and W. Lin, "Scalable image quality assessment based on structural vectors," in *Proc. IEEE Int. Workshop MMSP*, Rio de Janeiro, Brazil, Oct. 5–7, 2009, pp. 1–6.

[51] U. Engelke, M. Kusuma, H. J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Process. Image Commun.*, vol. 24, no. 7, pp. 525–547, Aug. 2009.

[52] A. Moorthy and A. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[53] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *Proc. IEEE ICIP*, 2007, pp. II-169–II-172.

[54] E. Larson, C. Vu, and D. Chandler, "Can visual fixation patterns improve image quality assessment?" in *Proc. IEEE ICIP*, 2008, pp. 2572–2575.

[55] Q. Ma and L. Zhang, "Image quality assessment with visual attention," in *Proc. ICPR*, Dec. 8–11, 2008, pp. 1–4.

[56] U. Engelke, V. X. Nguyen, and H. Zepernick, "Regional attention to structural degradations for perceptual image quality metric design," in *Proc. ICASSP*, 2008, pp. 869–872.

[57] J. You, A. Perkis, M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. ACM Int. Conf. MM*, Beijing, China, Oct. 19–24, 2009, pp. 561–564.

[58] P. Gastaldo, S. Rovetta, and R. Zunino, "Objective quality assessment of MPEG-2 video streams by using CBP neural networks," *IEEE Trans. Neural Netw.*, vol. 13, no. 4, pp. 939–947, Jul. 2002.

[59] P. Callet, V. Christian, and B. Dominique, "A convolutional neural network approach for objective video quality assessment," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1316–1327, Sep. 2006.

[60] A. Bouzerdoum, A. Havstad, and Beghdadi, "Image quality assessment using a neural network approach," in *Proc. 4th IEEE Int. Symp. Signal Process. Inf. Technol.*, 2004, pp. 330–333.

[61] P. Carrai, I. Heynderickz, P. Gastaldo, R. Zunino, and P. Monza, "Image quality assessment by using neural networks," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 5, pp. V-253–V-256.

[62] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1/2, pp. 245–271, Dec. 1997.

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

[64] E. Larson and D. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, p. 011 006, Jan. 2010.

[65] M. Sendashonga and F. Labeau, "Low complexity image quality assessment using frequency domain transforms," in *Proc. Int. Conf. Image Process.*, 2006, pp. 385–388.

[66] T. Falk and W. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[67] U. Engelke, H. Zepernick, and M. Kusuma, *Wireless Imaging Quality Database*, 2010. [Online]. Available: http://www.bth.se/tek/rcg.nsf/pages/wiq-db

[68] F. Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of H.264/Avc video sequences transmitted over a noisy channel," in *Proc. 1st Int. Workshop QoMEX*, San Diego, CA, Jul. 2009, pp. 204–209.

[69] C. Yang, "Inverted pattern approach to improve image quality of information hiding by LSB substitution," *Pattern Recognit.*, vol. 41, no. 8, pp. 2674–2683, Aug. 2008.

[70] S. Channappayya, A. Bovik, and R. Heath, "Rate bounds on SSIM index of quantized images," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1624–1639, Sep. 2008.

[71] Y. Huang, T. Ou, P. Su, and H. Chen, "Perceptual rate-distortion optimization using structural similarity index as quality metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1614–1624, Nov. 2010.

[72] A. Aznaveh, A. Mansouri, F. Azar, and M. Eslami, "Image quality measurement besides distortion type classifying," *Opt. Rev.*, vol. 16, no. 1, pp. 30–34, Jan. 2009.

[73] M. Narwaria and W. Lin, "Objective image quality assessment with singular value decomposition," in *Proc. 5th Int. Workshop VPQM*, Scottsdale, AZ, Jan. 13–15, 2010.

[74] A. Mansouri, A. Aznaveh, F. Azar, and J. Jahanshahi, "Image quality assessment using the singular value decomposition theorem," *Opt. Rev.*, vol. 16, no. 2, pp. 49–53, Mar. 2009.

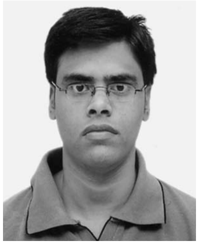
[75] F. Autrusseau, *Subjective Quality Assessment-Fourier Subband Database*, 2009. [Online]. Available: http://www.irccyn.ec-nantes.fr/~autrusse/Databases/FourierSB/

[76] W. Lin and C. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, May 2011.

[77] K. Okarma, "Combined full-reference image quality metric linearly correlated with subjective assessment," in *Proc. 10th Int. Conf. Artif. Intell. Soft Comput. Part I*, vol. 6113, *LNAI*, 2010, pp. 539–546.

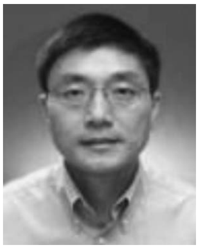
[78] K. Okarma, "Color image quality assessment using the combined full-reference metric," in *Advances in Intelligent and Soft Computing, Computer Recognition Systems 4*. New York: Springer-Verlag, 2011, pp. 287–296.

[79] X. Zhu and P. Milanfar, "Automatic parameter selection for denoising algorithms using a no-reference measure of image content," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3116–3132, Dec. 2010.



Manish Narwaria received the B.Tech. degree in electronics and communication engineering from Amrita Vishwa Vidyapeetham, Coimbatore, India, in 2008. He is currently working toward the Ph.D. degree with the School of Computer Engineering, Nanyang Technological University, Singapore.

His research interests include image/video and speech quality assessment, pattern recognition, and machine learning.



Weisi Lin (M'92–SM'98) received the B.Sc. degree in electronics and the M.Sc. degree in digital signal processing from Zhongshan University, Guangzhou, China, in 1982 and 1985, respectively, and the Ph.D. degree in computer vision from King's College, London University, London, U.K., in 1992.

He taught and researched in Zhongshan University, Shantou University, Shantou, China, Bath University, Bath, U.K., National University of Singapore, Singapore, Institute of Microelectronics, Singapore, and Institute for Infocomm Research, Singapore. He also served as the Lab Head of Visual Processing and the Acting Manager with the Department Media Processing (with 50+ research scientists), Institute for Infocomm Research. He is currently an Associate Professor with the School of Computer Engineering, Nanyang Technological University, Singapore, and also the Double Degree (business and computing) Program Coordinator in the same university. He authored over 190 scholarly publications and is an inventor of 14 patents/filings and a recipient of over S$3.8 M in research grant funding, since 1997. His areas of expertise include image processing, video compression, perceptual visual and audio modeling, computer vision, and multimedia communication.

Dr Lin is a Chartered Engineer and a Fellow of The Institution of Engineering and Technology (IET). He currently serves as Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE SIGNAL PROCESSING LETTERS and *Journal of Visual Communication and Image Representation*. He is also on four IEEE Technical Committees, and on the Technical Program Committees of a number of international conferences. He has maintained an active long-term working relationship with a number of companies. He organized special sessions in the 2006 IEEE International Conference on Multimedia and Expo (ICME06), the 2007 IEEE International Workshop on Multimedia Analysis and Processing (IMAP07), the 2010 IEEE International Symposium on Circuits and Systems (ISCAS10), the 2009 Pacific-Rim Conference on Multimedia (PCM09), the 2010 Visual Communications and Image Processing (VCIP10), and the 2011 Asia Pacific Signal and Information Processing Association (APSIPA11). He gave invited/keynote/panelist/tutorial talks in the 2007 IEEE International Conference on Computer Communications and Networks (ICCCN07), the 2009 International Workshop on Video Processing and Quality Metrics (VPQM06), VCIP10, PCM07, PCM09, IEEE ISCAS08, IEEE ICME09, APSIPA10, and the 2010 IEEE International Conference on Image Processing (ICIP10), with different topics on perceptual modeling, visual processing, and multimedia communication.