Pattern Recognition 45 (2012) 299–313

Contents lists available at ScienceDirect

Pattern Recognition

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

Scalable image quality assessment with 2D mel-cepstrum and machine learning approach

Manish Narwaria [a], Weisi Lin [a,*], A. Enis Cetin [b]

[a] School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore
[b] Department of Electrical and Electronics Engineering, Bilkent University, Ankara 06800, Turkey

ARTICLE INFO

Article history:
Received 24 February 2011
Received in revised form
20 April 2011
Accepted 21 June 2011
Available online 19 July 2011

Keywords:
Image quality assessment
Machine learning
Feature extraction
2D mel-cepstral features

ABSTRACT

Measurement of image quality is of fundamental importance to numerous image and video processing applications. Objective image quality assessment (IQA) is a two-stage process comprising of the following: (a) extraction of important information and discarding the redundant one, (b) pooling the detected features using appropriate weights. These two stages are not easy to tackle due to the complex nature of the human visual system (HVS). In this paper, we first investigate image features based on two-dimensional (2D) mel-cepstrum for the purpose of IQA. It is shown that these features are effective since they can represent the *structural* information, which is crucial for IQA. Moreover, they are also beneficial in a reduced-reference scenario where only partial reference image information is used for quality assessment. We address the second issue by exploiting machine learning. In our opinion, the well established methodology of machine learning/pattern recognition has not been adequately used for IQA so far; we believe that it will be an effective tool for feature pooling since the required weights/parameters can be determined in a more convincing way via training with the *ground truth* obtained according to subjective scores. This helps to overcome the limitations of the existing pooling methods, which tend to be over simplistic and lack theoretical justification. Therefore, we propose a new metric by formulating IQA as a pattern recognition problem. Extensive experiments conducted using six publicly available image databases (totally 3211 images with diverse distortions) and one video database (with 78 video sequences) demonstrate the effectiveness and efficiency of the proposed metric, in comparison with seven relevant existing metrics.

1. Introduction

Images and videos produced by different imaging and visual communication systems can be affected by a wide variety of distortions during the process of acquisition, compression, processing, transmission and reproduction. This generally leads to visual quality degradation due to the added noise or loss of image information. Therefore there is need to establish a criteria to measure the perceived image quality. Since the opinion of human observers is the ultimate benchmark of quality, subjective assessment is the most accurate and reliable way of assessing visual quality, if the number of subjects is sufficiently large. The International Telecommunication Union Recommendation (ITU-R) BT.500 [92] has formally defined subjective assessment as the most reliable way of IQA. However, subjective assessment is cumbersome, expensive, and unsuitable for in-service and real-time applications. Furthermore, since it is also affected by the mood and environment of the subjects, it may give less consistent results when the subject pool is not big enough. With the prospects to overcome these limitations, objective IQA has attracted significant attention over the past decade and has widespread applications. For instance, measuring image quality enables to adjust the parameters of image processing techniques in order to maximize image quality or to reach a given quality in applications like image coding [79]. Another practical use of IQA can be found in the area of information hiding [1] where secret messages are embedded into images so that an unauthorized user cannot detect the hidden messages. Because such an embedding process will degrade 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 the choice of parameters in the processing algorithm. For example, the well-known IQA metric SSIM (structural similarity index measure) [11] has been recently used as the optimization criterion in H.264 video coding algorithm [90,91].

However, objective IQA is a challenging problem [12,60,61,77,85,88] due to the inherent complex nature of the HVS and the

* Corresponding author. Tel.: +65 67906651; fax: +65 67926559.
E-mail addresses: mani0018@e.ntu.edu.sg (M. Narwaria), wslin@ntu.edu.sg (W. Lin), cetin@bilkent.edu.tr (A.E. Cetin).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.06.023

combined effect of multiple factors involved in it. The Peak signal to Noise Ratio (PSNR) is still the most widely used IQA metric but is often criticized [87] for its inability to match HVS's perception. As a result significant research effort has been put into devise alternatives to PSNR. The reader is referred to [12,60,85,88] for recent review of developments in the field of visual quality assessment.

Broadly speaking, objective IQA can be handled [12,60,85] by two approaches: (i) the vision modeling approach and (ii) the signal processing based approach. The vision modeling approach, as the name implies, is based on modeling various components of the human visual system (HVS). The HVS-based metrics aim to simulate the processes of the HVS from the eye to the visual cortex. These metrics are intuitive and appealing since they attempt to account for the properties of the HVS relevant to perceptual quality assessment. The first image and video quality metrics were developed by Mannos and Sakrison [31] and Lukas and Budrikis [4]. Later the well-known HVS-based metrics are the Visual Differences Predictor (VDP) [6], the Sarnoff JND (just noticeable difference) metric [7], Moving Picture Quality Metric [8] and Winkler's perceptual distortion metric [10]. Although the HVS-based metrics are attractive in theory, they may suffer from some drawbacks [11]. The HVS comprises of many complex processes, which work in conjunction rather than independently, to produce visual perception. However, the HVS-based metrics generally utilize results from psychophysical experiments, which are typically designed to explore a single dimension of the HVS at a time. In addition, these experiments usually use simple patterns such as spots, bars, and sinusoidal gratings, which are much simpler than those occurring in real images. For instance, psychophysical experiments characterize the masking phenomenon of the HVS by superposing a few simple patterns. In essence, these metrics suffer from drawbacks, which mainly stem from the use of simplified models describing the HVS. Moreover these metrics generally depend on the modeling of the HVS characteristics, which are not fully understood yet. While our knowledge about the HVS has been improving over the years, we are still far from a complete understanding of the HVS and its intricate mechanisms. Moreover, due to the complex and highly non-linear nature of the HVS, these metrics can be complicated and time-consuming to be used in practice. Their complexity may lead to high computational cost and memory requirement, even for images of moderate size. Owing to these limitations, the second type namely the signal processing based approach has gained popularity during recent years [60,85]. In the following sections of this paper, we will first discuss the signal processing based approach in more detail and then propose a new IQA metric based on it.

2. Signal processing based approach for IQA

The signal processing based approach [60] is based on the extraction and analysis of features in images. Feature extraction exploits various signal processing techniques to obtain suitable image representation for image quality assessment. These can be either structural image elements such as contours, or specific distortions that are introduced by a particular processing step, compression technology or transmission link, such as blocking, blurring and ringing artifacts. Metrics developed with this approach can also take into account the relevant psychophysical aspects of the HVS. With the signal processing based approach, IQA can be considered as a two stage process: (a) feature extraction and (b) feature pooling. As we have already stated, both these issues are not straightforward owing to the complex and highly non-linear nature of the HVS as well as the relatively limited understanding of its intricate mechanisms.

Regarding the issue of feature extraction, several methods/features have been proposed in literature including local variance and correlation [11], the Singular Value Decomposition [14–16], frequency domain transforms like DCT and wavelets [86], wave atoms transform [19], discrete orthogonal transforms [20], contourlet transform [21], Riesz transform [22], etc. In contrast to this, the issue of feature pooling is a relatively less investigated topic. Currently, methods like simple summation based fusion, Minkowski combination, linear or weighted combination, etc. are still widely used. These pooling techniques, however, impose constraints on the relationship between the features and the quality score. For example, a simple summation or averaging of features implicitly constraints the relationship to be linear. Similarly, the use of Minkowski summation for spatial pooling of the features/errors implicitly assumes that errors at different locations are statistically independent. Hence, there has been some research into developing alternative pooling schemes. The method presented in [24] 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, these assumptions are arbitrary and lack justification. The visual attention (VA) model has also been explored [27] for feature pooling and is based on the premise that certain regions in images attract more eye attention than the 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 automatically find regions that attract attention. Furthermore, improvement in quality prediction using VA is not yet clearly established and still open to scrutiny [26,27]. Overall, feature pooling is done largely using ad-hoc methods and therefore calls for further investigation and analysis. In our opinion, machine learning is an attractive alternative for feature pooling. Today the field of machine learning and pattern recognition finds applications not only in the traditional fields like speech recognition [29,67] but also in new and emerging research areas (for example, isolated word recognition [23] using lip reading). Machine learning has also been used for many image processing applications such as image classification [56]; image segmentation [3,52], which is often used in many video and computer vision applications such as object localization/tracking/recognition, signal compression, and image retrieval [47]; image watermarking [54]; handwriting recognition [9]; age estimation from facial images [17]; object detection [59]; sketch recognition [69]; texture classification [75], etc. We refer the reader to [43,80] for comprehensive reviews of the applications of machine learning in image processing.

In summary, while the existing features have demonstrated reasonable success for IQA, some of them lack clear physical meaning while others may not be able to tackle a wide range of image distortions. Therefore, there is need to explore new image features for more efficient IQA. In addition the existing feature pooling methods also suffer from drawbacks as already mentioned. To overcome the aforesaid problems, in this paper we propose a new IQA metric. Firstly, we explore the 2D mel-cepstrum based image features and provide a comprehensive analysis to justify their use for IQA. Secondly, given the strong theoretical foundations and proven success of machine learning in numerous applications, we employ it for feature pooling. Because 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. As opposed to the existing pooling methodologies, which usually make apriori assumptions about the mapping (relationship) between features and quality score, the related model parameters can be estimated in a more convincing manner with the use of machine learning. Stated differently, use of machine learning in

an IQA metric can help to avoid assumptions on the relative significance and relationship of different distortion statistics (i.e. feature changes), since the weight adjustment would be done via proper training with substantial ground truth.

The remainder of this paper is organized as follows. Section 3 of this paper discusses the details of the proposed visual quality metric detailing the feature extraction and pooling procedure with proper analysis and justification. We describe the databases, the training and test methodology in Section 4. Substantial experimental results and the related analysis are presented in Section 5. We explore the possibilities for reduced-reference IQA in Section 6. Finally, Section 7 gives the concluding remarks.

3. The proposed visual quality metric

In this section, we describe the details of the proposed metric whose block diagram is shown in Fig. 1. The first step is to extract the 2D mel-cepstral features from both the reference and distorted images. Then, the difference (or similarity) is computed between the two feature vectors. Finally, machine learning is used to fuse the elements of the feature vector into a single number that represents the objective quality score. Thus, we formulate IQA as a supervised pattern recognition problem.

3.1. Feature extraction using 2D mel-cepstral features

An error (or distortion) in a different context may not have the same perceptual impact on quality. For example, low pass filtering (i.e. blur) has lesser effect on the smooth areas in an image while it has a higher impact on edges. Due to this, it is important to distinguish/differentiate error in different image components. This is the reason why the PSNR (or related metrics like MSE) is less effective: it does not separate/differentiate the signal components and assigns equal weights to all the pixel errors irrespective of their perceptual impact. Therefore, the motivation behind feature extraction for IQA is to separate/differentiate the image signal into its components since their contribution to the perceived quality is different. This is a crucial step towards effective IQA because the separation of the components will then allow us to treat (i.e. weigh) them appropriately according to their perceptual significance. In this paper, we use the mel-cepstral analysis for images to extract meaningful components from the image signal.

Mel-cepstral analysis is one of the most successful and widely used feature extraction techniques in speech processing applications including speech and sound recognition [67]. Inspired by its success in various areas of audio/speech processing, we propose its exploitation to assess the quality of images objectively. The 2D mel-cepstrum has been proposed recently [68]. We now describe the feature extraction with 2D mel-cepstrum and outline its possible advantages in the context of IQA. The proposed scheme is the first attempt in the existing literature to explore the 2D mel-cepstrum for IQA.

The 2D cepstrum $\hat{c}(p,q)$ of a 2D image $y(n_1,n_2)$ is defined as

$$\hat{c}(p,q) = F_2^{-1}(\log(|Y(u,v)|^2)) \quad (1)$$

where (p,q) denotes 2D cepstral frequency [5] coordinates, $Y(u,v)$ is the 2D Discrete Fourier transform (DFT) of the image $y(n_1,n_2)$ (size N by N) and defined as

$$Y(u,v) = \frac{1}{N}\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1} y(n_1,n_2)e^{-j2\pi(un_1+vn_2/N)}$$

F_2^{-1} denotes the 2D Inverse Discrete Fourier transform (IDFT) given by

$$F_2^{-1} = \frac{1}{N}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} Y(u,v)e^{j2\pi(un_1+vn_2/N)}$$

Energy of natural images drops at high frequencies (i.e. natural images have more low frequency as compared to high frequency). Due to this, the effect of high frequency components is suppressed as the bigger values of low frequency coefficients will tend to dominate. Furthermore, the number of coefficients is very large (equal to image size). Therefore the direct use of frequency coefficients will be less effective in determining image quality. To overcome this we use 2D mel-cepstrum in which non-uniform weighting is employed to group the frequency coefficients. Specifically, in 2D mel-cepstrum the DFT domain data are divided into non-uniform bins in a logarithmic manner and the energy of each bin is computed as

$$G(m,n) = \sum_{k,l \in B(m,n)} w(k,l)\ Y(k,l) \quad (2)$$

where $B(m,n)$ is the (m,n)th cell of the logarithmic grid corresponding to weight $w(k,l)$. Cell or bin sizes are smaller at low frequencies compared to high-frequencies. A representative grid diagram is shown in Fig. 2(a) where cell sizes can be taken to represent the weights $w(k,l)$. As can be seen, cell sizes are smaller at lower frequencies compared to the higher frequencies, which are assigned higher weights. The equivalent diagrammatic representation of the non-uniform normalized weighting is shown in Fig. 2(b) where white means weight is 1 and black denotes weight is 0. The smallest value used in Fig. 2(b) is 0.005.

This approach is similar to the mel-cepstrum computation in speech processing where the weights are assigned using a mel scale in accordance with the perception of the human ear. In our earlier work [34], we have demonstrated the effectiveness of mel features for the quality assessment of noise suppressed speech. Although the weights used in case of speech signals (1D signal) are not the same as for the image (2D signal), nevertheless both are similar in concept. Like speech signals, most natural images contain more low frequency information. Therefore, as mentioned, there is more signal energy at low-frequencies compared to high

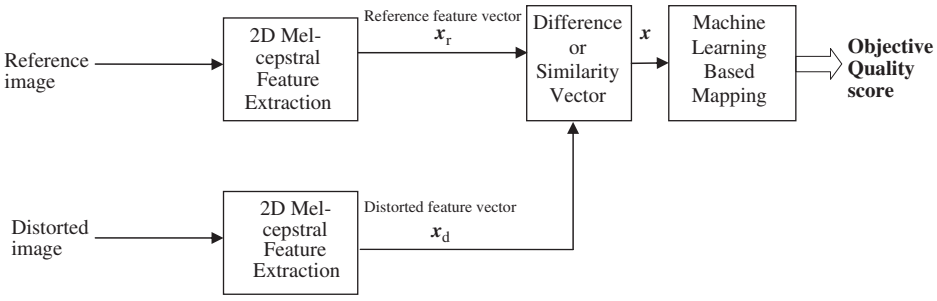

Fig. 1. Block diagram of the proposed scheme.

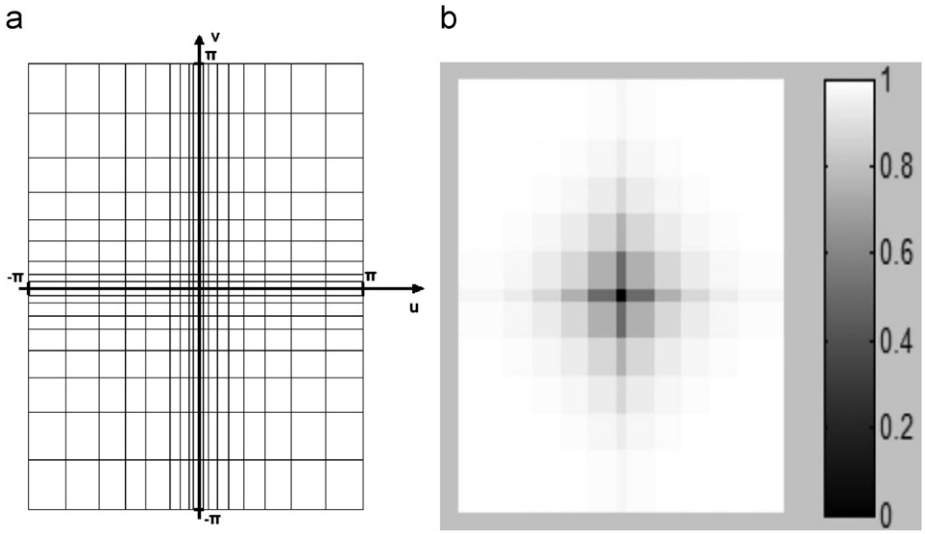

Fig. 2. (a) Non-uniform grid representation with smaller cell sizes at low frequencies compared to high frequencies and (b) representation of the normalized weights for emphasizing high frequencies (white corresponds to 1 and black corresponds to 0).

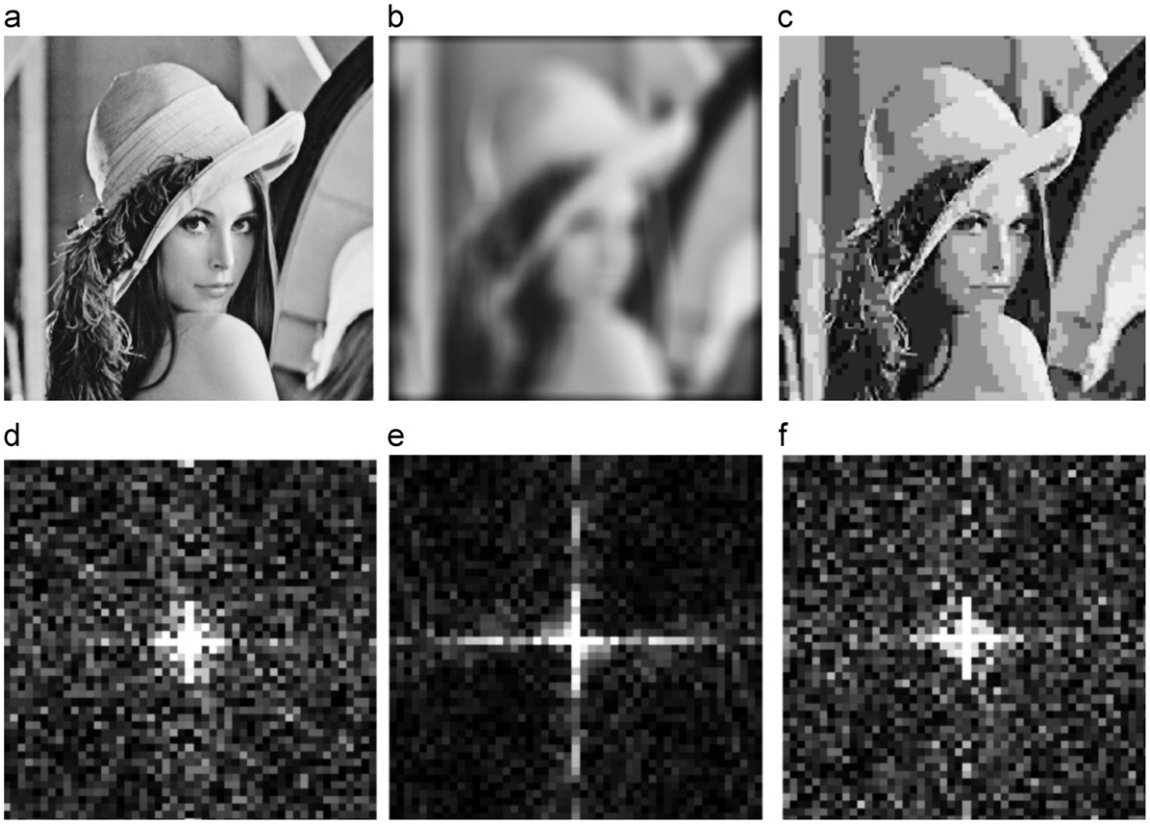

Fig. 3. (a) Original Lena image, (b) blurred image, (c) JPEG compressed image, (d) 2D mel-cepstrum of (a), (e) 2D- mel-cepstrum of (b) and (f) 2D mel-cepstrum of (c). White indicates a value of 1 (the highest strength) whereas black corresponds to 0 (zero strength).

frequencies. Logarithmic division of the DFT grid emphasizes high frequencies. Finally, the 2D mel frequency cepstral coefficients $\hat{c}(p,q)$ are computed using DCT or inverse DFT (IDFT) as

$$\hat{c}(p,q) = F_2^{-1}(\log(|G(m,n)|^2)) \quad (3)$$

Note that in Eq. (3) we use the absolute value of the bin energy $G(m,n)$ (magnitude) and discard phase for reasons given later in Section 6B.

We now analyze why the 2D mel-cepstral features form a good image representation for quality assessment. Psychovisual studies have shown that edges, texture and smooth components in images have different influence on the HVS's perception. The HVS is more sensitive to image areas containing edges [81,82]. Further, image content recognition is widely believed to rely on the perception of image details, such as sharp edges, which are conveyed by high spatial frequencies [83,84]. Therefore edges and other higher frequency components are perceptually significant

[13,50,51,71]. Due to this such features have also been used for IQA. For instance, the well known metric SSIM has been improved [28] by incorporating edge information. Some other IQA metrics based on edge information can be found in [30,32,48,64,73]. Recently image contours/edges have also been explored for image utility assessment [2], which is related to IQA. It follows that edges/contours are more important for HVS's perception i.e. any change in the high frequency components is expected to have a larger impact on perceived image quality. Therefore, accounting for the difference in perception of edges, texture and smooth components by the HVS is beneficial for IQA. The 2D mel-cepstrum precisely achieves this using unequal weights for different frequency components as shown in Fig. 2. As a result, high frequency components (which correspond to strong edges) can be further emphasized. Apart from this, the lower frequency components (like weak edges and texture), which have relatively less influence on the HVS are assigned smaller weights. The said non-uniformity therefore results in better representative image features. It also accounts for the masking property of the HVS: in the presence of a strong edge, the weaker edge is masked i.e. its influence is reduced. In other words, the stronger edges tend to dominate, i.e. they have higher weight or impact. This is also the reason why the 2D mel-cepstrum representation is suitable for face recognition [33,35] (since it highlights edges and other facial features in the face image).

To give an illustration, we show the original 'Lena' image, its blurred version and JPEG compressed version in Fig. 3(a), (b) and (c), respectively. The corresponding 2D mel-cepstrum of the images is shown below the respective images. We observe that blurring mainly damages the high frequency components. This can be visualized through its 2D mel-cepstrum where the strength of high frequency components is reduced. We can also see that the strength of lower frequency components is increased since blur makes the image more uniform. In the extreme case, if all pixels have the same value then we will see only one white spot exactly in the center of the 2D mel-cepstrum (i.e. the DC component). The case of JPEG compression is different in that it causes blockiness and can introduce false structure or edges in the image. This can again be captured from the 2D mel-cepstrum features because the strength/magnitude of frequency components changes due to distortions. Therefore, a comparison between the 2D mel-cepstrum features of the reference and distorted image is expected to give a good indication of change in image spatial content (or structural change).

To summarize, the following are the major advantages of the 2D mel-cepstrum, which can be exploited for IQA:

- The non-uniform weighting is consistent with the edge masking property of the HVS. Because it is possible to emphasize the high frequency components apart from retaining the lower frequency ones, a more informative and comprehensive representation can be obtained. Specifically, it provides more details about features like edges and contours, which are important for the HVS's perception of image quality. Therefore, it is more effective compared to other transforms since more discriminatory and meaningful image signal components can be extracted.
- Since several DFT values are grouped together in each bin, the resultant 2-D mel-cepstrum sequence computed using the IDFT has smaller dimensions than the original image. It can therefore be viewed as a *perceptually* motivated dimension reduction tool, which can preserve image structure. That is, it can be considered as a good tradeoff between retaining important image information and achieving dimensionality reduction. In other words, perceptually important frequencies are enhanced and the feature size is also reduced. For an N by N image, using the 2D mel-cepstrum we can obtain the dimension reduced data M by M with $M < N$.
- We obtain decorrelated features, so the redundant information is discarded. This results in a compact numerical representation of the image signal to characterize its quality. Thus, the advantage of 2D mel features is that they produce representations that are statistically independent and comprise an orthogonal space.
- Another advantage of 2D mel-cepstral feature is that small change in the features corresponds to small change in perceptual quality and vice-versa. This implies that they can also capture small changes or differences of pixel intensity (magnitude) more efficiently. This property is especially crucial for quality prediction of images with near threshold (i.e. just noticeable) distortions as will be demonstrated later in Section 5 of the paper.
- The reader will notice from Eq. (3) that 2D mel-cepstrum involves the logarithms of the squared bin energies denoted by $|G(m,n)|^2$. This reduces the dynamic range of the values and is consistent with the so-called "suprathreshold effect" of the HVS. Suprathreshold effect [55,57,72,76] means that the ability to perceive variations in the distortion level decreases as the degree of distortion increases. The logarithm operation essentially accomplishes this desirable property as elaborated later in Eq. (5) and shown graphically in Fig. 4.
- The 2D mel-cepstrum is also associated with clearer physical meaning because it essentially works in the Fourier (frequency) domain, which is a well established method for image analysis. However, in the Fourier or DCT domain one usually discards the higher frequency components (for example JPEG compression) in order to achieve dimension reduction. By contrast in 2D mel-cepstrum, the high frequency DFT and DCT coefficients are not discarded in an ad-hoc manner. Instead the high frequency component cells of the 2D DFT grid are multiplied with higher weights as compared to the low frequency component bins in the grid resulting in more suitable image representation for IQA.
- The non-uniform weighting shown in Fig. 2 is perceptually meaningful and can be further exploited to design a reduced-reference metric as discussed later in Section 6.

Let $\boldsymbol{x}_r$ and $\boldsymbol{x}_d$ denote the 2D mel-cepstral features of the reference and distorted images, respectively. The vectors $\boldsymbol{x}_r$ and $\boldsymbol{x}_d$ can be thought to represent the *timbral texture space* [66] of the two image signals and we use them to quantify perceived similarity between them. This is similar at the concept level to tasks like computing music similarity [63], genre classification [65], etc. in the field of

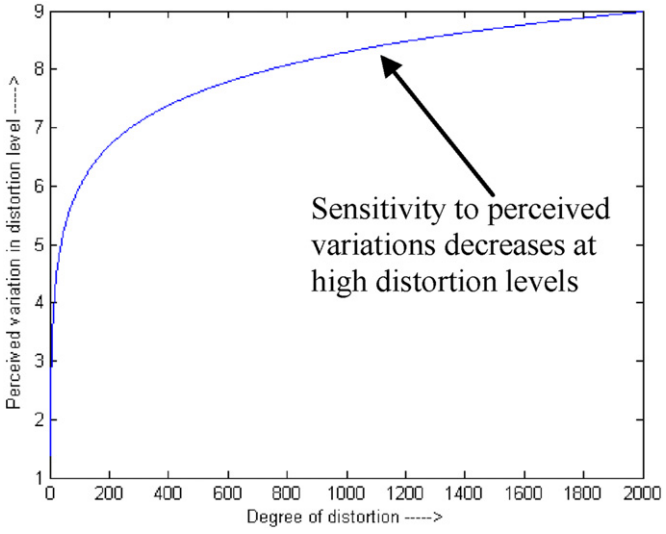

Fig. 4. Illustration of the suprathreshold effect.

audio/speech processing. Because our aim is to compute quality of the distorted image with respect to the reference image, we use the absolute difference between the two feature vectors for computing quality of the distorted image and define

$$\boldsymbol{x} = |\boldsymbol{x}_r - \boldsymbol{x}_d| \tag{4}$$

We can see that the elements of $\boldsymbol{x}$ represent the absolute difference between the 2D mel-cepstrum coefficients of the reference and distorted images. This lends $\boldsymbol{x}$ a better physical meaning since its elements can be thought as the change in frequency components of the reference image due to distortion, i.e. it accounts for the loss of image spatial information. Therefore, (4) defines the feature vector of the distorted image, which will be used to compute its quality.

As aforesaid, suprathreshold effect implies that the same amount of distortion becomes perceptually less significant as the overall distortion level increases. Researchers have previously modeled suprathreshold effect using visual impairment scales [18] that map error strength measures through concave non-linearities, qualitatively similar to the logarithm mapping, so that they emphasize the error at higher quality. The definition of feature vector in (4) accounts for this effect and can be explained as follows. Eq. (4) can be written as

$$\begin{aligned}\boldsymbol{x} = |\boldsymbol{x} - \boldsymbol{x}_d| &= |F_2^{-1}(\log(|G_r(m,n)|^2)) - F_2^{-1}(\log(|G_d(m,n)|^2))| \\ &= \left| F_2^{-1}\left[\log\left\{\frac{|G_r(m,n)|^2}{|G_d(m,n)|^2}\right\}\right]\right| \end{aligned} \tag{5}$$

where $G_r(m,n)$ and $G_d(m,n)$ denote the bin energies from reference and distorted images, respectively. We can observe from (5) that the ratio of the squares of absolute bin energies can be regarded as the distortion measure on which suprathreshold function (logarithm) has been applied. For a simple intuitive explanation, consider the two quantities $\log\{60/40\} = 0.4055$ and $\log(1020/1000) = 0.0198$. As we can see, the difference between the numerator and denominator in the two cases is the same (it is 20). However, the perceived change is smaller in the second case. That is lesser sensitivity to changes at larger amplitudes, which is the suprathreshold effect or the saturation effect as visually exemplified in Fig. 4.

As mentioned before, high frequencies are assigned more weight. Therefore $\boldsymbol{x}$ is expected to be an effective feature vector characterizing the loss of image structure. To illustrate this point further, we show 7 images in Fig. 5. In this, image (a) is the original image taken from the LIVE image database (details of the database are given later). Images (b)–(d) have been obtained by blurring the original image with increasing blur levels. On the other hand, images (e)–(g) have been generated by JPEG compression of the original image with increasing compression levels. As can be seen, the increasing blurring reduces the high frequency content of the original image and destroys its spatial information. Similarly in JPEG compression the high-frequency components are largely removed owing to non-uniform quantization and these result in blockiness as shown in the second row of Fig. 5. We also computed the feature vector for each distorted image

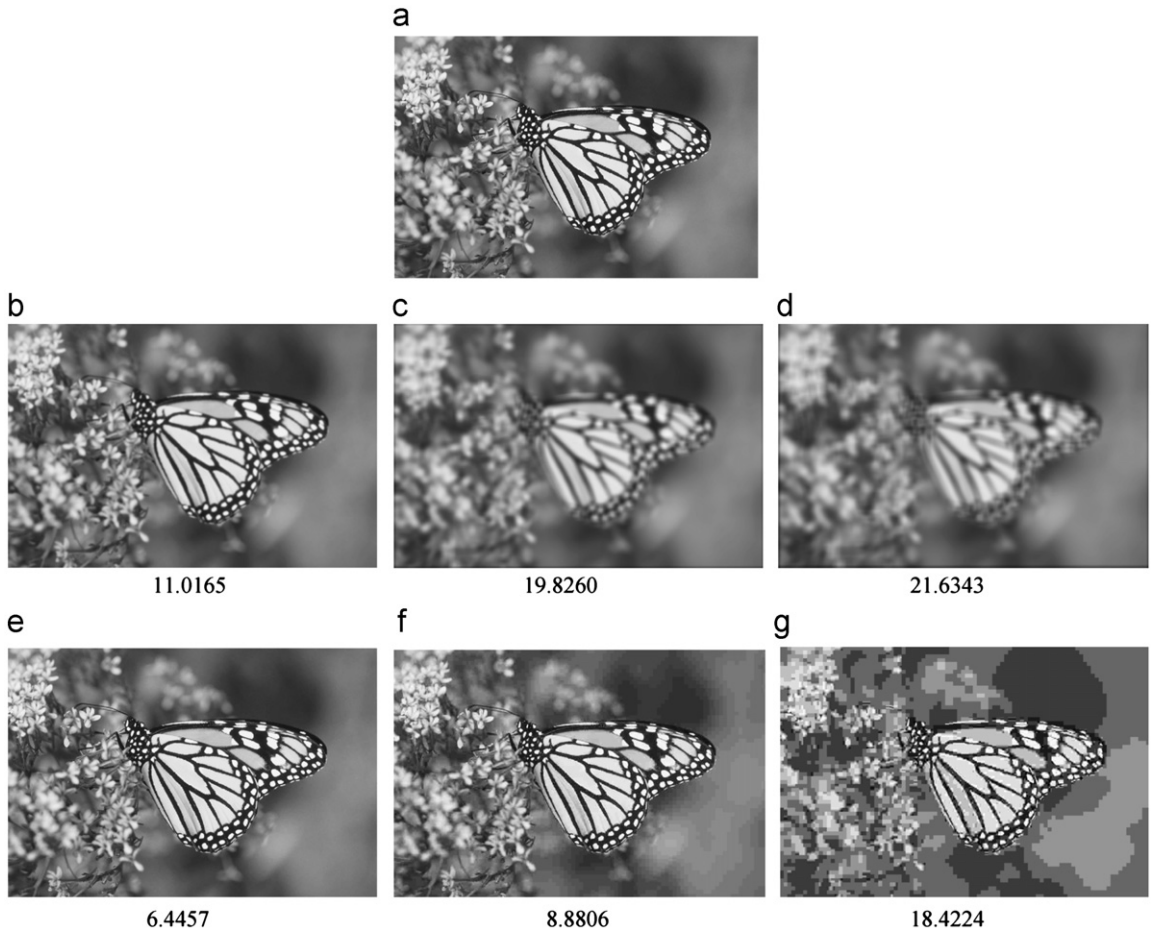

Fig. 5. (a) Original image, (b) low blurring, (c) medium blurring, (d) high blurring, (e) low JPEG compression level, (f) medium JPEG compression level and (g) high JPEG compression level. The number below each respective image denotes the sum of the elements of the feature vector defined in (4). A higher number denotes more loss of spatial information i.e. higher distortion.

with respect to the original image as done in (4). Next, we obtained the sum of the elements of the respective feature vectors for each image and the same has been indicated below each respective image. We find that the sum is large for the heavily blurred image (Fig. 5(d)) i.e. large loss of spatial information, while it is small for the less blurred image. A similar trend can be seen for the JPEG distorted images. That is, we get an indication of the loss of spatial information due to artifacts like blur and JPEG, which can damage image structure. Of course a simple summation of the elements of feature vector alone will be insufficient for determining overall quality especially in case of complex and diverse distortion types. Nevertheless, this analysis indicates that the feature vector $\boldsymbol{x}$ defined in (4) can be expected to be effective for assessing the extent of structure damage or the change in image spatial information due to the external perturbation (distortion). Based on the foregoing analysis, we conclude that $\boldsymbol{x}$ accounts for perceptual properties such as sensitivity to loss of structure, edge masking and the suprathreshold effect. Furthermore, $\boldsymbol{x}$ can be used to assess quality independent of the distortion or image content and the reason is as follows. Different types of distortions affect visual quality in a largely similar fashion: by introducing structural changes (or change in spatial contents) that lead to different extents of perceived quality degradation. That is, even though $\boldsymbol{x}$ does not take into account the effects of different distortions explicitly, the perceptual annoyance introduced by them is expected to be captured reasonably well. Due to the existence of the underlying common patterns associated with quality degradation, machine learning can be exploited to develop a general model by learning through examples as will be demonstrated by extensive experimental results in Section 5. Hence $\boldsymbol{x}$ can be used to compute quality in general situations. Of course, we must still combine/pool the elements of $\boldsymbol{x}$ with proper weights for which we use machine learning as explained in the next section.

3.2. Combining features into a perceptual quality score

Appropriate feature pooling is an essential step for perceptual IQA but there is lack of physiological and psychological knowledge for the convincing modeling (the psychophysical studies that have been conducted in the related field are for a single or at most two visual stimuli (e.g. in frequency, orientation, etc.)), while real-world images are with many stimuli simultaneously. Therefore, we use machine learning to tackle the complex issue of feature pooling.

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

$$Q = f(\boldsymbol{x}) \tag{6}$$

To estimate f we use a machine learning approach, which is expected to give a more reasonable estimate compared to the existing pooling approaches, especially when the number of features to be pooled is large. In this work, we use the Support Vector Regression (SVR) to map the high dimensional feature vector into a perceptual quality score, by estimating the underlying complex relationship among the changes in cepstral features and the perceptual quality score. Although other choices of machine learning techniques are possible, in this paper, we have used SVR because it is a popular and well established technique.

The goal of SVR is to find f, based on training samples. Suppose that $\boldsymbol{x}_i$ is the feature vector of the ith image in the training image set ($i=1, 2, \ldots l$; l is the number of training images). In the $\in$−SV regression [36,78] the goal is to find a function $f(\boldsymbol{x}_i)$ 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 [36]. The function to be learned is $f(\boldsymbol{x})=\boldsymbol{W}^T\varphi(\boldsymbol{x})+b$, where $\varphi(\boldsymbol{x})$ is a non-linear function of $\boldsymbol{x}$, $\boldsymbol{W}$ is the weight vector and b is the bias term. We find the unknowns $\boldsymbol{W}$ and b from the training data such that the error

$$|s_i - f(\boldsymbol{x}_i)| \le \in \tag{7}$$

for the ith training sample $\{\boldsymbol{x}_i, s_i\}$. In SVR, a kernel function $\phi(\boldsymbol{x})$ is employed to map the data into a higher dimensional space. We solve the following optimization problem

$$\min_{W,b,\xi,\xi^*} \frac{1}{2}\boldsymbol{W}^T\boldsymbol{W} + C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$$
$$\text{subject to } \begin{cases} s_i-(\boldsymbol{W}^T\phi(x_i)+b) \le \in +\xi_i \\ (\boldsymbol{W}^T\phi(x_i)+b)-s_i \le \in +\xi_i^* \\ \xi_i,\xi_i^* \ge 0 \end{cases} \tag{8}$$

where ξ_i is the upper training error (ξ_i^* is the lower training error) subjected to the $\in$ insensitive tube $|y-(\boldsymbol{W}^T\phi(\boldsymbol{x})+b)| \le \in$ with $\in \nearrow$ being a threshold; $(1/2)\mathbf{W}^T\mathbf{W}$is the regularization term to smooth the function $\boldsymbol{W}^T\varphi(\boldsymbol{x})+b$ in order to avoid overfitting; $C>0$, being the penalty parameter of the error term. Eq. (8) can be solved using the dual formulation to obtain the solution ($\boldsymbol{W}$,b).

It has been shown in [36] that

$$W = \sum_{i=1}^{n_{sv}}(\eta_i^*-\eta_i)\varphi(\boldsymbol{x}_i) \tag{9}$$

where η_i^* and η_i ($0 \le \eta_i^*, \eta_i \le C$) are the Lagrange multipliers used in the Lagrange function optimization, C is the tradeoff error parameter and nsv is the number of support vectors. For data points for which inequality (7) is satisfied, i.e. the points, which lie within the $\in$ tube, the corresponding η_i^* and η_i will be zero so that the Karush Kuhn Tucker (KKT) conditions are satisfied [36]. The samples that come with nonvanishing coefficients (i.e. non-zero η_i^* and η_i) are support vectors, and the weight vector $\boldsymbol{W}$ is defined only by the support vectors (not all training data). The function to be learned then becomes

$$f(\boldsymbol{x}) = \boldsymbol{W}^T\varphi(\boldsymbol{x}_i)+b = \sum_{i=1}^{n_{sv}}(\eta_i^*-\eta_i)\varphi(\boldsymbol{x}_i)^T\varphi(\boldsymbol{x})+b$$
$$= \sum_{i=1}^{n_{sv}}(\eta_i^*-\eta_i)K(\boldsymbol{x}_i,\boldsymbol{x})+b \tag{10}$$

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 $\{\boldsymbol{x}_i, s_i\}$, and the unknowns $\boldsymbol{W}$ and b are estimated to obtain the desired function (10). 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). In this paper, we have used the Radial Basis Function (RBF) as the kernel, 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.

4. Databases and metric verification

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 [38]. We have therefore conducted extensive experiments on totally 7 open databases. As will be shown in Section 5 of this paper, a metric performing well on 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 procedure adopted to verify the proposed approach.

4.1. Database description

The LIVE image database [39] includes 29 original 24-bits/pixel color images. Totally it consists of 982 images (779 distorted images and 203 reference images). Five types of distortions were introduced to obtain the distorted images: (1) JPEG-2000 compression, (2) JPEG compression, (3) White Gaussian Noise (WGN), (4) Gaussian blurring and (5) Rayleigh-distributed bit stream errors of a JPEG-2000 compressed stream or Fastfading distortions (FF). Subjective quality scores for each image are available in the form of Differential Mean Opinion Scores (DMOS).

The IRCCyN/IVC database [40] consists of 10 original color images with a resolution of 512×512 pixels from which 185 distorted images have been generated, using 4 different processes: (1) JPEG compression, (2) JPEG2000 compression, (3) LAR (locally adaptive resolution) coding and (4) blurring. Subjective quality scores are available in the form of Mean Opinion Scores (MOS).

In the A57 database [41], 3 original images of size 512×512 are distorted with 6 types of distortions and 3 contrasts. These result in 54 distorted images (3 images $\times$ 6 distortion types $\times$ 3 contrasts). The distortion types used are: (1) quantization of the LH subbands of a 5-level DWT of the image using the 9/7 filters, (2) additive Gaussian white noise, (3) baseline JPEG compression, (4) JPEG-2000 compression, (5) JPEG-2000 compression with the Dynamic Contrast-Based Quantization algorithm of which applies greater quantization to the fine spatial scales relative to the coarse scales in an attempt to preserve global precedence and (6) blurring. The subjective scores have been made available in the form of DMOS.

The Tampere Image Database (TID) database [42] involves 25 original reference color images (resolution 512×384), which have been processed by 17 different types of distortions: additive Gaussian noise, additive noise in color components, spatially correlated noise, masked noise, high frequency noise, impulse noise, quantization noise, Gaussian blur, image denoising, JPEG compression, JPEG2000 compression, JPEG transmission errors, JPEG2000 transmission errors, non-eccentricity pattern noise, local block-wise distortions of different intensity, mean shift (intensity shift) and contrast change. There are 4 distortion levels and thus it consists of 1700 ($25\times17\times4$) distorted images; there are 100 images for each distortion type. Subjective quality scores are reported in the form of MOS.

The Wireless Imaging Quality (WIQ) database [53] consists of 7 undistorted reference images, 80 distorted test images, and quality scores rated by human observers that have been obtained from two subjective tests. In each test, 40 distorted images along with the 7 reference images were presented to 30 participants. The quality scoring was conducted using a Double Stimulus Continuous Quality Scale (DSCQS). The difference scores between reference and distorted image were then averaged over all 30 participants to obtain a DMOS for each image. The test images included in the WIQ database consist of wireless imaging artifacts, which are not considered in any of the other publicly available image quality databases.

A publicly available video database [44] was also used in this study and we refer to this database as the EPFL video database. Six original video sequences at CIF spatial resolution (352×288 pixels) were encoded with H.264/AVC. For each encoded video sequence, 12 corrupted bit streams were generated by dropping packets according to a given error pattern. To simulate burst errors, the patterns have been generated at six different packet loss rates (0.1%, 0.4%, 1%, 3%, 5% and 10%) and two channel realizations have been selected for each packet loss rate. The packet loss free sequences were also included in the test material, thus finally 78 video sequences were rated by 40 subjects. Subjective scores have been made available as MOS.

Finally, we used another publicly available image database [49]. It is different from all the databases discussed above, with respect to the distortion type since the distortion is due to watermarking. It consists of 210 images watermarked in three distinct frequency ranges. The watermarking technique basically modulates a noise-like watermark onto a frequency carrier, and additively embeds the watermark in different regions of the Fourier spectrum. The subjective scores are reported as MOS.

4.2. Test procedure and evaluation criteria

We evaluate the performance of the proposed scheme in two different ways. Firstly, we have employed the k-fold cross validation (CV) strategy [45] for each database separately: the data was 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 presenting 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 a same original image. As an example, consider the TID database, which consists of 25 original images. In this case, the first chunk included all the distorted versions of the first five original images. The second chunk consisted of distorted versions of the next five original images and so on. Thus, in this case there were a total of five chunks each of which comprised different image contents. With the similar splitting procedure we obtained 10 chunks for IVC database, seven chunks for WIQ database and three chunks for A57 database. In this way, it was ensured that images appearing in the test set are not present in the training set. As the second way of performance assessment, we have used the cross database evaluation: the proposed system was trained from the images in one database and images from the remaining databases formed the test set.

A 4-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 [46], 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 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 Mean Squared Error (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 [15], VSNR [55], VIF [57] and PSNR-HVS-M [74]. For VSNR, VIF, IFC and SSIM implementation, we have used the publicly accessible Matlab package that implements a variety of visual quality assessment algorithms [58]; they are the original codes provided by the image quality assessment algorithm designers. For PSNR-HVS-M, we used the code provided by its authors and is publicly available at [70]. The publicly available LibSVM software package [78] was used to implement the SVR algorithm. In addition, the results for another recent metric presented in [20] are also reported; however, since the code is not publicly available, we derive the results directly from their paper [20] for only the databases, which had been used in [20]; also, since two metrics were proposed using geometric moments [37] with one using Tchebichef moments and the other using Krawtchouk moments, we only compare with the best results among the two. Since we have used all publicly accessible

softwares and databases in this paper as far as possible, the results reported in this paper can be reproduced for any future research.

Most of the existing visual quality metrics work only with the luminance component of the image/video. Therefore, all 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).

5. Experimental results and analysis

5.1. Performance evaluation

The results for the k-fold CV tests (denoted by Q) for the individual image databases are given in Table 1. Furthermore, for an overall comparative performance, the averaged results over the 5 image databases are given in Table 2. We computed the average values for two cases. In the first case, the correlation scores were directly averaged, while in the second case, a weighted average was computed with the weights depending on the number of distorted images in each database (refer to Section 4.1 for such numbers). We can see that the proposed Q performs better than the other IQA schemes. Recall that for Q we made sure that the images used for training did not appear in the test set. It was also found that in general, the proposed scheme performed well for individual distortion types. We can also observe from Table 2 that the proposed metric gives the better overall performance in both averaging cases for the three evaluation criteria.

Another observation from Table 1 is that some existing metrics are less consistent since they do not perform well for all the databases. For instance VSNR does well on A57 but its performance is relatively low for other databases. VIF performs well on 3 databases but performs rather poorly on A57. By contrast, the proposed scheme is more consistent in its performance. To gain more insights into such behavior of quality metrics, we perform additional analysis using the TID database. In our opinion, the variation in performance of quality metrics over the different databases is partly due to the distortion levels. For instance, A57 database mainly contains images with near-threshold distortions i.e. image quality degradation is just noticeable. On the other hand, databases like LIVE and IVC consist of images with supra-threshold distortions i.e. image quality degradation could be severe and more noticeable to the human eye. We conducted further tests to verify this. We observed the performance of different metrics for the 4 distortion levels of the TID database. The first level (Level 1) denotes just noticeable or near threshold distortion while the fourth level (Level 4) indicates higher distortions. With a total of 1700 distorted images and 4 distortion levels, there are 425 images for each distortion level. Table 3 presents the C_P values for the prediction performance of different metrics on the 4 distortion levels. The C_S and RMSE values are not presented here since they lead to similar conclusion as C_P values. We can see that MSVD, VIF, VSNR and PSNR-HVS-M perform relatively better for the fourth distortion level (i.e. higher amount of distortion) while they are relatively poor for lower distortion

Table 1
Experimental results for the 5 image databases individually. The three best metrics have been highlighted by bold font for quick glance.

Criteria	Metric	LIVE	A57	WIQ	IVC	TID
C_P	SSIM	0.9473	0.8033	0.7876	**0.9018**	0.7756
	MSVD	0.8880	0.7099	0.7433	0.7975	0.6423
	VIF	**0.9655**	0.6139	0.7559	**0.8966**	**0.8049**
	VSNR	0.9520	**0.9210**	0.7623	0.8025	0.6820
	PSNR	0.9124	0.6273	0.7601	0.7196	0.5677
	PSNR-HVS-M	0.9432	0.8896	0.8191	0.8902	0.5784
	Ref. [16]	0.9253	0.6799	–	0.8776	–
	Q	**0.9684**	**0.9021**	**0.9048**	**0.9511**	**0.8092**
	Q_{TID}	0.9519	**0.9019**	**0.8489**	0.8772	–
	Q_{LIVE}	–	0.8944	0.8473	0.8784	0.7859
	Q_{IVC}	**0.9554**	0.9008	0.8472	–	0.7840
	$Q_{watermark}$	0.9552	0.9011	**0.8480**	0.8794	**0.7881**
C_S	SSIM	**0.9500**	0.8103	0.7261	**0.9017**	**0.7792**
	MSVD	0.9102	0.6485	0.6362	0.7734	0.6520
	VIF	**0.9735**	0.6223	0.6918	**0.8964**	0.7491
	VSNR	0.9400	**0.9355**	0.6558	0.7993	0.7000
	PSNR	0.9056	0.6189	0.6257	0.6885	0.5773
	PSNR-HVS-M	0.9372	0.8962	0.7568	0.8832	0.5952
	Ref. [16]	0.9216	0.7255	–	0.8952	0.6740
	Q	**0.9599**	**0.8586**	0.8064	**0.9171**	**0.7848**
	Q_{TID}	0.9383	**0.8561**	**0.8410**	0.8677	–
	Q_{LIVE}	–	0.8532	**0.8396**	0.8690	0.7732
	Q_{IVC}	0.9442	0.8496	**0.8420**	–	0.7645
	$Q_{watermark}$	0.9433	0.8551	0.8389	0.8688	**0.7690**
RMSE	SSIM	8.0553	0.1914	13.8160	**0.5303**	0.8511
	MSVD	10.6315	0.1731	15.3228	0.7739	1.0285
	VIF	**6.0174**	0.1940	14.9964	**0.5239**	**0.7945**
	VSNR	7.0804	**0.0957**	14.8864	0.7269	0.9851
	PSNR	9.0864	0.6189	14.8856	0.8460	1.1047
	PSNR-HVS-M	8.0564	0.1156	13.1412	0.5550	**1.0947**
	Ref. [16]	–	–	–	–	–
	Q	**5.5731**	**0.0988**	**7.6384**	**0.3649**	**0.7930**
	Q_{TID}	7.0830	**0.1062**	**12.1058**	0.5849	—
	Q_{LIVE}	–	0.1099	**12.1305**	0.5823	0.8296
	Q_{IVC}	**6.8303**	0.1068	12.1688	–	0.8331
	$Q_{watermark}$	6.8430	0.1066	12.1658	0.5800	**0.8261**

Table 3
C_P values for the 4 distortion levels in TID database. Level 1 indicates lower distortion while Level 4 corresponds to high distortion. The best three metrics have been highlighted by bold font for quick glance.

Metric	Level 1	Level 2	Level 3	Level 4
SSIM	**0.7564**	0.6102	0.6326	0.6766
MSVD	0.4811	0.5844	0.3869	0.6050
VIF	0.5355	0.5197	**0.8146**	**0.8851**
VSNR	0.6180	**0.6402**	0.4687	0.6492
PSNR	0.5742	0.3241	0.3601	0.3601
PSNR-HVS-M	0.4232	0.5036	0.4657	0.5114
Q	**0.7649**	**0.6464**	**0.6882**	**0.7655**
$Q_{watermark}$	**0.7579**	**0.6376**	**0.6723**	**0.7401**

Table 2
Average performance of metrics over the 5 image databases. The two best metrics have been highlighted by bold font for quick glance.

Type of average	Criteria	SSIM	MSVD	VIF	VSNR	PSNR	PSNR-HVS-M	Q	$Q_{watermark}$
Direct averaging	C_P	0.8431	0.7562	0.8074	0.8240	0.7174	0.8241	**0.9071**	**0.8744**
	C_S	0.8335	0.7241	0.7866	0.8061	0.6832	0.8145	**0.8654**	**0.8550**
	RMSE	4.6888	5.5839	4.4988	4.7547	5.3083	4.5926	2.8936	4.1043
Weighted averaging	C_P	0.8404	0.7362	**0.8584**	0.7842	0.6961	0.7290	**0.8743**	0.8520
	C_S	0.8418	0.7435	0.8279	0.7877	0.6936	0.7369	**0.8522**	**0.8356**
	RMSE	3.5225	4.5175	**2.8547**	3.3182	4.0592	3.6430	**2.5008**	3.0691

levels. Also we find that there is large variation in prediction accuracies for MSVD, VIF and PSNR-HVS-M as we go from Levels 1 to 4. On the other hand, SSIM, VSNR and Q are more consistent for the 4 levels with Q being better than the two. Therefore, Q, in general, not only performs better for each distortion level but is also more stable and consistent for the 4 levels. We believe this to be a reason for the better performance of the proposed metric for all the databases. That is, it achieves a better tradeoff for the performance on near-threshold and supra-threshold distortions. This confirms the point we made earlier in Section 3: 2D mel-cepstrum features can tackle near threshold distortions more efficiently. Overall, the proposed metric performs consistently better across databases and this is an important advantage over the existing metrics.

5.2. Cross database validation

Since the proposed scheme involves training, we further present the results for the cross-database evaluation in Table 1 where Q_{TID}, Q_{LIVE}, Q_{IVC} and $Q_{watermark}$ denote that training is done with TID, LIVE, IVC and watermarked image databases, respectively, while the remaining databases form the test sets. Since the training and test sets come from different databases, the cross database evaluation helps to evaluate the robustness of the proposed scheme to untrained data. We can again see that the proposed scheme performs quite well with all the 3 test criteria (C_P, C_S and RMSE). It is also worth pointing out that Q_{IVC} achieves good results for the TID database since in this case the training set size (185 images) is relatively smaller than the test set (1700 images). Similar comments can also be made for $Q_{watermark}$ where training set consists of 210 images.

As mentioned before, we also used 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, $Q_{watermark}$ denotes the training with watermarked image database. As can be seen, $Q_{watermark}$ performs quite well. This further confirms our claim that quality degradation due to different distortion types can be assessed by exploiting the underlying common patterns characterized by the *structure* loss. We have also presented the results for $Q_{watermark}$ for the 2 averaging cases mentioned before (see Table 2) and we find that it performs very well especially given the relatively small training set size and training content being only watermark distortion.

As the last test in cross database evaluation, we test the performance of the proposed scheme for a video database. The trained system is used to predict the quality score of each individual frame and the overall quality score of the video is determined as the average of the scores all the frames in the video. The same procedure was also adopted for evaluating the other metrics. We present the results in Table 4. We can see that Q_{TID}, Q_{IVC}, Q_{LIVE} and $Q_{watermark}$ all perform better than the existing metrics under comparison. Note that the videos in this database have been distorted due to H.264/AVC, which is obviously not present in the image databases. Since the training is done with image databases, the good performance of the proposed metric is again indicative of its generalization ability to new visual/distortion content. The better performance of the proposed metric for this video database is also important since H.264/AVC is a recent video coding standard, which is fast gaining industry appreciation. Although video quality assessment may also involve temporal factors for quality estimation, the aforesaid procedure of using the average of frame level quality as the overall video quality score is still a popular and widely used method. Accounting for the temporal factors for video quality assessment is out of the scope of this paper and is a potential future work. Moreover, in this paper, we used the video database primarily to evaluate the proposed metrics performance for untrained contents.

Table 4
Experimental results for EPFL video database. The three best metrics have been highlighted by bold font for quick glance.

Criteria/metric	C_P	C_S	RMSE
SSIM	0.6878	0.7080	0.9790
MSVD	0.8554	0.8508	0.6987
VIF	0.7519	0.7524	0.8892
VSNR	0.8838	0.8631	0.6310
PSNR	0.6910	0.6869	0.9750
PSNR-HVS-M	0.8865	0.8760	0.6240
Q_{TID}	0.9390	0.9293	0.4640
Q_{LIVE}	**0.9426**	**0.9321**	**0.4502**
Q_{IVC}	**0.9411**	**0.9311**	**0.4562**
$Q_{watermark}$	**0.9394**	**0.9304**	**0.4626**

5.3. Metric efficiency evaluation

An important criterion to judge the performance of an IQA metric is its efficiency in terms of computational time required. The practical utility of a metric will reduce significantly if it is slow and computationally expensive in spite of its high prediction accuracy. In this section, we compare the efficiency (i.e. computational complexity) of different metrics. 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 5 shows the average time required per image (s), with all the codes implemented in Matlab. We can see that the proposed metric takes less time than all the metrics except PSNR and SSIM. This is because the feature extraction stage in the proposed metric takes the advantage of the Fast Fourier Transform (FFT) algorithm during the DFT computation. Note that DFT normally requires $O(N^2)$ operations to process N samples but for FFT this number is only $O(N\log(N))$. Hence the proposed metric is reasonably efficient in terms of execution time required (in addition to better prediction accuracies) and as a result more suitable for real time IQA.

6. Analysis for reduced-reference scenario and further discussion

6.1. Reduced-reference IQA

Objective IQA metrics can be classified into 3 categories based on the amount of information used for predicting quality: (1) full-reference (FR) metrics, which uses complete reference image

Table 5
Average execution time (ss/image) for different metrics. The three best metrics have been highlighted by bold font for quick glance.

Metrics	SSIM	MSVD	VIF	VSNR	PSNR	PSNR-HVS-M	Ref. [95]	Proposed
Time (s)	**0.0454**	0.6036	3.4829	0.4452	**0.0037**	2.5586	5.9276	**0.3268**

information, (2) reduced-reference (RR) metrics, which uses only partial information from the reference image and (3) no-reference (NR) metrics, which do not use any reference image information. FR metrics are generally more accurate while NR metrics can be used when the reference image is not available. RR metrics are essentially a tradeoff between these two since only partial information of the reference image is required. Literature survey shows that there has been more progress in developing FR IQA while RR and NR IQA have been relatively unexplored.

Obviously an RR IQA allows lower requirement of memory, bandwidth and computations. In a practical context of RR IQA, within an image transmission service, the reference image information (RRI) is sent along with the image to be transmitted. The compression of the RRI can be achieved by lossless coding. At the receiver end, one uses the RRI and compares it with the features of the decoded/received image. From this comparison one determines the objective quality score of the image received.

In the proposed metric, the length of the feature vector is M^2 for an N by N image where the reduced data is M by M with $M<N$. In our case we used $M=49$ as in [33]. Therefore, we only need $49^2=2401$ coefficients of the reference image to perform quality assessment. Thus, for $N=512$ (i.e. 512×512), we need only 0.92% of the actual reference image size. For an image with size 512×384 (as in TID database), we need to have only 1.2% for the amount of data in comparison with the actual reference image size. Therefore, even in its original form, the proposed metric can be considered an RR metric.

We now explore further possibility of using the proposed metric in reduced-reference scenario. A block diagram is shown in Fig. 6 where a new block "dimension reduction" has been used reduce the number of features as required in RR IQA. We now outline the "dimension reduction" procedure. From Eq. (2) we find that each bin energy $G(m,n)$ (a complex number in general) is a weighted sum of the frequency components where the non-uniform weights can be visualized through the grid or the weight diagram representation shown in Fig. 2. Therefore, the energy in each bin has a pre-defined contribution from each frequency component. Now it is a fact that higher frequency components are more important for quality assessment, and to use this to our advantage, we retain only those bin energies corresponding to the higher frequency components and discard the lower frequency components. Effectively this will mean ignoring the effect of the lower frequency components for the benefit of achieving further dimension reduction. We use the 2D mel-cepstrum features $\hat{c}(p,q)$ defined in Eq. (3) on which we apply the said dimension reduction procedure to obtain $\hat{c}_R(p,q)$ where we used the subscript R to distinguish it from $\hat{c}(p,q)$. We then define the new feature vector as

$$\boldsymbol{x}^{(new)}=|\boldsymbol{x}_r^{(new)}-\boldsymbol{x}_d^{(new)}| \tag{11}$$

where $\boldsymbol{x}_r^{(new)}$ and $\boldsymbol{x}_d^{(new)}$ are the features from the reference and distorted images, respectively, with reduced dimension $R<M$. We note that similar to $\boldsymbol{x}$ defined in Eq. (4), $\boldsymbol{x}^{(new)}$ also accounts for the perceptual properties like sensitivity to loss of structure, edge masking and the suprathreshold effect. However, unlike $\boldsymbol{x}$ it lacks information regarding the changes corresponding to lower frequency components. To illustrate the usefulness of the said dimension reduction procedure, we present the experimental results with $R=500$ as an example i.e. we retain only 500 coefficients out of 2401, which corresponds to using only 20.82% of the total number of coefficients. This in turn means that we need to transmit only R ($=500$ in this example) coefficients from the reference image to compute the quality of the transmitted image. For notations regarding the RR metric, we use the superscript R with all the previously defined symbols. For instance, $Q_{watermark}^{(R)}$ denotes the system trained only with the watermarked image database with R coefficients. So the superscript in $Q_{watermark}^{(R)}$ distinguishes it from the symbol $Q_{watermark}$, which corresponds to the FR case. We follow a similar notation for Q, Q_{TID}, Q_{LIVE}, $Q_{watermark}$ and Q_{IVC}. We present the experimental results for the RR case in Table 7. We find that though the prediction accuracies decrease they are acceptable and compare favorably to the case when no dimension reduction is employed. We can also observe that the prediction performance is quite competitive with the existing full-reference metrics, which use the complete reference image

Table 6
Demonstration of scalability on TID and LIVE databases. The prediction accuracy is presented in terms of C_S for the metric $Q_{watermark}^{(R)}$ for different values of R.

R	C_S for LIVE database	C_S for TID database	Percent savings relative to $R=2401$	ΔC_S for LIVE database	ΔC_S for TID database
2401	0.9433	0.7697	–	–	–
2000	0.9402	0.7624	16.70	0.0031	0.0073
1500	0.9421	0.7592	37.53	0.0012	0.0105
1000	0.9387	0.7502	58.35	0.0046	0.0195
500	0.9379	0.7435	79.80	0.0090	0.0262
400	0.9270	0.7388	83.34	0.0163	0.0309
300	0.9276	0.7378	87.51	0.0157	0.0319
200	0.9135	0.7268	91.67	0.0145	0.0397
100	0.9208	0.7240	95.84	0.0225	0.0457

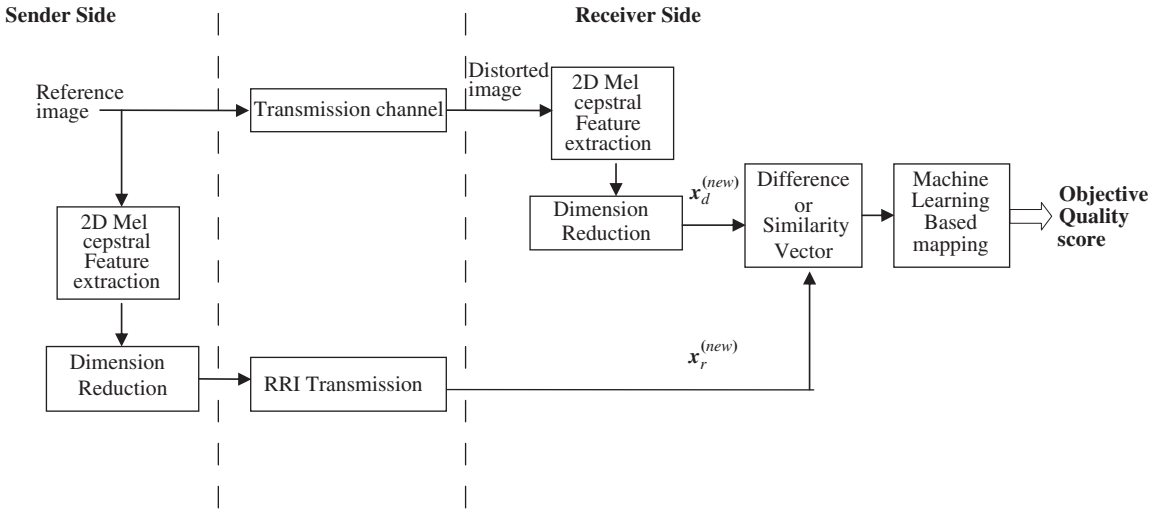

Fig. 6. Block diagram of the proposed RR metric.

Table 7
Performance of the proposed metric for reduced-reference scenario with $R=500$.

Criteria	Metric	LIVE	A57	WIQ	IVC	TID	EPFL video database
C_P	$Q^{(R)}$	0.9287	0.8491	0.8678	0.9381	0.7664	–
	$Q^{(R)}_{TID}$	0.9108	0.8814	0.8164	0.8739	–	0.9244
	$Q^{(R)}_{LIVE}$	–	0.8612	0.8163	0.8785	0.7607	0.9186
	$Q^{(R)}_{IVC}$	0.9267	0.8646	0.8194	–	0.7675	0.9246
	$Q^{(R)}_{watermark}$	0.9350	0.8742	0.8219	0.8757	0.7697	0.9252
C_S	$Q^{(R)}$	0.9115	0.7898	0.7527	0.9002	0.7442	–
	$Q^{(R)}_{TID}$	0.9256	0.8386	0.8126	0.8620	–	0.9107
	$Q^{(R)}_{LIVE}$	–	0.8156	0.8110	0.8700	0.7431	0.9068
	$Q^{(R)}_{IVC}$	0.9232	0.8299	0.8170	–	0.7433	0.9086
	$Q^{(R)}_{watermark}$	0.9379	0.8248	0.8164	0.8647	0.7435	0.9103
RMSE	$Q^{(R)}$	7.8902	0.1055	8.5038	0.3993	0.8613	–
	$Q^{(R)}_{TID}$	9.5445	0.1161	13.2294	0.5923	–	0.5145
	$Q^{(R)}_{LIVE}$	–	0.1249	13.2326	0.5820	0.8710	0.5330
	$Q^{(R)}_{IVC}$	8.6877	0.1235	13.1308	–	0.8602	0.5139
	$Q^{(R)}_{watermark}$	7.2549	0.1193	13.0467	0.5884	0.8567	0.5117

information. We further present the variation in performance (denoted by ΔC_S) of $Q^{(R)}_{watermark}$ for TID and LIVE databases with different R values in Table 6. We have included the results only for these two databases since they are the biggest in terms of the number of images and distortion types. One can observe that the performance is quite robust and there is graceful degradation in metric performance as R decreases. We also present the amount of information (%) saved with decreasing R relative to $R=2401$. Given that 2401 is itself a small number compared to the typical image size, the savings made are significant. Similar observations were also made for the other databases but not presented here for the sake of brevity. For the same reason, we have omitted the results for $Q^{(R)}$, $Q^{(R)}_{TID}$, $Q^{(R)}_{LIVE}$ and $Q^{(R)}_{IVC}$ in Table 6.

The presented analysis indicates the potential of achieving effective reduced—reference quality assessment with the proposed metric. Since we can select the amount of RRI to be sent based on the available resources (like bandwidth), the proposed metric is scalable. Scalability is referred to the ability of a quality metric to perform in accordance with the available resources (like bandwidth, computational power, memory capacity, etc.) of a practical system with a graceful and reasonable degradation in metric performance due to the resource constraints. Such scalability offers more flexibility to the proposed scheme in comparison to FR metrics, which require the entire reference image for quality computation. There are two reasons, which contribute to the resulting scalability. Firstly, the discarded coefficients in essence correspond to lower frequencies, which basically represent weak edges and texture. Due to the masking properties of the HVS weak edges are masked or their effect is reduced. So removing such coefficients has lesser impact on the prediction performance. The second reason is the use of SVR. Since the weights for the pooling stage are determined via sufficient training with subjective scores, it further reduces the impact of these coefficients on the overall quality. This in turn minimizes the loss of prediction accuracy and results in more robust quality prediction. Scalability is an important and desirable feature of the proposed RR IQA metric because it can achieve good tradeoff between the prediction accuracy and the amount of RRI.

We also compared $Q^{(R)}_{watermark}$ with a recently developed Weibull statistics based RR metric [95] (hereafter we denote is as WSRRM). We obtained the software code for WSRRM from its authors. The reader may recall that $Q^{(R)}_{watermark}$ implies training with contents that are different from those in the test databases. We present the results for different databases in Table 8. In this case, we have used $R=6$ to make a fair comparison with WSRRM, which uses 6 scalars [95] for RR quality computation. As can be seen, $Q^{(R)}_{watermark}$ performs better than (we obtained similar conclusions for $Q^{(R)}_{TID}$, $Q^{(R)}_{LIVE}$ and $Q^{(R)}_{IVC}$) than WSRRM for A57, IVC and TID databases and is competitive for LIVE and WIQ databases. Furthermore, WSRRM suffers from the following drawbacks, which are alleviated in our RR metric:

Table 8
Comparison with WSRRM [95]. The better metric has been highlighted by bold font for quick glance.

Criteria	Metric	LIVE	A57	WIQ	IVC	TID
C_P	WSRRM	**0.8849**	0.5830	**0.8244**	0.5267	0.5536
	$Q^{(R)}_{watermark}$	0.8642	**0.5836**	0.7996	**0.5882**	**0.6313**
C_S	WSRRM	**0.8827**	0.5621	**0.8076**	0.4512	0.5415
	$Q^{(R)}_{watermark}$	0.8598	**0.5842**	0.7905	**0.5881**	**0.6366**
RMSE	WSRRM	**10.7670**	0.1997	**12.9649**	1.0357	1.1176
	$Q^{(R)}_{watermark}$	11.6313	**0.1992**	13.7560	**0.9853**	**1.0407**

- Low efficiency with regards to its execution speed as well as the higher computational costs. On an average it takes about 5.92 s per image. The reason for this is that it uses multi-scale image decomposition using the steerable pyramid decomposition. In contrast, the proposed RR metric takes only 0.32 s per image.
- It lacks scalability while the proposed RR metric being scalable offers more flexibility as already discussed.

6.2. Further discussion

We have three points, which deserve further discussion and are explained in what follows. First, the reader will recall that we used only the magnitude of the bin energy $G(m,n)$ in Eq. (3). Note that $G(m,n)$ will be a complex number in general, which we denote as $Ae^{j\alpha}$ with magnitude A and phase α. The 2D mel-cepstrum computation involves the logarithm of $G(m,n)$, so we have $\log(G(m,n))=\log(Ae^{j\alpha})=\log(A)+j\alpha$. Now both A and α should be continuous functions for them to have a valid Fourier transform. However, since $\alpha\in[-\pi,\pi]$ we must first unwrap the phase so that it becomes continuous. The major problem is that unwrapping the phase in 2-D is very difficult [89] due to two reasons. First, a typical image may contain thousands of individual phase wraps. Some of these wraps are genuine, while others may be false and are caused by the presence of noise and sometimes by the phase extraction algorithm itself. The process of differentiating between genuine and false phase wraps is extremely difficult and this adds complexity to the phase unwrapping problem. A second reason that complicates the phase unwrapping problem is its accumulative nature. The image is processed sequentially on a pixel-by-pixel basis. If a single genuine phase wrap between two neighboring pixels is missed due to noise, or a false wrap appears in the phase map, an error occurs in unwrapping both pixels. This kind of error then propagates throughout the rest of the image. In addition, phase unwrapping will be computationally expensive step and potentially a major bottle neck in the use of the proposed metric for real-time applications. Therefore, we used only the magnitude and discarded the phase.

The second point is regarding the use of multiple databases in this paper. It ensures that the proposed system is tested for its robustness to a wide variety of image and distortion contents on

which the proposed system is not trained. Besides, it also helps in more comprehensive metric testing since as discussed in Section 5, a metric performing well for one database may not do well on another. In addition, it facilitates the cross database evaluation, which provides a strong and convincing demonstration of the proposed system's ability to predict the quality well for untrained data. It may be mentioned here that for the cross database evaluation, we did not do any parameter optimization towards the test database. For instance consider $Q_{watermark}$. In this case, once we learn the model using all the images and associated subjective scores of the watermarked image database, we use the same model for testing LIVE, A57, TID, WIQ, IVC and EPFL (video database) databases. That is, we used the same kernel function namely RBF and the other parameters (i.e. radius of Gaussian function ρ, the tradeoff error C and regression tube width) were all kept constant when testing other image databases. Similar comments can be made for Q_{TID}, Q_{LIVE}, Q_{IVC} and $Q^{(R)}_{watermark}$. The performance improves further if we train a model specifically for each test database separately. It is also worth pointing out that the proposed metric is pretty robust to the different SVR parameters in that small changes in them does not cause large change in the prediction performance.

Finally, as demonstrated the proposed scheme is more consistent and stable in its performance across multiple databases than the existing metrics. This highlights that the selected features based on the 2D mel-cepstrum are effective. In addition, they convey a clearer physical meaning. The exploitation of 2D mel-cepstral features for IQA is novel and interesting since originally mel-cepstrum analysis was formulated for speech/audio signals. Since audio and visual signals have certain similarity as natural signals therefore it is not surprising that a similar approach can be used for analyzing them. The theory of natural signal statistics [62] also confirms that natural signals (including images and sounds) share statistical properties (for instance natural signals are highly structured). These features are also of interest for pattern recognition applications since they allow representing the spectra by points in a multidimensional vector space. The feature pooling via SVR is more convincing and reasonable since a quantitative data-driven modeling procedure is employed for the complex mapping of the feature vector to the desired output. In summary, the novelty of the proposed scheme in comparison to the existing IQA metrics is due to the following reasons:

- We used the 2D mel-cepstrum features, which to our knowledge have not been exploited in the literature for IQA. From the point of view of pattern recognition, they are also effective for dimension reduction. Essentially, we used them to quantify the perceptual similarity between the spectral envelopes of reference and distorted images. We have given proper analysis and reasoning behind using them for IQA and also outlined how they can account for the HVS properties like sensitivity to structure and suprathreshold effect in connection with IQA. The presented analysis provides new insights and can be useful for related applications like image utility assessment [2], image similarity assessment, etc.
- We employed machine learning technique for more systematic feature pooling. The proposed methodology demonstrates the effectiveness of machine learning in avoiding unrealistic assumptions currently imposed in the existing feature pooling methods. It is therefore an attractive alternative to bridge the gap between the psychophysical ground truth and the realistic engineering solution.
- Since we could discard some coefficients based on their perceptual significance, we arrived at an RR IQA metric. The reduced number of coefficients was selected in a way that reduces the information required while still maintaining good performance. This further confirms the analysis presented in this paper and provides evidence in favor of the validity of the theoretical points made. The reduced-reference prospects and the associated scalability make the proposed metric more attractive and useful.
- The proposed metric is more efficient than many existing metrics in terms of execution time needed and thus suitable for real-time deployment.

7. Conclusions

In this paper, we have explored the 2D mel-cepstrum features and SVR image quality assessment, and formulated the task of image quality prediction as a pattern recognition problem, to enable the use of more sophisticated pooling techniques like the SVR to achieve robust, accurate, consistent and scalable quality prediction. This helps to overcome the limitations of the existing pooling methods in image quality assessment (IQA). We provided in-depth analysis and justification of the 2D mel-cepstrum features to be employed for IQA. A thorough and extensive experimental validation using seven independent and publicly available image/video databases with diverse distortion types provides strong ground for the usefulness of the proposed metric. The experimental results confirm the effectiveness of the proposed feature selection and pooling method towards more effective and consistent IQA. We have also compared the performance of the proposed metric with seven relevant existing metrics and shown that the proposed metric performs consistently better across all the databases. In addition, we also explored the possibility for reduced-reference situations and demonstrated good performance as well.

Acknowledgment

The authors wish to thank Prof. Xuanqin Mou (Xi'an Jiaotong University) for providing the code for WSRRM used in this paper. This work is partially supported by MoE AcRF-Tier 1 (2007) funding, and AcRF-Tier 2 funding (T208B1218), Singapore. A.E. Cetin's research was funded by the European Commission, FP7-ENV-244088 "FIRESENSE—Fire Detection and Management through a Multi-Sensor Network for the Protection of Cultural Heritage Areas from the Risk of Fire and Extreme Weather" and FP7-PEOPLE-247091 "MIRACLE-Microscopic Image Processing, Analysis, Classification and Modelling Environment".

References

[1] C. Yang, Inverted pattern approach to improve image quality of information hiding by LSB substitution, Pattern Recognition 41 (2008) 2674–2683.
[2] D. Rouse, S. Hemami, Natural image utility assessment using image contours, in: Proceedings of the International Conference on Image Processing, 2009, pp. 2217–2220.
[3] S. Lee, H. Koo, N. Cho, Image segmentation algorithms based on the machine learning of features, Pattern Recognition Letters 31 (2010) 2325–2336.
[4] F.X.J. Lukas, Z.L. Budrikis, Picture quality prediction based on a visual model, IEEE Transactions on Communications 30 (7) (1982) 679–1692.
[5] A. Oppenheim, R. Schafer, From frequency to quefrency: a history of the cepstrum, IEEE Signal Processing Magazine 21 (5) (2004) 95–106.
[6] S. Daly, The visible differences predictor: an algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 179–206.
[7] J. Lubin, D. Fibush, Sarnoff JND Vision Model, T1A1.5Working Group Document #97-612, ANSI T1 Standards Committee, 1997.
[8] C.J. van den Branden Lambrecht, O. Verscheure, Perceptual quality measure using a spatio-temporal model of the human visual system, in: Proceedings of the SPIE digital video compression: algorithms and technologies, San Jose, CA, vol. 2668, January 28–February 2, 1996, pp. 450–461.
[9] Y. Kessentini, T. Paquet, A. Hamadou, Off-line handwritten word recognition using multi-stream hidden Markov models, Pattern Recognition Letters 31 (2010) 60–70.

[10] S. Winkler, A perceptual distortion metric for digital color video, in: Proceedings of the SPIE Human Vision and Electronic Imaging, San Jose, CA, vol. 3644, January 23–29, 1999, pp. 175–184.
[11] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (4) (2004).
[12] S. Winkler, Perceptual video quality metrics—a review, in: H.R. Wu, K.R. Rao (Eds.), Digital Video Image Quality and Perceptual Coding, CRC Press, Boca Raton, FL, 2005 (Chapter 5).
[13] M.C. Morrone, D.C. Burr, Feature detection in human vision: a phase-dependent energy model, Proceedings of the Royal Society of London B 235 (1280) (1988) 221–245.
[14] M. Narwaria, W. Lin, Scalable image quality assessment based on structural vectors, in: Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP'09), Rio de Janeiro, Brazil, October 5–7, 2009.
[15] A. Eskicioglu, A. Gusev, A. Shnayderman, An SVD-based gray-scale image quality measure for local and global assessment, IEEE Transactions on Image Processing 15 (2) (2006) 422–429.
[16] M. Narwaria, W. Lin, Objective image quality assessment based on support vector regression, IEEE Transactions on Neural Networks 21 (3) (2010) 515–519.
[17] S. Choi, Y. Lee, S. Lee, K. Park, J. Kim, Age estimation using a hierarchical classifier based on global and local facial features, Pattern Recognition 44 (2011) 1262–1281.
[18] A. Watson, L. Kreslake, Measurement of visual impairment scales for digital video, Proceedings of the SPIE, Human Vision Visual Processing and Digital Display 4299 (2001) 79–89.
[19] Z. Haddad, A. Beghdadi, A. Serir, A. Mokraoui, Image quality assessment based on wave atoms transform, in: Proceedings of the International Conference on Image Processing, 2010, pp. 305–308.
[20] C. Wee, R. Paramesran, R. Mukundan, X. Jiang, Image quality assessment by discrete orthogonal moments, Pattern Recognition 43 (2010) 4055–4068.
[21] M. Liu, X. Yang, Image quality assessment using contourlet transform, Optical Engineering 48 (10) (2009) 107201.
[22] L. Zhang, L. Zhang, X. Mou, RFSIM: a feature based image quality assessment metric using Riesz transforms, in: Proceedings of the International Conference on Image Processing, 2010, pp. 321–324.
[23] J. Shin, J. Lee, D. Kim, Real-time lip reading system for isolated Korean word recognition, Pattern Recognition 44 (2011) 559–571.
[24] Z. Wang, X. Shang, Spatial pooling strategies for perceptual image quality assessment, in: Proceedings of the IEEE International Conference on Image Processing, (ICIP), 2006, pp. 2945–2948.
[26] A. Ninassi, O. Lemeur, P. Callet, D. Barba, Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric, in: Proceedings of the IEEE International Conferencing on Image Processing (ICIP), 2007, pp. 169–172,.
[27] J. You, A. Perkis, M. Hannuksela, M. Gabbouj, Perceptual quality assessment based on visual attention analysis, in: Proceedings of the ACM International Conference on Multimedia (MM'09), Beijing, China, October 19–24, 2009.
[28] C. Li, A. Bovik, Content-partitioned structural similarity index for image quality assessment, Signal Processing: Image Communication 25 (2010) 517–526.
[29] S. Scanzio, S. Cumani, R. Gemello, F. Mana, P. Laface, Parallel implementation of artificial neural network training for speech recognition, Pattern Recognition Letters 31 (2010) 1302–1309.
[30] G. Chen, C. Yang, S. Xie, Gradient-based structural similarity for image quality assessment, in: Proceedings of the IEEE International Conference on Image Processing, October 2006, pp. 2929–2932.
[31] J.L. Mannos, D.J. Sakrison, The effects of a visual fidelity criterion on the encoding of images, IEEE Transactions on Information Theory 20 (4) (1974) 525–536.
[32] L. Liang, S. Wang, J. Chen b, S. Mac, D. Zhao, W. Gao, No-reference perceptual image quality metric using gradient profiles for JPEG2000, Signal Processing: Image Communication 25 (2010) 502–516.
[33] S. Cakir, A. Cetin, Image feature extraction using 2D mel-cepstrum, in: Proceedings of the International Conference on Pattern Recognition (ICPR), 2010, pp. 674–677.
[34] M. Narwaria, W. Lin, I.V. McLoughlin, S. Emmanuel, C.L. Tien, Non-intrusive speech quality assessment with support vector regression, in: Proceedings of the 16th International Conference on Multimedia Modeling, Lecture Notes in Computer Science, vol. 5916, 2010, pp. 325–335.
[35] S. Cakir, A. Cetin, Mel-cepstral methods for image feature extraction, in: Proceedings of the International Conference on Image Processing, 2010, pp. 4577–4580.
[36] B. Scholkopf, A.J. Smola, Learning with Kernels, MIT Press, 2002.
[37] C. Wee, R. Paramesran, R. Mukundan, Fast computation of geometric moments using a symmetric kernel, Pattern Recognition 41 (7) (2008) 2369–2380.
[38] T. Sylvain, A. Florent, Z. Parvez, H. Yuukou, Impact of Subjective dataset on the performance of image quality metrics, in: Proceedings of the IEEE International Conference on Image Processing, 2008.
[39] H.R. Sheikh, Z. Wang, A.C. Bovik, L.K. Cormack, Image and Video Quality Assessment Research at LIVE [online]. Available from: 〈http://live.ece.utexas.edu/research/quality/〉.
[40] P. Le Callet, F. Autrusseau, Subjective Quality Assessment IRCCyN/IVC Database. 〈http://www2.irccyn.ec-nantes.fr/ivcdb/〉.
[41] A57 dataset: 〈http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html〉.
[42] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, F. Battisti, Color image database for evaluation of image quality metrics, in: Proceedings of the International Workshop on Multimedia Signal Processing, 2008, pp. 403–408.
[43] M. Peterson, D. Ridder, H. Handels, Image processing with neural networks—a review, Pattern Recognition 35 (2002) 2279–2301.
[44] F. Simone, M. Naccari, M. Tagliasacchi, F.C. Dufaux, S. Tubaro, T. Ebrahimi, Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel, in: Proceedings of the First International Workshop on Quality of Multimedia Experience (QoMEX 2009), San Diego, California, July 2009.
[45] P. Bartlett, S. Boucheron, G. Lugosi, Model selection and error estimation, Machine Learning (2002) 85–113.
[46] VQEG, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Phase II August 2003 online. Available from: 〈http://www.vqeg.org〉.
[47] K. Lu, J. Zhao, Y. Wu, Hessian optimal design for image retrieval, Pattern Recognition 44 (2011) 1155–1161.
[48] Z. Sazzad, Y. Kawayoke, Y. Horita, No reference image quality assessment for JPEG2000 based on spatial features, Signal Processing: Image Communication 23 (2008) 257–268.
[49] F. Autrusseau, Subjective quality assessment—Fourier subband database [online]. Available from: 〈http://www.irccyn.ec-nantes.fr/~autrusse/Databases/FourierSB/〉, 2009.
[50] D. Marr, E. Hildreth, Theory of edge detection, Proceedings of the Royal Society of London B 207 (1167) (1980) 187–217.
[51] X. Ran, N. Farvardin, A perceptually-motivated three-component image model—Part I: description of the model, IEEE Transactions on Image Processing 4 (4) (1995).
[52] X. Wang, T. Wang, J. Bu, Color image segmentation using pixel wise support vector machine classification, Pattern Recognition 44 (2011) 777–787.
[53] U. Engelke, H.-J. Zepernick, M. Kusuma, Wireless Imaging Quality Database (2010) 〈http://www.bth.se/tek/rcg.nsf/pages/wiq-db〉.
[54] H. Tsai, C. Liu, Wavelet-based image watermarking with visibility range estimation based on HVS and neural networks, Pattern Recognition 44 (2011) 751–763.
[55] D.M. Chandler, S.S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Transactions on Image Processing 16 (9) (2007).
[56] A. James, S. Dimitrijev, Inter-image outliers and their application to image classification, Pattern Recognition 43 (2010) 4101–4112.
[57] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2) (2006) 430–444.
[58] M. Gaubatz, Metrix MUX Visual Quality Assessment Package, 〈http://foulard.ece.cornell.edu/gaubatz/metrix_mux/〉.
[59] X. Zhao, Y. Satoh, H. Takauji, S. Kaneko, K. Iwata, R. Ozaki, Object detection based on a robust and accurate statistical multi-point-pair model, Pattern Recognition 44 (2011) 1296–1311.
[60] S. Winkler, P. Mohandas, The evolution of video quality measurement: from PSNR to hybrid metrics, IEEE Transactions on Broadcasting 54 (3) (2008) 660–668.
[61] A. Maeder, The image importance approach to human vision based image quality characterization, Pattern Recognition Letters 26 (2005) 347–354.
[62] O. Schwartz, E. Simoncelli, Natural signal statistics and sensory gain control, Nature Neuroscience 4 (2001) 819–825.
[63] J. Aucouturier, F. Pachet, Music similarity measures: what's the use? in: Proceedings of the International Symposium on Music Info. Retrieval (ISMIR), 2002.
[64] D.O. Kim, H.S. Han, R.H. Park, Gradient information-based image quality metric, IEEE Transactions on Consumer Electronics 56 (2) (2010) 930–936.
[65] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on Speech and Audio Processing 10 (5) (2002) 293–302.
[66] H. Terasawa, M. Slaney, J. Berger, A timbre space for speech, Proceedings of the Interspeech (2005).
[67] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Inc., Upper Saddle River, NJ, 1993.
[68] S. Cakir, A. Cetin, Mel-cepstral feature extraction methods for image representation, Optical Engineering 49 (9) (2010) 097004.
[69] R. Arandjelović, T. Sezgin, Sketch recognition by fusion of temporal and image-based features, Pattern Recognition 44 (2011) 1225–1234.
[70] 〈http://www.ponomarenko.info/psnrhvsm.htm〉.
[71] D. Marr, Vision, W. H. Freeman and Company, New York, 1980.
[72] D. Chandler, S. Hemami, Suprathreshold Image Compression based on Contrast Allocation and Global Precedence, Proceedings of the Human Vision and Electronic Imaging (2003).
[73] G. Cheng, J. Huang, C. Zhu, Z. Liu, L. Cheng, Perceptual image quality assessment using a geometric structural distortion model, in: Proceedings of the IEEE International Conference on Image Processing, September 2010, pp. 325–328.
[74] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, V. Lukin, On between-coefficient contrast masking of dct basis functions, in: Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2007.
[75] J. Melendez, D. Puig, M. Garcia, Multi-level pixel-based texture classification through efficient prototype selection via normalized cut, Pattern Recognition 43 (2010) 4113–4123.
[76] S. Hemami, M. Ramos, Wavelet coefficient quantization to produce equivalent visual distortion in complex stimuli, in: Human Vision and Electronic Imaging V, Proceedings of the SPIE, vol. 3959, 2000, pp. 200–210.

[77] Z. Wang, A.C. Bovik, L. Lu, Why is image quality assessment so difficult? in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.
[78] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, 2001 [online]. Available from: 〈http://www.csie.ntu.edu.tw/\verb~cjlin/libsvm〉.
[79] H. Noda, M. Niimi, Local MAP estimation for quality improvement of compressed color images, Pattern Recognition 44 (2011) 788–793.
[80] S. Antani, R. Kasturi, R. Jain, A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video, Pattern Recognition 35 (2002) 945–965.
[81] V. O'Brien, Contour perception, illusion and reality, Journal of the Optical Society of America 48 (1958) 112–119.
[82] J.H. Elder, S.W. Zucker, Evidence for boundary-specific grouping in human vision, Vision Research 38 (1) (1998) 143–152.
[83] R.L. De Valois, K.K. De Valois, Spatial Vision, Oxford University Press, New York, 1990.
[84] W.K. Pratt, Digital Image Processing: PIKS Inside, 3rd ed., Wiley-Interscience, New York, 2001.
[85] W. Lin, M. Narwaria, Perceptual image quality assessment: recent progress and trends, in: Proceedings of the SPIE, vol. 7744, July 2010, p. 774403.
[86] M. Sendashonga, F. Labeau, Low complexity image quality assessment using frequency domain transforms, in: Proceedings of the International Conference on Image Processing, 2006, pp. 385–388.
[87] B. Girod. What's wrong with mean-squared error? in: A.B. Watson (Ed.), Digital Images and Human Vision, The MIT Press, 1993, pp. 207–220.
[88] W. Lin, C. Kuo, Perceptual visual quality metrics: a survey. Journal of Visual Communication and Image Representation, in press. Available from: 〈http://www.sciencedirect.com〉.
[89] D. Ghiglia, M. Pritt, Two-dimensional Phase Unwrapping: Theory, Algorithms and Software, John Wiley & Sons, 1998.
[90] S.S. Channappayya, A.C. Bovik, R.W. Heath, Rate bounds on SSIM index of quantized images, IEEE Transactions on Image Processing 17 (2008) 1624–1639.
[91] Y.H. Huang, T.S. Ou, P.Y. Su, H.H. Chen, Perceptual rate–distortion optimization using structural similarity index as quality metric, IEEE Transactions on Circuits Systems and Video Technology 20 (11) (2010) 1614–1624.
[92] Recommendation ITU-R BT.500-11, Methodology for the subjective assessment of the quality of television pictures, June 2002.
[95] W. Xue, X. Mou, Reduced reference image quality assessment based on Weibull statistics, in: Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX 2010), vol. 1, 2010, pp. 1–6.

Manish Narwaria received his B.Tech. degree in Electronics and Communication Engineering from Amrita Vishwa Vidyapeetham, India. He is currently pursuing his PhD degree in the School of Computer Engineering at Nanyang Technological University, Singapore. His research interests include image/video and speech quality assessment, pattern recognition and machine learning.

Weisi Lin received his Ph.D. degree in computer vision from King's College, London University, London, UK, in 1992. Currently, he is an Associate Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. His areas of expertise include image processing, perceptual modeling, video compression, multimedia communication, and computer vision.

A. Enis Cetin got his Ph.D. degree in systems engineering from the University of Pennsylvania, Philadelphia. Currently, he a Professor in Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey. His research areas include Human-Computer Interaction using vision and speech, Audio-Visual Multimedia Databases, Speech Processing, Digital Coding of Waveforms.