SIViP
DOI 10.1007/s11760-013-0445-2

ORIGINAL PAPER

A new psychovisual paradigm for image quality assessment: from differentiating distortion types to discriminating quality conditions

Ke Gu · Guangtao Zhai · Xiaokang Yang · Wenjun Zhang

Received: 15 November 2011 / Revised: 18 May 2012 / Accepted: 13 October 2012

Abstract This paper investigates the impacts of image quality level on the prediction accuracy of image quality metrics. While many state-of-the-art perceptual image quality assessment methods have achieved fairly well performances in terms of the correlation between the quality predictions and the subjective scores, none of them took into account the effects of the quality levels of those test images on prediction accuracy of the quality metrics. In this work, inspired by the mechanism of human perception under high- and low-quality conditions, we propose a new image quality assessment paradigm based on image quality level classification. Our investigation on TID2008 and other three publicly available databases (LIVE, CSIQ and Toyama-MICT) results in two valuable findings. First, the performances of major well-known image quality assessment methods are significantly affected by image quality level. Second, through combining different quality metrics for different quality levels, superior performance can be achieved as compared to some of the best image quality metrics, e.g., SSIM, MS-SSIM, VIF and VIFP. Experiments and comparative studies are provided to confirm the effectiveness of the proposed new paradigm by differentiating quality levels for image quality assessment.

This work was supported in part by NSERC, NSFC (61025005, 60932006, 61001145), SRFDP (20090073110022), postdoctoral foundation of China 20100480603, postdoctoral foundation of Shanghai 11R21414200 and the 111 Project (B07022).

K. Gu (✉) · G. Zhai · X. Yang · W. Zhang
The Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, China
e-mail: gukesjtuee@gmail.com

K. Gu · G. Zhai · X. Yang · W. Zhang
Shanghai Key Laboratory of Digital Media Processing and Transmissions, Shanghai, China

Keywords Image quality assessment (IQA) · Image quality classification · Different perception (DIP) mechanism · Human psychovisual perception · Free energy

1 Introduction

Perceptual image quality assessment (IQA) plays an important part in many areas of digital image processing, such as the development and optimization of image compression, storage, transmission and reproduction algorithms. Existing IQA approaches fall into two categories: subjective assessment and objective assessment. Although the subjective assessment approach should be the ultimate quality gauge for images, it is usually time-consuming, expensive and impractical for real-time image processing systems. Therefore, there had been an increased interest in developing objective IQA metrics. According to the availability of reference images to be compared with during the tests, objective IQA methods can be further classified into three categories. Most approaches are known as full-reference methods, assuming the reference image is completely known. In many practical applications, however, the reference image is not available, and a no-reference IQA algorithm is then desirable. The third type is referred to as reduced-reference IQA algorithm, which is applied to the situation where the reference image is only partially available and some extracted features are made available as side information to help to evaluate the quality of the distorted image. In this work, we concentrate on full-reference IQA approach.

The mean-squared error (MSE) and its relative peak signal-to-noise ratio (PSNR) are still the most widely used objective quality metrics, both due to their convenience and due to their clear physical meaning as distortion/fidelity measures. However, it has been widely recognized that MSE

Published online: 16 March 2013

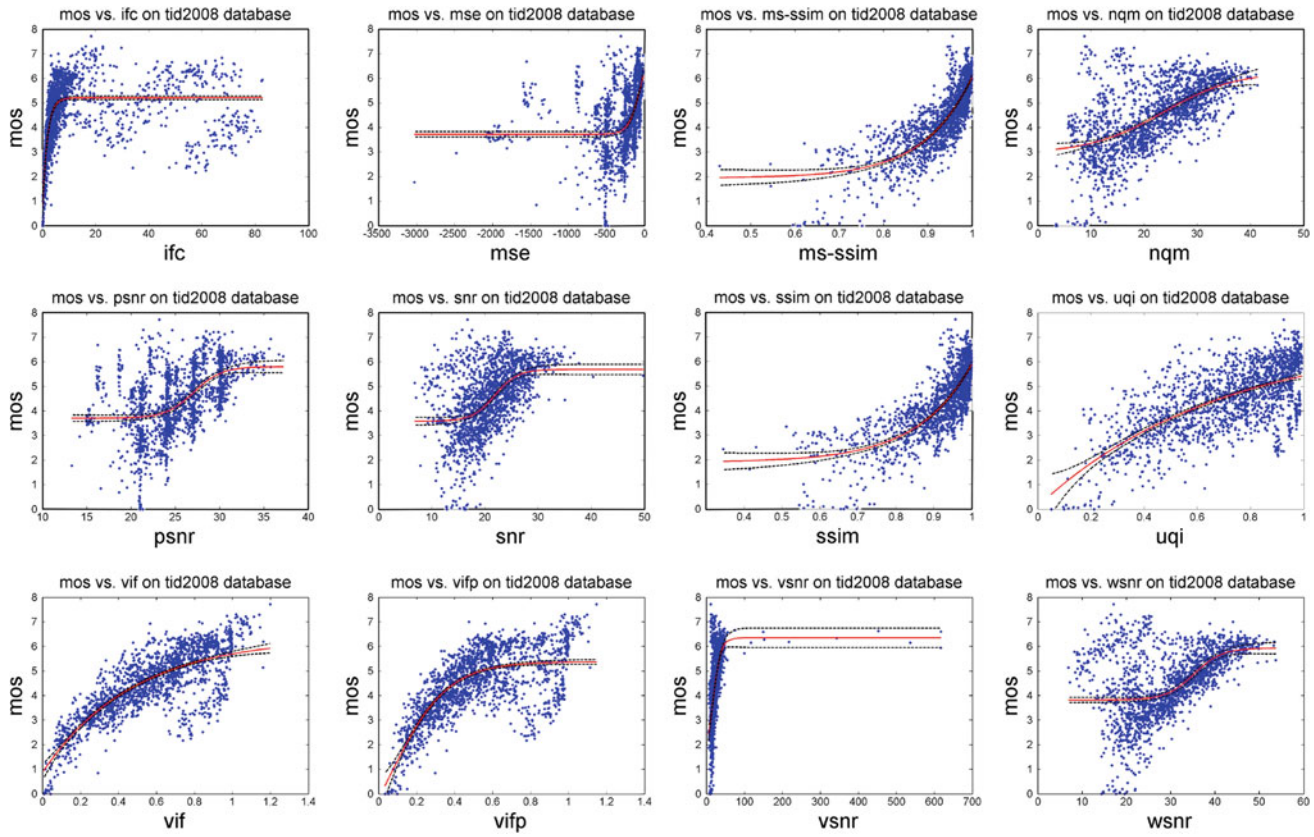

Fig. 1 Scatter plots of MOS versus IFC, MSE, MS-SSIM, NQM, PSNR, SNR, SSIM, UQI, VIF, VIPF, VSNR and WSNR, respectively. The *(red) lines* are curves fitted with the logistic function, and the *(blue) dash lines* are 95 % confidence intervals (color figure online)

and PSNR are not well correlated with human judgment of quality, i.e., the Mean Opinion Score (MOS) [1]. A large set of the so-called cognitivist methods, inspired by a classical cognitivist paradigm of psychology [2], have been proposed thorough the years. Structural SIMilarity (SSIM) index [3], the most popular cognitivist method, focuses on structural information substantially. Various other cognitivist algorithms, such as multi-scale SSIM (MS-SSIM) [4], visual information fidelity (VIF) [5] and a pixel-based version of VIF (VIFP) [5], have been proposed later for better quality prediction.

During the last decade, much effort have been devoted to incorporating the properties of the human visual system (HVS) into image quality metrics and many new IQA methods have been proposed [6–8]. Liu et al. put human beings' saliency map into PSNR and SSIM metrics by locally weighting the corresponding distorting map, like the combination strategy in [9]. Thus, to combine PSNR/SSIM and human beings' saliency map, the WPSNR/WSSIM [10] is proposed.

Very recently, many researchers in the area of IQA realized the importance of distortion classification. For example, the Blind Image Quality Indices (BIQI) [11] is a two-stage method: Images are first explicitly classified into different distortion categories using distorted image statistics (DIS) [12]. And then, the quality of each image is predicted based on the distortion-specific quality assessment (DSQA). Another method called the virtual cognitive model (VICOM) [13] also tries to differentiate images with respect to their impairment type first to overcome the obstacle of uneven response to common impairment sources. As a consequence, the discrimination of distortion types has been becoming a major new direction for current research of IQA.

Despite the abundant literature on IQA, however, very little effort [14] has been devoted to the study of the influence of image quality level on prediction accuracy of IQA metrics. It is noticed in [14] that fixations largely depend on the amount of distortion (near- and supra-threshold) for spatially localized distortion. Inspired by this valuable testing result, it has been further testified in our research that the image quality level can have significant influences on the performances of quality metrics. Interestingly, we found that SSIM/MS-SSIM is more accurate for assessing images with high quality and VIF/VIFP performs better for images with low quality. These observations can be explained by the fact that human perception mechanisms are different under different quality conditions, which may be also explained by the near- and suprathreshold principles like fixations. Based on those observations, we propose a novel DIfferent Perception

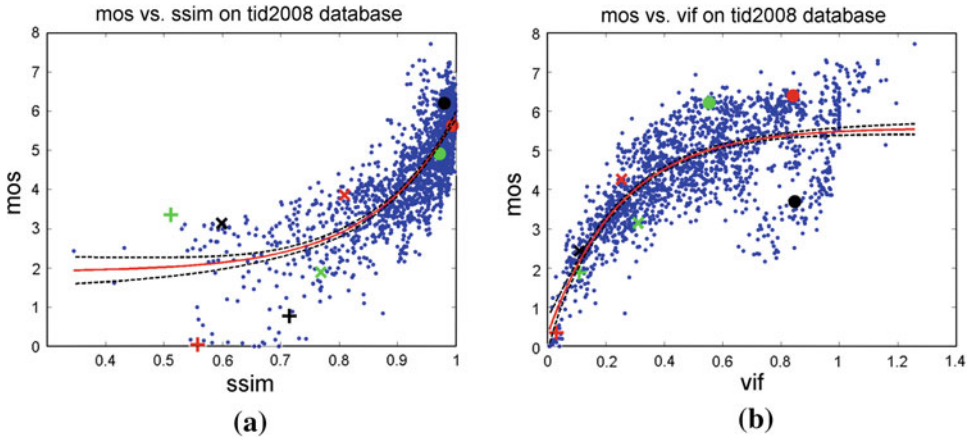

Fig. 2 Scatter plots of MOS versus **a** SSIM with eight distorted images (*non-blue dots*); **b** VIF with seven distorted images (*non-blue dots*) on TID2008 (color figure online)

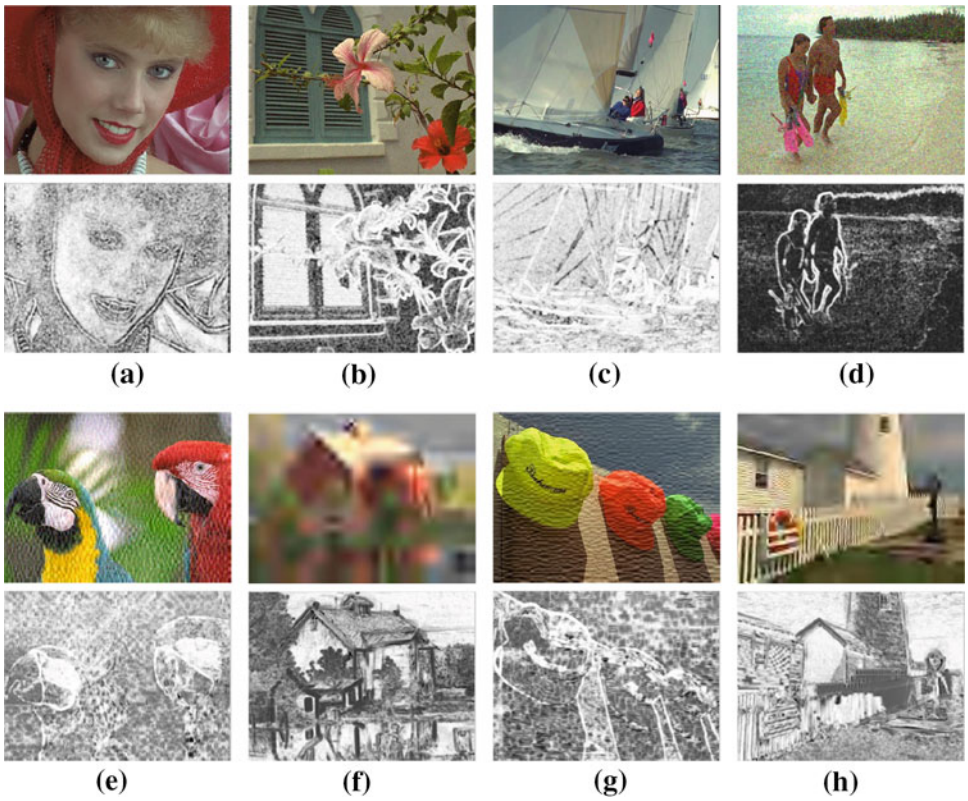

Fig. 3 Eight random distorted images and their monochrome SSIM maps (on the *lower row*): **a** red ·; **b** green ·; **c** black ·; **d** red x; **e** green x; **f** black x; **g** red +; **h** green + in Fig. 2a (color figure online)

(DIP) mechanism inspired IQA method, making full use of strength of SSIM/MS-SSIM and VIF/VIFP. The DIP method also has two steps: discrimination of the quality level using initial quality score and prediction of the final quality by combining quality scores of the component IQA metrics.

The remainder of this paper is organized as follows. Section 2 introduces some mainstream IQA metrics and analyzes their performances before presenting human different perception mechanism to be explored in this paper. Section 3 describes our proposed DIP paradigm in detail. In Sect. 4, experimental results using the four well-known databases, Tampere Image Database 2008 (TID2008) [15], Laboratory for Image and Video Engineering (LIVE) database [16], Categorical Image Quality (CSIQ) database [17] and Toyama-MICT database [18], are reported and analyzed. Finally, conclusion is drawn and future work is given in Sect. 5.

Table 1 Consistency degree of distorted images and their SSIM maps

High quality		
Image number	Fig. 3a–c	
Consistent degree	Good	
Low quality		
Image number	Fig. 3d, e, g	Fig. 3f, h
Consistent degree	Poor	Bad

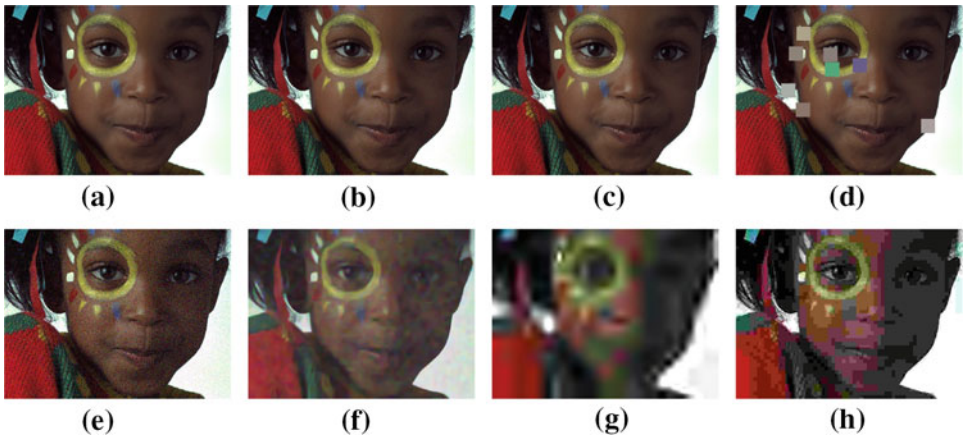

Fig. 4 A randomly chosen reference image and its corresponding seven distorted images (also randomly taken): **a** reference image; **b** red ·; **c** green ·; **d** black ·; **e** red x; **f** green x; **g** black x; **h** red + in Fig. 2b (color figure online)

Table 2 Consistency degree of reference and distorted images

High quality		
Image number	Fig. 4b–d	
Consistent degree	Bad	
Low quality		
Image number	Fig. 4e	Fig. 4f–h
Consistent degree	Good	Excellent

2 Different perception mechanism

2.1 Performance analysis of mainstream IQA metrics

The analysis in this work is mainly performed on TID2008 [15], the largest IQA database developed for the verification of full-reference quality metrics. TID2008 contains 25 reference images and 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, noneccentricity pattern noise, local blockwise distortions of different intensity, mean shift (intensity shift) and contrast change.

Twelve mainstream IQA metrics [19]: IFC [20], MSE, MS-SSIM [4], NQM [21], PSNR, SNR, SSIM [3], UQI [22], VIF, VIFP, VSNR [23] and WSNR [24], have been tested using nonlinear regression with a four-parameter logistic function as suggested by VQEG [25]

$$q(s) = \frac{\beta_1 - \beta 2}{1 + \exp(-(s - \beta_3)/\beta_4)} + \beta_2 \tag{1}$$

with s being the input score and $q(s)$ the mapped score, and β_1 to β_4 are free parameters to be determined during the curve-fitting process. The scatter plots of difference IQA metrics and regression results are illustrated in Fig. 1.

Two important observations can be made from Fig. 1: (1) The convergence trend of some IQA metric scores versus MOS is very unclear, e.g., MSE, NQM, PSNR and WSNR; (2) several algorithms, such as MS-SSIM, SSIM, VIF, VIFP and VSNR, have quite convergent results in part of the scatter plots. And by convergent, we mean high correlation seems to exist between subjective and objective scores. For example, in Fig. 1, when VIF value of image is less than about 0.65, the corresponding dots are very close to the fitted curve of Eq. (1). In other words, VIF has better performance for images with low quality. It is also noticed that IFC and VIFP have the similar property. On the contrary, data points of SSIM and MS-SSIM versus MOS are quite close to the respective fitted curves when their values are higher than about 0.9, indicating better performances on images with high quality.

2.2 Perception mechanism in different duality conditions

Why is there such a distinction between these IQA metrics in their performances under different quality conditions? Using SSIM and VIF as examples, we will analyze the perception mechanism of the HVS under different quality conditions.

SSIM: The basic spatial domain SSIM algorithm [3] is based on separated comparisons of local luminance, contrast and structure between a distorted image and its reference image. The luminance, contrast and structural similarities between two local image patches extracted from the reference and distorted images are evaluated as

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{2}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{3}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{4}$$

where μ_x, σ_x and σ_{xy} represent the mean, standard deviation and cross-correlation evaluations, respectively, and C_1 to C_3

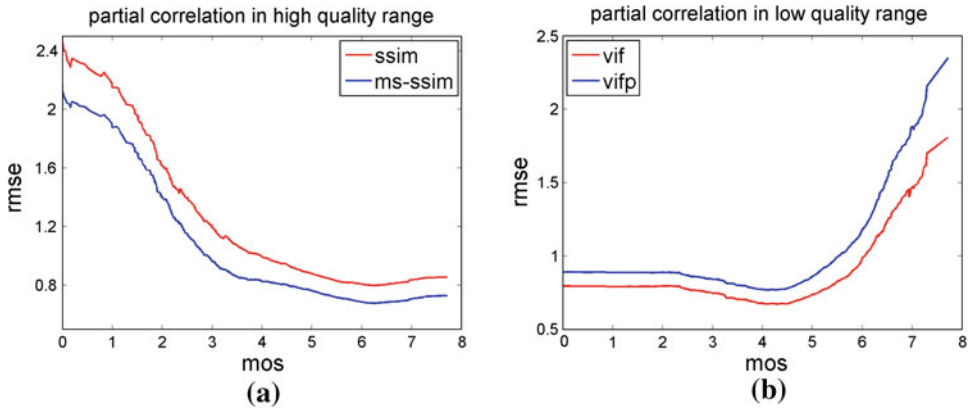

Fig. 5 Two smallest partial correlation of quality metrics. **a** SSIM and MS-SSIM in the high-quality range, **b** VIF and VIFP in the low-quality range

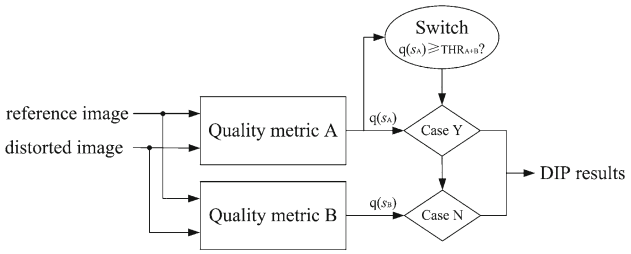

Fig. 6 Illustration of two steps of the DIP metric

are small constants. The SSIM_MAP is defined as the product of the three components,

$$\text{SSIM\_MAP}(x_i, y_i) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where x_i and y_i are the image contents at the ith local window. The SSIM index evaluating the overall image quality is defined by

$$\text{SSIM}(X, Y) = \frac{1}{L}\sum_{i=1}^{L} \text{SSIM\_MAP}(x_i, y_i) \tag{6}$$

where X and Y are the reference and distorted images, respectively. L is the number of local windows in the image.

Admitting the fact that the accuracy of SSIM depends on the agreement between the distorted image and the computed SSIM_MAP image, eight distorted images were randomly selected from TID2008 and have been illustrated in Fig. 3. As shown in Fig. 2a, corresponding data points of the eight images distribute widely on the scatter plots, where in the range of images with high quality, red, green and black "·" indicate accurate prediction. The visible distortions within those images as manifested in the SSIM maps are highly consistent with human visual perception, as shown in Fig. 3a–c. However, things are completely different in low-quality range. On low-quality images shown in Fig. 3d–h (marked by red "x", green "x", black "x", red "+" and green "+" in Fig. 2a), SSIM all have low performance in that the perceptual distortion and the computed SSIM map are poor correlation.

This variation of performance for SSIM may be explained from the viewpoint of human psychovisual perception as follows: SSIM extracts structural information from visual scenes. Structural information can be extracted easily and accurately when images are clear enough, i.e., with high quality. However, for images with low quality, extraction of the structural information is affected by the type of distortions. For example, Fig. 3d, e, g with distortions of high-frequency noise and JPEG2000 transmission error helps to demonstrate that the structural information cannot be effectively computed by SSIM when the qualities of these images are poor. And JPEG2000 compression distortion exerts little influence on structural information even when it corrupts the images seriously, as illustrated in Fig. 3f, h. Therefore, it is generally difficult to extract structural information effectively when images are of bad quality. Table 1 tabulates the consistency between the perceived distortion and their SSIM maps.

VIF: Based upon the definitions in [5], the VIF is computed as

$$\text{VIF} = \frac{\sum_{j\in\text{subbands}} I(\vec{C}^{N,j}; \vec{F}^{N,j}|s^{N,j})}{\sum_{j\in\text{subbands}} I(\vec{C}^{N,j}; \vec{E}^{N,j}|s^{N,j})} \tag{7}$$

where $I(\vec{C}^{N,j}; \vec{F}^{N,j}|s^{N,j})$ is the mutual information between a distorted image and its reference image, defined by

$$I(\vec{C}^{N}; \vec{F}^{N}|s^{N}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1 + \frac{s_i^2\lambda_k}{\sigma_n^2}\right) \tag{8}$$

and $I(\vec{C}^{N,j}; \vec{E}^{N,j}|s^{N,j})$ indicates information content of the reference image, given by

$$I(\vec{C}^{N}; \vec{E}^{N}|s^{N}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{M}\log_2\left(1 + \frac{g_i^2 s_i^2\lambda_k}{\sigma_v^2 + \sigma_n^2}\right) \tag{9}$$

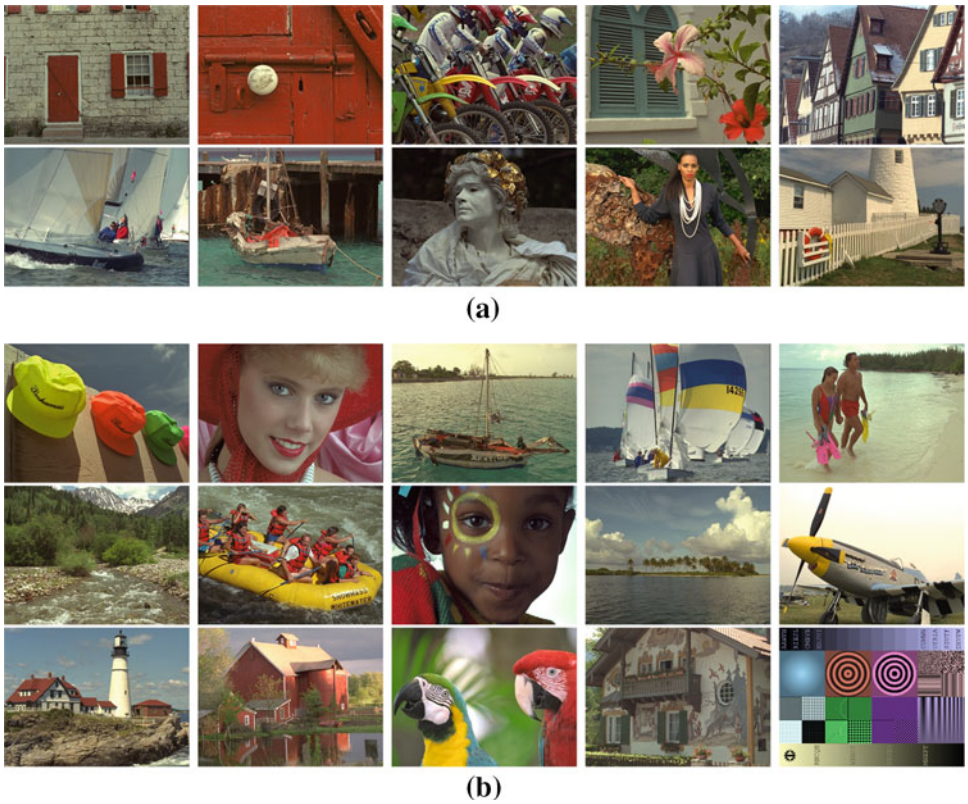

Fig. 7 Illustration of the training and testing groups: **a** Group 1; **b** Group 2

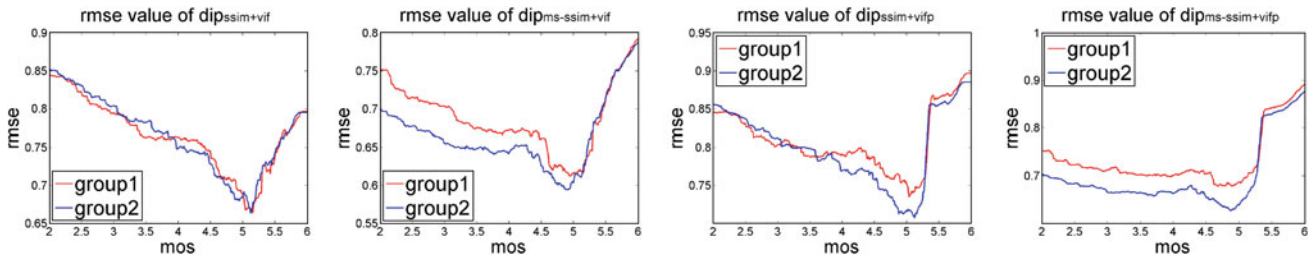

Fig. 8 Changes of RMSE values of $DIP_{SSIM+VIF}$, $DIP_{MS\text{-}SSIM+VIF}$, $DIP_{SSIM+VIFP}$ and $DIP_{MS\text{-}SSIM+VIFP}$ on TID2008 database

Table 3 The THR_{A+B} values with different combinations of A and B on TID2008 database

	Group 1 (680 images)		Group 2 (1,020 images)	
	$RMSE_{min}$	THR_{A+B}	$RMSE_{min}$	THR_{A+B}
A:SSIM B:VIF	0.6637	5.1600	0.6631	5.1300
A:MS-SSIM B:VIF	0.6125	4.9600	0.5941	4.9000
A:SSIM B:VIFP	0.7344	5.0400	0.7071	5.1300
A:MS-SSIM B:VIFP	0.6781	4.6900	0.6260	4.8800

where the specific meanings of some symbols can be found in [5].

We also randomly selected seven distorted images from TID2008 with the same reference image as shown in Fig. 4. Their corresponding data dots scatter widely in Fig. 2b. Note that the computed information content in the formulation of VIF is the same since the reference images are the same, so the prediction accuracy of VIF is fully determined by the agreement between the perceptual quality and the mutual information extracted. When the images are of low quality, as those shown in Fig. 4e–h, the difference between the reference and distorted images is obvious and the mutual information can be easily estimated, see points marked by red "x", green "x", black "x" and red "+" in Fig. 2b. However, in the

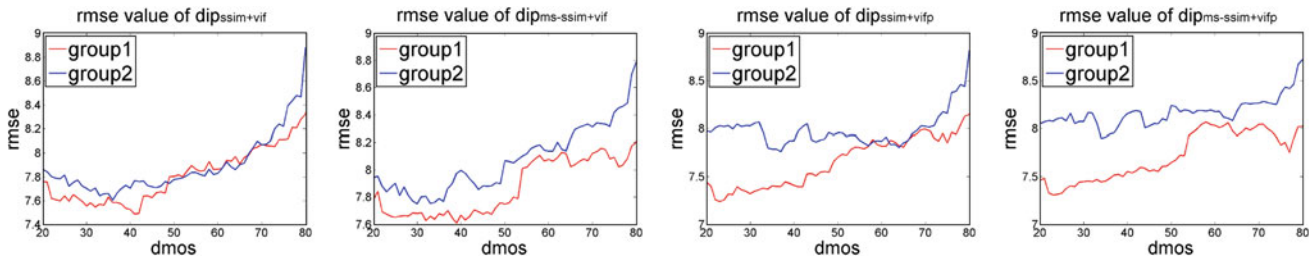

Fig. 9 Changes of RMSE values of $DIP_{SSIM+VIF}$, $DIP_{MS\text{-}SSIM+VIF}$, $DIP_{SSIM+VIFP}$ and $DIP_{MS\text{-}SSIM+VIFP}$ on LIVE database

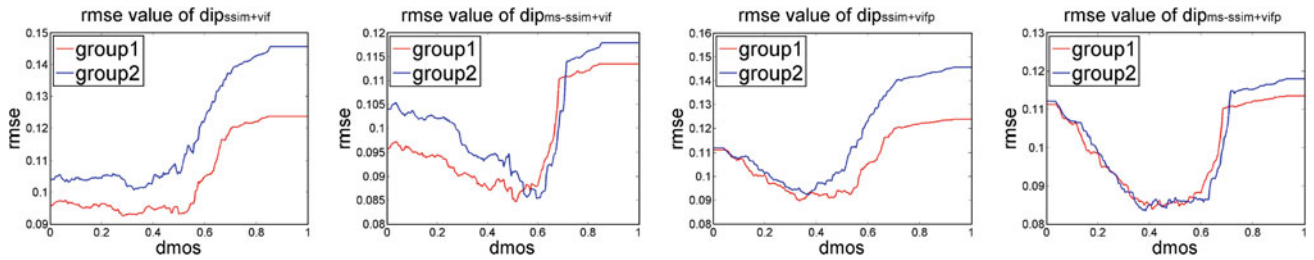

Fig. 10 Changes of RMSE values of $DIP_{SSIM+VIF}$, $DIP_{MS\text{-}SSIM+VIF}$, $DIP_{SSIM+VIFP}$ and $DIP_{MS\text{-}SSIM+VIFP}$ on CSIQ database

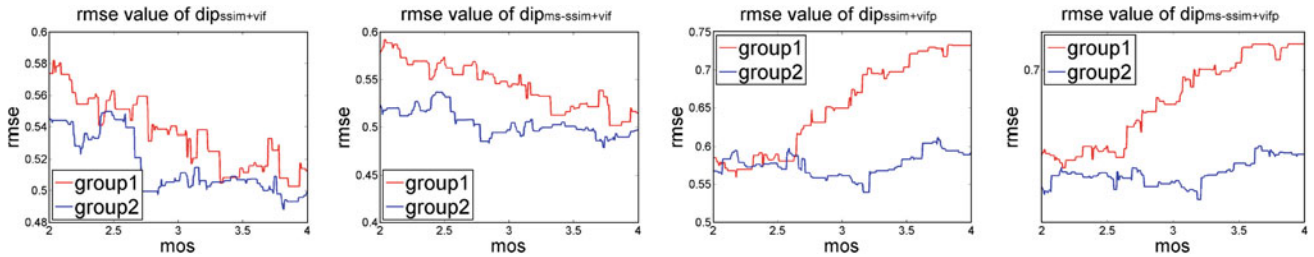

Fig. 11 Changes of RMSE values of $DIP_{SSIM+VIF}$, $DIP_{MS\text{-}SSIM+VIF}$, $DIP_{SSIM+VIFP}$ and $DIP_{MS\text{-}SSIM+VIFP}$ on Toyama-MICT database

region of high quality, as marked by red, green and black "·" in Fig. 2b, the HVS hardly perceives any difference between the reference and distorted images, i.e., their corresponding mutual information seldom exists, as shown in Fig. 4b–d. Consequently, the inaccuracy of the mutual information estimation leads to distinct departure between the objective and subjective quality scores.

The behavior of VIF can also be reasoned from the perspective of human vision: VIF quantifies information content and mutual information for image quality assessment. For images with low quality, both mutual information and information content are easy to capture for VIF. However, for images with high quality, VIF tends to fail since there can be very little difference between mutual information and information content. The consistency degree of distorted and reference images is presented in Table 2.

In summary, for image with high quality, i.e., details in the image can be clearly identified, the HVS extracts structure information as predicted by the generic IQA framework [3], where SSIM can be computed faithfully from the original and distorted images. However, for images with low quality, where objects or scenes in the image cannot be reliably represented, the perceptual quality can be approximated by the uncertainty of the unfaithful part of the scene, according to the free energy principle in [26]. Under the situation of low image quality, VIF establishes a link between the uncertain parts in the distorted image and their corresponding parts in the reference image by measuring the mutual information. And as a consequence, SSIM and VIF perform well in high and low image quality range, respectively. Thus, it would be natural to combine the merits of SSIM and VIF type of methods toward better IQA results, which is the topic of the next section.

3 Image quality assessment by differentiating quality levels

Though the idea of combining the strength of the two types of algorithms looks quite attractive, we are still facing two major difficulties: (1) which IQA metrics to use and (2) how to combine the metrics. To solve the first problem, we provide more empirical studies on the partial correlation between each IQA algorithms and the subjective scores. And by

Table 4 The THR_{A+B} values with different combinations of A and B on LIVE database

	Group 1 (379 images)		Group 2 (400 images)	
	$RMSE_{min}$	THR_{A+B}	$RMSE_{min}$	THR_{A+B}
A:SSIM B:VIF	7.4884	41.000	7.6054	36.000
A:MS-SSIM B:VIF	7.6088	39.000	7.7499	30.000
A:SSIM B:VIFP	7.2386	23.000	7.7618	37.000
A:MS-SSIM B:VIFP	7.3074	23.000	7.8976	34.000

Table 5 The THR_{A+B} values with different combinations of A and B on CSIQ database

	Group 1 (461 images)		Group 2 (405 images)	
	$RMSE_{min}$	THR_{A+B}	$RMSE_{min}$	THR_{A+B}
A:SSIM B:VIF	0.0924	0.2850	0.1006	0.3250
A:MS-SSIM B:VIF	0.0847	0.5100	0.0854	0.5950
A:SSIM B:VIFP	0.0897	0.3350	0.0922	0.3700
A:MS-SSIM B:VIFP	0.0838	0.4100	0.0834	0.3850

Table 6 The THR_{A+B} values with different combinations of A and B on Toyama-MICT database

	Group 1 (72 images)		Group 2 (96 images)	
	$RMSE_{min}$	THR_{A+B}	$RMSE_{min}$	THR_{A+B}
A:SSIM B:VIF	0.5027	3.8550	0.4881	3.8150
A:MS-SSIM B:VIF	0.5016	3.7800	0.4790	3.7300
A:SSIM B:VIFP	0.5590	2.1800	0.5390	3.1650
A:MS-SSIM B:VIFP	0.5715	2.1600	0.5295	3.2000

partial, we mean that only images within a specific quality range are used. At first, we sort the quality scores of every IQA metric according to MOS values from low to high, or vice versa, i.e., $q(s_1), \ldots, q(s_K)$. K is the total number of images. Then, the root mean-squared error (RMSE) of the nonlinear regression being used as a partial correlation measure is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\mathrm{MOS}_i - q(s_i))^2} \tag{10}$$

Table 7 The final chosen values of THR_{A+B} on four databases

THR_{A+B}	TID2008	LIVE	CSIQ	Toyama-MICT
A:SSIM B:VIF	5.1450	38.500	0.3050	3.8350
A:MS-SSIM B:VIF	4.9300	34.500	0.5525	3.7550
A:SSIM B:VIFP	5.0850	30.000	0.3525	2.6725
A:MS-SSIM B:VIFP	4.7850	28.500	0.3975	2.6800

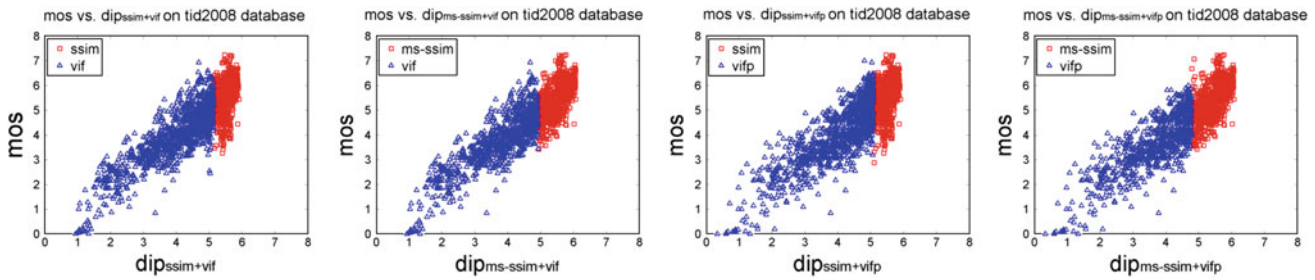

Fig. 12 Scatter plots of MOS versus $DIP_{SSIM+VIF}$, $DIP_{MS\text{-}SSIM+VIF}$, $DIP_{SSIM+VIFP}$ and $DIP_{MS\text{-}SSIM+VIFP}$ (after nonlinear regression) on TID2008 database

Table 8 Five groups of PLCC, SRCC, KRCC, AAE and RMSE values (after nonlinear regression) of IFC, MSE, MS-SSIM, NQM, PSNR, SNR, SSIM, UQI, VIF, VIFP, VSNR, WSNR, IW-SSIM, $DIP_{SSIM+VIF}$, $DIP_{MS\text{-}SSIM+VIF}$, $DIP_{SSIM+VIFP}$ and $DIP_{MS\text{-}SSIM+VIFP}$ on TID2008, LIVE, CSIQ and Toyama-MICT databases

Metrics	PLCC	SRCC	KRCC	AAE	RMSE
TID2008 (1,700 images) [15]					
IFC [20]	0.7170	0.5690	0.4256	0.7612	0.9355
MSE	0.5689	0.5531	0.4027	0.8395	1.1036
MS-SSIM [4]	0.8404	0.8542	0.6568	0.5659	0.7272
NQM [21]	0.6096	0.6236	0.4600	0.7796	1.0637
PSNR	0.5643	0.5531	0.4027	0.8467	1.1079
SNR	0.5288	0.5235	0.3744	0.8768	1.1389
SSIM [3]	0.7715	0.7749	0.5768	0.6588	0.8537
UQI [22]	0.6632	0.5851	0.4255	0.8141	1.0043
VIF [5]	0.8051	0.7491	0.5861	0.6069	0.7960
VIFP [5]	0.7481	0.6539	0.4945	0.7202	0.8904
VSNR [23]	0.6817	0.7045	0.5340	0.6904	0.9813
WSNR [24]	0.5383	0.4877	0.3930	0.8054	1.1309
IW-SSIM [27]	0.8488	0.8559	0.6636	0.5543	0.7095
$DIP_{SSIM+VIF}$	0.8593	0.8182	0.6324	0.5219	0.6649
$DIP_{MS\text{-}SSIM+VIF}$	0.8878	0.8636	0.6784	0.4721	0.6030
$DIP_{SSIM+VIFP}$	0.8384	0.7877	0.5950	0.5819	0.7213
$DIP_{SM\text{-}SSIM+VIFP}$	0.8725	0.8562	0.6632	0.5182	0.6509
LIVE database (779 images) [16]					
IFC [20]	0.9250	0.9248	0.7561	8.5649	10.389
MSE	0.8584	0.8755	0.6865	11.192	14.026
MS-SSIM [4]	0.9402	0.9511	0.8043	7.4382	9.3122
NQM [21]	0.9128	0.9093	0.7430	8.5183	11.164
PSNR	0.8701	0.8755	0.6865	10.540	13.473
SNR	0.8591	0.8649	0.6738	10.933	13.991
SSIM [3]	0.9383	0.9478	0.7961	7.5251	9.4508
UQI [22]	0.8984	0.8941	0.7100	9.4233	12.006
VIF [5]	0.9594	0.9633	0.8273	6.2323	7.7102
VIFP [5]	0.9594	0.9618	0.8249	6.1186	7.7143
VSNR [23]	0.9228	0.9271	0.7610	8.0616	10.531
WSNR [24]	0.9145	0.9159	0.7502	8.1651	11.059
IW-SSIM [27]	0.9425	0.9567	0.8175	7.4405	9.1317
$DIP_{SSIM+VIF}$	0.9601	0.9647	0.8305	6.2206	7.6472
$DIP_{MS\text{-}SSIM+VIF}$	0.9598	0.9632	0.8272	6.3007	7.7233
$DIP_{SSIM+VIFP}$	0.9600	0.9633	0.8283	6.0782	7.6578
$DIP_{SM\text{-}SSIM+VIFP}$	0.9594	0.9626	0.8260	6.1186	7.7143
CSIQ database (866 images) [17]					
IFC [20]	0.8358	0.7671	0.5897	0.1130	0.1441
MSE	0.8030	0.8058	0.6084	0.1175	0.1565
MS-SSIM [4]	0.8979	0.9133	0.7393	0.0875	0.1156
NQM [21]	0.7422	0.7412	0.5653	0.1334	0.1759
PSNR	0.7998	0.8058	0.6084	0.1195	0.1576
SNR	0.7821	0.7995	0.6004	0.1255	0.1636
SSIM [3]	0.8594	0.8756	0.6907	0.1008	0.1342
UQI [22]	0.8294	0.8098	0.6188	0.1124	0.1467

Table 8 continued

VIF [5]	0.9253	0.9195	0.7537	0.0753	0.0996
VIFP [5]	0.9043	0.8807	0.6969	0.0909	0.1121
VSNR [23]	0.8005	0.8109	0.6248	0.1161	0.1573
WSNR [24]	0.7703	0.7730	0.5989	0.1218	0.1674
IW-SSIM [27]	0.9025	0.9213	0.7529	0.0867	0.1131
$\text{DIP}_{\text{SSIM+VIF}}$	0.9300	0.9257	0.7630	0.0735	0.0970
$\text{DIP}_{\text{MS-SSIM+VIF}}$	0.9358	0.9344	0.7763	0.0704	0.0933
$\text{DIP}_{\text{SSIM+VIFP}}$	0.9388	0.9348	0.7729	0.0714	0.0915
$\text{DIP}_{\text{MS-SSIM+VIFP}}$	0.9491	0.9474	0.7937	0.0652	0.0844
Toyama-MICT database (168 images) [18]					
IFC [20]	0.8403	0.8354	0.6370	0.5371	0.6784
MSE	0.4421	0.4433	0.3644	0.9794	1.1225
MS-SSIM [4]	0.8920	0.8874	0.7029	0.4368	0.5657
NQM [21]	0.8892	0.8871	0.7049	0.4405	0.5726
PSNR	0.6355	0.6132	0.4443	0.7832	0.9662
SNR	0.5963	0.5725	0.4131	0.8330	1.0045
SSIM [3]	0.8877	0.8794	0.6939	0.4451	0.5762
UQI [22]	0.7164	0.7028	0.5227	0.6961	0.8731
VIF [5]	0.9136	0.9077	0.7315	0.4033	0.5087
VIFP [5]	0.8471	0.8479	0.6587	0.4969	0.6649
VSNR [23]	0.8705	0.8608	0.6745	0.4653	0.6160
WSNR [24]	0.7990	0.7988	0.5988	0.6040	0.7525
IW-SSIM [27]	0.9243	0.9202	0.7537	0.3696	0.4775
$\text{DIP}_{\text{SSIM+VIF}}$	0.9214	0.9155	0.7422	0.3895	0.4975
$\text{DIP}_{\text{MS-SSIM+VIF}}$	0.9214	0.9151	0.7416	0.3894	0.4974
$\text{DIP}_{\text{SSIM+VIFP}}$	0.8974	0.8887	0.7068	0.4269	0.5613
$\text{DIP}_{\text{MS-SSIM+VIFP}}$	0.9011	0.8970	0.7165	0.4200	0.5522

where MOS_i and s_i are the MOS value and quality score of ith test image. T indicates the number of test images. SSIM/MS-SSIM and VIF/VIFP were selected as candidates, due to the fact that they have the two smallest RMSE values for a wide range of k, as shown in Fig. 5.

For the integration of the methods shown in Fig. 6, we can combine two types of IQA metrics under the paradigm of differentiating quality levels to develop new metric $\text{DIP}_{\text{A+B}}$, i.e., use metric A for assessing images with high quality and metric B for images with low quality:

$$\text{DIP}_{\text{A+B}} = \begin{cases} q(s_{\text{A}}) & \text{if } q(s_{\text{A}}) \geq \text{THR}_{\text{A+B}} \\ q(s_{\text{B}}) & \text{otherwise} \end{cases} \tag{11}$$

where s_{A} and s_{B} indicate the prediction values of the two IQA methods for testing reference and distorted image pair s, and $\text{THR}_{\text{A+B}}$ represents the most reliable partition threshold of $\text{DIP}_{\text{A+B}}$. This threshold $\text{THR}_{\text{A+B}}$ is calculated by solving the following optimization problem:

$$\text{THR}_{\text{A+B}} = \arg\min_{\text{THR}} \{\text{RMSE}([B(s_1, \ldots, s_k), \text{A}(s_{k+1}, \ldots, s_K)] | k < \text{THR} \leq k+1, k \in \text{Interger}\}. \tag{12}$$

To find the $\text{THR}_{\text{A+B}}$ value, we randomly divide all the images in TID2008 into two groups with respect to various reference images, as illustrated in Fig. 7. And the two sets are individually trained by finding the respective $\text{THR}_{\text{A+B}}$ using Eq. (12) and then comparing the $\text{THR}_{\text{A+B}}$ values with each other.

With respect to various partition thresholds, the changes of RMSE values of four DIP methods are illustrated in Fig. 8. The $\text{THR}_{\text{A+B}}$ values, and their corresponding minimum RMSE, with different choices of A and B are listed in Table 3. Due to the fact that $\text{THR}_{\text{A+B}}$ of each group for every combination is almost the same, we simply report the average and set $\text{THR}_{\text{SSIM+VIF}} = 5.1450$, $\text{THR}_{\text{MS-SSIM+VIF}} = 4.9300$, $\text{THR}_{\text{SSIM+VIFP}} = 5.0850$ and $\text{THR}_{\text{MS-SSIM+VIFP}} = 4.7850$ in this paper.

Table 9 Direct and database size-weighted average results of PLCC, SRCC, KRCC, AAE and RMSE values (after nonlinear regression) in Table 8

Metrics	PLCC	SRCC	KRCC	AAE	RMSE
Direct average					
IFC [20]	0.8295	0.7741	0.6021	2.4941	3.0368
MSE	0.6681	0.6694	0.5155	3.2821	4.1022
MS-SSIM [4]	0.8926	0.9015	0.7258	2.1321	2.6802
NQM [21]	0.7885	0.7903	0.6183	2.4680	3.2441
PSNR	0.7174	0.7119	0.5355	3.0724	3.9262
SNR	0.6916	0.6901	0.5154	3.1921	4.0745
SSIM [3]	0.8642	0.8694	0.6894	2.1825	2.7537
UQI [22]	0.7769	0.7480	0.5693	2.7615	3.5075
VIF [5]	0.9009	0.8849	0.7247	1.8295	2.2786
VIFP [5]	0.8647	0.8361	0.6688	1.8567	2.3454
VSNR [23]	0.8189	0.8258	0.6486	2.3334	3.0714
WSNR [24]	0.7555	0.7439	0.5852	2.4241	3.2775
IW-SSIM [27]	0.9045	0.9135	0.7469	2.1128	2.6080
$DIP_{SSIM+VIF}$	0.9177	0.9059	0.7417	1.8014	2.2267
$DIP_{MS\text{-}SSIM+VIF}$	0.9260	0.9191	0.7559	1.8082	2.2293
$DIP_{SSIM+VIFP}$	0.9087	0.8932	0.7251	1.7896	2.2580
$DIP_{MS\text{-}SSIM+VIFP}$	0.9205	0.9156	0.7496	1.7805	2.2505
Database size-weighted average					
IFC [20]	0.7983	0.7095	0.5495	2.3211	2.8244
MSE	0.6847	0.6816	0.5145	2.9639	3.7365
MS-SSIM [4]	0.8792	0.8918	0.7120	1.9657	2.4724
NQM [21]	0.7229	0.7285	0.5604	2.3201	3.0611
PSNR	0.6936	0.6898	0.5183	2.8139	3.6088
SNR	0.6677	0.6696	0.4984	2.9194	3.7420
SSIM [3]	0.8357	0.8431	0.6591	2.0336	2.5695
UQI [22]	0.7589	0.7146	0.5409	2.5445	3.2262
VIF [5]	0.8741	0.8462	0.6879	1.7135	2.1438
VIFP [5]	0.8382	0.7874	0.6255	1.7515	2.2009
VSNR [23]	0.7735	0.7876	0.6134	2.1726	2.8783
WSNR [24]	0.6914	0.6679	0.5328	2.2592	3.0768
IW-SSIM [27]	0.8864	0.8974	0.7240	1.9572	2.4190
$DIP_{SSIM+VIF}$	0.9021	0.8817	0.7135	1.6687	2.0652
$DIP_{MS\text{-}SSIM+VIF}$	0.9170	0.9056	0.7386	1.6616	2.0512
$DIP_{SSIM+VIFP}$	0.8929	0.8673	0.6953	1.6674	2.0965
$DIP_{MS\text{-}SSIM+VIFP}$	0.9120	0.9040	0.7338	1.6437	2.0728

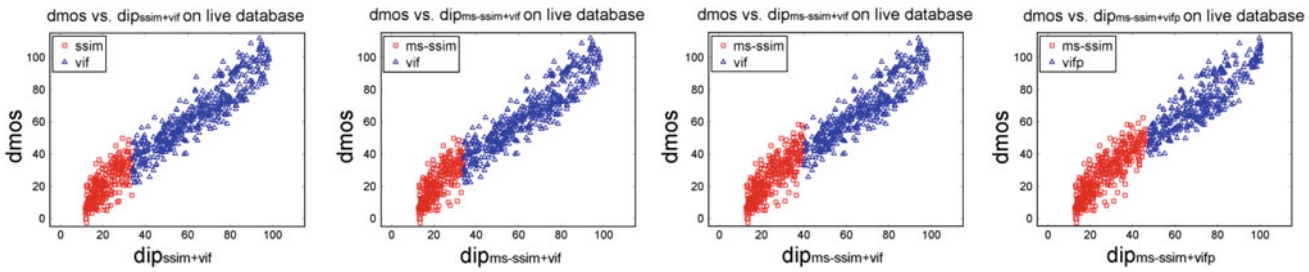

Fig. 13 Scatter plots of DMOS versus $DIP_{SSIM+VIF}$, $DIP_{MS\text{-}SSIM+VIF}$, $DIP_{SSIM+VIFP}$ and $DIP_{MS\text{-}SSIM+VIFP}$ (after nonlinear regression) on LIVE database

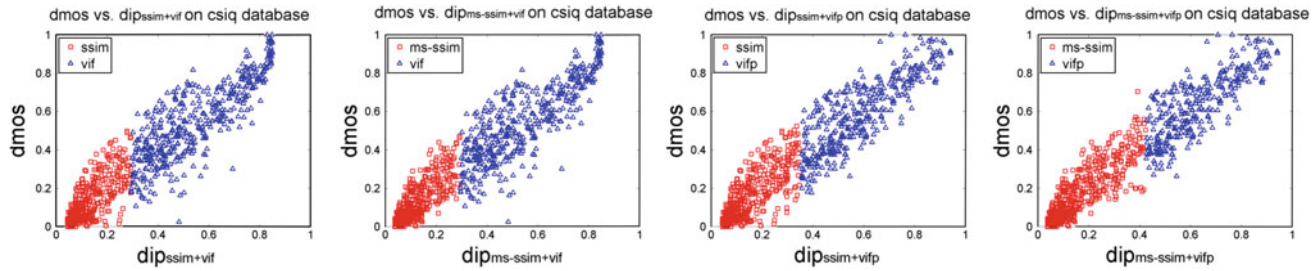

Fig. 14 Scatter plots of DMOS versus $\text{DIP}_{\text{SSIM+VIF}}$, $\text{DIP}_{\text{MS-SSIM+VIF}}$, $\text{DIP}_{\text{SSIM+VIFP}}$ and $\text{DIP}_{\text{MS-SSIM+VIFP}}$ (after nonlinear regression) on CSIQ database

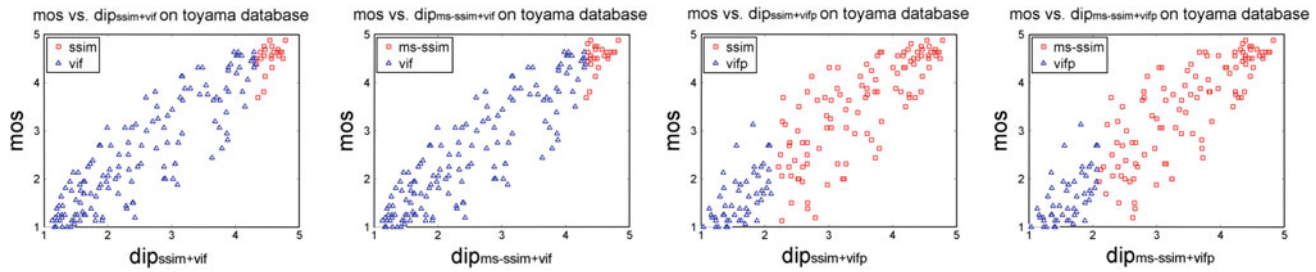

Fig. 15 Scatter plots of MOS versus $\text{DIP}_{\text{SSIM+VIF}}$, $\text{DIP}_{\text{MS-SSIM+VIF}}$, $\text{DIP}_{\text{SSIM+VIFP}}$ and $\text{DIP}_{\text{MS-SSIM+VIFP}}$ (after nonlinear regression) on Toyama-MICT database

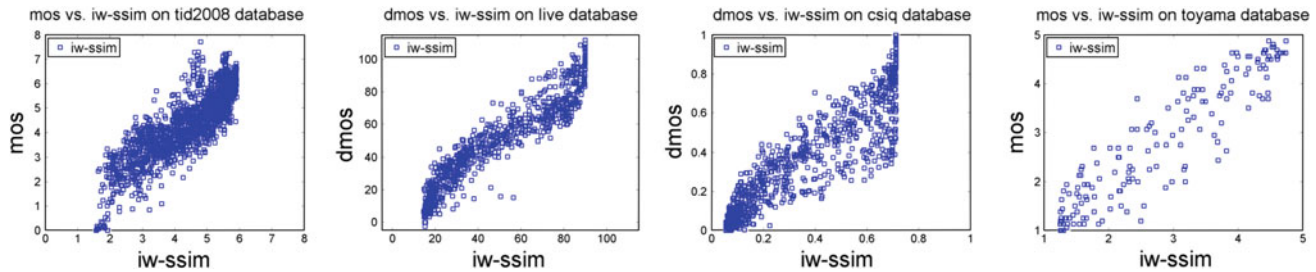

Fig. 16 Scatter plots of MOS/DMOS versus IW-SSIM (after nonlinear regression) on the four databases

For the other three databases (LIVE, CSIQ and Toyama-MICT), the changes of RMSE values of DIP algorithms with respect to various partition thresholds are illustrated in Figs. 9, 10, and 11. Similar approaches were applied to find the reliable $\text{THR}_{\text{A+B}}$ values for all the four combinations, as tabulated in Tables 4, 5 and 6. Table 7 presents the average $\text{THR}_{\text{A+B}}$ as the final chosen threshold values.

4 Experimental results

Figure 12 presents the scatter plots between the MOS and our proposed approaches, on the test database of TID2008. As shown in Fig. 1, these four DIP-based IQA methods have clearly achieved inspiring improvements. Moreover, five commonly used performance metrics as suggested by VQEG [25] are employed to further evaluate the competitive DIP-based IQA metrics and the thirteen mainstream methods, namely IFC, MSE, MS-SSIM, NQM, PSNR, SNR, SSIM, UQI, VIF, VIFP, VSNR, WSNR and recently proposed IW-SSIM [27], on TID2008, LIVE, CSIQ and Toyama-MICT databases. The first metric is the Pearson linear correlation coefficient (PLCC) between MOS and the objective scores after nonlinear regression. It can be defined by

$$\text{PLCC} = \frac{\sum_i (q_i - \bar{q}) * (o_i - \bar{o})}{\sqrt{\sum_i (q_i - \bar{q})^2 * (o_i - \bar{o})^2}} \tag{13}$$

where o_i is the subjective score of the ith image. The second metric is the Spearman rank-order correlation coefficient (SRCC), computed as

$$\text{SRCC} = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{14}$$

where d_i is the difference between the ith image's ranks in subjective and objective evaluations. It is a nonparametric rank-based correlation metric, independent of any monotonic nonlinear mapping between subjective and objective scores. The third metric, Kendall's rank-order correlation coefficient

(KRCC), is another nonparametric rank correlation metric given by

$$\mathrm{KRCC} = \frac{N_\mathrm{c} - N_\mathrm{d}}{\frac{1}{2}N(N-1)} \tag{15}$$

where N_c and N_d are the numbers of concordant and discordant pairs in the data set, respectively. Average absolute prediction error (AAE) is the fourth metric, which is calculated using the converted objective scores after the nonlinear mapping of Eq. (1):

$$\mathrm{AAE} = \frac{1}{N}\sum |q_i - o_i| \tag{16}$$

and the final metric RMSE has been already described in Eq. (10) by letting T equal to K.

All the five groups of corresponding PLCC, SRCC, KRCC, AAE and RMSE values, as well as their average results inspired by [27] over these four databases, are illustrated in Tables 8 and 9 ($\mathrm{THE}_{\mathrm{A+B}}$ values of other three databases are also individually trained by the same method for TID2008). The two groups of average results are direct average and average based on their size (1700 for TID2008 [15], 799 for LIVE [16], 866 for CSIQ [17] and 168 for Toyama-MICT [18]). Besides, the scatter plots of MOS/DMOS versus the proposed approaches on the other three databases as well as MOS/DMOS versus IW-SSIM on all the four databases are displayed in Figs. 13, 14, 15 and 16. It can be seen that our proposed four DIP-based IQA methods basically have led to better results than the thirteen mainstream metrics on these four popular image quality databases. Despite this, it still can be noticed that the performance gain of the proposed four algorithms are also different. $\mathrm{DIP}_{\mathrm{MS\text{-}SSIM+VIF}}$ method often has the largest gain. This can be explained by the fact that MS-SSIM and VIF are computed over multiple scales of the reference and distorted image patches by subband decomposition, which is closer to the behavior of the HVS in the real world.

Although $\mathrm{DIP}_{\mathrm{MS\text{-}SSIM+VIF}}$ approach generally has the best performance, it is easy to imagine that the higher accuracy comes with the cost of higher computational complexity. As shown in Fig. 6, metric A (SSIM/MS-SSIM) is firstly computed and then used to judge whether metric B (VIF/VIFP) is needed. According to our statistics using Figs. 12, 13, 14 and 15, the probability of invoking metric B is around 0.6, so the overall computational complexity of the proposed DIP method is about the sum of complexity of metric A plus 0.6 times of the complexity of metric B. It is well-known that SSIM has very low computational complexity, and its multiple version of MS-SSIM is with about 3–4 times more complex than SSIM. Meanwhile, VIF is based on wavelet decomposition and therefore has much higher complexity than SSIM. Accordingly, we suggest using $\mathrm{DIP}_{\mathrm{MS\text{-}SSIM+VIF}}$ only when the computational power allows. Under resource deficient conditions, other simpler forms of DIP type of metric can be employed.

Besides, it can also be easily observed from Figs. 8, 9, 10 and 11 that the thresholds do not have the same values for different DIP measure across the image databases. We believe these phenomena can be explained by the following reasons. According to the survey in [28], external factors should be taken into account in the research of image quality assessment, such as ambient illumination, image size and viewing distance. Consequently, the threshold values of DIP method for the same image database are generally quite close, while they can be quite different across databases with distinctive external factors.

Finally, it is worth emphasizing that our DIP inspired IQA paradigm, which may be explained by the fact that near- and suprathreshold principles, is not purely a new full-reference image quality metric, but a model for IQA, including full-reference, reduced-reference and no-reference conditions. In other words, through the application of various combinations of metrics A and B in Eq. (11), we can achieve different DIP methods. And moreover, with the development of researches in IQA, new quality approaches may be processed by DIP model to obtain much higher prediction accuracy. Again, despite the DIP-based full-reference image quality paradigm in this paper, our model also can have an immensely valuable application for no-reference or reduced-reference image quality assessment.

5 Conclusion

In this paper, we propose a new IQA paradigm through combining SSIM/MS-SSIM and VIF/VIFP under different image quality conditions. This work is inspired by two valuable findings of the performances of existing IQA metrics. First, the performances of major well-known IQA methods are considerably affected by the image quality level. Second, combining two metrics based on image quality level classification can lead to consistent performance improvement of the IQA metrics. Due to the fact that the significant $\mathrm{THR}_{\mathrm{A+B}}$ is trained by two independent groups, our method can perform well in the practical application, but it may have a little higher computational complexity because of the possible computation on both metrics A and B. Experimental results verify that the performances of the proposed methods are generally better than SSIM, MS-SSIM, VIF, VIFP and other mainstream IQA algorithms. Future work will be devoted to the studies of the relationship between images in various databases and the corresponding threshold values, and the influence of external factors during the subjective experiments on prediction accuracy of image quality metrics.

Acknowledgments This work was supported in part by postdoctoral foundation of Shanghai 11R21414200, postdoctoral foundation of China 20100480603, 201104276, NSERC, NSFC (61025005, 60932006, 61001145), SRFDP (20090073110022), the 111 Project (B07022) and STCSM (12DZ2272600).

References

1. Eskicioglu, A.M., Fisher, P.S.: Image quality measures and their performance. In: IEEE Trans. Commun. **43**, 2959–2965 (1995)
2. Miller, G.A.: The cognitive revolution: a historical perspective. Trends Cogn. Sci. **7**(3), 141–144 (2003)
3. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
4. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multi-scale structural similarity for image quality assessment. In: IEEE Asilomar Conference on Signals, Systems and Computer, vol. 2, pp. 1398–1402 (2003)
5. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
6. Lu, Z., Lin, W., Yang, X., Ong, E.P., Yao, S.S.: Modeling visual attentions modulatory aftereffects on visual sensitivity and quality evaluation. IEEE Trans. Image Process. **14**(11), 1928–1942 (2005)
7. Sadaka, N.G., Karam, L.J., Ferzli, R., Abousleman, G.P.: A no-reference perceptual image sharpness metric based on saliency weighted foveal pooling. In: IEEE International Conference on Image Processing, pp. 369–372 (2008)
8. Noorthy, A.K., Bovik, A.C.: Visual importance pooling for image quality assessment. IEEE J. Sel. Top. Signal Process. **3**(2), 193–201 (2009)
9. Ninassi, A., Meur, O.L., Callet, P.L., Barba, D.: Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In: IEEE International Conference on Image Processing, pp. 169–172 (2007)
10. Liu, H., Heynderickx, I.: Visual attention in objective image quality assessment: based on eye-tracking data. IEEE Trans. Circuits Syst. Video Technol. **21**(7), 971–982 (2011)
11. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. IEEE Signal Process. Lett. **17**(5), 513–516 (2010)
12. Moorthy, A.K., Bovik, A.C.: Statistics of natural image distortions. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 962–965 (2010)
13. Capodiferro, L., Jacovitti, G., Di Claudio, E.D.: Two-dimensional approach to full reference image quality. Assessment based on positional structural information. IEEE Trans. Image Process. **21**(2), 505–516 (2012)
14. Vu, C.T., Larson, E.C., Chandler, D.M.: Visual fixation patterns when judging image quality: effects of distortion type, amount, and subject experience. In: IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 73–76 (2008)
15. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID2008-A database for evaluation of full-reference visual quality assessment metrics. Adv. Mod. Radioelectron. **10**, 30–45 (2009)
16. Sheikh, H.R., Seshadrinathan, K., Moorthy, A.K., Wang, Z., Bovik, A.C., Cormack, L.K.: Image and video quality assessment research at LIVE. [Online] Available: http://live.ece.utexas.edu/research/quality/
17. Larson, E.C., Chandler, D.M.: Categorical image quality (CSIQ) database. [Online] Available: http://vision.okstate.edu/csiq
18. Horita, Y., Shibata, K., Kawayoke, Y., Sazzad, Z.M.P.: MICT image quality evaluation database. [Online] Available: http://mict.eng.u-toyama.ac.jp/mict/index2.html
19. Gaubatz, M.: Metrix MUX visual quality assessment package. http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
20. Sheikh, H.R., Bovik, A.C., de Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. Image Process. **14**(12), 2117–2128 (2005)
21. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on a degradation model. IEEE Trans. Image Process. **9**, 636–650 (2000)
22. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Process. Lett. **9**, 81–84 (2002)
23. Chandler, D.M., Hemami, S.S.: VSNR: a wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans. Image Process. **16**(9), 2284–2298 (2007)
24. Mitsa, T., Varkur, K.: Evaluation of contrast sensitivity functions for the formulation of quality measures incorporated in halftoning algorithms. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 301–304 (1993)
25. VQEG: Final report from the video quality experts group on the validation of objective models of video quality assessment. (2000). http://www.vqeg.org/
26. Zhai, G., Wu, X., Yang, X., Lin, W., Zhang, W.: A psychovisual quality metric in free-energy principle. IEEE Trans. Image Process. **21**(1), 41–52 (2012)
27. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Process. **20**(5), 1185–1198 (2011)
28. Lin, W., Jay Kuo, C.-C.: Perceptual visual quality metrics: a survey. J. Vis. Commun. Image R. **22**(4), 297–312 (2011)