Limitation and Challenges of Image Quality Measurement

Fan Zhang, Songnan Li, Lin Ma, King Ngi Ngan
The Chinese University of Hong Kong, N. T., Hong Kong, China SAR

ABSTRACT

Subjectively-rated image databases have become increasingly popular in the evaluation of image quality measurement algorithms. Several groups recently have improved their metrics' performance in matching these databases, using particular HVS (human visual system) properties or image statistical models. However, it is difficult to know whether these improvements are due to progress towards mimicking the perceptual properties, or are due to matching some characteristics of the databases. This paper demonstrates an inherent limitation in using such databases, showing that our very simple metric, built on the contrast masking effect, is able to perform as good as many state-of-the-art metrics. It is also argued that existent databases neither contain enough images with particularly biased distortions to test the significance of single HVS property, nor cover diverse distortion types to reflect the requirement of emerging applications.

Keywords: Image quality measurement, Subjective database

1. INTRODUCTION

Image quality measurement is a difficult computation problem [1]. Up till now, no single metric can completely represent the human response in visual quality assessment. The core problem is that each kind of image signal difference contributes differently to perceptual visual distortion. For example, we can hardly notice a slight change in: (1) scale, orientation, lighting, and contrast of an image, (2) the spatial distribution of a homogenous texture surface, and (3) the position of objects; but are easily aware of even small impairments like: (1) sharpness of contours, (2) deformation of faces, words and familiar symbols, (3) other artifacts in smooth or sensitive regions [2].

Exploiting our brain's solution to visual quality assessment requires understanding the HVS (human visual system) and an accurate but tractable computational model of visual perception. On the other hand, recent proposal of information theoretic approach has to be based on an ideal statistical model of images for the computational tractability. Nevertheless, the state-of-the-art metrics become more and more complicated.

A major challenge, however, lies in how to assess the performance of such "advanced" models. Ideally, objective metric should evaluate the image quality in a way that is consistent with our visual systems, but it is unclear how to evaluate the progress towards this goal. In practice, this amounts to create a subjectively-rated database to test the performance improvement by the new proposal. For example, at least five subjective databases have been shared on internet [3-7], including VIF, IVC, Toyama, TID2008, and A57. In these databases, many distorted images, generated to ostensibly represent the distortions encountered in the real world, were evaluated by human subjects. The subjective scores were finally regarded as "ground truths" about the image quality. Every improvement of the correlation between the subjective scores and the objective ones given by the metric is taken as a critical proof of the success of new proposal.

We should be cautious when using the distorted image databases, as it is not clear to what extent such databases actually engage the perceptual properties of the HVS in quality assessment. Although the databases contain a large number of images, with variations in distortion type and intensity; some distortion features are poorly defined, for example, the distortions in the smooth or rough regions, the dark or bright regions, the interested or uninterested regions, etc. To explore this issue, we find that a simple metric based on MSE (mean squared error) can achieve a comparable performance as the state-of-the art metrics.

Visual Communications and Image Processing 2010, edited by Pascal Frossard,
Houqiang Li, Feng Wu, Bernd Girod, Shipeng Li, Guo Wei, Proc. of SPIE Vol. 7744,
774402 · © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.863083

2. SIMPLE SOLUTION

The proposed metric computes the MSE for each distorted image block at a low frequency subband and weights the MSE with a contrast mask quantified by the block variance. This is a simple model, because it is a quadratic signal difference, without consideration of higher-order image statistics or any HVS properties other than the contrast mask. Specifically, it contains no explicit mechanisms related to counting chromatic channels, isolating the image structure, estimating the image entropy, calculating JND (just noticeable difference) [8], making a multi-scale analysis or discriminating the regions of interest.

The definition of the simple metric is given by (1). X and Y represent the reference and the distorted images respectively. The metric is based on 5×5 non-overlapping blocks. Let $\bar{x}_i$ and $\bar{y}_i$ denote the means of i-th block inside the reference and the distorted images respectively. Suppose both images have a total of n blocks. We use the operator var(·) to calculate the variance, i.e., var(X) is the image variance, var(x_i) is the block variance of x_i, and $\mathrm{var}(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{x}_i - \bar{X}\right)^2$ where $\bar{X}$ is the image mean given by $\bar{X} = \sum_{i=1}^{n}\bar{x}_i$. Then, the metric is defined as:

$$D(X,Y) = \frac{\mathrm{var}(\bar{x})}{\mathrm{var}(X)}\sum_{i=1}^{n}\frac{(\bar{x}_i - \bar{y}_i)^2}{\sqrt{\mathrm{var}(x_i)}+20} \tag{1}$$

Note that $D(X, Y)$ can be further normalized by the number of pixels, if the image resolutions of the database are not identical. In (1), “20” is a constant to avoid dividing by zero. This constant, 20, and the block size, 5, are the only two parameters whose values are determined by training on the databases. However, the metric performance is not sensitive to these two values.

The simple metric can be regarded as a weighted MSE, where all the weights are determined by the reference image but are independent of the distorted image. Since the metric is not symmetric, i.e., $D(X, Y) \neq D(Y, X)$, it is not a real distance metric. Nevertheless, it is quite simple due to its quadratic form with respect to Y.

We found that this simple metric is remarkably accurate on the five databases, outperforming PSNR, DCTune [9], SSIM [10], VSNR [12], and PSNR-HVS [13], and achieving comparable performance as VIF [11]. The results are analyzed in the next section.

3. EXPERIMENTS RESULTS

The correlation between the objective scores given by the metrics and the subjective ones given by the database are calculated to evaluate the performance of the metrics. The correlation is quantified by the linear correlation coefficient (LCC), root-mean-squared error (RMSE) and Spearman rank order correlation coefficient (SROCC). Before calculating the LCC, the objective scores are nonlinearly regressed according to [11]:

$$\mathrm{Score}(x) = \beta_1\left[\frac{1}{2} - \frac{1}{1+\exp\{\beta_2(x-\beta_3)\}}\right] + \beta_4 x + \beta_5 \tag{2}$$

The monotonic mapping (2) does not change the order of the original objective scores but probably improves the LCC between the objective scores and the subjective ones. We used the entire dataset from the databases, except for TID 2008. For TID 2008, we selected its distortion types of Gaussian noise, blur, JPEG compression, and JPEG 2000 compression, because only the four distortion types are covered by other four databases. The scores of the five databases are plotted in Fig. 1. The vertical coordinate is the subjective score (DMOS or MOS) while the horizontal coordinate is the score given by the simple metric. The blue curves in the figure are the optimized nonlinear regression results. The simple metric’s score is increasing with the distortion; however, the subjective scores in LIVE and A57 database are increasing with the distortions but the ones in IVC, Toyama and TID 2008 are decreasing, since the grading strategies of subjects were different for the five databases. As a result, some of the regressed curves are increasing while others are decreasing. The simple metric is also compared with PSNR, DCTune, SSIM, PSNR-HVS, VSNR and VIF, as listed in Table 1.

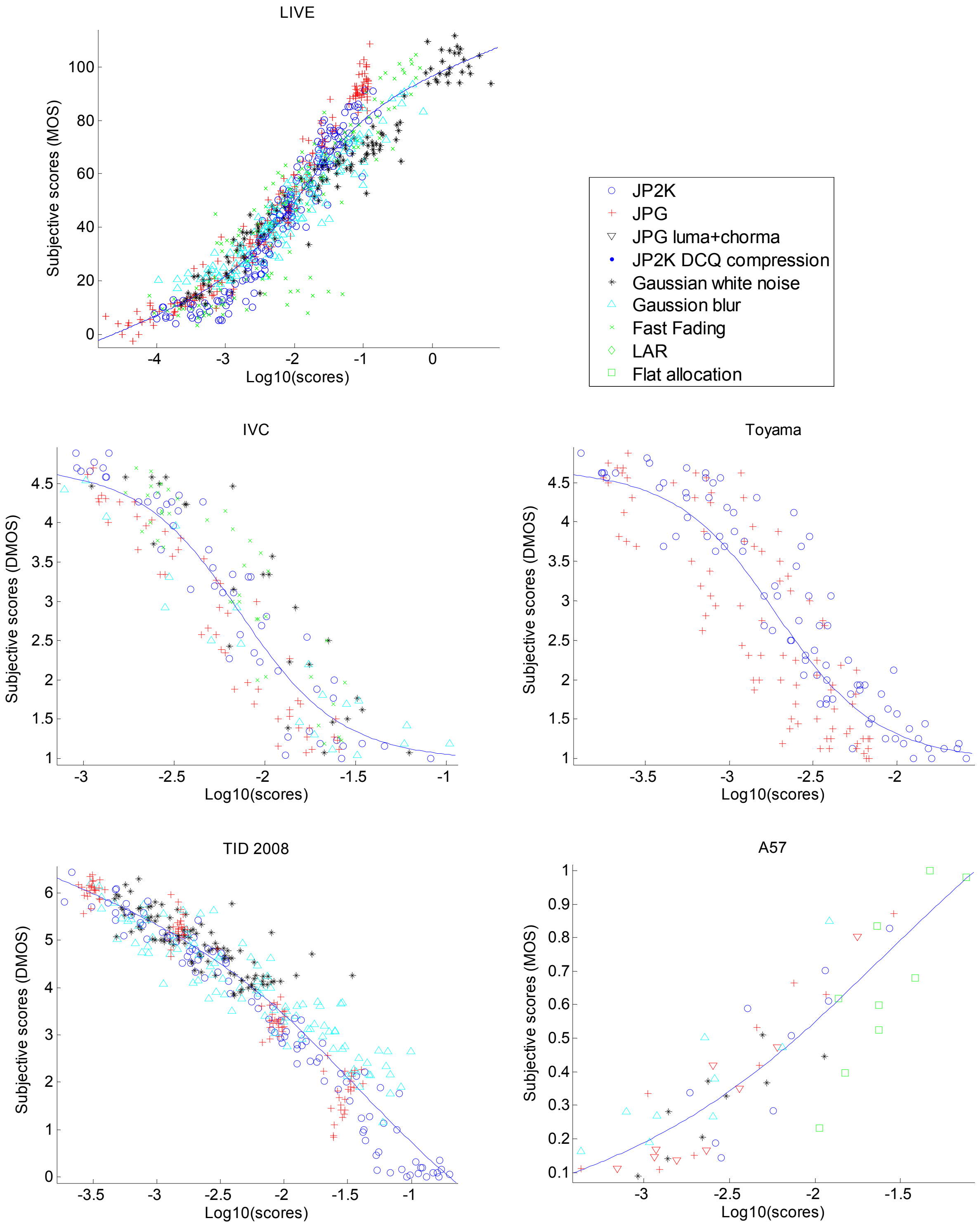

Fig. 1 Performance of the simple metric on the five databases.

Table 1. Correlations with Subjective Rating on Popular Distortions

		PSNR	DCTune	SSIM	VIF	PSNR-HVS	VSNR	Simple
LIVE	LCC	0.872	0.833	0.904	**0.960**	0.897	0.637	0.927
	RMSE	13.36	14.29	11.68	**7.649**	12.08	21.13	10.24
	SROCC	0.856	0.841	0.910	**0.964**	0.930	0.648	0.942
IVC	LCC	0.704	0.680	0.776	0.903	0.890	0.803	**0.915**
	RMSE	0.866	0.858	0.769	0.524	0.555	0.726	**0.491**
	SROCC	0.679	0.695	0.778	0.896	0.883	0.799	**0.905**
Toyama	LCC	0.632	0.605	0.718	**0.914**	0.858	0.862	0.886
	RMSE	0.970	0.992	0.872	**0.508**	0.643	0.634	0.581
	SROCC	0.612	0.594	0.787	**0.908**	0.848	0.861	0.884
TID2008 (subset)	LCC	0.804	0.678	0.899	0.946	0.921	0.900	**0.949**
	RMSE	0.940	1.163	0.692	0.512	0.618	0.690	**0.498**
	SROCC	0.851	0.748	0.885	0.935	0.943	0.906	**0.947**
A57	LCC	0.644	0.458	0.415	0.618	0.916	**0.950**	0.871
	RMSE	0.192	0.206	0.224	0.193	0.099	**0.077**	0.121
	SROCC	0.570	0.437	0.407	0.622	0.896	**0.935**	0.856

Table 2. Correlations with Subjective Rating on Particular Distortions in TID 2008

Distortion	Correlation	PSNR	DCTune	SSIM	VIF	PSNR-HVS	VSNR	Simple
Mean shift	LCC	**0.685**	0.648	0.588	0.595	0.684	-[*]	0.610
	RMSE	**0.420**	0.438	0.465	0.463	0.420	-	0.456
	SROCC	**0.696**	0.661	0.607	0.510	0.693	-	0.618
Block-wise different intensity	LCC	0.628	0.234	0.052	**0.827**	0.645	0.273	0.272
	RMSE	0.516	0.644	0.661	**0.372**	0.506	0.637	0.637
	SROCC	0.619	0.231	0.118	**0.832**	0.624	0.195	0.268

4. DISCUSSION

If any HVS properties are mimicked in the simple metric, only lowpass filtering and contrast masking can be considered. Calculating the difference of the blocks' means is equivalent to a lowpass filtering before subtraction. The higher frequency subbands are neglected when calculating the difference. On the other hand, the contrast masks in (1) can be divided into two parts. One is the local contrast embodied in the denominator "$\sqrt{\mathrm{var}(x_i)}+20$", which evaluates the contrast of each block inside an image. For instance, Fig. 2 shows the local contrast map of image *Lighthouse* and *Parrot* from LIVE database. The higher the local contrast, the darker is the region of the map. The other contrast mask is a global smoothness quantified by the multiplier "$\mathrm{var}(\bar{x})/\mathrm{var}(X)$", which provides a smoothness scalar for each image.

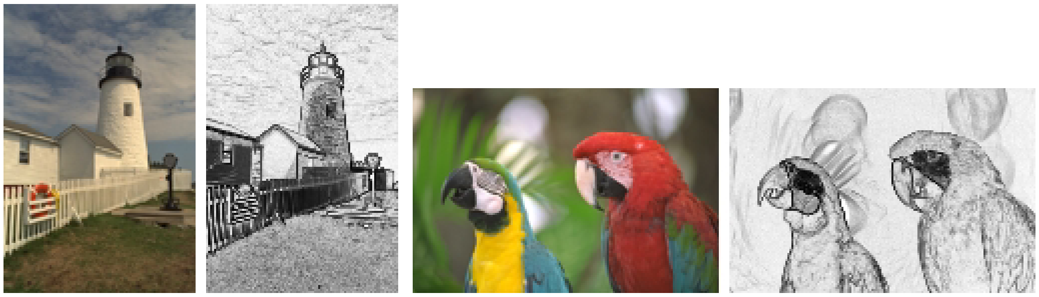

Fig. 2 Local roughness of *Lighthouse* and *Parrot*. Left: the color image; right: its local roughness map.

[*] The source code of metric [7] failed to return some valid scores when testing on the distorted images.

In order to explain the global smoothness, we demonstrate the case of 1D signal in Fig. 3. The left figure in Fig. 3 is a descending staircase signal with two smooth steps and the right figure is a horizontal sawtooth signal. The left signal looks smoother in the local regions than the right signal, despite of its larger variance in a global scale. As the result depicted by the global smoothness with respect to a window width of 5, the left signal has a global smoothness of 1 and the right signal of 0.05, that is, the left signal is considered to be smoother. The global smoothness of LIVE images are sorted in Fig. 4. The larger the global smoothness of the image is, the higher the image locates. The images with large smooth regions have higher global smoothness while the images with lots of texture have lower ones. As expected, the global smoothness of the random noise is as low as 0.04.

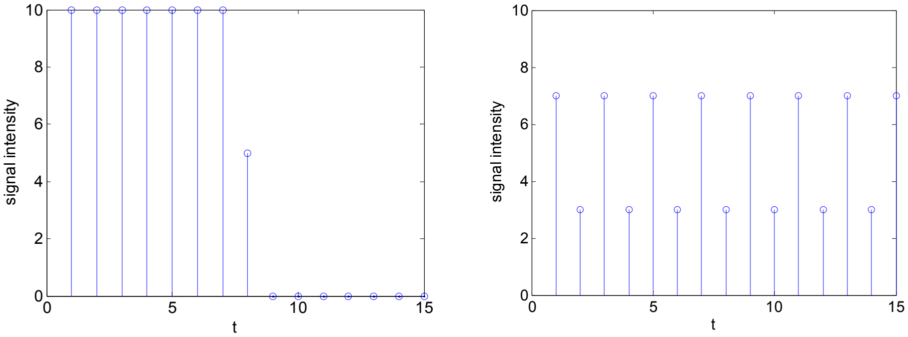

Fig. 3, Global smoothness demonstrated for 1D signal. Left: 1; right: 0.05 (window width = 5).

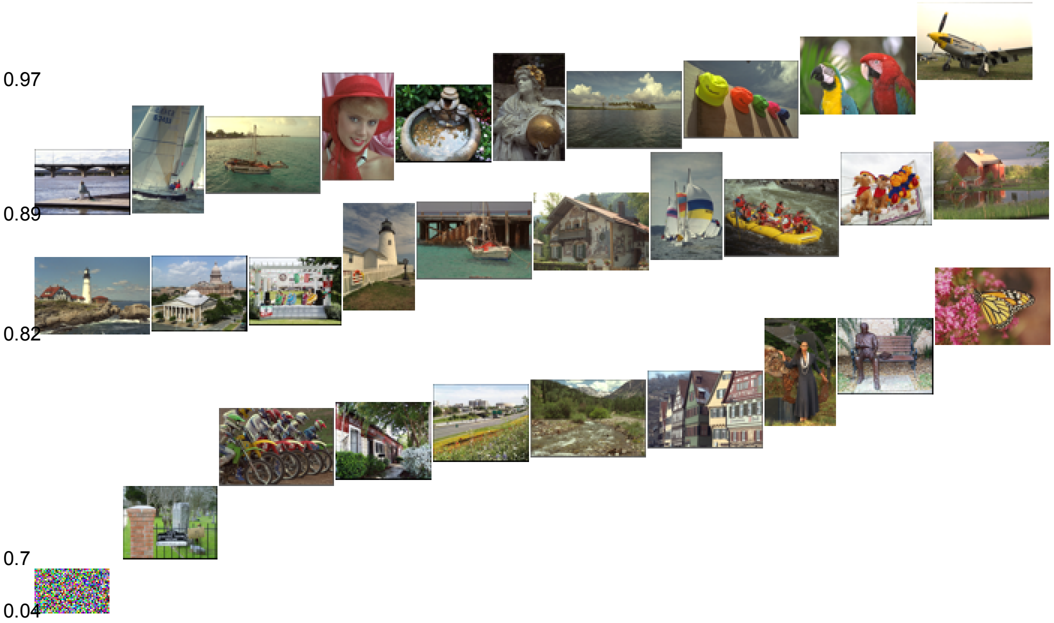

Fig. 4 Global smoothness of the reference images in LIVE database

The unexpected finding drove us to rethink current methodology of quality measurement. We do not think most HVS properties play trivial roles in our subjective quality evaluation. Most properties are unimportant and can be absent in the metric, only because the databases do not take into account the HVS properties when the distorted images are generated. Actually, their distortions are often unbiased in the features which the HVS is sensitive to. For instance, the distribution of the distortions is generally independent of visual salience, color space, scale or local luminance of the reference image. As shown on the left of Fig. 5, the logarithmic MSE of 779 distorted images from LIVE database are plotted in RGB color space. Every MSE vector consists of three color components and concentrates around the line $R = G = B = t$. It means that all distortions are even in every color channel without bias. Another example, shown on the right of Fig. 5, gives the MSE vectors of LIVE database. Every MSE vector consists of three components in the finest scale (scale 0), the second finest one (scale 1), and the third finest one (scale 2). We perform multi-scale analysis on images by means of Laplacian pyramid. All the MSE vectors in the scale space also concentrate around a line. It means that all distortions are unbiased in the scale space too. Due to the unbiasedness of the databases, a simple metric like (1) can be effective enough in comparison with the state-of-the-art metrics.

However, good performances on the popular databases do not mean that the problem of quality measurement has been solved completely. One proof is presented in Table 2. The distortion of mean shift is difficult to evaluate by all the state-of-the-art metrics. It also implies that more specially distorted image databases need to be created, if we want to test the significance of the particular HVS properties. The special database should be set up in two steps. Firstly, given a noise power, images should be impaired by the visible and invisible noises respectively, under the guidance of the HVS properties to be tested. Secondly, each distortion type should be accompanied by a significance factor, which describes the occurrence probability of the distortion in certain real-world application. Only with the completion of the two steps above will the evaluation of new HVS properties be effective according to the subjective results obtained from the databases.

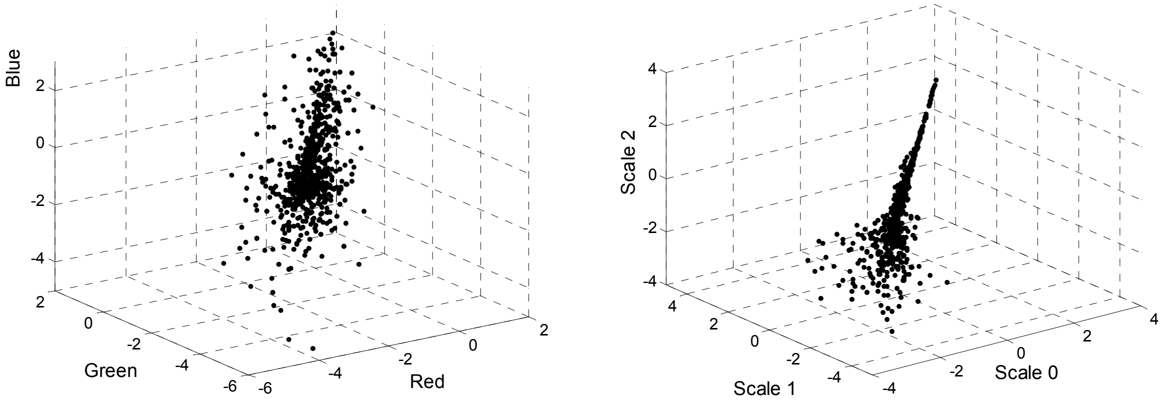

Fig. 5 Unbiasedness of LIVE database in color space (left) and scale space (right)

It is worth mentioning that up to 17 types of distortions was taken into account in the TID2008 database. Several special distortions, e.g., high frequency noise and masked noise, are generated to test the particular HVS properties [14]. The metric PSNR-HVS, which consider(s) the CSF (contrast sensitive function) and contrast masking effect did well in the corresponding distortions [13]. However, some distortions are not common in real-world applications. Table 2 shows the results on the distortion, “local block-wise distortions of different intensity”. None of the metrics can evaluate this distortion well. It is also conceivably difficult for any metric to tackle too many types of distortions properly. The significance factor is therefore helpful to indicate the impact of the distortions on image quality in the application and to make the problem tractable. Another approach is to define the distortion subset. A group of typical distortion types are counted in a subset for a particular application. A57 provides a good example for the subset definition [14].

Another problem with the subjective databases is that their testing conditions seem varied. A metric which works well on one database, may perform less well on others. For example, VSNR and VIF did best on the databases (i.e., A57 and LIVE database respectively) set up to test them, but their performances on other four databases drop, sometimes significantly. It is not convincing to show results on just one database.

Currently, a popular trend is to refine a metric by adjusting its parameters according to the training results on certain database. Given enough number of parameters, it is not difficult to find an optimized metric with an excellent performance. However, it is doubtful whether the metric will still perform well on other databases.

In summary, we must be cautious when using current subjective databases to test an objective quality metric exploiting HVS properties. The distortion types contained in the databases cover only a narrow range, and thus are probably not suitable to test most HVS properties. A metric trained on a single database possibly will not perform well on other databases. Moreover, the more empirically-trained parameters a metric contains, the more unstable it is when used on a new dataset. On the contrary, a simple solution may perform better than we expect.

5. NEW CHALLENGES

Although our simple metric has achieved a good performance on current databases, it does not imply that the image quality measurement problem is straightforward. We assert that current subjective databases do not cater for all the applications, especially some emerging new challenges for image quality measurement.

(a) Image inpainting

The image inpainting includes many image editing operations, like filling in the image impairment with adjacent texture or inferred structure, removing an object from the scene flawlessly, fusing two image patches seamlessly, etc. How to assess the quality of an inpainted image is nontrivial. The metric design can partly learn from the distortion criteria embodied in the problem formulation, i.e., the continuity of gradient optical flow field, edges, contours and structures. However, there are still more features to be exploited. Progress in quality metric is certainly helpful in the improvement of inpainting algorithms.

(b) Advanced image coding

The state-of-the-art image coding algorithms have already employed many advanced techniques, such as spatio-temporal JND [15], object-oriented coding, shape/texture coding, fovea coding, etc. A new rate-distortion theory has been requested to guide the optimization of the coding scheme. Visual attention is crucial for quality measurement, since the regions of interest are allocated the majority of bits during coding. Texture similarity is also a critical component of the metric, because traditional metric can hardly assess the quality of the textured region. An intelligent coding technique fro texture may save bits further.

(c) Image watermarking and covert communication

Image perceptual watermarking is an interesting field, where lots of HVS properties are considered in order to hide more payload of invisible watermark. Watermark is actually a flexible and smart noise, concentrating in very low or high frequency subbands, distributing in very bright or dark regions, as well as being non-uniform in other features which the HVS is sensitive to. For the application of covert communication, a looser distortion constraint can be accepted. The covert message is embedded as a watermark. Once it does not incur the eavesdropper's suspicion (i.e. without semantic distortions or abnormal flaws), the watermarked image will be well acceptable even its traditional quality is still low. A metric which exhaustively exploits the HVS properties will lead to a milestone in watermarking and covert communication.

(d) Image adaptation

Image adaptation is a hot topic which facilitates the universal multimedia access (UMA) on the communication network. After the adaptation process, the image resolution and aspect ratio may be changed, but the leading foreground must not be deformed and need be displayed with an adequate size. As a result, the image may suffer from a non-uniform geometrical distortion. If the quality degradation due to image adaptation can be assessed adequately, it could be used as a similarity metric for scene retrieval.

6. CONCLUSION

During the past decade, novel quality metrics, subjective evaluation standards, and database benchmarks came to the forth, and the problem of image quality measurement was defined and classified more and more clearly. However, the problem is still far from satisfactorily solved, since it is very difficult to design a subjective database that properly defines what constitutes "improvement". It deserves more attention and effort to establish the subjective databases with more special distortion types designed deliberately and corresponding significance factors ranked experimentally.

Acknowledgement

This work was partially supported by a grant from the Chinese University of Hong Kong under the Focused Investment Scheme (Project 1903003).

REFERENCES

[1] Z. Wang, A. C. Bovik and L. Lu, "Why is image quality assessment so difficult?" IEEE International Conference on Acoustics, Speech, & Signal Processing, (2002).
[2] S. Li and K. N. Ngan, "Influence of the Smooth Region on the Structural Similarity Index," proceeding of Pacific Rim Conference on Multimedia, Bangkok, Thailand, (2009).
[3] LIVE image quality assessment database: http://live. ece.utexas.edu/research/quality/.
[4] IRRCyN/IVC database: http://www.irccyn.ecnantes.fr/ivcdb/.
[5] Toyama Image quality evaluation database: http:// mict.eng.u-toyama.ac.jp/mict/index2.html.
[6] Tampere image database TID2008: http://www. ponomarenko.info/ tid2008.htm
[7] A57 database: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html
[8] W. Lin, "Computational models for just-noticeable difference," Digital Video Image Quality and Perceptual Coding, H. R. Wu and K. R. Rao, Eds. FL: CRC, (2005).
[9] A. B. Watson, "DCT quantization matrices visually optimized for individual images," Human Vision, Visual Proc., and Digital Display IV, Proc. SPIE, 1913: 202-216, (1993).
[10] Z. Wang and A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., 13 (4): 600-612, (2004).
[11] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Process., 15 (11): 3411-3452, (2006).
[12] D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," IEEE Trans. Image Process., 16(9): 2284-2298, (2007).
[13] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-07, Scottsdale, Arizona, USA, 25-26 (2007).
[14] N. Ponomarenko, F. Battisti, K. Egiazarian, et. al. "Metrics performance comparison for color image database", Fourth international workshop on video processing and quality metrics for consumer electronics, Scottsdale, Arizona, USA. Jan. 14-16, (2009).
[15] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain," IEEE Trans. Circuit Syst. Video Tech., 19 (3): 337-346, (2009).