On Monotonicity of Image Quality Metrics

Guangtao Zhai and Xiaolin Wu
ECE Department, McMaster University, Ontario, Canada, L8S 4K1
Email: xwu@ece.mcmaster.ca

***Abstract*—Perceptual image quality assessment (IQA) is an important research topic of visual signal processing both in its own right and for its utility in designing various optimal image processing and coding algorithms. This work is concerned with an issue that has been largely overlooked by the research community of IQA, that is, the monotonicity, or lack of it, between the subjective scores and the predictions of image quality metrics (IQM) for images with compression artifacts. We analyze the data of several well-known databases for IQA and expose among them a large number of instances of non-monotonicity between subjective and objective quality scores. Further, we observe that a nonlinear dynamical model of 3D cusp catastrophe can well explain the intricate relationship between the subjective and objective quality scores. Our findings identify an inherent flaw of current signal-distance or fidelity-based IQMs, which neglect the psycho-physiological aspect of human visual perception. This research suggests a new direction of IQA research and it also sheds light on the design of subjective quality evaluation process.**

***Index Terms*—image quality assessment; visual communication; catastrophe theory; nonlinear dynamical system**

I. Introduction

In digital visual communication, bits should be spent to maximize perceptual quality of the image/video delivered to the intended viewer. A subjectively meaningful quality metric of images/videos is a prerequisite for optimal design of visual signal compression and communication systems. Existing visual communication systems are designed under the mean square error criterion (i.e., PSNR). But the perceptual validity of PSNR has long been questioned. Numerous alternative image quality metrics (IQMs) have been proposed for improved measurement of subjective visual quality. [1]–[3]. In 1974, the international telecommunication union (ITU) made a series of recommendations ITU-R BT.500 [4] on the methodologies for subjective quality assessment of television pictures. And following those recommendations, open subjective quality databases [5]–[7] were made available to facilitate the research of image quality assessment (IQA). Although IQMs have improved over the years in terms of increased correlation between objective and subjective quality scores [8], IQA remains a very difficult and challenging research area.

In our view, a minimum requirement of an IQM is the following monotonicity. That is, the score of the IQM has to be a monotonic function of the subjective opinion score given by human viewers. However, we have discovered that all IQMs in the literature fail to satisfy the above monotonicity over a large number of instances drawn from the well-known IQA databases. Fig. 1 plots subjective scores vs. IQMs predictions for test image 'woman' (shown in Fig. 2) undergone JPEG compression and non-eccentric patter distortion. The distorted images and their subjective scores are taken from the widely used LIVE [6] and TID2008 [7] databases. Note that subjective scores can be in the form of either opinion score (MOS) or difference mean opinion score (DMOS). Three IQMs, PSNR, SSIM [2] and VIF [3], are examined. The data points in Fig. 1 are linked according to the rank order of MOS/DMOS scores, and they show that the IQM scores are not monotone in MOS/DMOS.

In fact, the monotonicity has been regarded an important performance measure of the IQMs by the video quality expert group (VQEG) [9]. VQEG suggested using a 3-parameter logistic function $\mathrm{S}_p = a_1 / \left(1 + e^{-a_2(\mathrm{S}-a_3)}\right)$ to map the IQM score S to predicted scores S_p so as to account for the non-uniformity between objective and subjective scores [10]. After the nonlinear mapping, the Pearson linear correlation and the Spearman rank order correlation are computed between S and S_p respectively, as measurements of prediction *accuracy* and *monotonicity* of the IQMs. In the example in Fig. 1, the Pearson correlation for the IQMs are around 0.95 while the Spearman's rho or Kendall's tau correlations are significantly lower ($0.6 \sim 0.8$) due the obvious non-monotonicity. The example here is by no means exceptional, for the JPEG data sets of LIVE database [6], non-monotonicity between DMOS and predictions of PSNR, SSIM and VIF is found on almost 80% of the test images. And for the JPEG2000 data set of Toyama database [5], roughly 64% of the images exhibit non-monotonicity between MOS and PSNR,SSIM and VIF scores.

The fluctuation of performances of IQMs in terms of the Spearman rank order correlation was noticed by Sheikh and Bovic [8] and they conjectured that sometimes "the nature of the data is such that Spearman rank order correlation is not a good measure of IQM performances", especially "when image quality is perceptually the same for many distortion strengths". This paper shows that the complicated relationship between subjective scores and IQM scores can be satisfactorily fit by a nonlinear dynamical model, specifically, a cusp catastrophe. The non-monotonicity corresponds to the bifurcation behavior of the cusp catastrophe. As a consequence, the non-monotonicity, being an inherent feature of the cusp catastrophe, cannot be modeled by any of the existing IQA methods.

The rest of this paper is organized as follows: Section II gives a brief introduction of the catastrophe theory and the cusp catastrophe model. The cusp perceptual quality model and the choice of control parameters are discussed in Section III. Concluding remarks and directions for future research are given in Section IV.

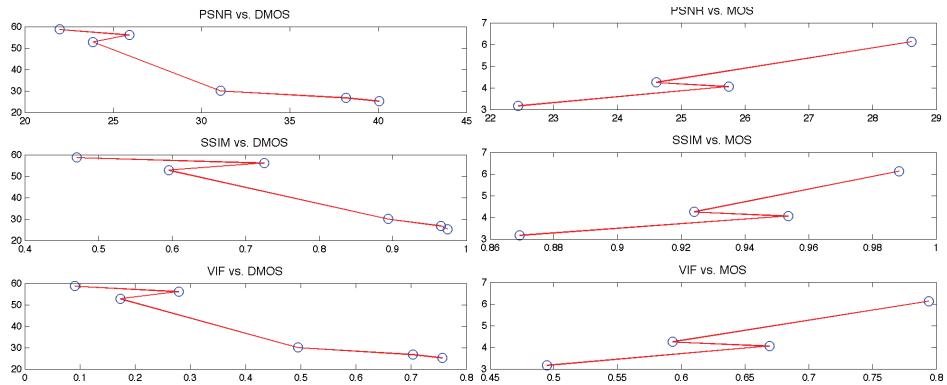

Fig. 1. DMOS/MOS vs. PSNR, SSIM and VIF for test image 'woman' from JPEG compression set of LIVE database [6] (left) and from non-eccentric pattern artifact set of TID2008 database [7] (right).

Fig. 2. Test images from the Kodak database, from left to right: 'woman', free energy = $4.96bpp$; 'monarch', free energy = $3.52bpp$; 'stream', free energy = $5.82bpp$.

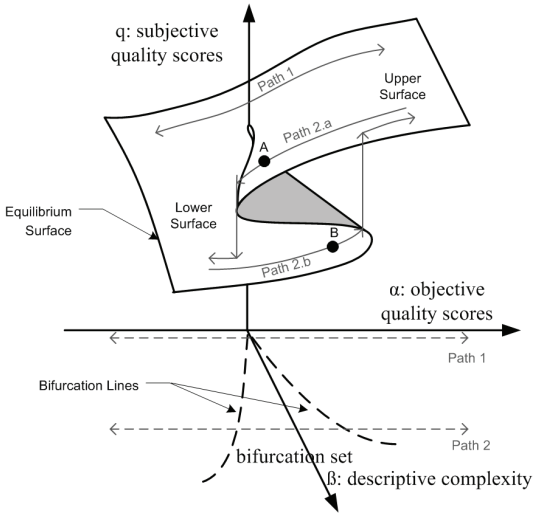

Fig. 3. The cusp catastrophe model for image quality assessment.

II. The Catastrophe Model

Consider a gradient dynamical system

$$\frac{\partial \mathbf{x}}{\partial t} = -\frac{\partial V(\mathbf{x};\mathbf{c})}{\partial \mathbf{x}}, \mathbf{x} \in \Re^p, \mathbf{c} \in \Re^k \tag{1}$$

where $\mathbf{x}$ is the state variable and $\mathbf{c}$ is the control variable of the system. A basic premise of (1) is that the system under study is driven toward some equilibrium state. In other words, if the system is at some non-equilibrium state, as time changes, the potential function $V(\mathbf{x};\mathbf{c})$ tends to be minimized with respect to system state $\mathbf{x}$, and the system will return to states $\partial V(\mathbf{x};\mathbf{c})/\partial \mathbf{x} = 0$. And the stable state is also called the equilibrium of the system. Obviously, the equilibrium corresponds to both the maxima and minima of the potential function $V(\mathbf{x};\mathbf{c})$. When the potential is at a local minimum, the system is said to be in a stable state. When the potential is at a local maximum, the system is in an unstable state and given any external perturbation, it will return to a stable state due to the mechanism of (1). Since virtually no practical system is totally isolated from the rest of the world, most of the gradient dynamical systems we encounter in the real world are in stable states. The states where the Hessian matrix of $V(\mathbf{x};\mathbf{c})$ has zero eigenvalues are called the degenerate equilibrium points. And at those degenerate points, the system has bifurcation behavior as control variables change.

If we pose the problem of perceptual quality assessment as the dynamical system in (1), the system can be simplified as

$$\frac{\partial q}{\partial t} = -\frac{\partial V(q;\mathbf{c})}{\partial q}, \mathbf{c} \in \Re^k \tag{2}$$

where q is a scalar quantifying the subjective quality of an image, whose various statistical and psychovisual features are packed in the control vector $\mathbf{c}$. The dynamics of the system (2) suggests that the subjective judgement of image quality are related to finding extrema of the potential function $V(q;\mathbf{c})$.

Catastrophe theory, being a branch of bifurcation theory and singularity theory, studies and classifies the sudden change of behaviors in the neighborhood of degenerate points of the potential function for the dynamical system. According to the study of Thom [11], if the potential function has two or fewer state variables and four or fewer control variables, then the system can be characterized by only seven generic forms (universal unfoldings) which are named elementary catastrophes. In many disciplines of natural and social science, e.g. physics, biology, economics, psychology, and etc, the most widely used catastrophe model is the so-called cusp model as it is the simplest catastrophe model capable of transiting between equilibrium states. The potential function of a cusp

catastrophe is

$$-V(q;\alpha,\beta) = \alpha q + \frac{1}{2}\beta q^2 - \frac{1}{4}q^4 \tag{3}$$

where α and β are two control variables and the coordinates (α, β) form a control plane for the system. The equilibrium state of the potential function (3) is given by

$$\frac{\partial V(q;\alpha,\beta)}{\partial q} = \alpha + \beta q - q^3 = 0. \tag{4}$$

The number of solutions of the cubic function in (4) is determined by the Cardano's discriminant $\delta = 27\alpha^2 - 4\beta^3$: if $\delta < 0$ then the function has three solutions; and if $\delta \geq 0$, the function has one solution. And on the control plane, the set of points (α, β) satisfying $27\alpha^2 - 4\beta^3 = 0$ constitutes the bifurcation lines and the set of points (α, β) satisfying $27\alpha^2 - 4\beta^3 < 0$ forms the bifurcation set. A graphical demonstration of the cusp catastrophe is given in Fig. 3.

III. The Cusp Perceptual Quality Model

To model the relationship between subjective and IQMs scores using the the cusp catastrophe model, we have to specify the control parameters α, β first. In Fig. 3, it is noticed that for a given β the state parameter q traverses from lower surface to upper surface as α increases. Higher α value generally indicates better subjective quality. We define the control variable α as a function of the IQM scores, e.g., $\alpha = b_1 S + b_2$. Using a function of IQM score rather than the score itself reconciles a technicality: parameter α in the cusp catastrophe can be negative but most IQM scores are by definition non-negative.

However, the choice of control parameter β needs more thought because of the complex relationship between q and β given α. In fact, the tendency of q versus β is affected by the relationship between α and β. Considering that α and β are orthogonal in the control plane, ideally they should carry mutually exclusive information on the image. As α is defined as a function of IQM score assigned to a degraded image, β should relate to some intrinsic properties of the original image. We empirically discover that the cusp catastrophe model can satisfactorily fit the non-monotonicity instances in existing IQA databases such as LIVE, Toyama and TID2008, if β is chosen as a linear function of the descriptive complexity of the image. The IQA databases show that images of simple contents tend to satisfy the monotonicity between MOS/DMOS and IQM scores, whereas for images of higher complexity, the monotonicity is more likely to be violated. For example, Fig. 4 plots DMOS vs. PSNR, SSIM and VIF for JEPG2000 compressed images 'monarch' (simple type) and 'stream' (complex type) (the originals shown in Fig. 2).

To measure the descriptive complexity of the image, we much question into the neurological and psychophysical mechanism of human vision. Particularly, we resort to the principle of free energy recently proposed by Friston *et. al* [12] for a understanding and approximation of the process of visual perception. The free energy principle was proposed to unify the brain theories in biological and physical sciences about human action, perception and learning. Generally speaking, the rationale behind free energy principle is that all adaptive biological agents resist the natural tendency to disorder in an ever-changing environment, as predicted by the second law of thermodynamics. Therefore, the free energy principle suggests that biological agents can somehow violate the second law of thermodynamics by keeping their internal states at low entropy level so as to maintain themselves within some physiological bounds. And this goal is realized through avoiding encountering 'surprise' under different environments. Although the 'surprise' cannot be measured or avoided by a biological agent directly, it can be upper bounded by a term called 'free energy'. As such, the minimization of free energy implicitly minimizes 'surprise'. The generative model exclusively defines the system and the quality of the free energy bound on 'surprise'. Models with higher descriptive power tend to explain the inputs better and keep the free energy (and therefore 'surprise') lower. More importantly, the free energy can be evaluated by a biological agent using its internal (generative) model and external (sensory) states. Given fixed inputs, the minimization is essentially a process of fitting the internal generative model to the external sensory states. This process echoes the Bayesian brain hypothesis [13], which is a belief that our brain works with uncertainties using optimal rules as studied in Bayesian statistics. Therefore, the visual perception is an inference process of the brain that actively predicts and explains sensations using internal generative models.

For visual perception, the free energy principle and Bayesian brain hypothesis both lead to the conjecture that brain has an internal generative model for the scenes we lay our eyesight on. The free energy principle suggests that the brain always seeks the most 'logical' explanation of each given scene by tuning its internal generative models. The gap between the external input and its generative-model-explainable part should therefore be related to the complexity of the given scene. In other words, the psychovisual complexity of an image can be explicitly defined as the agreement between the scene itself and the output of the internal generative model that best describes the scene. And the image descriptive complexity can be mathematically quantified by the uncertainty of the residuals between the image and its predicted version by the internal generative model. Mathematical formulations of 'surprise' and 'free energy' for visual perception can be found in [14], where variational approximation of the internal generative model with an linear model is also provided. And the free energy values for the test images 'woman', 'monarch' and 'stream' are also provided in Fig. 2, where 'stream' with much details of rocks and bushes has the highest free energy value, whereas 'monarch' with largely blurred background has the lowest free energy value. The average free energy is 4.84 for test images in the JPEG subset of the LIVE database with non-monotonicity between subjective and IQM scores of PSNR, SSIM and VIF. While for images without the non-monotonicity in the same data set, the average free energy is only 3.43.

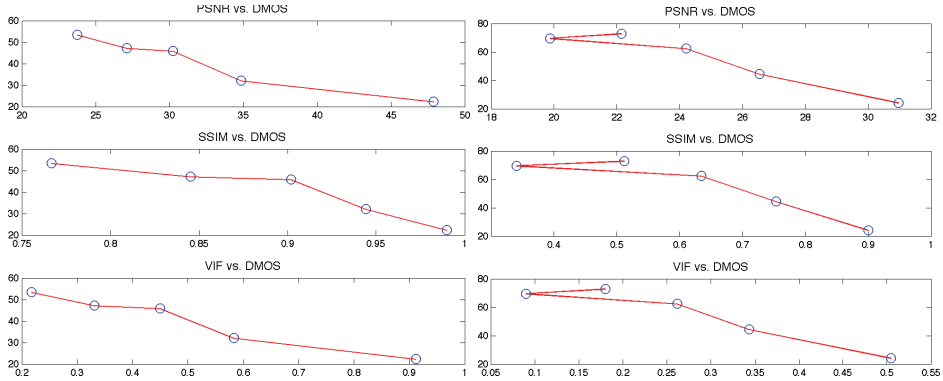

Fig. 4. DMOS/MOS vs. PSNR, SSIM and VIF for test image 'monach' (left) and 'stream' (right) in JPEG2000 compression set from LIVE database [6].

With the cusp catastrophe model for IQA, as illustrated in Fig. 3, the discussed monotonic or non-monotonic relationships between subjective and objective quality scores can be well explained. For images of lower complexity, the IQM scores are near Path 1 in Fig. 3, and the subjective scores tend to change monotonically with IQM scores. For images of higher complexity, the IQM scores approximately lie on Path 2. On Path 2 there exist data points such that images of higher (lower) IQM scores counterintuitively have lower (higher) subjective scores (see point A on Path 2.a and point B on Path 2.b in Fig. 3). This phenomenon is due to the fact that given an image, its MOS or DMOS value assigned by a subject depends on the order in the set of test images by which this image is presented to him/her. In the subjective viewing test, for images of high free energy, if a subject is asked to assess the quality of image instances with decreasing objective quality (e.g., with decreasing PSNR values) starting from a high initial value, then the resulting points (α, β, q) move along Path 2.a; on the contrary, if the subject is shown image instances with increasing objective quality, the collected data points move along Path 2.b.

Unfortunately, neither ITU nor VQEG has any suggestions on a proper order of objective quality of image instances used in the subjective viewing test. Most of the tests that resulted existing IQA databases were conducted with random orders of presentation. This causes the above mentioned non-monotonicity between subjective and objective scores. Since this non-monotonicity is inherent to existing IQA databases, the performances of IQMs cannot be accurately quantified in terms of rank order correlation as suggested by VQEG. And the cusp catastrophe based quality model also explains the fluctuations of the Spearman rank order performances of IQMs across different images and IQA databases as noticed in [8].

IV. Conclusion

Non-monotonicity between subjective and objective scores of popular databases for image quality assessment research, although being counterintuitive, occurs with significant probability in some popular databases for image quality assessment research. This intriguing phenomenon can be explained by a cusp catastrophe model. This research opens up a new line of enquiry into the design of fair subjective viewing tests, and it also highlights the importance of accounting for nonlinear psycho-physical properties of the human visual system when designing subjectively meaningful image quality metrics.

Acknowledgment

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

References

[1] S. Winkler, *Digital Video Quality - Vision Model and Metrics.* John Wiley & Sons, January 2005.

[2] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[3] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[4] ITU, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation, International Telecommunication Union/ITU Radiocommunication Sector*, 2009.

[5] Z. M. P. Sazzad, Y. kawayoke, and Y. Horita, "Image quality evaluation database," 2000. [Online]. Available: http://mict.eng.u-toyama.ac.jp/database_toyama/

[6] H. Sheikh, Z.Wang, L. Cormack, and A. Bovik, "Live image quality assessment database release 2." [Online]. Available: http://live.ece.utexas.edu/research/quality

[7] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, pp. 30–45, 2009.

[8] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440 –3451, nov. 2006.

[9] VQEG, "VQEG final report of FR-TV phase II validation test," 2003. [Online]. Available: www.vqeg.org

[10] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press, 1991.

[11] R. Thom and D. Fowler, *Structural Stability and Morphogenesis: An Outline of a General Theory of Models*. W.A. Benjamin, Michigan, 1975.

[12] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, pp. 127–138, 2010.

[13] D. C. Knill and A. Pouget, "The bayesian brain: the role of uncertainty in neural coding and computation," *Trends in Neurosciences*, vol. 27, pp. 712–719, 2004.

[14] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free energy principle," *IEEE Transactions on Image Processing, in press*, 2011.