VISUAL SALIENCY DETECTION VIA IMAGE COMPLEXITY FEATURE

Min Liu[†‡]*, Ke Gu*[§]*, Guangtao Zhai*[†]*, and Patrick Le Callet*[‡]

[†]Insti. of Image Commu. & Infor. Proce., Shanghai Jiao Tong University, China
[‡]Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, France
[§]School of Computer Engineering, Nanyang Technological University, Singapore

ABSTRACT

In this paper we propose a novel bottom-up visual saliency detection model by analysis of image complexity. Compared with existing works, we emphasize the important impact of image complexity on saliency detection. Inspired by the free energy theory, a hybrid parametric and non-parametric model is used to estimate the complexity of a visual signal. Taking the image complexity as a new feature, this paper constructs a heuristic framework to systematically combine two different types of saliency detection models, separately using local and global features, in order to predict human fixation points more accurately. In contrast to classical and modern models, our algorithm has achieved noticeably superior results. And furthermore, it is worthy to stress that the proposed saliency detection method can also help to facilitate the performance of image quality metrics on popular image databases.

Index Terms— Saliency detection, image complexity, free energy, image quality assessment

1. INTRODUCTION

By removal of redundant information, saliency detection plays a critical role in promoting various kinds of image processing and computer vision applications [1]. For instance, as for image quality assessment (IQA), the distortions appeared in salient areas are easily arousing attentions and thus severally degrade the visual quality of an image. To this aim, several eye-tracking experiments are conducted to obtain the real visual saliency map for weighting, and gained promising results compared with the case without considering visual saliency [2, 3]. In the pursuit of efficacy and efficiency, many IQA approaches have incorporated computed visual saliency algorithms for similarity measure or weighting [4-6]. In addition, visual saliency also helps to facilitate the study of automatic image contrast enhancement algorithm [7-9]. So a fast and faithful computational model is highly desired for visual saliency detection in the encountered scene.

The last 25 years have witnessed the emergence of over hundreds saliency detection techniques [10]. This number is estimated to be continually growing in the future. In terms of different attentional mechanisms, existing methods can roughly fall into two types. The first one is top-down task-dependent methods requiring prior knowledge about visual content. The second one is bottom-up stimulus-driven methods relying on the self information of the image itself.

In this paper we focus our attention on exploring bottom-up models. Many of this type of techniques were developed to seek for locations by maximizing local saliency with biologically motivated local features [11-15]. These features, which mainly consist of intensity, edge, texture, color and orientation, are inspired by neural responses in lateral geniculate nucleus and V1 cortex. Part of other technologies rely on global features [16-19], in which salient areas are detected from a visual signal in transform domains. On the basis of global considerations, this type of methods have the advantages of quickly and precisely finding visual "pop-outs" by locating possible salient objects. Very recently, it was found that there exist some limitations in only local or global features for visual saliency detection. As thus, an increasing number of researches have concentrated on taking account of both two types of features [20-22]. The majority of them were developed based on complementary strategies, thereby obtaining remarkable high performance.

In spite of numerous saliency detection models explored during the last decades, very few efforts were applied to investigate the influence of image complexity. In fact, image complexity was found to possess strong impact on the visual saliency detection techniques. Simply speaking, for comparatively low-complexity images of obviously separated foreground objects and background regions, their salient areas can be easily detected with properly downsampling before local comparisons. On the contrary, for comparatively high-complexity images that do not have definite foreground and background layers, global-consideration-based saliency detection models are more appropriate for predicting the salient areas. Enlightened from this, we in this paper propose a novel complexity-weighted saliency detection model, dubbed as CWS, which uses image complexity feature as a weighting to systematically fuse local and global features.

Compared with existing works for visual saliency detection, this paper has made four major contributions as given below: 1) to the best of our knowledge, this work is the first one considering image complexity features into building the

saliency detection model; 2) our saliency detection technique has achieved the best fixation prediction performance to date; 3); the proposed strategy is general to be extended to merging different types (global and local) of saliency detection methods towards better performance; 4) our model can be also inserted into our recent IQA framework leading to the currently highest performance with human subjective ratings.

The remainder of this paper are organized as follows: Section 2 details the CWS saliency detection technique. In Section 3, comparative tests validate the superiority of our CWS algorithm over classical/state-of-the-art relevant models. In Section 4, a new CWS-based IQA metric is calibrated and its performance is conducted on the popular image databases to prove its higher accuracy than state-of-the-art competitors. Finally, we conclude the whole paper in Section 5.

2. SALIENCY DETECTION MODEL

Most of existing saliency detection techniques are based on the features of intensity, edge, texture, color, orientation, etc. However, we find that image complexity, as an essential attribute of visual signals, has a vital influence on visual saliency detection. As for comparatively low-complexity images, they usually have definitely separated foreground objects and background regions. Therefore we are able to search for their saliency maps through first downsampling the input image signal properly and then conducting local comparisons. Conversely, it is not easy to clearly distinguish foreground and background layers in images of comparatively high complexity. In this regard, it was found that using global-based strategy is more suitable for detecting salient regions. The proposed CWS technique is established based on this inspiration, namely computing image complexity feature to weight and combine local and global features.

Given an input visual signal, we first introduce how to estimate image complexity. According to our previous works [23-25], the free energy based brain theory is proved to have a very close relation to human sensation of visual saliency and quality. As a consequence, the free energy principle is used for measuring the image complexity. To be more concretely, the free energy principle can be regarded as a unified brain theory combining popular brain theories in biological and physical sciences [26]. Essentially, this principle makes a basic premise that the human cognitive process is controlled by a so-called internal generative model in the brain. Using this internal generative model, the brain is capable of analyzing and predicting the given scene in a constructive way. In other ways, we can also consider this way to be a probabilistic model consisting of a likelihood term and a prior term. In the following, the HVS is to interpret the posterior possibilities of the given scene. It is obvious that, since the generative model is unable to work well everywhere, there must be a "surprise" gap of the input image and the brain's explanation, which we suppose is closely connected to image complexity.

In reality, free energy measures the difference (the error map) between an image and its output best prediction inferred by the internal generative model. In the computed error map, high-value pixels stand for what cannot be well explained by the generative model, namely "surprise". On the opposite, low-value pixels stand for what can be easily predicted. We can obtain this error map via free-energy minimization. As analyzed in [27], minimizing free energy is quite similar to the predictive coding, and it can be as the entropy of the residual error map of the input image and its explained one.

Though the autoregressive model (AR) is simple and can simulate a wide range of natural scenes [23, 24], it is not always stable at image edges. Another commonly used operator, the non-parametric bilateral filter (BL), is also introduced for the sake of its good edge-preserving ability. So we assign the internal generative model to be a hybrid parametric and non-parametric model (HPNP) through integrating the parametric AR model and the non-parametric BL filter.

To be more specific, we conduct the AR operator in each proper-size local patch, as defined to be

$$y_i = \Gamma_\theta(y_i)\mathbf{u} + \hat{e}_i \tag{1}$$

where y_i represents the pixel value at the location x_i of the input image I; $\Gamma_\theta(y_i)$ defines the θ member neighborhood vector, $\mathbf{u} = (u_1, u_2, ..., u_\theta)^T$ is a vector of AR parameters; $\hat{e}_i$ indicates a difference term between input pixel values and output predictions. In order to determine $\mathbf{u}$, we can express the linear system in a matrix way as

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\mathbf{y} - \mathbf{Y}\mathbf{u}\|_2 \tag{2}$$

where $\mathbf{y} = (y_1, y_2, ..., y_\theta)^T$; $\mathbf{Y}(i,:) = \Gamma_\theta(y_i)$. This linear system can be solved with the least square method as $\hat{\mathbf{u}} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{y}$. The BL filter is a classical non-linear model and we can easily construct and compute it [28]. The BL filter can be expressed by

$$y_i = \Gamma_\theta(y_i)\mathbf{v} + \tilde{e}_i \tag{3}$$

where $\mathbf{v} = (v_1, v_2, ..., v_\theta)^T$ is a vector of BL filter coefficients; $\tilde{e}_i$ is an error term. The vector $\mathbf{v}$ consists of two controlling factors: (i) the spatial Euclidean distance between x_i and x_j; (ii) the photometric distance between y_i and y_j. On this basis, BL filter coefficients can be defined as

$$v_j = \exp\Big(\frac{-\|x_i - x_j\|^2}{2\sigma_x^2} + \frac{-(y_i - y_j)^2}{2\sigma_y^2}\Big) \tag{4}$$

where σ_x and σ_y are fixed as constants for balancing the relative importance of the spatial and photometric distances.

The error map of the input visual signal and the output best explanation is obtained by integrating the AR model and BL filter. The pixel value $\tilde{y}_i$ at the location x_i in the error map can be computed by

$$\tilde{y}_i = \frac{\Gamma_\theta(y_i)\hat{\mathbf{u}} + w\Gamma_\theta(y_i)\mathbf{v}}{1 + w} \tag{5}$$

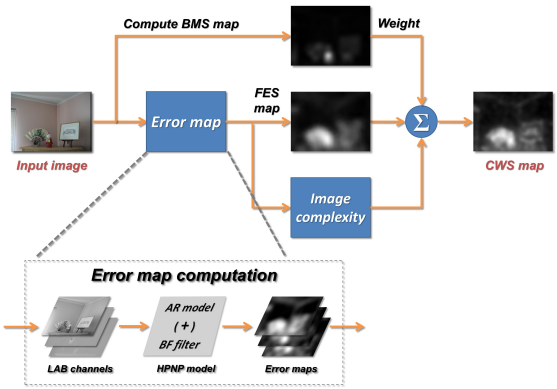


Fig. 1. The framework of our saliency detection model.

where w is a fixed positive number to adjust the relative importance of above two components. The resultant image complexity can be estimated as the entropy of the error map

$$I_c = -\int_i p(i)\log p(i)di \tag{6}$$

where $p(i)$ indicates the probability density of grayscale i in the error map.

After computing the image complexity, it will be used as weights to combine local and global features for visual saliency detection. Note that, on one hand, the proposed model is an heuristic general framework, not merely for some features, and on the other hand, we wish the interested readers are able to easily copy our idea into their researches towards better fixation prediction performance. Therefore, we instead choose the recently proposed two models, namely local-based FES [15] and global-based BMS [19], for the sake of their attributes and high accuracy.

The proposed CWS saliency detection technique employs image complexity to combine BMS and FES models, as depicted in Fig. 1. Specifically, assuming an input visual signal I, we first separately generate saliency maps via BMS and FES strategies. Then we estimate its image complexity I_c to discriminate high- and low-complexity images for discrimination before choose the better saliency detection scheme:

$$S_{cws} = \frac{S_{fes} + \gamma S_{bms}}{1+\gamma} \tag{7}$$

where ξ_{thr} is a fixed positive constant; S_{fes} and S_{bms} stand for saliency maps detected using FES and BMS models; γ is computed by

$$\gamma = \phi^{F_s(I_c - \xi_{thr})} \tag{8}$$

where F_s is used to obtain the sign; ϕ is a constant base number greater than 1. In this case, γ is a continuously changing parameter towards better integrating two different types of saliency techniques.

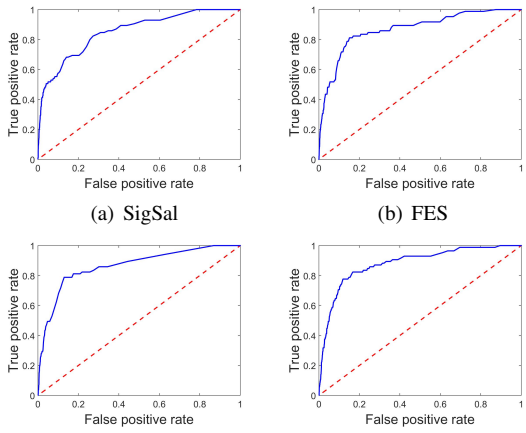
3. EXPERIMENTAL RESULTS

The fixation prediction performance of the proposed FES saliency detection technique is validated using three popular databases including Toronto dataset [14], FIFA dataset [13] and MIT dataset [20]. We elaborately choose up to nine algorithms, which are composed of classical Itti [11], AIM [14], Judd [20] models, and prevailing SigSal [17], HFT [18], LG [21], CAS [22], FES [15], BMS [19] models.

The quantified comparison is conducted against existing relevant technologies. For each image, the shuffled ROC Area Under the Curve (sAUC) score is first computed to testify the consistency between a particular saliency map and a group of fixations, as used in recent works [15, 17, 19]. Note that, as for sAUS, the chance level is 0.5 while the perfect prediction is 1.0. We take an example in Fig. 2, in which the blue curve shows the ROC curve computed on the sample image in Fig. 1 with our CWS model and three modern techniques, while the red diagonal dash line indicates the chance level. Results verified that the area under the blue curve of each of popular SigSal, FES and BMS models has achieved very promising sAUS scores, respectively 0.8517, 0.8693 and 0.8639. By comparison with the proposed CWS algorithm, however, the sAUC score of our model is up to 0.8804, the highest one across the all and remarkably superior to competitors.

We furthermore compute and illustrate the sAUC scores of the nine testing saliency detection models and the proposed CWS technique on three eye tracking datasets and their direct averages in Table 1. In the first classical Toronto dataset, our approach achieves the top performance across all testing algorithms, with a gain over the second FES and BMS of larger than 2%. In the second FIFA dataset concentrating on the face detection, our CWS method is also of noticeable superiority to other testing ones. The performance gain is about 0.8% relative to the second-performer FES and 3.4% relative to the third-performer BMS. The results may tell one merit of our approach regarding its good ability to help promote the scientific research and practical application of face-related technologies. In the last large-scale MIT dataset, our technique still obtains the first place, exceeds the second-place BMS with a gain about 1.6%. Finally, we compute and compare the average performance across all testing methods, and the results demonstrate the superiority of our CWS approach over state-of-the-art competitors.

One important point should be laid stress on at last. By introducing a novel feature – image complexity, we have proposed a general framework to properly combine image complexity, local and global features for visual saliency. It is natural that the proposed CWS saliency detection technique is able to perform the best across all the testing relevant methods in fixation prediction. A successful technology had better be inserted into some applications and bring remarkable performance gains. In the next section, we will illustrate how to apply our CWS to derive a high-accuracy IQA metric.



(a) SigSal (b) FES (c) BMS (d) CWS

Fig. 2. The blue curve shows the ROC curve of four models on the sample image given in Fig. 1, with the red reference dash line indicating the chance level of 0.5. The area under the blue curve of SigSal, FES and BMS models are 0.8517, 0.8693 and 0.8639. In contrast, that of our CWS technique is up to the highest 0.8804.

Table 1. Comparison of Shuffled-AUC scores.

Models	Toronto	FIFA	MIT	Average
Itti [11]	0.6565	0.6727	0.6424	0.6572
AIM [14]	0.6822	0.7234	0.6674	0.6910
Judd [20]	0.6838	0.7083	0.6588	0.6836
SigSal [17]	0.7053	0.7281	0.6682	0.7005
HFT [18]	0.6897	0.6987	0.6513	0.6799
LG [21]	0.6880	0.6737	0.6717	0.6778
CAS [22]	0.6919	0.7104	0.6681	0.6902
FES [15]	0.7197	0.7541	0.6871	0.7203
BMS [19]	0.7197	0.7352	0.6915	0.7155
CWS (Pro.)	0.7305	0.7601	0.7023	0.7310

Table 2. Performance comparison of distinct IQA methods.

SRCC	LIVE	TID2008	CSIQ	TID2013	Average
GSI [34]	0.956	0.850	0.910	0.794	0.878
IGM [35]	0.958	0.890	0.940	0.809	0.900
GMSD [36]	0.960	0.890	0.957	0.804	0.903
LTG [29]	0.958	0.906	0.959	0.883	0.927
Proposed	0.965	0.910	0.963	0.887	0.931

4. APPLICATION TO QUALITY ASSESSMENT

It has been argued that visual saliency is the product of brain activity and a good relevant model should simulate the brain process of perceiving visual signals, e.g. for saliency detection or quality assessment. In reality, on the one hand visual saliency and quality have shared several application scenarios, such as image/video coding and tone mapping, and on the other hand they still strongly interact each other. For example, distortions in salient areas wield more influence on the overall image quality score than those in other areas and this has been widely used in dozens of IQA methods [2, 3, 4]. Taking account of the close connection between visual saliency and perceived quality, in this paper we insert the proposed CWS saliency detection model into a IQA framework given in [29], in order to contribute a new more effective quality metric and simultaneously to indirectly demonstrate that the effectiveness of our CWS technique for the tasks of saliency detection and quality assessment.

In our recent work [29], a very fast and faithful IQA metric, dubbed as LTG, has been proposed and its extremely high performance has been fully validated on popular four large-scale image databases. The LTG model is composed of three components, which separately consider the global degradation, particular degradation with intensity distributed non-uniformly and chrominance information distortion. Details about the three terms (G_g, G_l and $P_m Q_m$) can be found in [29]. The influence of visual saliency on the visual quality is, however, overlooked in our previous LTG algorithm. To address this issue, the LTG model is improved by incorporating the saliency maps computed using CWS on the reference and distorted images to be compared as the fourth component. To specify, we consider using the proposed CWS model to compute the saliency maps of the reference and distorted images, denoted by S_r and S_d, and define the fourth component as the similarity of S_r and S_d, $S = \frac{\min(S_r,S_d)+C}{\max(S_r,S_d)+C}$, where C is a fixed constant for increasing the stability. We finally combine the above-mentioned four components together, $QS = \frac{G_l}{G_g} \cdot S \cdot P_m \cdot Q_m$, leading to our new IQA metric.

The new IQA metric is verified using the four large-scale databases, LIVE [30], TID2008 [31], CSIQ [32] and TID2013 [33]. We select recently designed GSI [34], IGM [35], GMSD [36] and LTG [29] for comparison. As suggested by the VQEG [37], we first map the objective predictions of each testing IQA technique to subjective scores using nonlinear regression with a five-parameter logistic function. Then, the most representative performance measure, SROCC, is used in this work. As given in Table 2, the indices on each database and the direct average across all the four databases are provided. It can be found that the proposed IQA metric has achieved very promising performance. On each testing database and on average, our IQA approach constantly performs the best.

5. CONCLUSION

In this paper we have proposed a new computational visual saliency detection technique. By introducing a new important feature of image complexity, we provide a simple general framework to systematically combine global and local feature for visual saliency functions. Experimental results on popular datasets demonstrate the superiority of our CWS saliency detection model for fixation prediction tasks. Furthermore, the proposed CWS model also can be effectively embedded into a recent IQA method towards the presently highest performance in terms of human subjective ratings.

6. REFERENCES

[1] P. Le Callet and E. Niebur, "Visual attention and applications in multimedia technologies," *Proc. IEEE*, 2013.

[2] O. Meur *et al.*, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *SPIC*, 2010.

[3] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *TCSVT*, 2011.

[4] K. Gu, G. Zhai, X. Yang, L. Chen, and W. Zhang, "Nonlinear additive model based saliency map weighting strategy for image quality assessment," *in Proc. IEEE Workshop on Multimedia Sig. Process.*, pp. 313-318, Sept. 2012.

[5] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *TIP*, 2011.

[6] X. Min, G. Zhai, Z. Gao, and K. Gu, "Visual attention data for image quality assessment databases," *in Proc. IEEE Int. Symp. Circuits and Systems*, pp. 894-897, Jun. 2014.

[7] K. Gu, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "Automatic contrast enhancement technology with saliency preservation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 9, pp. 1480-1494, Sept. 2015.

[8] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement," *IEEE Trans. Cybernetics*, vol. 46, no. 1, pp. 284-297, Jan. 2016.

[9] S. Wang, K. Gu, S. Ma, W. Lin, X. Liu, and W. Gao, "Guided image contrast enhancement based on retrieved images in cloud," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 219-232, Feb. 2016.

[10] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *TPAMI*, 2013.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *TPAMI*, 1998.

[12] O. L. Meur *et al.*, "A coherent computational approach to model bottom-up visual attention," *TPAMI*, 2006.

[13] M. Cerf *et al.*, "Predicting human gaze using low-level saliency combined with face detection," *NIPS*, vol. 20, 2008.

[14] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vision*, 2009.

[15] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Visual saliency detection with free energy theory," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1552-1555, Oct. 2015.

[16] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *CVPR*, 2007.

[17] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *TPAMI*, 2012.

[18] J. Li *et al.*, "Visual saliency based on scale-space analysis in the frequency domain," *TPAMI*, 2013.

[19] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: A boolean map approach," *TPAMI*, 2016.

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *ICCV*, 2009.

[21] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," *CVPR*, 2012.

[22] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *TPAMI*, 2012.

[23] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50-63, Jan. 2015.

[24] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "No-reference image sharpness assessment in autoregressive parameter space," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3218-3231, Oct. 2015.

[25] K. Gu, G. Zhai, W. Lin, X. Yang, and W. Zhang, "Learning a blind quality evaluation engine of screen content images," *Neurocomputing*, vol. 196, pp. 140-149, Jul. 2016.

[26] K. Friston, "The free-energy principle: A unified brain theory?" *Nature Rev. Neuroscience*, vol. 11, pp. 127-138, 2010.

[27] H. Attias, "A variational bayesian framework for graphical models," *NIPS*, 2000.

[28] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *ICCV*, 1998.

[29] K. Gu, G. T. Zhai, X. K. Yang, and W. J. Zhang, "An efficient color image quality metric with local-tuned-global model," *in Proc. IEEE Int. Conf. Image Process.*, pp. 506-510, Oct. 2014.

[30] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "LIVE image quality assessment Database Release 2," [Online]. Available: http://live.ece.utexas.edu/research/quality

[31] N. Ponomarenko *et al.*, "TID2008-A database for evaluation of full-reference visual quality assessment metrics," *AMR*, 2009.

[32] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *JEI*, 2010. Online at: http://vision.okstate.edu/csiq

[33] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *SPIC*, 2015.

[34] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *TIP*, 2012.

[35] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism," *TIP*, 2013.

[36] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *TIP*, 2014.

[37] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Mar. 2000, http://www.vqeg.org/.