Seven Challenges for Image Quality Research

Damon M. Chandler, and Md Mushfiqul Alam, and Thien D. Phan

Laboratory of Computational Perception and Image Quality
School of Electrical and Computer Engineering
Oklahoma State University, Stillwater, OK 74078 USA

ABSTRACT

Image quality assessment has been a topic of recent intense research due to its usefulness in a wide variety of applications. Owing in large part to efforts within the HVEI community, image-quality research has particularly benefited from improved models of visual perception. However, over the last decade, research in image quality has largely shifted from the previous broader objective of gaining a better understanding of human vision, to the current limited objective of better fitting the available ground-truth data. In this paper, we discuss seven open challenges in image quality research. These challenges stem from lack of complete perceptual models for: natural images; suprathreshold distortions; interactions between distortions and images; images containing multiple and nontraditional distortions; and images containing enhancements. We also discuss challenges related to computational efficiency. The objective of this paper is not only to highlight the limitations in our current knowledge of image quality, but to also emphasize the need for additional fundamental research in quality perception.

Keywords: Image quality, quality assessment, visual masking, suprathreshold perception, visual distortion

1. INTRODUCTION

Digital images are subject to numerous forms of processing as they are captured, stored, transmitted, and ultimately displayed to the consumer. Because such processing can alter the image's appearance, there is a need to assess the impacts of the processing on the resulting visual quality. To meet this need, numerous algorithms for image quality assessment (IQA) have been researched and developed over the last several decades, with particularly explosive growth over the last 10 years. Today, IQA research has emerged as an active subdiscipline of image processing, and many of the resulting techniques and algorithms have proved useful for a variety of electronic imaging applications.[1–9]

Research on image quality assessment can be traced back to the early research on quality evaluation of optical systems and analog television broadcast/display systems.[10–21] This early work not only laid the foundation for our current understanding of the factors that affect quality—e.g., luminance and spatial resolution, contrast and color range, gradation, brilliance, flicker, noise—but it also stressed the need to take into account properties of the human visual system during the quality-assessment process.

Due in large part to research efforts within the HVEI community, IQA research has particularly benefited from improved models of visual perception. Such vision models have been instrumental for providing IQA algorithms both the ability to take into account key properties of visual perception, and the ability to take mimic the biological processing strategies which underlie the quality-assessment process. However, over the last decade, the focus of quality-assessment research in the signal-processing community has largely shifted from the previous broader objective of gaining a better understanding of the human visual system, to the current limited objective of better fitting the available ground-truth subjective data. The end result is that we have gained a substantial amount in the form of new algorithms, but significantly less in the form of new fundamental knowledge on how humans perceive artifacts in images.

In a recent review paper,[8] we provided a summary of past and present IQA algorithms and outlined seven challenges for future IQA research. In that paper, rather than present just another review of the IQA successes,

Further author information: (Send correspondence to D.M.C)
D.M.C: E-mail: damon.chandler@okstate.edu, Telephone: 1 405 744 9924

Human Vision and Electronic Imaging XIX, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Huib de Ridder, Proc. of SPIE-IS&T Electronic Imaging, Proc. of SPIE Vol. 9014, 901402
 · doi: 10.1117/12.2047117

the aim was to shed light on the limitations in the current knowledge of image quality in the hopes of opening doors for further studies. Here, we will reiterate and discuss these challenges in the context of new psychophysical results obtained after the first publication. The seven challenges, which we will pose here in the form of questions, are as follows:

- *Challenge 1—How to improve masking models for use on natural images?* Computational neural models of masking are crucial for determining whether distortions are visible, a neccessary first step in IQA. However, current V1 models of masking have yet to be extensively tested in terms of their ability to predict detection thresholds in the presence of natural-image masks. In Section 2 we evaluate and discuss the performance of one such model on this task.

- *Challenge 2—How to deal with suprathreshold distortions?* Although masking models are instrumental for determining whether distortions are visible, a common criticism of such models is that they are not applicable for distortions which are well beyond the threshold of visibility—i.e., for distortions which are *suprathreshold*. In Section 3 we evaluate the effectiveness of masking for local quality predictions as the distortions become increasingly suprathreshold.

- *Challenge 3—How to model the effects of distortions on the image's appearance?* Distortions can be perceived to overlay on top of an image, or they can be perceived to interact with an image's objects. As we and other have advocated, these different percepts require different computational approaches in terms of IQA. In Section 4 we discuss how local phase information may be able to predict the occurences of these different percepts.

- *Challenges 4, 5, and 6—How to deal with multiple distortions, geometric changes, and image enhancements?* Images can be subject to multiple simultaneous forms of distortions, including both photometric and geometric changes, and a combination of distortions and enhancements. In Section 5 we discuss two common applications which can give rise to such changes, but for which existing IQA algorithms are largely ill-equipped to handle.

- *Challenge 7—How to improve runtime and memory performance?* Although a great deal of research on IQA has focused on improving prediction accuracy, much less research has addressed algorithmic and microarchitechtural efficiency. In Section 6 we discuss the results of a recent study designed to investigate how IQA algorithms interact with the hardware on a modern desktop computing platform.

In addition, we will propose the block diagram shown in Figure 1, which illustrates how a future IQA system might operate to handle some of these challenges; we will refer to portions of this diagram throughout the paper. The inputs to system are two corresponding local image regions from the original and distorted images. As we will discuss, both the visibility of the distortions and the impact of the distortions on quality can be highly dependent on the recognizability of the image content. Accordingly, the very first stage in the system is a stage which quantifies that recognizability, yielding an entropy masking[22] index (EMI). This EMI and the image regions are then analyzed by a masking model to determine whether the distortions are visible (suprathreshold).

As we will discuss, suprathreshold distortions can be viewed either as overlay-type distortions which are perceived to overlay on top of the image, or as object-degrading distortions which are perceived to interact with and degrade the image's objects. Thus, if the distortions are deemed visible, the system must determine whether the suprathreshold distortions are overlay-type distortions or object-degrading distortions. For overlay-type distortions, existing models of perceived contrast can potentially be used to gauge the apparent contrast of the distortions (e.g., Ref. 23). However, such models would need to take into account how this apparent contrast is influenced by the image.[24,25]

For object-degrading distortions, the system uses different IQA strategies depending on the type of image content. For textures, it is well known that point-by-point comparisons cannot be used; rather, a more statistical approach (e.g., Ref. 26) is warranted. For edges, attributes such as sharpness, contrast, and alignment are perhaps most relevant to quality, and thus a specific edge-based IQA strategy is warranted. For other structures, a hybrid or perhaps a completely different approach is warranted. It could also be that the classification of

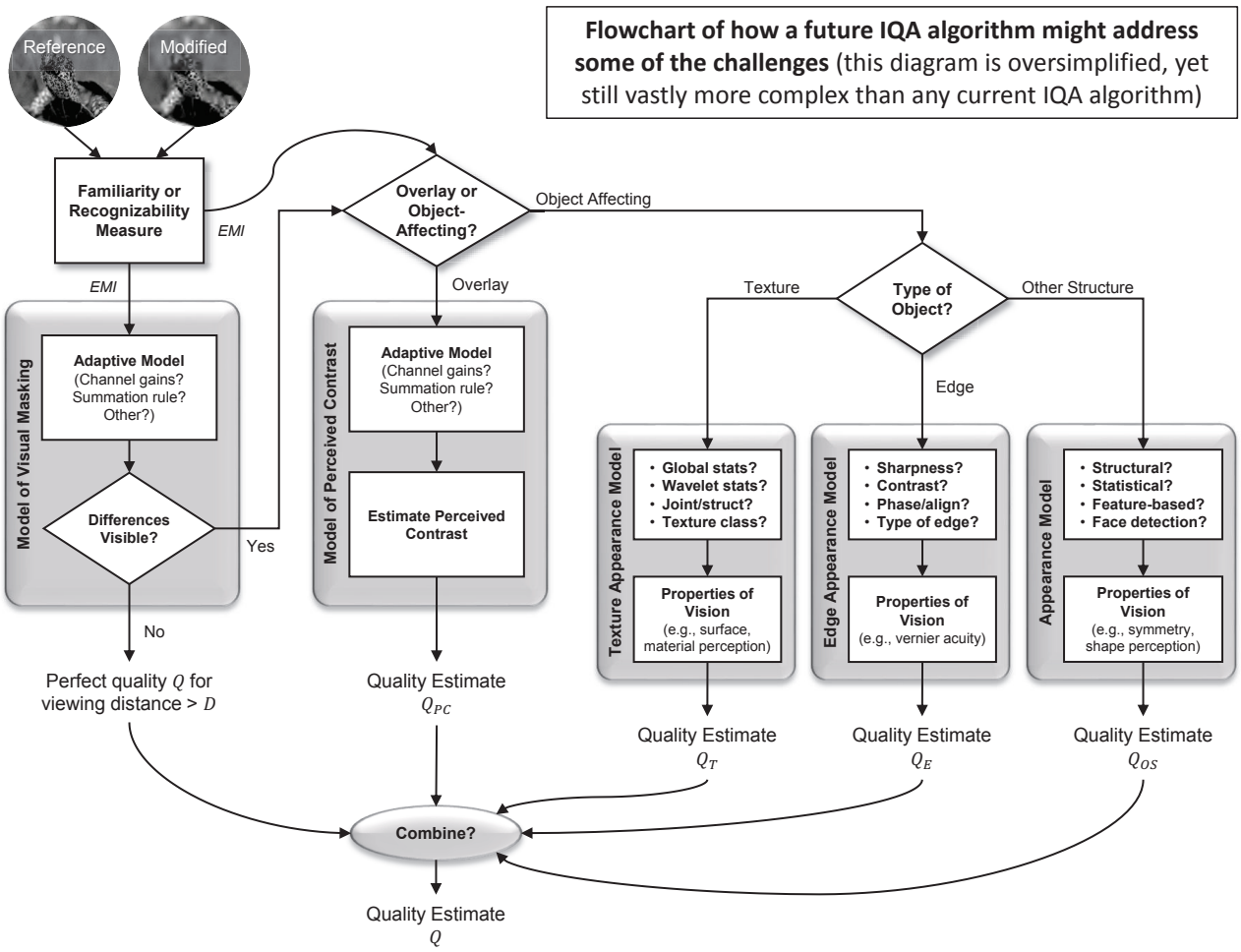

Figure 1. Block diagram of how a future IQA system might deal with some of the challenges mentioned in this paper (see text for details).

the local image content should be a soft classification, and thus all of the approaches might be used to varying extents based on the class proportions.

The quality estimate yielded by the system could either be a scalar based on combination of the individual quality estimates from the separate stages, or perhaps a scalar does not make sense, and thus only a vector-based quality estimate is appropriate. It should also be noted that the system could very well have interconnections from each block to all other blocks. And, all stages of the sytem could be influenced by higher-level or task-based effects (e.g., artistic intent) which would somehow need to be taken into account.

In short, the system shown in Figure 1 is, on the one hand, extremely oversimplified; and, on the other hand, much more complex than any existing IQA algorithm. And, this is a full-reference IQA system—it does not even attempt to address reduced-reference or no-reference quality assessment.

As we discuss the seven challenges in this paper, it is important to note that the challenges by no means represent an exhaustive list. Most of these challenges were chosen due their strong links to visual perception, and notable progress has been made on some of these challenges. Nonetheless, the seven challenges remain largely unsolved. By highlighting these challenges, our hope is to raise awareness of not only the current limitations in IQA knowledge, but also the need for additional psychophysical and computational studies beyond those commonly cited, and the need for alternative theories and techniques beyond those commonly employed.

2. CHALLENGE 1: HOW TO IMPROVE MASKING MODELS FOR USE ON NATURAL IMAGES?

For high-quality images, in which the distortions are near the threshold of detectability, a successful IQA algorithm must be able to predict for which image regions the distortions are detectable, and then use those predictions to estimate local and/or global quality. A common approach toward predicting the visibility of local distortions is to use a computational model of V1 which accounts for visual masking. However, as discussed in Ref. 8, due to the lack of ground-truth data, such V1 models have never been extensively tested on a large database of detection thresholds measured for natural images.

To address this issue, at HVEI 2012, we described a psychophysical study designed to obtain local contrast detection thresholds for a database of natural images.[27] Via a three-alternative forced-choice experiment, we measured thresholds for detecting 3.7 cycles/degree vertically oriented log-Gabor targets placed within each 85×85-pixel patch (1.9 degrees) of 15 natural images from the CSIQ image database.[28] Thus, for each image, we obtained a *masking map* in which each entry in the map denotes the RMS contrast threshold for detecting the log-Gabor target at the corresponding spatial location in the image. We have now extended the database to include thresholds for all 30 images in the entire CSIQ database. Thus, we have detection thresholds for 1080 patches, and here we briefly discuss the results of applying a standard V1 model of masking[29] to predict these thresholds. (Please refer to Ref. 29 for details of the model implementation.)

Figure 2 shows the ground-truth masking maps and the model predictions. Overall, the model performed relatively well at predicting the thresholds, but with some notable failure cases which motivates the need for further research. For the 1080 thresholds taken as a whole, a correlation coefficient (CC of 0.77 and a RMS error of 5.5 dB was observed between the ground-truth thresholds and the model's predictions (after non-linear regression using a four-parameter logistic nonlinearity). On a per-image basis, the model performed best in terms of CC for images *Sunset Color* (CC=0.96) and *Cactus* (CC=0.95). Both of these images generally contain simplistic content: blank regions, single-orientation edges, and/or non-structured (highly stochastic) textures. The model performed worst in terms of CC for images *Foxy* (CC=0.51) and *Aerial City* (CC=0.56). These latter images contain more structured content: recognizable objects such as the fox's face in *Foxy* and the sides of buildings in *Aerial City*, structured texture such as leaves/grass in *Foxy*, and multiple-orientation edges such as the distant cityscape in *Aerial City*.

Overall, the performance of the model if both noteworthy and indicative of the need for further research on the interplay between recognition and masking. In general, the model performed worst for regions with recognizable structure. Figure 3 shows the 14 patches for which the model yielded the absolute worst predictions (largest RMS errors) in comparison to the ground-truth thresholds; the patches are shown with their contexts as was displayed in the experiment. Observe that nearly all of these patches contain recognizable content, and such recognizability has been argued to influence detection thresholds.[22,30,31] Watson *et al.* proposed the term "entropy masking" to describe this special form of masking in which thresholds are elevated due to the subject's unfamiliarity with the mask.[22]

For reference, IQA algorithms generally fail at predicting these thresholds. For example, we evaluated various techniques of applying both MS-SSIM[32] and MAD[33] toward predicting the thresholds. MS-SSIM yielded an average CC of 0.38, with the best and worst predictions on, respectively, images *Foxy* (CC=0.67) and *Native American* (CC=0.13). MAD yields an average CC of 0.64, with the best and worst predictions on, respectively, images *Bridge* (CC=0.86) and *Shroom* (CC=0.28). It is important to note, however, that MS-SSIM was never designed to measure masking, and MAD uses a simplistic spatial-domain-only local contrast model. Nonetheless, these subpar preformances underscore the importance of using a V1 masking model in order to accurately predict the visibility of distortions—a crucial task for IQA of very high-quality images.

Thus, in the block diagram in Figure 1, we propose that the first step in the IQA process should be to quantify the recognizability of the image region so as to yield, for example, an entropy masking index (EMI). This index should then be used within a V1 masking model to determine whether the distortions are visible within the given region. However, developing a computational model of entropy masking or even recognizability remains an elusive goal. Clearly, additional research is needed in order to develop improved models of masking, and thereby develop improved IQA algorithms for use in the very high-quality regime.

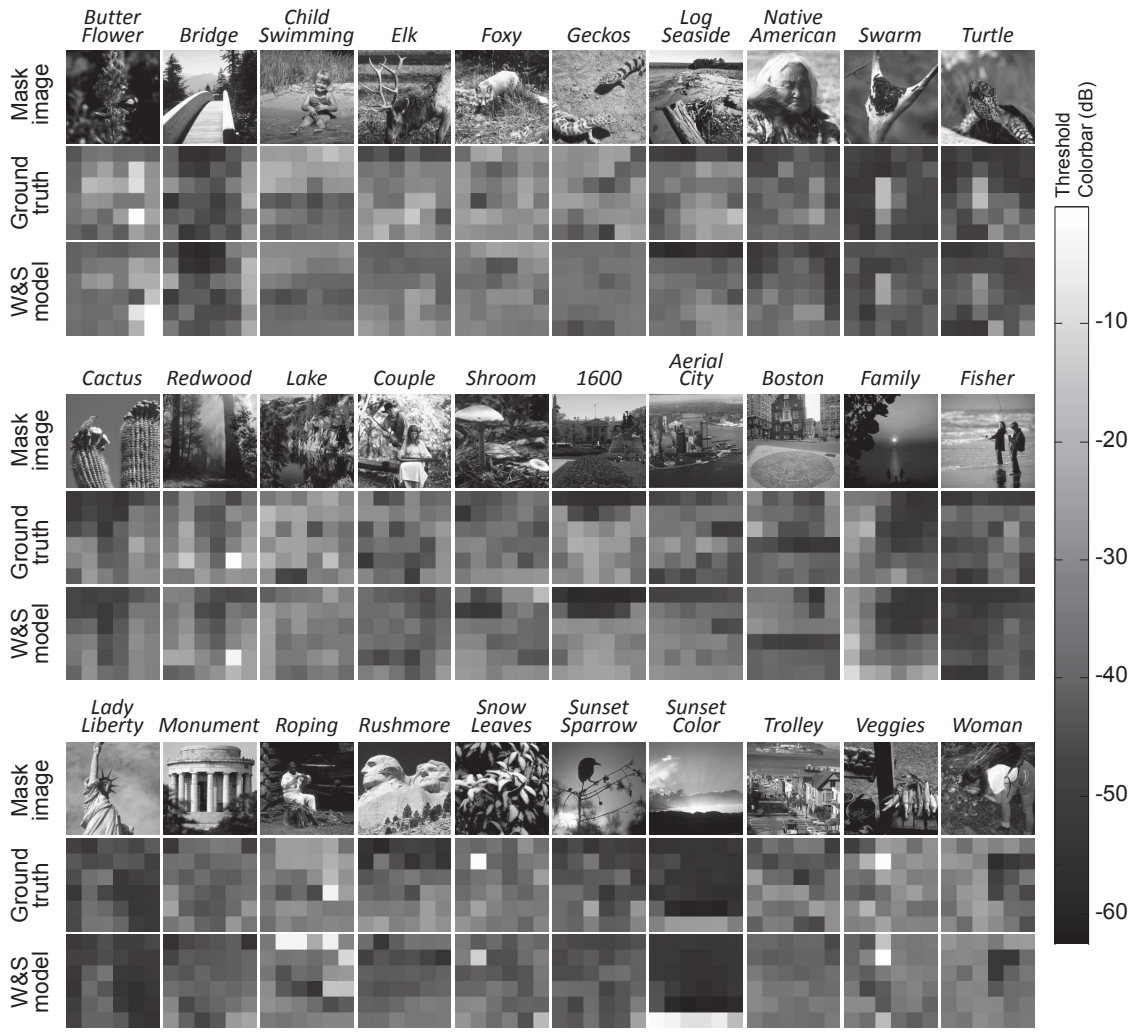

Figure 2. Local masked detection thresholds and model predicted thresholds for images from the CSIQ image-quality database (see text for details).

3. CHALLENGE 2: HOW TO DEAL WITH SUPRATHRESHOLD DISTORTIONS?

Although masking models are instrumental for determining whether distortions are visible, a common criticism of such models is that they are not applicable for distortions which are well beyond the threshold of visibility—i.e., for distortions which are *suprathreshold*. Indeed, current V1 models have most often been fitted to psychophysical data consisting of detection thresholds, and thus the applicability of such models for predicting the qualities of images containing suprathreshold distortions remains unclear.

Some have argued that this "suprathreshold problem" obviates the utility of vision modeling for IQA.[34] Others have demonstrated that V1 masking models can also work for IQA with proper parameter adjustments.[35] In Refs. 36 and 33, we demonstrated that masking models can outperform other IQA approaches as long as such models are adaptively combined with complimentary models of suprathreshold perception. Nonetheless, how humans judge quality at near-threshold vs. suprathreshold levels of distortion remains an open question.

One hypothesis for how quality judgments vary as a function of distortion level is that, at near-threshold distortion levels, IQA based on masking is most appropriate, whereas at suprathreshold distortion levels, IQA based on models of contrast discrimination or apparent contrast are more appropriate. For example, in Ref. 33, we proposed that the strategy used by the HVS adapts from a detection-based strategy at near-threshold distortion

Figure 3. Fourteen patches (shown with contexts) for which the V1 masking model yielded the worst predictions in terms of RMSE (shown below each patch). Observe that these patches generally contain recognizable structure.

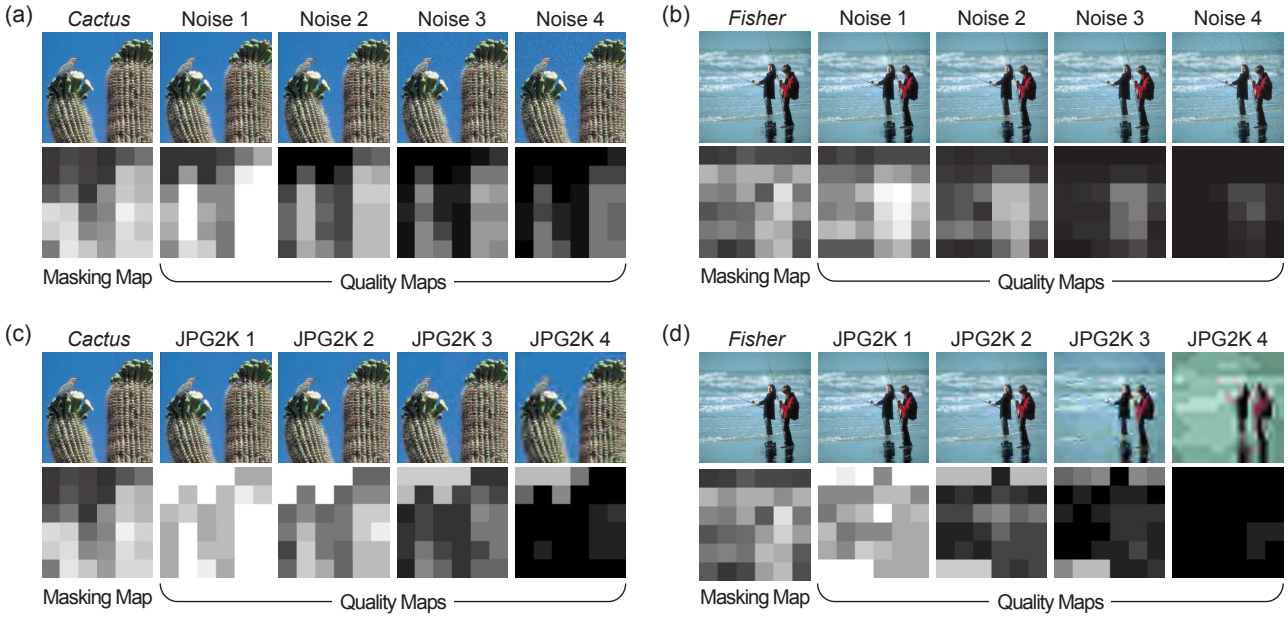

Figure 4. Local quality ratings for images distorted with additive white noise (a) and (b), and via JPEG2000 compression (c) and (d). Observe that the masking maps demonstrate some correspondence with the local quality maps, mostly at high-quality levels, but also into the mid-quality range for additive noise distortion. For mid-to-low-quality JPEG2000 distortion in image *Fisher*, masking appears to be less applicable, owing perhaps to the fact that JPEG2000 induces object-degrading distortions. However, the quality maps for image *Cactus* appear to be more consistent across distortion types, owing perhaps to the entropy masking from the cactus' textures. Note that these are relative quality ratings which have not been properly scaled for across-image and across-distortion-type comparisons.

levels, to an appearance-based strategy at clearly suprathreshold distortion levels. We accordingly developed an IQA algorithm, called *MAD: Most Apparent Distortion*, which used and adaptively combined separate IQA models (masking and appearance) for these two distortion-level regimes. However, this dual-strategy approach was never verified psychophysically.

To further investigate the possibility that humans employ a dual- or multiple-strategy approach, we have recently begun developing a database of local quality ratings for images. Using a subset of the same reference images and the same 85×85 blocks used in the masking database discussed in Section 2, we asked subjects to provide a quality rating for each block of each distorted image. Two distortion types were used: additive Gaussian white noise and JPEG2000 compression. Four distortion levels were used, ranging from very high-quality images

(near-threshold distortions) to extremely low-quality images (clearly suprathreshold distortions). Subjects were simultaneously shown the reference image, the distorted image, and a separate version of the distorted image on which a block marker was placed. Subjects were instructed to rate the quality of the indicated block by using a 1-5 discrete quality scale. The per-subject ratings were converted to z-scores and then averaged across subjects. Thus, this experiment yielded for each distorted image, a local quality map.

Figure 4(a) and (b) show representative results for two of the images (images *Cactus* and *Fisher*) for the white noise distortion. In each subfigure, the distorted images are shown in the first row, with increasing amounts of distortion from left to right. The corresponding quality maps are shown in the second row. For reference, the masking maps are also shown for each image. Observe from (a) and (b) that, for additive white noise, the quality maps are generally in agreement with the masking maps, particularly for the first two distortion levels. For the latter two (greater) distortion levels, there is less agreement between the quality maps and masking maps; however, some clear correspondences between the maps can still be observed. Thus, for white noise, it would appear that masking can play an important role in determining quality, even for mid-quality images. For example, for image *Cactus* at the highest noise level, even though the noise is visible within the body of the cactus, the quality ratings remain greater for these high-masking regions than for other regions. The ability of such high-masking regions to maintain quality even for suprathreshold distortions may perhaps be attributable to entropy masking; i.e., the distortions are visible yet not detrimental to quality due to the fact that the cactus is textured and thus the addition of noise does not destroy this general texture, despite the fact that pointwise differences are induced.

Figure 4(c) and (d) show representative results for images *Cactus* and *Fisher* for JPEG2000 distortion. Unlike additive white noise, JPEG2000 induces a variety of distortions (blurring, ringing, aliasing), all of which are spatially correlated with the objects in the images. Thus, JPEG2000 distortions will not appear in areas which are naturally smooth or otherwise devoid of objects (i.e., areas for which the wavelet coefficients are natively zero or near zero). Consequently, the quality remains relatively high in such regions, even through the third level of distortion. However, even if such undistorted blocks are ignored, there is much less correspondence between the masking map and the quality maps for JPEG2000 distortion as compared to the maps for white noise. For mid-to-low-quality JPEG2000 distortion in image *Fisher*, masking appears to be less applicable, owing perhaps to the fact that JPEG2000 induces object-degrading distortions. However, the quality maps for image *Cactus* appear to be more consistent across distortion types, owing perhaps to the entropy masking from the cactus' textures. As reported by several of the subjects in the experiment, white noise and JPEG2000 distortions give rise to different percepts: white noise appears as overlay-type distortions which do not interact with the image's objects, whereas JPEG2000 appears as object-degrading-type distortions which affect the phenomenal appearance of the image's content.

Clearly additional research is needed in order to answer the question of how quality judgments change as a function of distortion level. However, these preliminary results seem to suggest that, depending on the interaction of the distortion and image, and depending on the subject's familiarity with the objects (entropy masking), quality is judged based on different strategies. For overlay-type distortions, quality judgments appear to be based, at least in part, on masking, particularly for high-quality images. However, masking can also be valid for mid-to-lower-quality images in regions (e.g., textures) for which the familiarity is low (entropy masking is high). Thus, as shown in the block diagram in Figure 1, for overlay-type distortions, an IQA algorithm based on masking, perceived contrast, or a combination of these two could be effective throughout a wide quality range. However, for object-degrading distortions, masking is applicable only for low amounts of distortion (high-quality images). As the distortions become increasingly suprathreshold, the applicability of masking reduces, and becomes perhaps applicable only in regions which induce entropy masking. For the more familiar objects in the image, quality seems to be judged based on the perceived integrity of such objects (i.e., an appearance-based strategy).

4. CHALLENGE 3: HOW TO MODEL THE EFFECTS OF DISTORTIONS ON THE IMAGE'S APPEARANCE?

As we argued in Ref. 25, and as noted earlier by Goodman and Pearson,[37] a distinction must be made between distortions which appear to overlay on top of the image, and distortions which appear to degrade the image's

Figure 5. Original images (a) and (b), and hybrid images [(c) and (d)] created by combining the magnitude spectrum of (a) with the phase spectrum of (b) either globally (c) or locally (d). See text for details.

objects. Goodman and Pearson[37] used a multidimensional scaling (MDS) experiment to investigate the quality of TV pictures impaired both by additive-type distortions (e.g., noise, echo) and by coding- and transmission-type distortions (DPCM quantization artifacts and blurring). Based on their MDS analysis, they reported that one the multiple perceptual dimensions "*appears to be separating those impairments which cause the integrity of the objects in the picture to be destroyed from overlay types of impairment.*"

In the visual psychophysics literature, these two scenarios would fall under the aegis of *capture* and *transparency*,[38] which describe whether a target + background are perceived as one combined stimulus (captured), or whether they are perceived as two separate stimuli (transparent). In Ref. 39, we argued that overlay-type distortions are viewed as perceptually transparent from the image, whereas object-degrading distortions are perceptually captured as part of the image. Thus, for overlay-type distortions, subjects tend to judge quality based on the perceived contrast of the distortions (i.e., they attempt to visually ignore the image), whereas when the distortions are severe and spatially correlated with the image, viewers tend to base quality judgments on the interaction between the distortions and the image's objects. Properly determining and modeling the perceptual effects of this latter interaction is yet another challenge in IQA research.

4.1 Can Local Phase Information Predict Capture vs. Transparency?

In Ref. 39, we reported that object-degrading distortion are on average 4.4 times more detrimental to quality and overlay-type noise. Thus, for an IQA algorithm, it would seem that a logical first step would be to predict whether distortions are perceived as either overlay distortions or object-degrading distortions.

One promising approach toward this prediction is the use of a method based on local phase information, as demonstrated in Figure 5. In this figure, the two original images [shown in (a) and (b)] are combined into hybrid images [shown in (c) and (d)] by using the magnitude information from (a) and the phase information from (b). The hybrid image in (c) was generated by using the global (image-wide) magnitude and phase spectra. The hybrid image in (d) was generated by using the local magnitude and phase spectra of each 16×16 patch. When using the global spectra, as argued by Oppenheim and Lim,[40] the image in (c) appears most similar to the image whose phase spectrum was used. And, when using the local spectra, as argued by Morgan *et al.*,[41] the image in (d) appears most similar to the image whose local magnitude spectra were used. However, in terms of *capture and transparency*, observe that the image in (c) appears as a distorted version of (b) in which the distortions appear as *overlay-type* distortions—i.e., the distortions are perceived as transparent from the image. On the other hand, the image in (d) appears as a distorted version of (a) in which the distortions appear as *object-degrading* distortions.

Thus, based on the observations from Figure 5, it would seem possible to predict whether the distortions are perceived as transparent vs. captured by examining whether the local phase information is intact. If the local magnitude information is intact and the local phase information is distorted, then the resulting distortions should appear as object-degrading distortions. If the local magnitude information is distorted and the local

phase information is intact, then the resulting distortions could appear as either overlay or object-degrading distortions, depending on the specific manipulations made to the local magnitude spectra.

4.2 Different IQA Strategies for Captured vs. Transparent Distortions

Assuming that an IQA algorithm could indeed classify the distortions as overlay-type (transparent) or object-degrading (captured), a different IQA strategy could then be used for these two types of distortions. For overlay-type distortions, a model which can successfully estimate the perceived contrast of the distortions would seem appropriate. However, such a perceived contrast model must also be able to take into account how the perceived contrast of the distortions is affected by the image. Adaptation and masking experiments have both shown that the perceived contrast of suprathreshold targets can be influenced by the presence of a background image.[24, 25]

For object-degrading distortions, the proper IQA strategy is an open question. Perhaps the proper strategy should be based on local structural comparisons as used in SSIM.[34] Or, perhaps the proper strategy should be based on an information-theoretic model as used in VIF.[42] Or, perhaps one should employ local statistical comparisons between modeled neural responses as used in MAD.[33] It could also be that a combination of all of these techniques (or maybe none of these techniques) is warranted.

Referring back to the block diagram in Figure 1, the IQA strategy for object-degrading distortions depends on both the type of local image content and the subject's familiarity with that content. We propose that the local image content should be classified, e.g., as either a texture, an edge, or some other structure. For textures, local statistical comparisons would seem appropriate.[26] For edges, an approach based on sharpness and local structural comparions would seem appropriate. For other structures, perhaps a hybrid or different approach is warranted. It could also be that the classification of the local image content should be a soft classification, and thus all of the approaches might be used to varying extents based on the class proportions. Clearly, determining the proper IQA strategy to use for object-degrading distortions remains a open research challenge.

5. CHALLENGES 4, 5, AND 6: HOW TO DEAL WITH MULTIPLE DISTORTIONS, GEOMETRIC CHANGES, AND IMAGE ENHANCEMENTS?

Developers of IQA algorithms are well aware of the limitations of these algorithms. In particular, it is well known that current IQA algorithms have largely been geared toward "classical distortions"—e.g., various forms of additive noise, compression artifacts, blur, and other photometric changes—distortions which commonly arise in engineering applications. However, as new applications emerge, other types of "non-classical" image changes are becoming increasingly prevalent. In particular, images can be subject to multiple, simultaneous forms of distortions, including both photometric and geometric changes, and a combination of distortions and enhancements. Unfortunately, existing IQA algorithms generally cannot accurately estimate the qualities of such images.

In Ref. 8, we discussed two particular types of changes which existing IQA algorithms are ill-equipped to handle: (1) geometric changes, and (2) changes due to image enhancement techniques. Critics may argue that such changes are so radically different that IQA algorithms should not even be employed for these scenarios. While this may be true, there are an abundance of practical applications in which visual comparisons need to be made, and for which an IQA-based approach would seem the most logical. Here we will highlight two particular examples: (1) comparisons of textures in a watermarking application (example of geometric changes); and (2) IQA of images which have undergone an artifact-removal process (example of edge enhancements). Referring back to the block diagram in Figure 1, these examples will demonstrate that different IQA strategies are needed for textures vs. edges vs. other structures.

5.1 Texture Similarity in a Watermarking Application

Although a great deal of human vision research has been conducted to investigate the perceptual and neural mechanisms which underlie the visual appearance of texture (see Ref. 43 for a review), further research is needed on how to actually apply these findings to IQA of textures. In Ref. 44, Bénard *et al.* investigated the effects of fractalization on the visual quality of synthesized textures; they reported that the average co-occurrence error between gray-level co-occurrence matrices measured for the original and fractilized textures can perform well

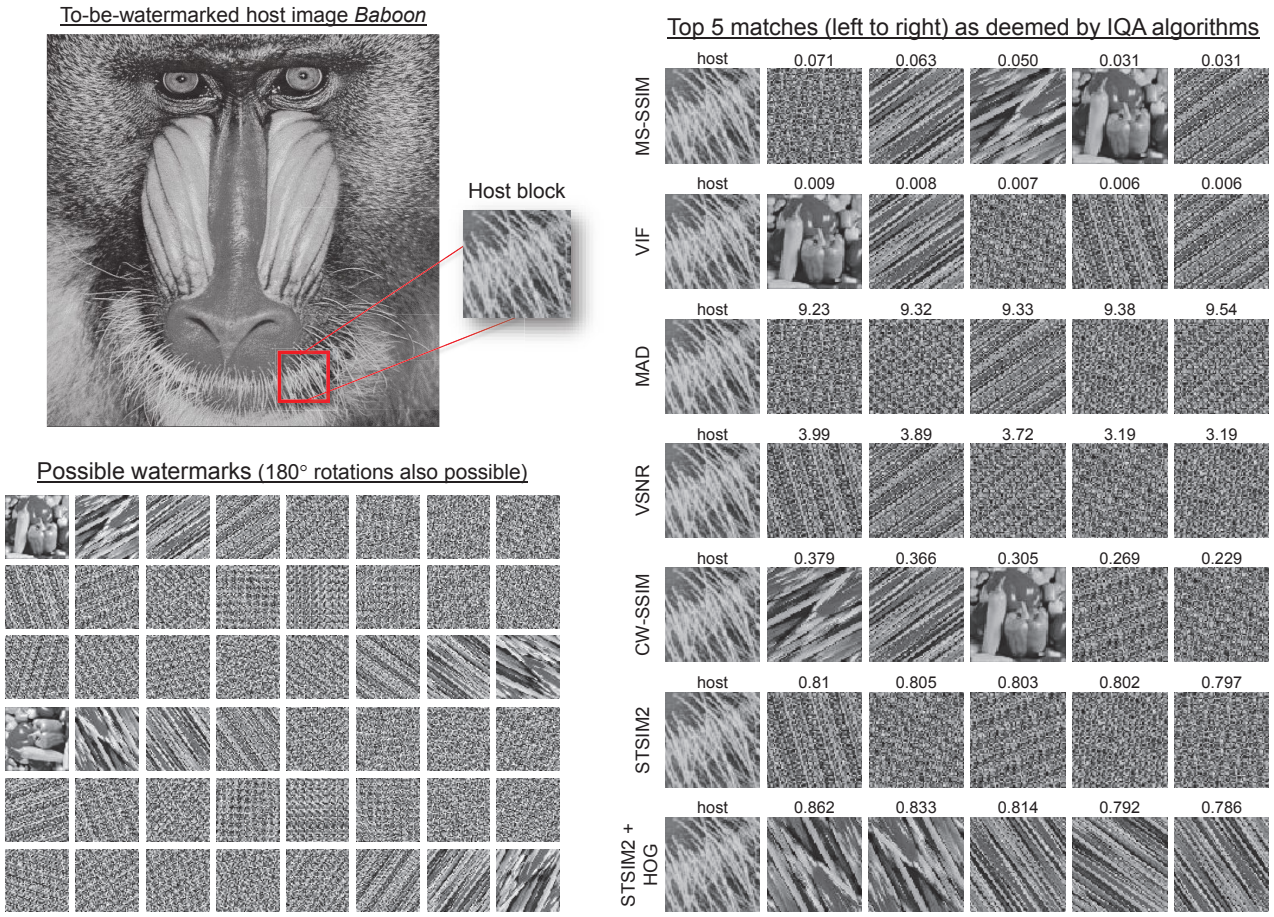

Figure 6. Most IQA algorithms fail when applied to non-classical IQA tasks such as determining the similiarity between a host block and texturized versions of a logo (see text for details).

in predicting the subjective ratings. More recently, in Ref. 26, Zujovic *et al.* addressed IQA of textures by designing a structural similarity index for textures.

We have recently developed a watermarking algorithm[45] in which a small logo serves as the watermark which is to be embedded within the textured regions of the host image. Rather than embedding the logo as-is, we adaptively texturize the logo to best match each block of the host image in which the texture is to be embedded. In order to determine this "best match," a natural approach would be to use an IQA algorithm to compare the texturized logo to the corresponding host block.

Figure 6 illustrates this scenario in which the logo is a small version of the image *Peppers*, which is to be embedded in the highlighted block of the host image *Baboon*. The texturized versions of the logo are shown in the lower left, and the task is to select the texturized version which is most visually similar to the host block. Although this task is straightforward for a human subject, it is extremely challenging for an IQA algorithm due to the fact that point-by-point comparisons can no longer be used. The top five matches as judged by various IQA algorithms (MS-SSIM,[32] VIF,[42] MAD,[33] VSNR,[36] CW-SSIM,[46] STSIM2[26]) are shown on the right, all of which are largely unsuccessful at this task. Indeed, most of these algorithms were not designed to measure such radical changes between images.

For this particular watermarking application, only texture-specific approaches (such as STSIM2 or the appearance stage of MAD) could be made successful via supplementation with a distortion measure based on differences in the histograms of oriented gradients (denoted by STSIM2+HOG in Figure 6). This example demonstrates

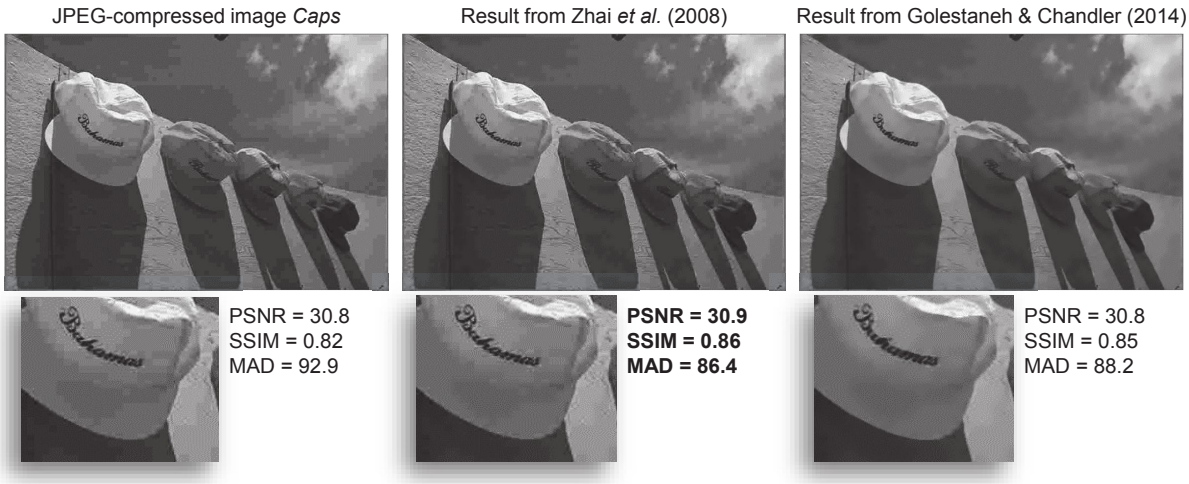

Figure 7. Full-size and close-ups of image *Caps* compressed with JPEG at a Q-factor of 10 (left), and two versions of the image processed via different artifact-removal algorithms (middle[48] and right[47]). The righmost image was generated via our recent artifact removal algorithm,[47] and this result was judged to be of greatest visual quality. However, PSNR, MS-SSIM,[32] and MAD[33] all incorrectly report the middle image to be of the greatest quality.

that a different IQA strategy must be used for textures as compared to other regions, thus motivating the need for the classification into texture vs. edge vs. other structure shown previously in Figure 1.

5.2 IQA of Images Subject to Artifact Removal

Another application in which existing IQA algorithms can fail to accurately estimate quality is for IQA of images which have undergone restoration followed by enhancement. As an example, we recently developed an algorithm for JPEG artifact removal which uses deblocking followed by an edge-regeneration process.[47] This edge-regeneration stage can produce edges which appear sharper than the compressed image's edges, but which may not perfectly align with the original (uncompressed) image's edges. Visually, however, these edges appear equivalent or in some cases superior to the original image's corresponding edges. Yet, due to other artifacts (such as blurring), which cannot be completely removed, the overall quality of the processed image is lower than the quality of the original image. Nonetheless, the processed image is always superior in visual quality to the JPEG-compressed version. Thus, it would seem logical to apply an IQA algorithm to quantify these differences in qualities (original vs. JPEG, and original vs. JPEG after artifact removal).

Figure 7 shows an example of this scenario. The JPEG-compressed image *Caps* from the LIVE image database[49] is shown on the left, and the images resulting from two artifact-removal algorithms are shown in the middle and on the right. The righmost image was generated via our recent artifact removal algorithm,[47] and this result was judged to be of greatest visual quality (see the close-ups). Also shown in Figure 7 are quality predictions from both MS-SSIM[32] and MAD,[33] both of which, along with PSNR, judge the middle image to be of greatest visual quality. Of course, there may certainly be other IQA algorithms that happen to succeed on these images. However, the fact that these two top-performing IQA algorithms cannot successfully predict the qualities of these images underscores the need for future research in this area.

6. CHALLENGE 7: HOW TO IMPROVE RUNTIME AND MEMORY PERFORMANCE?

Although a great deal of research on IQA has focused on improving prediction accuracy, much less research has addressed algorithmic and microarchitechtural efficiency. As IQA algorithms move from the research environment into more mainstream applications, issues surrounding efficiency—such as execution speed and memory bandwidth requirements—begin to emerge as equally important performance criteria. How to improve runtime and memory performance without sacrificing prediction accuracy is another open challenge in IQA research.

In collaboration with a computer engineer, we recently published a study to examine various IQA algorithms from the perspective of their interaction with the underlying hardware and microarchitectural resources.[50] We implemented four popular full-reference IQA algorithms (MAD, MS-SSIM, VIF, VSNR) and two no-reference algorithms (BLIINDS-II,[51] BRISQUE[52]) in C++ based on the code provided by their respective authors. We then conducted a hotspot analysis to identify sections of code that were performance bottlenecks and we performed a microarchitectural analysis to identify the underlying causes of these bottlenecks.

From the microarchitectural analysis, we found that for almost all of the IQA algorithms the primary cause of performance degradation was due to memory bottlenecks rather than core (i.e., computational) bottlenecks. Within the memory bottlenecks, the most common slowdown was due to limited L1, L2, and L3 (LLC) cache sizes; nearly all algorithms suffered from L1D and L2D replacements and LLC misses. To improve performance in such cases, the ideal solution is to have a large cache, but strategic coding techniques can also help (see Ref. 50). The next common memory bottleneck was overhead due to the use of a Data Translation Lookaside Buffer (DTLB). A DTLB is special cache which stores a subset of translations from virtual memory to physical memory. MAD, VIF, and MS-SSIM all showed performance degradation due to DTLB overhead. Another, less common bottleneck was 4K aliasing, which was observed in VSNR. This 4K aliasing stems from out-of-order execution of memory instructions in the processor.

Performance degradation due to slow computation (core bottlenecks) was less frequent, but nontheless observed. The most common core bottleneck was the overwhelming of the floating-point unit. All of the image transforms, filtering, and statistic calculations require floating-point operations and consequently a floating-point execution unit. Another core bottleneck was due to slow LEA instructions, which are instructions resulting from the complex addressing mode of the CISC Intel architecture. The final core bottleneck was the generation of micro assists, which occurred because the operands or results of an operation were denormals (extremely small, yet non-zero values).

Our microarchitectural analysis is only a first step toward understanding how IQA algorithms interact with the underlying hardware; the study was limited to one particular hardware platform (Intel Core i5 general-purpose machine) and six IQA algorithms. Furthermore, many of the observed bottlenecks can be circumvented only via hardware modifications, and even if a software-based solution is possible, such a solution may give rise to additional unexpected bottlenecks. Clearly, this is an area which could benefit from additional research.

7. CONCLUSIONS

In this paper, we have highlighted seven open challenges in image quality research. There are, of course, many more important challenges beyond those discussed here—for example the influence of regions of interest and saliency on image quality, the influence of task-based objectives on quality, the relations between image quality and image utility, and many issues surrounding artistic intent. As in Ref. 8, the objective of this discussion was not only to highlight the limitations in our current knowledge of image quality, but to also emphasize the fact that there is substantial room for alternative theories and techniques for quality assessment.

ACKNOWLEDGMENTS

The authors are grateful to the HVEI community for its continued fundamental research on visual perception and image quality. This work was supported by, or in part by, the National Science Foundation Awards 0917014 and 1054612, and by the U.S. Army Research Laboratory (USARL) and the U.S. Army Research Office (USARO) under contract/grant number W911NF-10-1-0015.

REFERENCES

[1] Beaton, R. J., "Quantitative models of image quality," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **27**(1), 41–45 (1983).

[2] Ahumada, Jr., A. J., "Computational image quality metrics: A review," in [*SID International Symposium Digest of Technical Papers*], **24**, 305–308 (1993).

[3] Keelan, B., [*Handbook of Image Quality: Characterization and Prediction*], Optical engineering, Taylor & Francis (2002).

[4] Seshadrinathan, K., Pappas, T., Safranek, R., Chen, J., Wang, Z., Sheikh, H., and Bovik, A., "Image quality assessment," in [*The Essential Guide to Image Processing*], Bovik, A., ed., *Electronics and Electrical*, Academic Press (2009).

[5] Hemami, S. S. and Reibman, A. R., "No-reference image and video quality estimation: Applications and human-motivated design," *Image Commun.* **25**, 469–481 (Aug. 2010).

[6] Seshadrinathan, K. and Bovik, A. C., "Automatic prediction of perceptual quality of multimedia signals–a survey," *Multimedia Tools Appl.* **51**, 163–186 (Jan. 2011).

[7] Lin, W. and Jay Kuo, C. C., "Perceptual visual quality metrics: A survey," *J. Vis. Comun. Image Represent.* **22**, 297–312 (May 2011).

[8] Chandler, D. M., "Seven challenges in image quality assessment: Past, present, and future research," *ISRN Signal Processing* **2013**(905685) (2013).

[9] Bovik, A. C., "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE* **101**(9) (2013).

[10] Horton, J., "The electrical transmission of pictures and images," *Proceedings of the Institute of Radio Engineers* **17**, 1540–1563 (Sept. 1929).

[11] Jesty, L. and Winch, G., "Television images: An analysis of their essential qualities," *Tran. Illum. Eng.* **2**, 316–334 (1937).

[12] Goldmark, P. and Dyer, J., "Quality in television pictures," *Proceedings of the IRE* **28**, 343–350 (Aug. 1940).

[13] Winch, G., "Colour television: some subjective and objective aspects of colour rendering," *Proceedings of the IEE—Part IIIA: Television* **99**(20), 854–860 (1952).

[14] Jesty, L., "Television as a communication problem," *Electrical Engineers, Journal of the Institution of* **1953**, 181–183 (Apr. 1953).

[15] Fellgett, P. B. and Linfoot, E. H., "On the assessment of optical images," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* **247**(931), 369–407 (1955).

[16] Dean, C., "Measurements of the subjective effects of interference in television reception," *Proceedings of the IRE* **48**, 1035–1049 (June 1960).

[17] Schmid, H., "Measurement of television picture impairments caused by linear distortions," *Journal of the SMPTE* **77**(3), 215–220 (1968).

[18] Sakrison, D. and Algazi, V., "Comparison of line-by-line and two-dimensional encoding of random images," *IEEE Transactions on Inf. Theor.* **17**, 386–398 (Sept. 1971).

[19] Budrikis, Z., "Visual fidelity criterion and modeling," *Proceedings of the IEEE* **60**, 771–779 (July 1972).

[20] Stockham, T.G., J., "Image processing in the context of a visual model," *Proceedings of the IEEE* **60**, 828–842 (July 1972).

[21] Schade, O., [*Image Quality: A Comparison of Photographic and Television Systems*], RCA Laboratories (1975).

[22] Watson, A. B., Taylor, M., and Borthwick, R., "Image quality and entropy masking," *Human Vision, Visual Processing, and Digital Display VIII, Proc. SPIE* **3016**, 2–12 (1997).

[23] Haun, A. M. and Peli, E., "Perceived contrast in complex images," *Journal of Vision* **13**(13) (2013).

[24] Webster, M. A. and Miyahara, E., "Contrast adaptation and the spatial structure of natural images," *J. Opt. Soc. Am. A* **14**, 2355–2366 (1997).

[25] Chandler, D. M. and Hemami, S. S., "Suprathreshold image compression based on contrast allocation and global precedence," in [*Proc. SPIE Human Vision and Electronic Imaging VIII*], Rogowitz, B. E. and Pappas, T. N., eds., *Proceeding SPIE Human Vision and Electronic Imaging* (2003).

[26] Zujovic, J., Pappas, T., and Neuhoff, D., "Structural texture similarity metrics for image analysis and retrieval," *IEEE Transactions on Image Processing* **22**, 2545–2558 (July 2013).

[27] Alam, M. M., Vilankar, K. P., and Chandler, D. M., "A database of local masking thresholds in natural images," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **8651** (Mar. 2013).

[28] Larson, E. C. and Chandler, D. M., "Categorical subjective image quality CSIQ database," (2009).

[29] Watson, A. B. and Solomon, J. A., "A model of visual contrast gain control and pattern masking," *J. Opt. Soc. Am. A* **14**, 2378–2390 (1997).
[30] Fahrenfort, J. J., Scholte, H. S., and Lamme, V. A., "Masking disrupts reentrant processing in human visual cortex.," *Journal of cognitive neuroscience* **19**(9), 1488–1497 (2007).
[31] Chandler, D. M., Gaubatz, M. D., and Hemami, S. S., "A patch-based structural masking model with an application to compression," *J. Image Video Process.* **2009**, 5:1–5:22 (Jan. 2009).
[32] Wang, Z., Simoncelli, E., and Bovik, A., "Multiscale structural similarity for image quality assessment," in [*Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*], **2**, 1398–1402 (Nov. 2003).
[33] Larson, E. C. and Chandler, D. M., "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging* **19**(1), 011006 (2010).
[34] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E., "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**, 600–612 (2004).
[35] Laparra, V., noz Marí, J. M., and Malo, J., "Divisive normalization image quality metric revisited," *J. Opt. Soc. Am. A* **27**, 852–864 (Apr. 2010).
[36] Chandler, D. M. and Hemai, S. S., "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing* **16**(9), 2284–2298 (2007).
[37] Goodman, J. S. and Pearson, D. E., "Multidimensional scaling of multiply-impaired television pictures," *Systems, Man and Cybernetics, IEEE Transactions on* **9**, 353–356 (June 1979).
[38] Morrone, M. C. and Burr, D. C., "Capture and trasparency in coarse quantized images," *Vis. Res.* **37**, 2609–2629 (1997).
[39] Chandler, D. M., Lim, K. H. S., and Hemami, S. S., "Effects of spatial correlations and global precedence on the visual fidelity of distorted images," in [*Proc. SPIE Human Vision and Electronic Imaging XI*], Rogowitz, B. E., Pappas, T. N., and Daly, S., eds. (2006).
[40] Oppenheim, A. V. and Lim, J. S., "The importance of phase in signals," *Proc. of the IEEE* **69**, 529–541 (1981).
[41] Morgan, M. J., Ross, J., and Hayes, A., "The relative importance of local phase and local amplitude in patchwise image reconstruction," *Biological Cybernetics* **65**(2), 113–119 (1991).
[42] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE Transactions on Image Processing* **15**(2), 430–444 (2006).
[43] Landy, M. S. and Graham, N., "Visual perception of texture," in [*The Visual Neurosciences*], 1106–1118, MIT Press (2004).
[44] Bénard, P., Thollot, J., and Sillion, F., "Quality assessment of fractalized npr textures: a perceptual objective metric," in [*Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*], *APGV '09*, 117–120 (2009).
[45] Andalibi, M. and Chandler, D. M., "Digital image watermarking via adative logo texturization," *submitted to IEEE Transactions on Image Processing* (2014).
[46] Sampat, M., Wang, Z., Gupta, S., Bovik, A., and Markey, M., "Complex wavelet structural similarity: A new image similarity index," *Image Processing, IEEE Transactions on* **18**, 2385–2401 (Nov. 2009).
[47] Alireza Golestaneh, S. and Chandler, D. M., "Algorithm for jpeg artifact reduction via local edge regeneration," *Journal of Electronic Imaging* **23**(1), 013018 (2014).
[48] Zhai, G., Zhang, W., Yang, X., Lin, W., and Xu, Y., "Efficient image deblocking based on postfiltering in shifted windows," *Circuits and Systems for Video Technology, IEEE Transactions on* **18**, 122–126 (Jan 2008).
[49] Sheikh, H., Z.Wang, Cormack, L., and Bovik, A., "LIVE image quality assessment database Release 2.."
[50] Phan, T. D., Shah, S., Chandler, D. M., and Sohoni, S., "Microarchitectural analysis of image quality assessment algorithms," *accepted to Journal of Electronic Imaging* (January 2014). (in press).
[51] Saad, M., Bovik, A., and Charrier, C., "A DCT statistics-based blind image quality index," *Signal Processing Letters, IEEE* **17**, 583–586 (June 2010).
[52] Mittal, A., Moorthy, A., and Bovik, A., "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing* **PP**(99), 1 (2012). in press.