Noname manuscript No.
(will be inserted by the editor)

Visual Quality Assessment Algorithms : What Does the Future Hold?

Anush K. Moorthy · Alan C. Bovik

Received: date / Accepted: date

Abstract Creating algorithms capable of predicting the perceived quality of a visual stimulus defines the field of objective visual quality assessment (QA). The field of objective QA has received tremendous attention in the recent past, with many successful algorithms being proposed for this purpose. Our concern here is not with the past however; in this paper we discuss our vision for the future of visual quality assessment research.

We first introduce the area of quality assessment and state its relevance. We describe current standards for gauging algorithmic performance and define terms that we will use through this paper. We then journey through 2D image and video quality assessment. We summarize recent approaches to these problems and discuss in detail our vision for future research on the problems of full-reference and no-reference 2D image and video quality assessment. From there, we move on to the currently popular area of 3D QA. We discuss recent databases, algorithms and 3D quality of experience. This yet-nascent technology provides for tremendous scope in terms of research activities and we summarize each of them. We then move on to more esoteric topics such as algorithmic assessment of aesthetics in natural images and in art. We discuss current research and hypothesize about possible paths to tread. Towards the end of this article, we discuss some other areas of interest including high-definition (HD) quality assessment, immersive environments and so on before summarizing interesting avenues for future work in multimedia (i.e., audio-visual) quality assessment.

Keywords Quality assessment · objective quality assessment · subjective quality assessment · perceived quality

Anush Moorthy
Dept. of Electrical and Computer Engg.,
The University of Texas at Austin
Tel.: 512-415-0213
E-mail: anushmoorthy@mail.utexas.edu

Alan Bovik
Dept. of Electrical and Computer Engg.,
The University of Texas at Austin
Tel.: 512-471-6530

1 Introduction

Man is a visual animal. Through evolution, man has always been fascinated by what he can see and imagine and has endeavored to re-create this world on canvas - from prehistoric cave paintings to modern-day films accompanied by sounds and visual effects. Although arguments on the reason for cave paintings persist, it is clear that at least in today's world, much of the created visual stimuli are for entertainment or informational purposes.

The urge to capture the world around us has lead to the creation of devices which are increasingly capable of doing so, and burgeoning demand has lead to increasing availability at ever-reducing costs. Transmission, storage and display of visual stimuli have also escalated. At the end of this chain is the receiver - the human observer. It should be obvious that with so many different devices capturing, storing, compressing, transmitting and displaying visual stimuli, the receiver is bound to receive stimuli of varying levels of palatability. Research in the area of visual quality assessment has aimed to algorithmically capture this palatability of visual stimuli.

Before we proceed, it is necessary to disambiguate certain terms that we will utilize throughout this paper. Visual stimulus is a generic term that encompasses 2D images, 2D videos, 3D images, 3D videos, immersive viewing environments and so on. Essentially, any captured stimulus that is incident upon the eyes is referred to as a visual stimulus.

Objective quality assessment (QA) refers to creation of algorithms that gauge the perceived quality of these visual stimuli. This definition necessitates the definition of what one means by 'perceived' quality. Since the ultimate receiver of any visual stimulus is a human observer, her opinion of quality is what is of import and is referred to as the perceived quality of the stimulus. An obvious question then is, how does one gauge this perceived quality? Such human perceived quality is gauged by conducting a large scale human study where a sizeable quantity of human observers are shown a series of visual stimuli whose quality they are asked to rate on a particular scale. The mean score of the these stimuli (after accounting for outlier subjects) is termed the mean opinion score (MOS) and is representative of the perceived quality of the stimuli. Such human assessment of quality is referred to as subjective quality assessment. Such large scale publicly available quality assessment databases exist for 2D image quality assessment (IQA) [70], [57] and 2D video quality assessment (VQA) [49], [67]. Recently, researchers have also proposed databases for 3D IQA [25] and 3D VQA [24]. These 3D QA databases are not 'quality' assessment databases in the sense used for the 2D case - we will have a lot to say about this later on.

The performance of any objective visual quality assessment algorithm is gauged by measuring its correlation with human perception. In order to do so, subjective studies as described above are carried out and the quality scores of the algorithm are correlated with MOS using a variety of statistical measures such as the Spearman's rank ordered correlation coefficient (SROCC), the linear correlation coefficient (LCC), root-mean-squared error (RMSE), outlier ratio (OR) and so on [72], between the objective and subjective scores. Since objective algorithms need not correlate linearly with the subjective scores, LCC and RMSE are generally computed after passing the objective scores through a logistic non-linearity [77]. SROCC and LCC values close to 1 and RMSE, OR values close to 0 indicate that the algorithm performs well. Further, statistical significance tests such as ANOVA [72] are used to gauge the significance of algorithm performance [77], [67], [70], [49]. Though these measures of performance are generally accepted and followed, researchers have also proposed various other measures of performance evaluation [87].

Objective algorithms themselves are generally categorized as (1) full-reference (FR), (2) no-reference (NR) and (3) reduced-reference (RR) algorithms [82] as seen in Fig. 1.

FR algorithms are those that need access to an original 'pristine' reference stimulus to produce a quality score that predicts the subjective judgment of the distorted stimulus.

At the other end of the spectrum are NR algorithms which require only the distorted stimulus to predict quality scores.

Finally, RR algorithms lie somewhere in-between these two extremes, where the algorithm requires *some* information about the reference stimulus (eg., watermarking [8], auxillary channel [37] etc.) but not the reference stimulus itself.

In making this classification, an assumption that we have made is that 'quality' is related to the presence or absence of distortions. In the 2-D case these definitions are perfectly valid, however, when one moves on to the 3-D case, the division between quality of experience and quality due to distortion becomes ambiguous. Here, we shall make this disambiguation explicit and hence when we refer to quality alone, we imply quality due to presence of distortion; quality of experience is explicitly referred to as quality of experience. By this definition, the aforementioned databases for 3D quality [24], [25], are quality of experience databases.

Now that we have journeyed through the various terms one is bound to encounter in the area of visual quality , we beg the reader's indulgence as we paint a rough picture of what the future holds for this exciting field of visual quality assessment. We begin at 2D IQA (Section 2.1 and make our way into the spatio-temporal world with 2D VQA (Section 2.2). We then continue onto 3D QA (Section 3.1) and cover portions of what we have termed 'quality of experience' (Section 3.2), which is especially relevant in 3D perception. Further, we explore the interesting realm of visual aesthetics and its relationship to quality (Section 4), where we also briefly touch upon how content influences quality/aesthetic ratings. Finally, although this chapter is focused on 'visual' quality, in section 5, we discuss multimedia quality assessment (a.k.a. multimodal QA) and mention in passing other interesting avenues of research including QA on hand-held devices, immersive environments and so on.

Before delving into the heart of the subject, we would like to emphasize that what we describe here is simply a rough outline of areas that we think will be relevant for research. Our lack of clairvoyant abilities implies that this list is by no means comprehensive. Further, we have chosen to focus on certain areas more than others, but this in no way negates the research opportunities that may present themselves in those areas (for example RR QA [37]).

2 2D Quality Assessment

2D quality assessment of images and videos has a rich history and the reader is directed to [66], [82], [48], [65] for descriptions of various algorithms. Descriptions of quality assessment databases which provide the reference stimuli, the distorted stimuli and

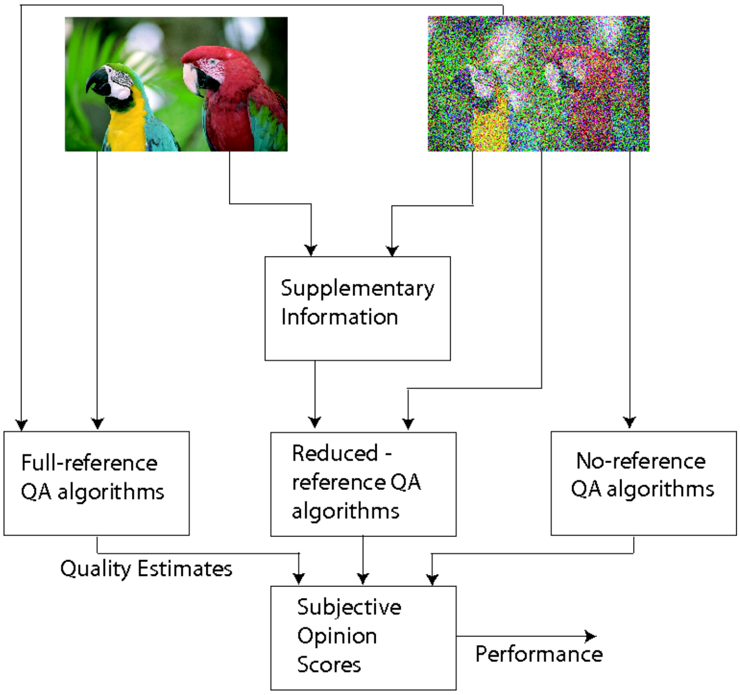


Fig. 1 Objective image quality assessment approaches: Full-reference - needs both the reference and distorted images, reduced-reference - needs the distorted image and some additional information regarding the original and no-reference - needs only the distorted image. All algorithms are evaluated for their performance using a human mean opinion score.

the associated human opinion scores in order to test algorithm performance can be found in [70], [1], [57] for images and in [49], [67] for videos. These descriptions also include performance evaluation of various leading quality assessment algorithms as well as links to the actual datasets, which are publicly available to researchers at no cost.

Our discussion on 2D QA is split into image quality assessment (IQA) and video quality assessment (VQA).

2.1 2D Image Quality

2.1.1 Full-reference algorithms

Approaches to objective 2D FR IQA include those based on explicit modeling of the HVS [27], [11], feature based approaches [73] and natural scene statistics based approaches [69]. Large scale studies analyzing algorithm performance have demonstrated that the structural similarity index (SSIM) [81] and the information-theoretic visual information fidelity (VIF) [69] index not only have statistically significant performance compared with the often criticized peak signal to noise ratio (PSNR) [23], [80], but are also the leading performers in terms of correlation with human perception compared with other approaches to IQA [69], [57]. Although many other perceptually motivated approaches to IQA have later been proposed [11], SSIM and VIF remain leading performers.

The simplicity of the SSIM index, coupled with its impressive performance has lead to a plethora of research aiming to better performance [14], [44], [35]. Although some of these techniques have demonstrated a slight improvement over SSIM, the gains achieved in performance are very small and statistically insignificant. Hence, it seems that the field of 2D FR IQA has currently plateaued. By this, we mean that additional gains achieved in performance would be minimal and the cost associated with achieving this performance would not justify the paltry gains. Having said that however, we believe that there exist many sub-fields which have not been explored as much.

One interesting direction of future research in 2D FR IQA would be efficient quality computation [58], [15]. Even though this may not be as relevant for IQA, lessons from this field could be applied to 2D VQA so as to enable real-time implementation of quality measures.

Again, although many successful approaches to QA have been proposed, only recently have these algorithms been used for benchmarking. Further, optimization of image processing algorithms using perceptual measures of visual quality as objective functions remains relatively unexplored. Some researchers have attempted to perform de-noising and linear filtering using quality indices as the optimality criterion instead of the popular mean squared error (MSE) [12], [13]. However, a lot needs to be done in this field. An IQA algorithm that correlates well with human perception could be used to perform optimal compression or filtering. Complicated models such as VIF do not lend themselves to a form that makes such operations easier, and future work would involve analyzing these measures to enable such optimal compression/filtering. These applications of QA algorithms aside, development of QA algorithms has an interesting future as well.

As we mentioned before, we believe that 2D FR IQA algorithms have achieved saturation in performance. However, it is our hope that as we understand the HVS better, QA models based on the HVS would indeed achieve significant improvement in performance. Further, investigation of eye-movements for the task of quality assessment is another exciting area for future work. Although some effort has gone towards understanding fixations and eye-movements [92], as well as their utility in quality assessment [44], [42], [7], [79], [78], [52], [46], we believe that this area has tremendous potential for creating better IQA algorithms.

In summary, we envision the future of 2D FR IQA algorithm research to draw greater inspiration from the HVS, leading to algorithms that perform the task of QA nearly as well as humans. Further, applications of such objective QA measures for various image processing tasks remain of interest.

2.1.2 No-reference algorithms

Traditionally, NR IQA algorithms have been distortion-specific. For example, there exist a plethora of algorithms that seek to assess the quality of blurred images [39], [61], [51], [93], [16], or those that assess the quality of compressed images - JPEG [41], [5], [86], JPEG2000 [56], [55], [71]. In these NR algorithms, distortion specific indicators of quality are computed and then 'mapped' onto the quality scale by using subjective opinion scores. Generally, most of these algorithms extract edge information and look for either edge-spread or distribution of edges to assess quality of blur or ringing (seen in JPEG2000 compression). In the case of JPEG compression, blockiness is measured at block-boundaries along with a measure of blur. These measures of distortion-specific indicators are then combined using a parametric form, the parameters of which are estimated using a subjective database.

Thus, although much work has gone into NR IQA, most of the work has centered around distortion-specific algorithms. Only recently have approaches that are distortion-independent been proposed in the literature [60], [45], both of which are based on the statistical properties of natural images. In [60], a DCT-domain-based approach is undertaken, where relevant features (including some inspired by [22]) are computed and then combined using a parametric approach to produce quality. On the other hand, in [45], a two-step framework for NR IQA has been proposed where the distortion-category is first identified (using techniques inspired from [88]) and then distortion-specific QA is performed. Since both these approaches are based on a train-test combination, these approaches are still not completely distortion-independent.

The astute reader would have already guessed where we think the future of NR IQA lies. Since distortion-specific QA algorithms are a dime-a-dozen, and many of these algorithms perform very well, we believe that distortion-independent NR IQA algorithms that are capable of assessing the perceived quality of *any* distorted image will be a major objective of future research; although research may continue in distortion-specific algorithms, seeking to better performance. Further, current NR IQA performance (distortion-specific or otherwise) does not match up to that of FR IQA algorithms. As with the FR case, we believe that incorporating perceptual mechanisms in NR IQA will benefit these algorithms, thereby pushing their performance to that of the theoretical null model [70].

There is an interesting aspect to IQA that generally remains un-noticed. Current algorithm performance is gauged on publicly available IQA databases. Although these databases incorporate various distortion categories[1] most of these categories tend to be singular distortions i.e., either compression or noise or blur. To the best of our knowledge there does not exist a database that specifically incorporates multiply distorted images and subjective quality scores. For example, would humans rate a noisy image (say caused by camera sensor noise) that is heavily compressed (for transmission) worse than images that are only noisy or only compressed? If yes, then how do these distortions interact with each other to affect perceived quality? Does one category of distortion mask the presence of the other? How do relative distortion strengths factor into perceived quality? Moving beyond just two distortions, one could imagine a case where the image was truly multiply distorted - eg., a noisy image, compressed and then transmitted over a lossy channel.

Although the area of multiply distorted images is an interesting one for research, not much work has been done in this direction [30], [21]. It is unclear if even leading FR IQA algorithms would correlate well with human perception in this case. Hence, creating such a multiply distorted image database accompanied by human subjective scores is an interesting area of future research. One would imagine that such databases will lead to new FR IQA algorithms that are capable of predicting perceived quality in these cases. It goes without saying that a holy grail of QA research would be an NR IQA algorithm that performs as well as (or better than) FR IQA algorithms even for such multiply distorted images.

[1] For eg., the LIVE IQA database has 5 distortion categories [70], while the TID2008 database incorporates 17 distortion classes [57]!

2.2 2D Video Quality Assessment

2.2.1 Full-reference algorithms

Simple approaches to 2D VQA involve application of a 2D IQA algorithm on a frame-by-frame basis. For example, one could compute the quality of a video using MSE or SSIM as a criterion, where MSE/SSIM is computed on each frame and the mean quality of the video is a Minkowski summation of the frame-level scores [84]. However, as many authors have pointed out, motion information is extremely important for VQA [84], [83], [64], [4], [43], [53]. Recently many successful algorithms that utilize motion information have been proposed and shown to perform well [64], [43]. A recent comprehensive study of leading FR VQA algorithms indicated that the motion-based video integrity evaluation (MOVIE) index [64] performed exceptionally well in terms of correlation with human perception, and that PSNR is one of the worst indicators of perceived quality.

Although algorithms that incorporate motion information and seek inspiration from the human visual system do well on the recently proposed LIVE video quality database [67], there is still room for improvement[2]. However, our opinion is that significant improvements can only be achieved by algorithms that employ advanced human visual system models.

As with FR IQA, efficient computation of FR VQA algorithms is an interesting direction of future research, especially when it comes to practical application of VQA algorithms. The challenge is to maintain acceptable performance in terms of correlation with human perception while still allowing for (near) real-time performance. Indeed, some researchers have already tackled this problem for the popular SSIM index [15] when applied to VQA. Apart from extending already proposed approaches for VQA, creating algorithms that are capable of real-time quality prediction remain of interest.

We discussed human attention and its use in the context of FR IQA; however, fixation prediction and human attention have even more relevance for VQA. Predicting human fixation locations accurately can help not only in efficient compression but also in developing algorithms that predict perceptual quality. In the case of FR VQA, one would imagine that extracting low-level features from the reference video would make possible improved prediction of fixations and hence of quality. Having said that however, there has been some evidence to show that distortions affect human fixations [46] and analyzing how distortions affect these fixations is another path that may be of relevance for VQA - especially in the case of NR VQA.

2.2.2 No-reference algorithms

NR VQA algorithms have not received much attention and most NR VQA algorithms are geared towards compressed videos [74], [90], [75], [91]. Generally, measures of blockiness or blur are evaluated and related to quality. Other approaches to VQA for videos transmitted through lossy networks include computing the number of packets lost, the type of packet lost and certain other features from the byte stream which are then mapped on to quality [50], [29]. A truly blind VQA algorithm that predicts the quality of any distorted video sequence is still to materialize.

[2] The best performing algorithm - MOVIE [64] has an SROCC of ~ 0.79 with human perception.

In order to achieve this goal, ideas from successful FR VQA approaches are relevant. We believe that NR VQA algorithms that incorporate perceptual models and motion information will be more likely to succeed in the task of quality assessment. Indeed, recent research on statistics of motion information [59] leads one to believe that analyzing motion from distorted videos would lead to a good indicator of quality. Further, as we discussed above, including human attentional mechanisms that are geared to predict human attention when viewing distorted videos remains of interest.

Although algorithms have been proposed for the task of NR VQA, a systematic evaluation of these algorithms, along the lines of those undertaken for FR VQA [67], [49] is missing. In order to compare algorithms in a fair manner, it is imperative that their performances be evaluated on a publicly available database using agreed-upon statistical analysis for performance evaluation. Future work could involve such a large scale comparison that can then be utilized to accurately measure the performance of an algorithm.

In the case of VQA, the temporal aspect has unique importance. Though much work has been done in terms of spatial pooling strategies for IQA [85], [44], little has been done in terms of temporal pooling strategies for VQA. Further, evaluating and incorporating memory effects for longer videos may lead to better temporal pooling strategies.

As with the IQA case, both FR and NR VQA algorithms have been rarely utilized for practical applications such as quality-based compression, or denoising [76]. We believe that the future of VQA algorithm research could revolve around such practical applications of algorithmic quality assessment. Other applications could include VQA algorithm-based rate-distortion optimization, motion-estimation etc.

3 3D Quality Assessment

Even though 3D stimuli have been around for a long time [40], and 3D movies were popular over 3 decades ago, interest in 3D stimuli had not caught on until recently. Hollywood's current fascination with 3D technology coupled with an increasing number of manufacturers churning out 3D TVs, 3D displays and 3D hand-held devices has made research in 3D quality assessment an extremely interesting area recently.

In the case of 3D, as we mentioned in the introduction, there is an added dimension to the problem of perceptual quality - the quality of 3D experience. Since the term 'quality assessment' has been traditionally applied to gauging the quality of impaired stimuli, we shall persist with that definition here. The term 'quality of experience assessment (QoE)' we shall reserve for algorithmically assessing the visuo-sensory 3D experience such as the perception of depth, naturalness of visual stimulus (i.e., avoiding the puppet theatre effect), visual comfort, ability to fuse the stereoscopic pairs and so on. Our discussion is split into 3D QA and 3D QoE assessment.

3.1 3D Quality Assessment

The field of 3D QA, although closely related to 2D QA, is a far more difficult problem to address. This is primarily because the algorithm does not have access to the perceived visuo-sensory 3D experience of the human created by the stereoscopic stimuli.

The algorithm simply has access to the stereo-pair of images and possibly associated disparity/depth (generally computed using an algorithm). The challenge then is to somehow utilize the available information and transcend this boundary between the stereo-pair and the visuo-sensory experience. Although many 3D QA algorithms have recently been proposed in literature [63], [26], [9], it is unclear if simple extensions of 2D QA measures, coupled with some disparity information, are sufficient to gauge perceived 3D quality. It is our opinion that future work in the area of 3D QA will aim to bridge this gap in order to better gauge perceived quality.

Further, at this point of time it is unclear how depth information affects/masks distortions. Although some researchers have claimed that the addition of depth information does not improve prediction quality of 3D QA algorithms [9], much work needs to be done in this area to understand stereoscopic masking and to apply this principle to 3D QA. Further, many related phenomena such as inter-ocular rivalry need to be evaluated in order to arrive at an accurate measure of quality. It is also unclear how much the perception of depth affects user ratings. For example, certain distortion categories such as noise do not destroy depth information; however other distortions such as compression/packet-loss may lead to a loss in perception of depth. In these cases, it is unclear how the subjective rating of quality is affected by such losses in depth. Psychovisual experiments need to be conducted in order to arrive at possible hypotheses.

Closely related to QA is the field of compression. For 3D QA, compression of the stereo-pair is extremely relevant, especially in low-bandwidth scenarios. There are some results which state that the human defaults to that quality score which is superior between the stereo pair [68]. This implies that one may compress one view with a much higher compression ratio than the other. Again, research needs to be undertaken in this field of asymmetric distortion - not only for compression but also for other categories of distortions.

Gauging the performance of any QA algorithm necessitates the presence of a database with human subjective scores. Although a few databases have been proposed in the literature [2], none of these databases are accompanied by precision depth-maps. The depth/disparity information is generally obtained using a disparity estimation algorithm. Since disparity estimation is still an ongoing area of research, such algorithmic disparity maps may hinder development of 3D QA algorithms. It is hence imperative that the subjective databases for 3D QA incorporate high-precision range maps along with the stereo-pair. We have recently developed such a database, and although the database may not be publicly available before this article goes to press, information about the LIVE 3D QA database may be found at [34]. This database consists of 20 reference and 400 distorted stereo pairs encompassing 5 distortion classes, accompanied by high-precision range maps obtained from a Reigel terrestrial range scanner. We believe that this database will aid researchers in developing high quality 3D QA algorithms.

The above discussion is equally valid for image and video quality assessment. However, when one moves on to 3D VQA, the interaction between motion and depth, temporal, spatial and stereoscopic masking will play an important role in the development of algorithms. We are unaware of any systematic study that attempts to understand these interactions and we believe that there is ample scope for research in this area.

3.2 3D Quality of Experience

Although 3D visual stimuli have come a long way since the 1960's, even today 3D viewing for extended periods of time can be accompanied by visual strain and discomfort [20], [40]. Understanding what causes this strain and how capture and display technologies may be modified to alleviate such strain is an interesting area of research. Such strain may be due to the actual stimuli or the rendering process (think cross-talk for example), causing an unnatural viewing experience, which reduces the quality of experience. Designing algorithms which can predict the human 3D experience and provide an index onto eye-strain or discomfort during viewing will be an invaluable asset to developers of 3D content. As with all visual quality, such algorithms will benefit from a better understanding of human stereo processing and advances in psychovisual research will help develop these QoE algorithms.

As with 3D QA, QoE requires databases, and researchers at EPFL have recently proposed two such databases - one for images [25], and the other for videos [24]. Although small, these databases represent a first step toward understanding 3D quality of experience and are the foundation for many such databases seeking to assess perceived 3D experience.

It is pertinent to note that 3D QoE is a blind problem - there is no reference stimulus against which the QoE can be measured. In this case, understanding the statistical properties of 3D natural environments and human attention in such environments may benefit QoE algorithms.

Finally, judging the overall quality of a (possibly distorted) 3D stimulus, would involve a confluence of lessons from 3D QA and 3D QoE assessment and would again be an interesting direction to pursue research.

4 Aesthetics, Quality and Content

Visual aesthetics is a measure of the perceived beauty of a visual stimulus and in the recent past, there has been some interest in algorithmically evaluating this perceived beauty [36], [17], [31], [38], [47]. Objective assessment of aesthetics is naturally 'blind' - i.e., given an image to be evaluated, there does not exist a corresponding image with 'perfect' aesthetics. Of course, this does not nullify the possibility of comparing the aesthetic appeal of two images. The goal of algorithmic aesthetic assessment then is to predict aesthetic scores (similar to quality scores) having a high correlation with human perception.

'Aesthetics' by itself is highly subjective and hence objective assessment of aesthetics is a far more difficult problem than that of quality assessment. Having said that however, recent approaches to this task have demonstrated appreciable success [17], [31], [38], [47]. The general flow of such objective assessment is to extract a series of features that are intuitive and/or borrowed from literature on photography and/or obtained from user studies. These features generally include low-level measures such as exposure (luminance distribution), contrast, colorfulness, color saturation and some ephemeral features such as the rule of thirds. In case of videos, additional motion based features are computed as well [38], [47]. Once these features are extracted from the stimulus, one can perform a regression onto subjective scores from datasets. However, as Datta et. al. point out, the subjectivity of the task makes such regression perform below par [18]. Hence, the current approach seems to be classifying stimuli into 'good' or 'bad' aesthetics. Even though this area has seen some progress, there are many avenues that need to be explored.

Currently, databases for image appeal have not been explicitly created and no large scale subjective study (such as [70]) has been conducted for image appeal. An exception is the one in [62], though no algorithmic assessment is performed there; further the database is not publicly available. The databases that are used for image appeal consist of images download from popular photo sharing websites such a Flickr or Photo.net [18]. Apart from the image itself, many of the websites provide a rating for the photo which is based on the scores given by users who have viewed the images at their homes. Since the scores are obtained from a variety of individuals who have viewed the image on a wide-range of (possibly uncalibrated) monitors with different screen resolutions, viewing distances and viewing conditions, the noise associated with these scores is very high [3]. Further, preference of a particular photographer and the relationship of the subject to the subject of the photograph as well as the photographer influences ratings. It is notable that most of the images that receive high aesthetic ratings are those of female nudes. Since there is little control on the source images, such a skew is natural. Thus, even though these datasets are large ([31] consists of a total of ≈ 6000 images, and [17] consists of a total of ≈ 3500 images) and are available publicly with average appeal ratings, these datasets are quite noisy.

In [18], the authors define the tasks that are of interest in the context of objective aesthetic prediction, discuss strategies for algorithm development and describe some datasets for performance evaluation. As the authors point out, datasets created by crawling websites are generally used since there is 'lack of theoretical grounding and controlled experimental data'. Although Photo.net has an aesthetic rating scale, our previous arguments regarding availability and noise hold for all the datasets analyzed in [18]. The authors conclude that 'a large, low-noise dataset based on controlled user studies will be a welcome addition'.

It should be clear that for image aesthetics, even though there appears to be a general agreement on the inherent subjectivity of the task and hence the prevalence of classification as against regression/correlation analysis; there does not exist a standard, controlled dataset with subjective ratings for performance evaluation. Further, the lack of agreed upon standard evaluation methods, such as those for IQA, implies that different researchers utilize different methods to measure performance on different datasets, making the task of comparing algorithms even more difficult. The reliance on datasets compiled by crawling photo-sharing websites adds noise to the provided ratings, thus rendering the ratings suspect. One important aspect of such crawled images is image quality.

Since many of the photo sharing websites have an upper limit on the size of the file that can be uploaded, most images on such websites have undergone compression prior to upload. This is true not only for images from consumer digital cameras but also those from professional SLRs which allow for images to be captured in an uncompressed format. For example, Fig. 2 (a) shows an image from the dataset in [31] - notice blocking artifacts which are visible at the top-left of the image. Compression at any stage of the image processing pipeline can cause distortion, and there are other image transformations that can introduce distortions altering the quality of the image, including those from the acquisition process. It is obvious that such variations in quality will affect subjective ratings. Further, the crawled datasets consist of artificially created images, computer-generated effects, images with artificially inserted borders or even those that have been heavily processed after acquisition - see Fig. 2

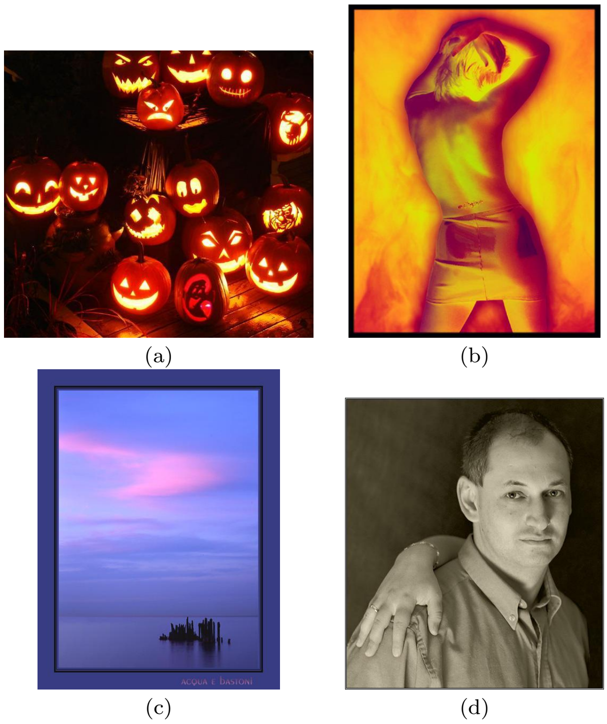
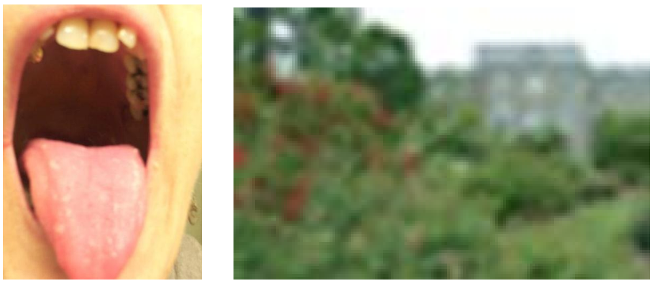


Fig. 2 Images with high aesthetic appeal: (a) compression artifacts, (b) computer-generated effects, (c) presence of border and (d) post-processed

(b)-(d) (images from the dataset in [17]). Some may argue that a good aesthetic prediction/classification algorithm must be able to predict the subjective appeal of such images as well; however, owing to the nascence of the field and the highly subjective nature of the task at hand, we believe that we must first endeavor to create algorithms that seek to predict the aesthetics of *natural* images, and once that problem has been surmounted, tasks of additional complexity may be added. This is due to the fact that natural images possess certain statistical regularities which may be modeled, thus making this seemingly insurmountable problem better-defined.

Finally, we note that the area of video aesthetics has seen little research and includes only the approaches in [38] and [47]. In [38], videos from YouTube were utilized for the study, and it is unclear how subjective ratings were obtained. In [47], the authors list the drawbacks of the approach in [38] and conduct a subjective study in a controlled environment to gauge subjective opinion on video aesthetics. Even though the proposed database in small in comparison with those for images and that proposed in [38] or [17], the approach in [47] represents the first systematic attempt in trying to gauge subjective opinion on video aesthetics.

Now that the reader has a taste for how quality assessment and aesthetics assessment is conducted, we would like to draw the reader's attention to an important aspect of subjective assessment which we have conveniently glossed over - the content of the visual stimuli. Currently, all controlled approaches at gauging quality or aesthetics require the user to simply rate the quality or the aesthetics of the stimulus without



(a) (b)

Fig. 3 Images with 'bad' aesthetic rating from [31]: (a)it is unclear if the poor aesthetic rating is influenced by the content and (b) it is unclear if the poor rating corresponds to poor quality or poor aesthetics

considering the actual stimulus content. Clearly, content of the stimulus will play an important role in the received ratings, and recently researchers have started to evaluate the effect of content on such subjective ratings [33]. However, a lot is left to be done.

Although the three categories for measurement of stimulus appeal (quality, aesthetics and content) could be disambiguated to some degree in theory, such disambiguation is lost when the user is asked to rate only one aspect of image appeal during the creation of databases to measure quality or aesthetics or content. In order to provide for accurate measurement of algorithm performance, it is clear that ratings from the subject that are used to measure correlation/classification accuracy must be primarily about the variable being assessed. We realize that completely disjoint ratings can never be obtained from human observers, however, a simple technique can be used in order to reduce the ambiguity.

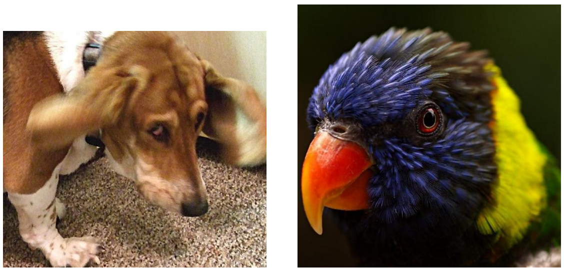
We propose that future human studies for visual appeal should allow for ratings on all the three scales - quality, aesthetics and content. Even though the final goal of the researcher may be quality assessment and not aesthetics assessment, such clear definition of the three aspects of visual appeal during the course of the subjective study will provide reliable ratings for each category. For example, Fig. 3 (a) shows an image from the database in [31], which is rated as aesthetically 'bad' by human observers. It is unclear if the rating is due to the nature of the content or truly because of the aesthetics. Fig. 3 (b) shows an image with bad aesthetics from [31], again, the image is of poor quality[3], and it is unclear if the subjective rating provided is for the quality or for the aesthetic appeal of the image. The proposed setup which requires ratings on all three scales will help to reduce such ambiguities.

As an example, consider Fig. 4 (a) which shows an image which may be considered to have high aesthetic rating, Fig.4 (b) is a noisy version of the same image - this image still should have high aesthetic rating, but a poor quality rating. In both these images, the content is simply a shack in the middle of a plain - for one of the authors of this chapter, this represents poor content; on the other hand, another author is interested in architecture and considers the content to be good. Now consider Fig. 5 (a) - not only does this image may have poor aesthetic value, but is also of poor quality (due to blur); however for the owner of the dog, the content may be perceived as being

[3] We define blurry images as those with poor quality, however blur can also be associated with positive aesthetics, such as an image of a softened, wrinkle-free face - future work needs to disambiguate such cases.

(a) (b)

Fig. 4 (a) An image with high aesthetic appeal and good quality. (b) The same image with high appeal but with poor quality.



(a) (b)

Fig. 5 (a) An image with poor aesthetics and poor quality, but possibly good content (for the dog owner for example). (b) An image with high aesthetic, quality and (possibly) content rating.

good. Finally, Fig. 5 (b) shows an image which for a majority of us has high aesthetic, quality and content value. Thus, utilizing the proposed 3-scale approach for visual appeal assessment - no matter what the goal of the algorithm may be - will allow for clear and relatively unambiguous rating for each of the three categories.

In the future, for all the three categories - quality, aesthetics, content - the goal of the algorithm should be to predict the subjective opinion score rather than binary classification. Such prediction will allow for systematic performance evaluation using techniques such as SROCC, LCC, RMSE and so on. Due to the highly subjective nature of visual content assessment, one could imagine algorithms that are personalized, in the sense that the algorithmic scores are valid for each subject individually. As the authors in [18] note, such personalization - either at the individual level, or at the *clique* level will be of interest for visual aesthetic assessment as well. Quality assessment may also benefit from such personalization - albeit to a lesser degree. Finally, even if the source images are compressed in nature (for example those obtained from photo-sharing sites), the subject now has the liberty to provide a high aesthetic rating to an image with poor quality (due to blocking, for example).

A large scale human study undertaken as described above will prove extremely useful to researchers who seek to develop algorithms to predict visual appeal. We believe that there is a dearth of work in the area of aesthetics and content influence on subjective ratings, and one important research problem would be objective prediction of visual appeal - where appeal includes quality, content and aesthetics.

5 Other Directions

Hopefully, by now, we have convinced the reader that the field of QA is vast and there exist a plethora of research opportunities - images, videos, 3D stimuli, art, aesthetics etc. We have tried to incorporate as many diverse fields as we could, however, as always there are many topics that we have failed to analyze in detail. Our discussion has mainly focused on areas which have received sizeable interest in recent times. In this section, we quickly skim over some of the topics that we failed to cover in details, but believe are relevant. This is an opportune moment to point out that this list is by no means comprehensive, indeed, it is not meant to be.

Advancing technologies have made high-definition (HD) televisions affordable, and in today's market HDTVs are a dime-a-dozen. Given their rising popularity and acceptance, it is no wonder that HD content is growing by the day as well. With the capability to now stream HD videos over the Internet onto HD-compatible displays at home or on the move, we come back to the bandwidth-quality compromise. Enter visual quality assessment algorithms. The case of large-screen displays is especially interesting due to its practical relevance and increasing bandwidth costs. Deciphering human attentional mechanisms would play an important role in HD quality assessment. Recently, this area has received some attention [10], however, a lot more needs to be done. Understanding user opinions on quality and user expectation is an important step in this process [32]. Utilizing the information from such studies in developing quality assessment algorithms remains of interest.

Another area that has received little attention is quality assessment in immersive environments - panoramic 3D displays, immersive dome-displays and so on. When a user is completely immersed in the audio-visual environment, judging the quality of the stimulus as well as the quality of experience will be a challenging task. The perceptual challenges in this case are similar to those in 3D QA/QoE assessment, where the algorithm needs to re-create the experienced environment. When one adds other modalities, such as auditory cues, tactile inputs, etc., the problem becomes gargantuan in complexity and hence tremendously attractive for researchers!

Moving away from large screen displays and immersive environments, the other end of the spectrum accommodates an increasingly important market - that of handheld devices. With the recent spurt in demand for smartphones, tablets and other handheld devices and improving wireless content-delivery channels, entertainment on-the-go is currently a commercial hotbed. With users demanding high quality entertainment and service providers crippled by the increasing amount of dataflow[4], maximizing quality of experience while bounded by bandwidth is an increasing relevant area. Development of QA algorithms for mobile devices is hence of significance, with recent in-roads being made in this field as well [89]. QA on mobile devices becomes even more interesting given the fact that, in the near future 3D on handheld devices will cease to simply be a possibility [5]. As one would imagine, the challenges associated with this field are numerous and include viewing distance, viewing angles, viewing conditions and so on. Predicting quality and quality of experience in these cases will surely be a challenging task.

---

[4] Indeed, AT&T recently announced a cap on the bandwidth that users could utilize on their 3G-enabled phones, and other service providers may follow suit [19].

[5] Recently Nintendo announced the Nintendo DS 3D gaming device, which utilizes an autostereoscopic display [54].

5.1 Multimodal Quality Assessment

In this writeup, we have extensively discussed visual quality assessment. This is partially due to the fact that our research interests lie this area. However, in terms of providing automated quality of experience indicies, which can be used for quality control or efficient compression, we have neglected an extremely important dimension - that of sound. In the case of videos, a wholesome experience involves not only the visuals but also the accompanying audio. Wholistic quality of experience measures must hence take into account the audio-visual quality of the stimulus; and this audio-visual quality is referred to as multimodal or multimedia quality assessment.

Multimodal quality assessment is not an easy problem to solve, since interaction between the visual and audio cues need to be understood in order to predict quality of experience. Currently, there seems to be a small body of work in this area of multimodal quality assessment [6]. It is our belief that as visual quality assessment measures improve, the area of multimedia quality assessment will capture the imagination of researchers. Assessing the quality of the audio alone has been of interest, with some researchers applying successful IQA algorithms for this purpose [28]. Once accurate measures of audio quality are developed, the combination of visual and audio quality measures will be an interesting avenue for research. Quality of experience may also involve measuring synchronization between the audio and video streams and modeling the effect that delays have on audio-visual perception.

6 Conclusion

In this paper we attempted to paint a picture of tomorrow for researchers in the area of quality assessment of visual stimuli. We described our views on current research in the area of image and video quality assessment and chiseled a rough road toward what we believe will make for good research in the near future. Our general philosophy through this writeup has been that of cautious optimism - we believe that a lot more needs to be done in the area of quality assessment and that each of the subfields described here (as well as those that were not) have great potential for growth. Our stance through the writeup has been that algorithmic quality assessment will heavily benefit from research in visual psychophysics. As we understand the human visual system better, quality assessment algorithms that incorporate these mechanisms will surely result in better performance.

Apart from quality assessment, we touched upon some esoteric topics such as aesthetics assessment and the effect that scene content can have on user opinion. The currently popular 3D quality assessment arena was explored as well. Finally, we had some opinions on how multi-modal (multimedia) quality assessment research would progress in the near future.

A famous African proverb goes ‘Tomorrow belongs to the people who prepare for it today’ and we hope that by setting our view of the future in print we have at least started preparing for this journey and that we have given the reader an idea of what the future of quality assessment research holds.

References

1. Toyama image database. *http://mict.eng.u-toyama.ac.jp/mict/index2.html.*

2. B. Alexandre, P. Le Callet, C. Patrizio, and C. Romain. Quality Assessment of Stereoscopic Images. *EURASIP Journal on Image and Video Processing*, 2009.
3. X. Amatriain, J. M. Pujol, and N. Oliver. I Like It, I Like It Not. *Proceedings Int. Conf. UMAP'09*, 2009.
4. M. Barkowsky, B. Eskofier J. Bialkowski and, R. Bitto, and A. Kaup. Temporal trajectory aware video quality measure. *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3(2):266–279, April 2009.
5. R. Barland and A. Saadane. A reference free quality metric for compressed images. *Proc. of 2nd Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2006.
6. J. G. Beerends and F.E. De Caluwe. The influence of video quality on perceived audio quality and vice versa. *Jnl. Aud. Engg. Soc.*, 47(5):355–362, 1999.
7. P. Le Callet, S. Perchard, S. Tourancheau, A. Ninassi, and D. Barba. Towards the next generation of video and image quality metrics: Impact of display, resolution, content and visual attention in subjective assessment. *Second International Workshop on Image media Quality and its Applications*, 83(73.05):10–51.
8. P. Campisi, M. Carli, G. Giunta, and A. Neri. Blind quality assessment system for multimedia communications using tracing watermarking. *IEEE Transactions on Signal Processing*, 51(4):996–1002, 2003.
9. P. Campisi, P. Le Callet, and E. Marini. Stereoscopic images quality assessment. In *Proceedings of 15th European Signal Processing Conference (EUSIPCO07)*, 2007.
10. G.W. Cermak. Consumer Opinions About Frequency of Artifacts in Digital Video. *IEEE Journal on Selected Topics in Signal Processing*, 3(2):336–343, 2009.
11. D. M. Chandler and S. S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, 2007.
12. S. S. Channappayya, A. C. Bovik, and R. W. Heath. A linear estimator optimized for the structural similarity index and its application to image denoising. *IEEE International Conference on Image Processing*, pages 2637–2640, 2006.
13. S.S. Channappayya, A.C. Bovik, C. Caramanis, and R.W. Heath. Design of linear equalizers optimized for the structural similarity index. *IEEE transactions on image processing*, 17(6):857–872, 2008.
14. G.H. Chen, C.L. Yang, and S.L. Xie. Gradient-based structural similarity for image quality assessment. *IEEE International Conference on Image Processing*, pages 2929–2932, 2006.
15. M. J. Chen and A. C. Bovik. Fast structural similarity index. *IEEE International Conferene on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.
16. M.J. Chen and A.C. Bovik. No reference Image Blur Assessment Using Multiscale Gradient. *1st International Workshop on Quality of Multimedia Experience (QoMEX)*, 2009.
17. R. Datta, D. Joshi, J. Li, and J.Z. Wang. Studying aesthetics in photographic images using a computational approach. *Lec. Notes. in Comp. Sci.*, 3953:288, 2006.
18. R. Datta, J. Li, and J.Z. Wang. Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition. *IEEE Intl. Conf. Image Proc.*, pages 105–108, 2008.
19. Larry Dignan. Tiered mobile data plans accelerate: Verizon Wireless to follow AT&T's lead. *ZDnet.com*, 2010.
20. Roger Ebert. Why I Hate 3-D (And You Should Too). *Newsweek*, May 2010.
21. M.C.Q. Farias, M.S. Moore, J.M. Foley, and S.K. Mitra. Perceptual contributions of blocky, blurry, and fuzzy impairments to overall annoyance. *Proceedings of the IS&T/SPIE Human Vision and Electronic Imaging IX*, 5292:109–120, 2004.
22. S. Gabarda and G. Cristobal. Blind image quality assessment through anisotropy. *J. Opt. Soc. Am. A*, 24:B42–B51, 2007.
23. B. Girod. What's wrong with mean-squared error?, Digital images and human vision, A. B. Watson, Ed. pages 207–220, 1993.
24. L. Goldmann, F. De Simone, and T. Ebrahimi. A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video. *Electronic Imaging (EI), 3D Image Processing (3DIP) and Applications*, 2010.
25. L. Goldmann, F. De Simone, and T. Ebrahimi. Impact of acquisition distortions on the quality of stereoscopic images. *Electronic Imaging (EI), 3D Image Processing (3DIP) and Applications*, 2010.
26. P. Gorley and N. Holliman. Stereoscopic image quality metrics and compression. *Proc. SPIE Stereoscopic Displays and Applications XIX*, 6803, 2008.
27. A. P. Hekstra, J. G. Beerends, D. Ledermann, F. E. De Caluwe, S. Kohler, R. H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss. PVQM–a perceptual video quality measure. *Signal Processing: Image Communication*, 17(10):781–798, 2002.

28. S. Kandadai, J. Hardin, and C. D. Creusere. Audio quality assessment using the mean structural similarity measure. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, pages 221–224, 2008.
29. S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan. Modeling packet-loss visibility in MPEG-2 video. *IEEE transactions on Multimedia*, 8(2):341–355, 2006.
30. V. Kayargadde and J.B. Martens. Perceptual characterization of images degraded by blur and noise: model. *Journal of the Optical Society of America A*, 13:1178–1188, 1996.
31. Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. *IEEE Conf. Comp. Vis. Pat. Recog.*, 1, 2006.
32. C. Keimel, T. Oelbaum, and K. Diepold. No-Reference Video Quality Evaluation for High-Definition Video. In *Proceedings of the International Conference on Image Processing, San Diego, CA, USA*, 2009.
33. P. Kortum and M. Sullivan. The Effect of Content Desirability on Subjective Video Quality Ratings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, May 2010.
34. Laboratory for Image and Video Engineering (LIVE). Quality Assessment Research at LIVE. *http://live.ece.utexas.edu/research/quality/*, 2010.
35. C. Li and A. C. Bovik. Three-Component Weighted Structural Similarity Index. *Proceedings of SPIE*, 7242:72420Q, 2009.
36. C. Li and T. Chen. Aesthetic Visual Quality Assessment of Paintings. *IEEE Jnl. Sel. Top. Sig. Proc.*, 3(2):236–252, 2009.
37. Q. Li and Z. Wang. Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation. *IEEE J. Selected Topics in Signal Proc.*, 3(2):202–211, 2009.
38. Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. *Eur. Conf. Comp. Vis.*, pages 386–399, 2008.
39. P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. Perceptual blur and ringing metrics: Application to JPEG2000. *Sig. Proc.: Image Comm.*, 19(2):163–172, 2004.
40. L. M. J. Meesters, W. A. IJsselsteijn, and P. J. H. Seuntiens. A survey of perceptual evaluations and requirements of three-dimensional TV. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(3):381–391, 2004.
41. L. M. J. Meesters and J. B. Martens. A single-ended blockiness measure for JPEG-coded images. *Signal Processing*, 82(3):369–387, 2002.
42. K. Miyata, M. Saito, N. Tsumura, H. Haneishi, and Y. Miyake. Eye movement analysis and its application to evaluation of image quality. In *Final Program and Proceedings of the Fifth IS&TSID Color Imaging Conference. Color Science, Systems and Applications*, pages 116–119, 1997.
43. A. K. Moorthy and A. C. Bovik. A motion-compensated approach to video quality assessment. *Proc. IEEE Asilomar Conference on Signals, Systems and Computers*, 2009.
44. A. K. Moorthy and A. C. Bovik. Visual importance pooling for image quality assessment. *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3(2):193–201, April 2009.
45. A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(2):587–599, May 2010.
46. A. K. Moorthy, W. S. Geisler, and A. C. Bovik. Evaluating the task dependence of eye movements for compressed videos. *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2010.
47. A. K. Moorthy, P. Obrador, N. Oliver, and A. C. Bovik. Towards computational models of visual aesthetic appeal of consumer videos. *European Conference on Computer Vision (ECCV)*, 2010.
48. A. K. Moorthy, K. Seshadrinathan, and A. C. Bovik. *Digital Video Quality Assessment Algorithms*. Springer, 2009.
49. A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik. Wireless video quality assessment: A study of subjective scores and objective algorithms'. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(4):513–516, April 2010.
50. M. Naccari, M. Tagliasacchi, F. Pereira, and S. Tubaro. No-reference modeling of the channel induced distortion at the decoder for H. 264/AVC video coding. In *Proceedings of the International Conference on Image Processing, San Diego, CA, USA*, 2008.
51. N.D. Narvekar and L.J. Karam. A No-refernece Perceptual Image Sharpness Metric based on Probability of Blur Detection. *1st International Workshop on Quality of Multimedia Experience (QoMEX)*, 2009.

52. A. Ninassi, O. Le Meur, P. Le Callet, D. Barba, and A. Tirel. Task Impact on the Visual Attention in Subjective Image Quality Assessment. In *The 14th European Signal Processing Conference*, 2006.
53. A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba. Considering temporal variations of spatial visual distortions in video quality assessment. *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3(2):253–265, April 2009.
54. Nintendo. Nintendo introduces glasses-free 3d gaming. *http://www.nintendo.com/whatsnew/detail/ MvsPLa99GTbqTKaOvwh93lOJHhSHJ_x2*.
55. E. P. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, and L. Jiang. No-reference JPEG-2000 image quality metric. In *International Conference on Multimedia and Expo*, volume 1, pages 6–9, 2003.
56. ZM Parvez Sazzad, Y. Kawayoke, and Y. Horita. No reference image quality assessment for JPEG2000 based on spatial features. *Signal Processing: Image Communication*, 2008.
57. N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti. Tampere image database. *http://www.ponomarenko.info/tid2008.htm*, 2008.
58. D.M. Rouse and S.S. Hemami. Understanding and simplifying the structural similarity metric. *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, pages 1188–1191, 2008.
59. M. A. Saad and A. C. Bovik. Natural motion statistics for no-reference video quality assessment. *International Workshop on Quality of Multimedia Experience*, pages 163–167, 2009.
60. M. A. Saad, A. C. Bovik, and C. Charrier. A DCT Statistics-based Blind Image Quality Index. *IEEE Signal Processing Letters*, 17(6):583–586, 2010.
61. N.G. Sadaka, L.J. Karam, R. Ferzli, and G.P. Abousleman. A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling. In *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, pages 369–372, 2008.
62. A. E. Savakis, S. P. Etz, and A. C. Loui. Evaluation of image appeal in consumer photography. *SPIE Proc., Human Vis. Elec. Img.*, pages 111–121, 2000.
63. Z. M. P. Sazzad, Y. Kawayoke, and Y. Horita. Spatial features based no reference image quality assessment for JPEG2000. *IEEE Int'l Conf. Image Proc.*, 3:517–520, 2007.
64. K. Seshadrinathan. *Video quality assessment based on motion models*. PhD thesis, The University of Texas at Austin, 2008.
65. K. Seshadrinathan and A. C. Bovik. *Video Quality Assessment in The Essential Guide to Video Processing*, chapter 14. Academic Press, 2009.
66. K. Seshadrinathan, R. J. Safranek, J. Chen, T. N. Pappas, H. R. Sheikh, E. P. Simoncelli, Z. Wang, and A. C. Bovik. *Image quality assessment in The Essential Guide to Image Processing*, chapter 20. Academic Press, 2009.
67. K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, June 2010.
68. P. Seuntiens, L. M. J. Meesters, and W. Ijsselsteijn. Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation. *ACM Transactions on Applied Perception (TAP)*, 3(2):109, 2006.
69. H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.
70. H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Tran. Image Proc.*, 15(11):3440–3451, 2006.
71. H.R. Sheikh, A.C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: JPEG 2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005.
72. D. Sheskin. *Handbook of parametric and nonparametric statistical procedures*. CRC Press, 2004.
73. A. Shnayderman, A. Gusev, and A.M. Eskicioglu. A multidimensional image quality measure using singular value decomposition. 5294:82–92, 2003.
74. O. Sugimoto, R. Kawada, M. Wada, and S. Matsumoto. Objective measurement scheme for perceived picture quality degradation caused by MPEG encoding without any reference pictures. *Proceedings of SPIE*, 4310:932, 2000.
75. S. Suthaharan. Perceptual quality metric for digital video coding. *Electronics Letters*, 39(5):431–433, 2003.

76. G. Varghese and Z. Wang. Video denoising based on a spatiotemporal gaussian scale mixture model. *IEEE Transactions on Circuits and Systems on Video Technology*, (to appear).
77. Video Quality Experts Group (VQEG). Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment phase I. *http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI*, 2000.
78. C. T. Vu, E. C. Larson, and D. M. Chandler. Visual Fixation Patterns when Judging Image Quality: Effects of Distortion Type, Amount, and Subject Experience. In *IEEE Southwest Symposium on Image Analysis and Interpretation, 2008. SSIAI 2008*, pages 73–76, 2008.
79. T. Vuori and M. Olkkonen. The effect of image sharpness on quantitative eye movement data and on image quality evaluation while viewing natural images. In *Proceedings of SPIE*, volume 6059, page 605903, 2006.
80. Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? - a new look at fidelity measures. IEEE Signal Processing Magazine, January 2009.
81. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error measurement to structural similarity. *IEEE Signal Processing Letters*, 13(4):600–612, April 2004.
82. Z. Wang and A.C. Bovik. *Modern Image Quality Assessment*, volume 2. Morgan & Claypool Publishers, 2006.
83. Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *Journal of the Optical Society of America*, 24(12):B61–B69, December 2007.
84. Z. Wang, L. Lu, and A. C. Bovik. Video quality assesssment based on structural distortion measurement. *Signal Processing: Image communication*, (2):121–132, February 2004.
85. Z. Wang and X. Shang. Spatial pooling strategies for perceptual image quality assessment. *IEEE international conference on Image Processing*, September 2006.
86. Z. Wang, H. R. Sheikh, and A. C. Bovik. No-reference perceptual quality assessment of JPEG compressed images. *IEEE Int'l Conf. Image Proc.*, 1:477–480, 2002.
87. Z. Wang and E. P. Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):1–13, September 2008.
88. Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E. H. Yang, and A. C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15(6):1680–1689, 2006.
89. S. Winkler and F. Dufaux. Video quality evaluation for mobile applications. *Proc. of SPIE Conference on Visual Communications and Image Processing, Lugano, Switzerland*, 5150:593–603, 2003.
90. S. Winkler, A. Sharma, and D. McNally. Perceptual video quality and blockiness metrics for multimedia streaming applications. In *Proceedings of the International Symposium on Wireless Personal Multimedia Communications*, pages 547–552, 2001.
91. K. C. Yang, C-C. Guest, K. El-Maleh, and P. K. Das. Perceptual temporal quality metric for compressed video. *IEEE Transactions on Multimedia*, 9(7):1528–1535, 2007.
92. A. L. Yarbus. *Eye Movements and Vision.* Plenum press, 1967.
93. X. Zhu and P. Milanfar. A no-reference sharpness metric sensitive to blur and noise. In *1st International Workshop on Quality of Multimedia Experience (QoMEX)*, 2009.