Blind Image Quality Assessment: Unanswered Questions and Future Directions in the Light of Consumers Needs

Michele A. Saad, Patrick Le Callet and Philip Corriveau

Motivation

Are proposed no-reference models accurate enough to be standardized for all use cases? What remains to be solved?

This past decade has seen significant progress in the field of image and video quality assessment. While full- and reduced-reference models (for images and videos), which emerged earlier than blind/no-reference ones, have managed to achieve significant quality prediction accuracy as measured by correlations with subjective quality ratings, there is still much progress to be made within the no-reference realm. In addition, the use cases covered by most standardization efforts are largely related to the content delivery chain, excluding acquisition and enhancement issues, and focusing more on compression or transmission impairments. The industry has been demanding the move towards blind assessment with the hope of being unshackled from requiring a reference. With the overwhelming ecosystem that now supports acquisition and consumption of media on a myriad of devices and context (e.g. viewing conditions) this move becomes even more urgent.

Indeed, some promising approaches to blind quality assessment have been proposed. These methods include, but are not limited to, LBIQ [Tang H, (2011)], CBIQ [Ye P. et al. (2011)], BLIINDS-II [Saad M. et al. (2012)], and NIQE [Mittal A. et al. (2013)]. These methods perform well on the databases on which they have been developed and on similar types of

images and distortions. The generalizability of their performance on many types of consumer images breaks however, for understandable reasons. These methods were developed and designed to achieve competitive performance on existing databases such as [Ponomarenko N. et al. (2013), Sheikh et al. (2005)], which are designed for quality assessment research; they should not be expected to perform well on images that are significantly different from images in those databases. These databases however, do not contain many of the distortions that are expected in many of the rapidly increasing consumer devices (most notably mobile devices as well as compact and higher end cameras). These databases that are the most widely used for algorithm design do not describe many of the very popular consumer usage models—the millions of images captured by mobile devices, for instance. Images and videos from consumer devices contain multiple distortions that are very complex in nature. These distortions may be a result of the optical system, the post processing that happens in the devices after the signal is captured, and the storage and display of the data. Simulating these distortions collectively is an extremely challenging task, and it would be necessary to create a comprehensive corpus of image distortions if one is to design algorithms for these types of images. Creating this database faces hurdles in and of itself due to privacy and sharing rights, and the requirement for a constantly new pool of test content to validate new models.

Existing subjective testing datasets for quality assessment are not suitable to validate NR models that test consumer devices.

Unanswered Questions

Going beyond fidelity: revising the methodologies

In full- and reduced-reference problems the question addressed is essentially that of fidelity: how close is a test image/video to a reference one. In blind assessment, on the other hand, prior to predicting quality, one needs to define what "better quality" is. This is all the more critical when

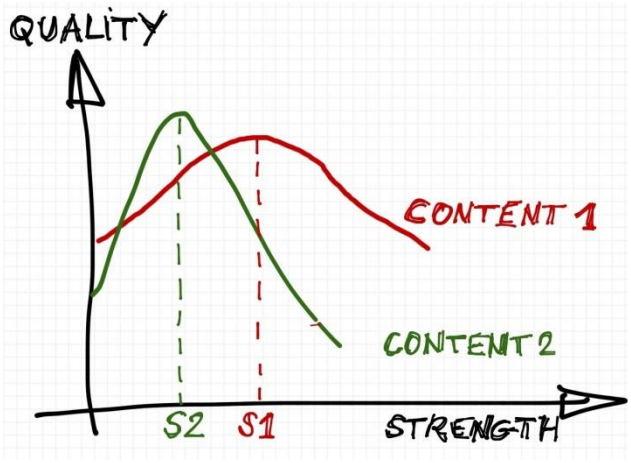

Figure 1. **The overshoot effect**: adjusting the strength for an enhancement processing (e.g. sharpening) may lead to quality improvement until a point that overshoot is reached, leading to a decrease of quality as the strength of the processing increase. Optimal strength is hard to estimate as it is often content dependent, as illustrated in this plot.

evaluating the effects of post processing, such as image enhancement, that should lead to an improvement in perceptual quality over the original image. Major challenges lie in trying to solve this problem, the primary one being that of content-dependency. No-reference predicted scores tend to be biased by content. Two images or videos of similar qualities should ideally get similar scores even if the content is highly different (for instance a low frequency sky scene versus a high frequency forest scene). The overshoot effect (see Figure 1) is another issue that needs to be accounted for by a blind quality evaluator. However, the decoupling of content from quality is a very challenging issue in blind quality assessment that still needs to be resolved.

Further, while certain methods can achieve relatively high correlations on databases that span a wide range of qualities from extremely bad (not necessarily always representative of what consumers encounter in real life) to excellent, how well proposed methods perform on a narrower range of qualities (typically a range in the higher quality end) is important for more realistic predictions on consumer content. This is illustrated in Figure 2, and is referred to as "the range effect". Addressing this issue might require revisiting subjective testing methodologies for NR model evaluation. For instance, usual subjective test methodologies such as ACR or DSCQS require a very large number of observers per condition before exhibiting statistically significant differences. Pair comparison methods might be good alternatives to achieve better sensitivity.

In full- and reduced-reference problems, the question addressed is essentially that of fidelity: how close is a test image/video to a reference one. In blind assessment on the other hand, prior to predicting quality, one needs to define what is meant by pristine or perfect quality.

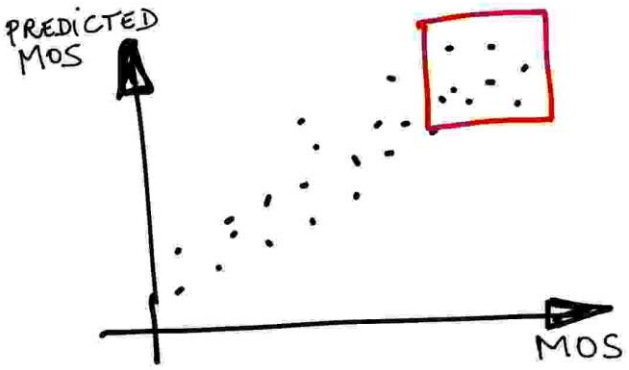

Figure 2. **The range effect**: on the overall quality range, MOS and predicted MOS seem well correlated; focusing on a particular range (e.g .points within red box), the correlation is lower.

Color is another domain where perceptual modeling for the purpose of quality assessment is lacking. The complexity of the human visual system's processing of color information has made understanding the effect of color aberrations (as opposed to only structural ones) difficult to model and predict in the no-reference quality prediction space.

Towards user profile

When it comes to image capture, another important factor has to be taken into account: the intention of the image taker. Blur, for instance, which is typically considered a distortion in image quality assessment, is often introduced on purpose by professional photographers. The distinction between artistic and undesirable effects of blur is a higher level problem that needs to be better understood and modeled. This also applies to other types of artistic effects such as film grain and motion blur.

What makes a good picture is highly subjective and very unique. Trying to capture all these effects through the prism of "king MOS" or a general quality metric may lead to a "grey car effect" (a situation in consumer science where simply averaging opinions may lead to a trade-off (but faulty) conclusion that only grey cars should to be produced since this averages out preference for black and white cars!) With this in mind, one might consider blind image quality tools tuned to specific user profiles or needs: one could imagine parameterized measures instead of general agnostic tools.

These are just a few of the issues pertaining to images. All of the mentioned "unanswered questions" hold for blind video quality assessment. Video on the other hand, exponentially increases the complexity of the search space. Motion modeling has for a long time now been an open area of research and

Michele A. Saad is a Senior Engineer and Researcher in perceptual image and video quality assessment at Intel. She received her Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 2013, the B.E. degree in computer and communications engineering from the American University of Beirut, Lebanon, in 2007, and the M.S. degree in electrical and computer engineering from the University of Texas at Austin in 2009. Her research interests include statistical modeling of images and videos, motion perception, design of perceptual image and video quality assessment algorithms, and statistical data analysis and mining and machine learning.

Patrick Le Callet is full professor at Ecole Polytechnique de l'Université de Nantes. Since 2006, he is the head of the Image and Video Communication lab at CNRS IRCCyN, a group of more than 35 researchers. He is mostly engaged in research dealing with the application of human vision modeling in image and video processing. He is currently co-chairs the "3DTV" activities and the "Joint-Effort Group", driving mostly High Dynamic Range topic in this latest.

Philip J. Corriveau is a Principal Engineer in Experience Development and Assessment in SMG at Intel. Philip received his Bachelors of Science Honors at Carleton University, Ottawa Canada in 1990. He immediately started his career at the Canadian Government Communications Research Center performing end-user subjective testing in support of the ATSC HD standard for North America. In January 2009 he was awarded a National Academy of Television Arts & Science, Technology & Engineering Emmy® Award for User Experience Research for the Standardization of the ATSC Digital System. He now directs a team of human factors engineers conducting user experience research across Intel technologies, platforms and product lines. Philip is currently on the board of directors for the School of Computing at Clemson University, on the UF CISE Industrial Advisory Board and an Adjunct Professor at Pacific University. He was a founding member of and still participates in VQEG.

understanding its effect on perceptual quality is yet to be better understood and modeled. Similar to the problem of image quality assessment, a few approaches have been proposed to assess the quality of video, but the generalizability of these approaches still has a way to go before we reach a solution generalizable enough to be standardized.

A new group within the Video Quality Experts Group has been formed so that advances can be made in these identified challenge areas. The focus is on understanding and driving solution spaces with blind and no-reference models. You all are encouraged to join, follow, and contribute to moving the needle on creating, validating, and standardizing these new models.

References

Mittal A. et al. (2013), "Making a 'Completely Blind' Image Quality Analyzer" in *IEEE Signal Process. Lett*., vol. 21, no. 3, pp. 209-212.

Ponomarenko N. et al. (2008, 2013), "Tampere Image Database", [online]: http://www.ponomarenko.info/tid2013.htm.

Saad M.A. et al. (2012), "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality" in *IEEE Trans. Image Process*., vol. 21, no. 8, pp. 3350-3364.

Sheikh H. R. et al. (2005), "LIVE Image Quality Assessment Database Release 2", [online]: http://live.ece.utexas.edu/research/quality.

Tang H., et al. (2011), "Learning a blind measure of perceptual image quality," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*., pp. 305-312.

Ye P. et al. (2011), "No-reference image quality assessment using visual codebook," *Proc. IEEE Int. Conf. Image Process*., pp. 3089-3092.