Revealing the Dark Side of a Subjective Study: Learnings from Noise and Sharpness Ratings

Michele A. Saad, Member, IEEE, Philip Corriveau, and Ramesh Jaladi,

Abstract—Significant improvement in no-reference image quality assessment (NR-IQA) methods has been demonstrated in recent years. The demonstrated prediction performance of proposed NR-IQA methods, in terms of correlations between predicted and subjective scores, has reached similar performance levels with full-reference image quality assessment (FR-IQA) measures, on popular image datasets. However, in our work we found that these correlations drop significantly when NR-IQA measures are applied on images that represent real consumer-capture scenarios. This is due to the fact that the datasets used thus far for NR-IQA research are not representative of these consumer scenarios. In an attempt to tackle this issue, we created a set of consumer photos and conducted a subjective experiment. In this paper we describe the subjective experiment, which asked participants for subjective ratings for sharpness, noise, and overall image quality, but yielded counterintuitive subjective ratings for noise due to the complexity of the interaction between sharpness, noise, and perceived quality in consumer content.

Index Terms—No-reference image quality assessment, subjective test, sharpness, noise.

I. INTRODUCTION

The past decade has seen a large effort in no-reference image quality assessment (NR-IQA) research. The need for reliable solutions to this important problem has continued to grow with the exponentially increasing volumes of visual multimedia traffic. Consumer devices, like phones and tablets with cameras in them, have become ever more ubiquitous with consumers demanding better quality of experience. This necessitates that the quality of images from these devices be objectively and accurately measured and quantified.

Numerous approaches to NR-IQA have been proposed [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. Researchers are gaining an increased understanding of how the Human Visual System (HVS) processes information and how the statistics of natural scenes behave as image quality changes from pristine to distorted. The proposed approaches have been shown to perform impressively well on the popular and publicly available databases that have been designed and created for the purpose of quality assessment research [11], [12], [13], [14], [15], [16], [17], [18]. In fact, the correlations obtained between subjective quality ratings and predicted scores (by those NR-IQA algorithms) are similar to the correlations obtained by full-reference image quality assessment (FR-IQA) approaches. Tables I and II show example Spearman rank order correlation coefficients (SROCC) for top performing full-reference and no-reference IQA methods respectively on the popular LIVE IQA database [11]. These SROCC results show that prediction performance of these NR-IQA methods is very close to the prediction performance of the top FR-IQA methods on this database.

Despite these impressive correlations, when we applied these methods on real consumer images taken by a number of mobile devices and higher end cameras, the predicted scores by those algorithms did not correlate well with human perceptual judgment of image quality. The work in [22], [23], [24] and [25], shows a significant drop in SROCC results for NR-IQA approaches when applied to images from consumer devices. The main reason for this drop in performance is the difference between the characteristics of the images in the datasets used to design and train the no-reference approaches and the characteristics of images from real image-capture scenarios. This is further discussed in [26].

In our attempt to tackle the problem for real-life consumer scenarios, we set about gathering representative consumer images with the purpose of understanding the performance of NR-IQA approaches on real consumer image capture...
scenarios; i.e., on images taken by consumer devices. We conducted a pilot subjective test to collect subjective ratings for perceived image noise, sharpness, and overall quality similar to the recent work in [22], [23] which asks for subjective ratings for a number of individual image features like sharpness, noise, lightness, and saturation as well as overall quality for consumer-type photographs.

The aim of our subjective test is to create a database of consumer images with realistic image impairments and associated subjective ratings for overall image quality as well as ratings for noise and sharpness. In our initial efforts we got surprising counterintuitive subjective ratings for noise on a few outlier images. Our aim in this work is to report the procedure that led to these surprising results, but which also served as significant learnings for our subsequent work.

II. IMAGE DATABASES FOR IQA RESEARCH

The popular databases that have been used for IQA design [11], [13], [15], [17], [18], typically contain single distortions (like Gaussian blur, additive white Gaussian noise, JPEG compression artifacts, JPEG2000 compression artifacts, LAR compression artifacts), or simulated, and not necessarily commonly occurring distortions. In other words, these datasets seem to have distortion characteristics different from the types of more subtle multiply-occurring distortions present in consumer content (images captured by mobile phones or higher end cameras). Furthermore, the range of qualities on these datasets is much larger than what one encounters in real consumer scenarios. This has effects on reported SROCC prediction results. This range effect is described in more detail in [26].

Consumer images typically exhibit complex image characteristics and subtle distortion effects due to device optics and complex post processing effects. Fig. 1, shows two images from two devices that have a perceived quality difference; The image on the top is considered much more noisy than the one on the bottom and yet the quality difference between the two is much narrower than typical quality ranges on the popular datasets. Further, the type of noise present in the images is significantly different from simulated noise on the popular datasets. Note that the noise in these images is very difficult to simulate as it includes the effects of the sensor as well as the post processing that happens after the sensor module. Typically the post processing that happens in consumer devices is proprietary and simulating every possible post processing scenario is extremely challenging.

Despite the high correlations achieved between predicted and subjective quality scores on popular image quality assessment datasets recent work in [23] and [27] has shown that these correlations drop from the high 0.9 correlation range to lower than 0.5 on consumer type photographs such as those shown in Fig. 1. In an effort to create more representative consumer content with associated subjective ratings, we set about conducting a pilot subjective study with the aim of collecting subjective ratings for sharpness, noise, and overall quality. Our initial interest was in individual measures of sharpness and noise. Work on overall image quality ensued.

III. LEARNINGS FROM A PILOT SUBJECTIVE STUDY

Before we proceed, we define what is meant by pilot study. In this case the term pilot study refers to the fact that the study is a path finding exercise to determine if the direction taken is valuable to be pursued further. Pilot studies can have as few as 8 participants all the way to a much higher \( n \), like in our experiment which we describe in the following sections.

Shortly after the pilot ratings were collected, we realized the study led to counterintuitive and unexpected noise ratings in certain outlier cases. In the following subsections we describe the pilot and show example images that yielded unexpected ratings. We will follow this with an explanation behind the counterintuitive ratings and the learnings derived from this pilot experiment.

A. The Pilot Study

The pilot study was conducted with 42 participants (23 male, 19 female), all of whom fit in the following categories a) imaging engineers, b) professional and amateur photographers, c) graphic designers, or d) frequent phone
camera users. All participants were screened for normal or corrected to normal visual acuity and normal color vision.

The database of images used was designed specifically to address characteristics of image capture typical of a consumer and of those not represented in other databases. Towards creating such a dataset we collected images from 27 different devices. The idea was to create a dataset of photos that does not contain artificial or simulated distortions. Instead the quality range of the images is dictated by the image quality produced by the 27 different devices. The devices were chosen so that they span a wide range of image qualities. The scenes spanned a range of illumination conditions from bright outdoor illumination (greater than 500 lux) to indoor normal (between 100 and 300 lux) and lower light conditions (lower than 100 lux), to night shots (around 20 lux). In an effort to have the captured content be repeatable portraits and in-scene motion were avoided. Another aspect of the database is that it does not contain reference images with their simulated distorted counterparts. Instead, since 27 different devices were used for the photo capture, there are 27 distinct angles of view for each photo due to the fundamental nature of differing aspect ratios, fields of view and shot angles. A total of 405 photos were evaluated (15 scenes from 27 capture devices).

Each participant was asked to provide three subjective ratings per image; 1) a rating for sharpness, 2) a rating for noise, and 3) an overall quality rating. In this study, we were primarily interested in understanding the effect of noise and sharpness in consumer photographs on overall perceived image quality. The standard single stimulus Absolute Category Rating (ACR) method was used to collect the subjective ratings on a five point Likert scale for each measure, (where 5 corresponds to Excellent, 4 Good, 3 Fair, 2 Poor and 1 Bad). Note if the image exhibits high sharpness, then it receives a high rating for sharpness (a value closer to 5 which corresponds to excellent), and if an image exhibits low noise it receives a high rating for noise (close to 5 which also corresponds to excellent). This was explained to the study participants. All image evaluations were completed on a Dell UltraSharp 24 inch Monitor (U2412M) with a resolution of 1920*1080 so that the capture device display was removed as a variable.

B. Counterintuitive Subjective Noise Ratings

During the analysis of the pilot subjective data that was obtained, we noticed surprising and unexpected subjective ratings for noise for a small number of images. While for most images in the dataset, the noise ratings seemed reasonable and in agreement with our expectation, the subjective noise ratings for certain other images were counter-intuitive and unexpected. These exceptions occurred only on a small number of images, specifically in a number of night scenes that were heavily under-exposed and contained very little detail. These images were characterized by very low perceived sharpness and consequently received very low subjective sharpness ratings and very low overall quality ratings from the study participants. Fig. 2 shows two of those images.

Subjects rated these images as very noisy even though they do not exhibit high noise quantitatively. They are however of low overall quality and exhibit very low sharpness. The subjects gave these photos low sharpness and overall quality scores but did not seem willing to give them good noise ratings. Consequently, they rated them as very noisy. This led us to conclude that these noise ratings (for certain outlier images) might be faulty or misleading, since the graininess of these images is extremely low.

![Fig. 2. Two images that were heavily under-exposed and low in sharpness and overall quality. Counterintuitively, subjects rated these as having high noise even though they are very smooth and exhibit very low graininess.](image)

Fig. 3 shows a comparison between an image which has high detail and one of our outlier, heavily under-exposed images that does not contain much detail. The comparison between these two images is interesting in that the photo at the bottom is clearly noisier (but sharper) than the image on the top. Yet subjects rated the image on the top as very noisy (it scored a similar noise value to the image on the bottom which we anticipated would be rated as noisier). Subjects seemed unwilling to give a good noise ratings when the overall sharpness was very poor due to heavy under-exposure, even when the noise or graininess level of the photo was actually extremely low.

C. Learnings from Counterintuitive Ratings

If one were to design an objective no-reference measure for noise assessment and test it on the noise ratings with outliers...
such as the images we described in the previous subsection, this would probably lead to inaccurate results (the expectation is that the algorithm should indicate that the image above in Fig. 3 is less noisy than the image below it). It is hence important that we understand why we obtained such ratings.

There are a number of hypotheses why the pilot subjective experiment described in Section III-A may have led to unexpected subjective noise ratings on certain outlier images.

1) It seemed that the very low quality of the outlier images for which unexpected subjective noise ratings were obtained, influenced subjects in the direction of penalizing even the noise value associated with these images. These images were very low in quality and exhibited such low sharpness and heavy underexposure that led the subjects to penalize even the noise score associated with these images.

2) Even though our subjects had some degree of imaging experience (which varied widely between subjects who ranged from imaging engineers to frequent camera users), the term noise may have been interpreted differently than what we were expecting. What if extreme-underexposure (where most of the image is black and lacking information) was interpreted as a form of noise, and hence influenced the noise rating accordingly?

3) From the consumer images that we collected, we notice that the higher end devices are able to produce images high in sharpness and low in noise simultaneously. Images that tended to exhibit low sharpness and detail, also exhibited high noise in most examples (except for the outlier images for which we obtained the unexpected noise ratings). This is unlike many of the images present in the popular IQA datasets in the literature where one parameter such as the image sharpness is held constant while noise is varied. The simultaneous increase in sharpness and noise quality on most images may have influenced the noise ratings on the outlier images that were heavily underexposed; as it seemed that the expectation from a low quality images is that it is most likely simultaneously low in sharpness and noisy.

Further studies are needed to understand and model the interactions between subjective ratings for multiple image quality-related quantities as well as their interaction on overall perceived quality. In this pilot study, we were primarily interested in noise and sharpness in consumer photographs.

IV. Conclusion

Defining specific quantities like noise and sharpness for study participants can be challenging since participants tend to come with predefined concepts. Very low image quality and features like sharpness have an influence of the subjective ratings of other features like noise and can lead to misleading subjective ratings. This can be corrected by potentially asking only for an overall image quality rating. However, further studies are needed to understand these interactions. The dark side of subjective testing has spawn a new line of thinking around how we do merger ratings of various image aspects effectively. Simply asking for subjective ratings for different quality related terms in the form of raw numbers (as we have done in this exercise) may not be the way to go. On the other hand, asking for some other forms of feedback might be. We will be considering these questions in future work and using our learnings to shape our next generation subjective test experimental designs.

REFERENCES


