# SUBJECTIVE AND OBJECTIVE QUALITY ASSESSMENT OF MOBILE VIDEOS WITH IN-CAPTURE DISTORTIONS

*Deepti Ghadiyaram*[*1], Janice Pan*[*1], Alan C. Bovik[1], Anush Moorthy[+2], Prasanjit Panda[3], and Kai-Chieh Yang[+4]*

[1]The University of Texas at Austin, [2]Netflix Inc., [3]Qualcomm Inc., [4]Aengin Inc.

## ABSTRACT

We designed and created a new video database that models a variety of complex distortions generated during the video capturing process on hand-held mobile capturing devices. We describe the content and characteristics of the new database, which we call the LIVE Mobile *In-Capture* Video Quality Database. It comprises a total of 208 videos that were captured using eight different smart-phones and were affected by six common in-capture distortions. We also conducted a subjective video quality assessment study using this data, wherein each video was assessed by 36 unique subjects. We evaluated several top-performing No-Reference IQA and VQA algorithms on the new database and find insights on how real-world *in-capture* distortions challenge both human subjects as well as automatic perceptual quality prediction models.

***Index Terms—*** mobile videos, in-capture video distortions, perceptual video quality, subjective quality assessment.

## 1. INTRODUCTION

The explosive growth of digital media has accelerated in recent years, owing to the ubiquitous availability of portable mobile devices for video capture and access. On YouTube alone, half of the billions of daily video views are received on mobile devices [1]. Every viewed digital video typically passes through several processing stages during and after its capture before ultimately reaching a human observer. Different forms of distortion can be introduced during video acquisition, transmission, and rendering processes, each of which in turn could impact an end user's quality of experience (QoE). The quality of a digital video (or a digital picture) as *perceived* by human observers is referred to as 'perceptual quality.'

Subjective Video Quality Assessment (VQA) studies, though time-consuming and cumbersome, are crucial for understanding humans' perceived quality of digital videos [2–6]. They also assist in the development of objective, automatic quality predictors, whose ultimate goal is to accurately predict perceived video quality. Subjective studies also provide valuable data that makes it possible to evaluate the performance of video quality predictors. Automatic quality predictors can be used to identify and cull low quality videos stored on digital devices and to prevent their occurrence using suitable quality correction processes during capture. More importantly, video quality predictors can be used to objectively measure and benchmark the camera and lens quality of emerging mobile devices and thereby drive "quality-aware" camera design strategies. These and other significant and potentially impactful benefits have greatly accelerated the development of objective video quality models.

**Relation to state-of-the-art:** Most of the existing video quality databases [2–5] model *post-acquisition* distortions such as encoding (compression) artifacts, transmission errors, and rebuffering events [7]. These databases contain videos captured using high-end cameras that have been impaired by one of a few *synthetically* introduced distortion types at a level of perceptual distortion chosen by video quality scientists. Though these databases have tremendously accelerated the development of VQA algorithms [8–13], they do not model *in-capture* distortions, such as texture distortions, artifacts due to exposure and lens limitations, focus, and color aberrations. We refer to these naturally occurring distortions as **authentic, in-capture distortions**. The majority of the mobile digital videos that are produced by casual, inexpert mobile users ostensibly suffer from in-capture, authentic distortions as opposed to single, separable post-capture distortions alone. Some of these distortions have previously been explored, using objective criteria, and there exist models that attempt to characterize some of these distortion effects, e.g., color [14], sharpness [15], noise and other artifacts [8, 16]. However, there does not yet exist a standardized methodology that can successfully evaluate multiple aspects of in-capture video distortions across diverse mobile devices. Our new database is a powerful tool for developing and evaluating such methodologies.

To the best of our knowledge, we are aware of only one other recently designed database [6] that models in-capture distortions. The videos in CVD2014 [6] were captured using 78 different cameras, all used in automatic mode. The

* Indicates equal contribution.

+Anush and Kai-Chieh were with Qualcomm Inc. at the time of the subjective study.

quality of the cameras used varied from low-quality mobile phone cameras to dedicated video and high-quality digital single lens reflex (DSLR) cameras. Videos from different devices were captured one at a time sequentially and were later edited by the database creators to be as similar as possible with respect to the video content.

While the CVD2014 database has great broad potential, capturing videos in completely uncontrolled settings, with no preset goal or information about the type of intended distortion will not allow us to deepen our understanding of the effects of specific distortion types and the performance of different mobile devices on an end user's QoE. Some distortions occur more frequently than others (exposure and color-related distortions, for instance). To design reliable models and algorithms for evaluating such distortions, databases are needed that contain videos that were specifically captured to represent those specific distortions.

Our goal was to strike a balance between a completely uncontrolled collection of *in-the-wild* videos (as done in [6]) and a systematic generation of singly distorted videos. Towards this end, we captured videos with intended "dominant" distortion types in mind, such as focus, color artifacts, and so on using a predetermined set of mobile camera devices. However, the captured videos could be afflicted by other unintended distortions, such as underexposure, or low-light noise if captured during the night. We did not try to avoid these unintended distortions; however we assigned each video to a group according to what we determined to be the most dominant distortion present in the video based on visual inspection. We will describe our choice of dominant distortion categories and their purpose with regards to our subjective study in Sec. 2.1. Further, we will describe the set-up we used to capture videos with significantly overlapping content, which allows for objectively comparing scenes across mobile devices and is another distinguishing feature of the proposed database over the CVD Database [6].

## 2. THE LIVE MOBILE IN-CAPTURE VIDEO QUALITY DATABASE

### 2.1. Video Sequences

Figure 1 show a sample of scenes from the LIVE Mobile In-Capture Video Quality Database. This database consists of a total of 208 videos, each of which is categorized into one of the following six video groups:

1. *Artifacts:* Videos afflicted by noise, blockiness, and other such distortions that are not part of the video content.
2. *Color:* Videos with incorrect or insufficient color representation.
3. *Exposure:* Videos containing over/under-exposed regions, making it difficult to see parts or the entirety of the scene.
4. *Focus:* Videos afflicted by autofocus related distortions, i.e., that are intermittently sharp or blurry over time.

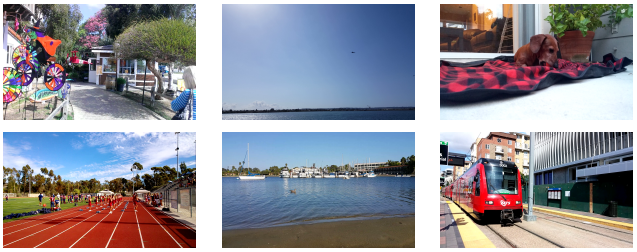

**Fig. 1**. Sample video frames from the LIVE Mobile In-Capture Video Quality Database.

5. *Sharpness* - Videos suffering from general unsharpness, i.e., lack of detail, texture, or sharpness.
6. *Stabilization* - Videos where the affects of camera shake overwhelm the content.

To be able to evaluate different camera devices across in-capture distortions, the new database was designed to meet the following requirements:(1) There should be an overlap of video content across different devices, and clips for any content should be spatially and temporally aligned; and (2) distortion severity and visibility should be perceptually separable for each content.

**Video capture process:** To satisfy these conditions, the videos were captured using a rig that holds four devices at once to simultaneously capture near-identical content under the same illumination conditions and from similar viewing angles. Since each phone's field-of-view is different, the phone positions on the rig were manually adjusted so that each group of four phones captured nearly identical scenes. The videos were captured using the default settings on the phones, and the touch-to-focus feature was never used. All videos were captured at 1080p resolution.

Segments of fifteen seconds in duration were selected (after removing the audio) and edited to length based on distortion relevance, content variability, and interest, while maintaining as much temporal alignment as possible. The contents providing adequate perceptual separability were selected and expertly categorized into one of the six categories based on the most dominant distortion apparent in the clips. The breakdown of capture devices and distortion categories for the videos are given in Table 1. The eight devices used for capture were some of the most widely-used mobile devices during the time of video content acquisition.

### 2.2. The Subjective Study

**Unbiased and biased subject groups:** The goal of our subjective study was two fold: (a) obtain overall subjective quality scores recorded by humans viewing naturally-occurring in-capture video distortions, and (b) for each of the six distortions under consideration, obtain opinion scores assessing how much each particular distortion affected a subject's perceived quality. Towards this end, the participating subjects were randomly divided into two groups: (a) a *biased* group

**Table 1**. Number of videos per phone and distortion (A: Artifacts; C: Color; E: Exposure; F: Focus; Sh: Sharpness; St: Stabilization)

| Phone | A | C | E | F | Sh | St | Total |
|---|---|---|---|---|---|---|---|
| Galaxy GS5 | 2 | 3 | 3 | 4 | 6 | 4 | 22 |
| Galaxy GS6 | 8 | 6 | 5 | 6 | 3 | 5 | 33 |
| HTC One VX | 8 | 6 | 5 | 6 | 2 | 5 | 32 |
| iPhone 5S | 1 | 3 | 4 | 3 | 6 | 4 | 21 |
| LG G2 | 7 | 6 | 5 | 5 | 3 | 5 | 31 |
| Lumia 1020 | 2 | 2 | 3 | 2 | 4 | 3 | 16 |
| Samsung Note 4 | 1 | 3 | 4 | 3 | 7 | 4 | 22 |
| Oppo Find 7 | 7 | 6 | 5 | 5 | 3 | 5 | 31 |
| Total | 36 | 35 | 34 | 34 | 34 | 35 | 208 |

and (b) an *unbiased* group. Subjects who were assigned to the biased group were informed *a priori* which type of distortion to focus their attention while viewing the videos (from the 6 distortions listed earlier). The biased subjects were asked to evaluate each video's perceptual quality, given that the video was afflicted with that particular distortion. This task, including the particular distortion the subject was to pay attention to, was displayed on the screen before each distortion-specific test session.

Subjects belonging to the unbiased group were not asked to attend to any particular distortion, i.e., no prior information about the afflicting dominant distortion was provided to them. Instead, they were asked to evaluate the overall perceptual quality of each video (presented at random). However, since these distortions are perceptually subtle, we also wanted to understand the distortion that dominated their quality perception, thereby their quality score. Therefore, once the subjects provided their opinion score for a video, another screen appeared where each subject had to select the distortion that dominated their judgment from a list of seven options (the six distortions mentioned earlier and a seventh option 'no dominant distortion was perceived').

**Study duration and display details:** Each subject completed three sessions of approximately 30 minutes each. Each biased session consisted of two 15-minute halves, each focusing on one of the six distortions. Each unbiased session was an uninterrupted 30 minute session, during which videos from any of the six distortions were randomly displayed. A subject was shown each test video in the database exactly once and the playlist order was randomized for each subject such that videos of the same content were not presented consecutively.

We designed the user interface using the XGL toolbox [17] with MATLAB 2015b on a Windows PC with an ATI Radeon X600 graphics card. Each video sequence was stored in raw YUV 4:2:0 format and loaded in its entirety into memory before displaying to subjects in order to avoid playback latencies. The video sequences were displayed on an ASUS VG248QE monitor in their native $1920 \times 1080$ resolution. All videos had a frame rate of 30 Hz and we set the monitor refresh rate to 60 Hz to avoid flicker artifacts, so each frame was displayed for two monitor refresh cycles.

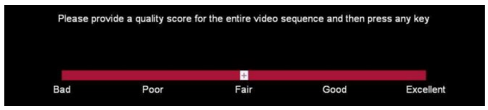


**Fig. 2**. The continuous rating bar that was displayed to the subjects after each video.

Following the presentation of each video, a rating screen with instructions on how to select and submit the quality score was displayed. This continuous rating bar (Fig. 2) is divided into five equal portions, labeled *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*, to reflect the ITU-R Absolute Category Rating (ACR) scale [18]. Subjects used a mouse to move the slider on the quality scale to their desired rating, and they were allowed to press any key, excluding the escape key, to submit their score. Biased subjects were instructed to take approximately 10 seconds to submit their scores, and unbiased subjects were instructed to take approximately 15 seconds to submit both their quality rating and their 'dominant distortion' selection. Subjects were not allowed to go back and change their score once they submitted it.

**Subjects and Training** A total of 39 subjects (undergraduate and graduate students at the University of Texas at Austin) were recruited and split into biased (19) and unbiased (20) groups. At the beginning of each session, each subject read instructions that included descriptions of the distortions they might see in the videos. The viewing distance was about 2 - 2.5 feet, which was deemed as a *comfortable* distance, and we asked subjects to approximately maintain their viewing position throughout the study.

Preceding each 15-minute distortion-focused biased session, subjects watched three to four training videos of a single content (captured using different devices), containing enough perceptual separability to help them understand how the relevant distortion manifests in a video. Unbiased subjects watched six to seven training videos preceding each of the 30-minute sessions. The videos were of random content and spanned the six distortion categories. The training sessions were primarily intended to help subjects become familiar with the testing procedure, the interface, and the distortions. The training videos were also 15 second long, were captured using the same rig and expertly categorized according to the dominant distortions afflicting them. At the end of all three sessions, each viewer answered a short questionnaire regarding what they found to be the most annoying, hardest to detect distortions, and their engagement levels during the study.

## 3. SUBJECTIVE DATA ANALYSIS

### 3.1. Processing the Subjective Scores

If we let $s_{ijk}$ denote the raw score assigned by subject $i$ to video $j$ during session $k$, we can compute Z-scores [19] $z_{ijk}$ to account for each subject's variability in their use of the

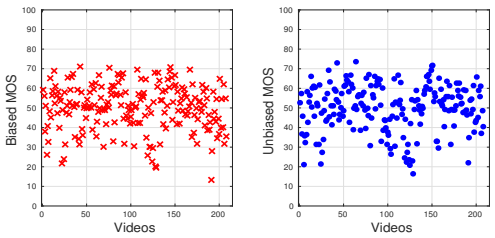


**Fig. 3**. Scatter plot of the MOS values obtained from the biased group (left) and the unbiased group (right) on all the videos in the LIVE In-Capture Mobile Video Database.

quality scale during each session:

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} s_{ijk}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} (s_{ijk} - \mu_{ik})^2}$$

$$z_{ijk} = \frac{s_{ijk} - \mu_{ik}}{\sigma_{ik}}$$

where $N_{ik}$ is the number of test videos seen by subject $i$ in session $k$. Note that $k = 3$ for both the biased and unbiased groups. Since biased subjects complete two 15-minute sessions back-to-back, we consider their use of the quality scale to be similar in those two sessions. We followed the subject rejection procedure detailed in the ITU-R BT.500-11 recommendation [18] using the Z-scores and rejected 3 subjects—1 biased and 2 unbiased, leaving the scores from 18 biased and unbiased subjects for analysis. As was done in [2], the Z-scores were linearly rescaled such that 99% of the scores would lie in the range [0,100]:

$$z'_{ij} = \frac{100(z_{ij}+3)}{6}.$$

Assuming that Z-scores are normally distributed, 99% of the scores should lie in the range [-3,3], which we found to be the case. Finally, we computed the Mean Opinion Score (MOS) of each video as the mean of the rescaled Z-scores for both the biased and the unbiased groups (Fig. 3).

**Table 2**. Performance of a few No-Reference IQA and VQA models on the proposed database. *Italicized* algorithms are IQA models.

| NR-Model | Median PLCC | Median SROCC |
|---|---|---|
| *BRISQUE* [20] | 0.4640 | 0.4579 |
| *NIQE* [21] | 0.4357 | 0.4333 |
| V-BLIINDS [12] | 0.4894 | 0.4933 |
| VIIDEO [13] | 0.1097 | 0.0677 |

### 3.2. Performance of Objective Quality Metrics

We evaluated the performance of some publicly available no-reference generic IQA and VQA algorithms on the LIVE Mobile In-Capture Video Quality Database. The entire dataset was divided into non-overlapping training and test data (80/20 split). To mitigate any bias due to the division of data, the process of randomly splitting the dataset was repeated 100 times. Since V-BLIINDS [12] is a learning-based model, in each iteration, we trained a model from scratch on 80% of the data and evaluated it on 20% of the data. For the other algorithms, no training was required, hence we report their performance on the test data alone. We used the publicly-available trained BRISQUE [20] and NIQE [21] models to predict the quality of each video frame and averaged the scores to achieve a final quality score for the video. Since these are generic NR IQA/VQA algorithms, we used the MOS scores obtained from the unbiased study group as ground truth scores. We note that the biased study ratings will prove valuable for evaluating distortion-specific quality assessment algorithms. Table 2 presents the median Pearson Linear Correlation Coefficient (PLCC) and Spearman's Rank Ordered Correlation Coefficient scores (SROCC) computed between predicted and ground truth quality scores across the 100 iterations. As may be observed, we found that existing blind IQA/VQA algorithms have significant room for improvement regarding their ability to accurately predict the quality of videos suffering from naturally-occurring, in-capture distortions.

### 3.3. Biased vs. Unbiased MOS

In order to understand if biasing a viewer to notice particular distortions could impact their perceived quality, we computed the correlation between the MOS obtained from the two subject groups for each distortion category. We found that the unbiased subjects generally agreed with the biased group in regards to distortion perception. An overall Spearman correlation of 0.76 was obtained between the ratings of the two groups (over all six distortions). A lowest Spearman correlation value of 0.69 was achieved on videos belonging to the color distortion category. This low correlation score is also supported by responses to one of the end-of-the-study questionnaire, where about 65% of the subjects in the unbiased group reported that videos with color-specific distortions were the hardest to detect.

## 4. FUTURE WORK

We designed a new database of videos captured using modern mobile camera devices exhibiting contemporary in-capture distortions. This new database of videos and associated subjective scores provide a valuable tool that may be used to address some of the limitations of current VQA databases [2–4] in regards to content diversity and distortion realism and variability. Building on our insights from this subjective study, we plan to explore the feasibility of developing powerful distortion-specific (from our biased study) and also unified generic (from our unbiased study) blind VQA models that perform well on videos afflicted by complex in-capture distortions. We are also interested in adapting such models to perceptually optimize mobile cameras and lenses.

## 5. REFERENCES


[1] YouTube.com, "YouTube Statistics," [Online]. Available: http://www.youtube.com/yt/press/statistics.html.

[2] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[3] F. De Simone, M. Tagliasacchi, M. Naccari, S. Tubaro, and T. Ebrahimi, "A H.264/AC video database for the evaluation of quality metrics," *IEEE Int. Conf. Acoustics, Speech, and Signal Process.*, pp. 2430–2433, March 2012.

[4] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *IEEE J. Selected Topics Signal Process.*, vol. 6, no. 6, pp. 652–671, Oct 2012.

[5] VQEG HDTV Group, "VQEG HDTV Database. Video Quality Experts Group (VQEG)," [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx.

[6] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Hkkinen, "CVD2014 - A database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3073–3086, 2016.

[7] D. Ghadiyaram, J. Pan, and A.C. Bovik, "A time-varying subjective quality model for mobile streaming videos with stalling events," in *SPIE Optical Engineering+ Applications*, 2015, pp. 959911–959911.

[8] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal process: Image commun.*, vol. 19, no. 2, pp. 121–132, 2004.

[9] Z. Wang, H. Sheikh, and A. C. Bovik, "Objective video quality assessment," *The Handbook of Video Databases: Design and Applications*, pp. 1041–1078, 2003.

[10] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb 2010.

[11] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 23, no. 4, pp. 684–694, April 2013.

[12] M. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, 2014.

[13] A. Mittal, M. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, 2016.

[14] S. Winkler, "Perceptual distortion metric for digital color video," *Proc. SPIE*, vol. 3644, pp. 175–184, 1999.

[15] J. Caviedes and S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis," in *Int. Conf. Image Process.*, 2002, vol. 3, pp. III–53–III–56 vol.3.

[16] A.B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *J. of Electronic Imaging*, vol. 10, no. 1, pp. 20–29, 2001.

[17] J. S. Perry, "XGL Toolbox," [Online]. Available: https://github.com/jeffsp/xgl, 2015.

[18] Int. Tellecommun. Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures ITU-R Recommendation BT.500-11, Tech Rep.," 2002.

[19] A. M. van Dijk, J.-B. Martens, and A. B.Watson, "Quality assessment of coded images using numerical category scaling," *Proc. SPIE Advanced Image and Video Communications and Storage Technologies*, 1995.

[20] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[21] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Sig. Proc. Letters.*, vol. 20, no. 3, pp. 209–212, 2012.