In-capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms


*The University of Texas at Austin, †Netflix Inc., ‡Qualcomm Inc., §Aengin Inc.

Abstract—Digital videos often contain visual distortions that are introduced by the camera hardware or processing software during the capture process. These distortions often detract from a viewer’s quality of experience. Understanding how human observers perceive the visual quality of digital videos is of great importance to camera designers. Thus, the development of automatic objective methods that accurately quantify the impact of visual distortions on perception has greatly accelerated. Video quality algorithm design and verification requires realistic databases of distorted videos and human judgments of them. However, most current publicly available video quality databases have been created under highly controlled conditions using graded, simulated, post-capture distortions (such as jitter and compression artifacts) on high-quality videos. The commercial plethora of hand-held mobile video capture devices produces videos often afflicted by a variety of complex distortions generated during the capturing process. These in-capture distortions are not well-modeled by the synthetic, post-capture distortions found in existing VQA databases. Towards overcoming this limitation, we designed and created a new database that we call the LIVE-Qualcomm Mobile In-Capture Video Quality Database, comprising a total of 208 videos, which model six common in-capture distortions. We also conducted a subjective quality assessment study using this database, where each video was assessed by 39 unique subjects. Furthermore, we evaluated several top-performing No-Reference IQA and VQA algorithms on the new database and studied how real-world in-capture distortions challenge both human viewers as well as automatic perceptual quality prediction models.

Index Terms—mobile videos, in-capture distortions, subjective quality assessment, objective algorithms, smartphone cameras.

I. INTRODUCTION

An explosive growth of digital media is still ongoing. An increasing number of applications for home video entertainment, internet video services such as Netflix and YouTube, as well as video-centric mobile applications such as Snapchat, Vine, Skype, and Instagram are fueling the demand for better quality video content on diverse display devices having varied resolutions. Given the ubiquitous availability of portable mobile devices for video capture and access, there has been a dramatic shift towards over-the-top video streaming and sharing of videos on social media. On YouTube alone, half of the billions of daily video views are already on mobile devices [1]. Furthermore, there is also an increasingly popular trend towards shooting music videos and films solely using cellphone cameras [2]–[4].

Every viewed digital video typically passes through several processing stages during and after its capture before ultimately reaching a human consumer. Different forms of distortions can be introduced during the video acquisition, transmission, and rendering processes. Although the limits imposed by current wireless network capacity are becoming increasingly tested, consumers have become quite knowledgeable and continue to demand higher quality video playback services. This has heightened concerns regarding end users’ satisfaction with transmitted video quality. Both content and network providers are deeply invested in finding better ways to monitor and control video quality. The quality of a digital video as perceived by human observers, which is what matters most, is referred to as ‘perceptual quality.’

Subjective Video Quality Assessment (VQA) studies, though time-consuming and cumbersome, provide deep insights into perceived video quality [5]–[10]. These studies provide valuable data for the development and evaluation of objective, automatic quality predictors, whose goal is to accurately predict perceived video quality [11]. 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. These predictors also have the potential to impact protocols for monitoring and controlling multimedia services on networks and devices and to improve the quality of visual signals by acquiring or transporting them via “quality-aware” processes. Such quality-aware processes could potentially be used to perceptually optimize the capture process and to modify transmission rates to ensure good quality across both wired and wireless networks. These strategies could in turn help ensure that end users enjoy satisfactory levels of quality of experience (QoE). More importantly, video quality predictors can be used to objectively measure and benchmark the camera and lens quality of emerging mobile devices and to thereby drive quality-aware camera design strategies. These and other significant and potentially impactful benefits have greatly accelerated interest in the development of objective video quality models.

A. Authentic, In-capture Video Distortions

Video quality assessment is a thriving area of research, and a number of video quality databases have been designed in the past decade [5]–[10]. All of these databases have been developed by first obtaining a small set of high-quality videos, then systematically distorting them in a controlled manner. These video distortion databases generally share four common properties. First, pristine source videos are often
filmed using professional-grade, high-end equipment, then carefully converted into digital format to ensure the creation of a set of reference videos that are effectively distortion-free. Second, these high-quality pristine videos are synthetically impaired by only one of several distortions. Third, the distortion severities and the parameter settings are carefully (but artificially) selected, typically for the purpose of modeling a wide range of distortions or for obtaining an observable degree of perceptual separation between videos distorted by the same process. Fourth, most of these databases only model post-capture distortions, such as compression and transmission errors, but not in-capture distortions, such as texture distortions, artifacts due to exposure and lens limitations, focus, and color aberrations.

The videos in the existing VQA databases are thus inauthentically distorted. Videos captured using high-end cameras and then impaired by one of a few synthetically introduced, post-capture distortions are very different from videos captured by real-world users and containing real-world distortions that are introduced by the highly diverse mobile cameras currently being marketed. We refer to these latter videos as authentically distorted [12], [13]. A vast majority of the mobile digital videos that are produced and consumed by social media are taken by casual, inexpert users, and the capture process is usually affected by delicate variables such as lighting, exposure, lens limitations, noise sensitivity, acquisition speed, in-camera processing, and camera shake, each of which can adversely affect a video’s perceived visual quality. Although state-of-the-art cameras often allow users to control some of the parameters of video acquisition, the unsure eyes and hands of most amateur camera users frequently lead to the presence of annoying video distortions during capture despite attempts to include corrective software in the camera devices.

A key aspect of real-world, authentically distorted videos captured by inexpert camera users is that these videos cannot be accurately described as suffering from single, separable distortions. Indeed, there currently does not exist any known way to categorize, characterize, or model the complex and uncontrolled combinations of video distortions that occur, and therefore, there is no systematic way of synthetically adding distortions to videos to accurately simulate authentically distorted videos.

Our goal was therefore to design a unique and challenging database containing a large number of authentically distorted videos, representative of real world distortions that frequently occur during the video capture process. But what makes a video quality database representative? Beyond capturing a wide variety of video content under varied illumination conditions of scenes that include animals, people, other objects, indoor and outdoor environments, etc., such a database should also reflect the current diversity of consumer video capture devices. More importantly, these videos should also exhibit a broad spectrum of frequently occurring real-world distortions (e.g., exposure or out-of-focus related distortions). The result of our effort is a new resource called the LIVE-Qualcomm Mobile In-Capture Video Quality Database that models only ‘in-capture’ distortions on a wide-variety of video contents. Since we did not synthetically introduce distortions on high quality videos, no pristine references are available. Hence the new database is only suitable for the design, testing, and comparison of no-reference (NR) or “blind” VQA models.

B. Contributions:

We proceeded to address the challenging problems of blind subjective and objective video quality assessment of real-world, authentic, mobile videos from the ground up and describe our contributions below:

1) First, we describe the content and characteristics of the new LIVE-Qualcomm Mobile In-Capture Video Quality Database, which consists of 208 videos that were captured using 8 different types of smart-phones and model six authentically-occurring distortion types. Each video was collected without artificially adding any distortions beyond those introduced during capture by the mobile device (Sec. III).

2) With an aim to gather rich, ground-truth human opinion scores on these videos, we designed a subjective study setup and used it to conduct a subjective video quality assessment study, wherein about 39 diverse observers recorded their judgments of quality on each video (Sec. IV).

3) With a motivation to better understand the quality of videos created using cameras in emergent mobile devices, we used the measured subjective opinion scores to objectively analyze the properties of the 8 mobile devices in regards to each of the 6 in-capture distortions modeled in the new database (Sec. VI-D and VI-E).

4) As a demonstration of the usefulness of the video collection and subjective study, we also conducted empirical studies on the performance of several top-performing NR IQA and VQA models (Sec. VII) on the new LIVE-Qualcomm Mobile In-Capture Video Quality Database.

Our results demonstrate that (1) state-of-the-art NR IQA and VQA algorithms have a lot of room for performance improvement on the LIVE-Qualcomm Mobile In-Capture Video Quality Database, and (2) although different phones capture and process videos differently with respect to each of the 6 distortions, a couple of phones stood out to have better overall video quality performance.

II. RELATED WORK AND MOTIVATION FOR DESIGNING A NEW VIDEO DATABASE

The LIVE VQA Database [5] remains one of the most comprehensive and widely used public-domain video quality databases among several later ones [6]–[10]. We direct the reader to [9] for a comprehensive list of existing VQA databases. The distortions in all of these databases pertain to post-acquisition distortions such as encoding (compression), transmission errors, and in the case of [10], buffering events. Though these databases have tremendously accelerated the development of VQA algorithms [14]–[20], none of them model authentic, in-capture distortions, such as texture distortions, artifacts due to exposure and lens limitations, focus, and color aberrations. Authentic distortions have been explored using objective criteria by a number of authors, and there exist
models that attempt to characterize some of these distortion effects, e.g., color [21], sharpness [22], noise and other artifacts [23], [24]. However, there are not yet any adequate quantitative models of the in-capture video distortions that occur in many flavors across the wide-diversity of modern mobile devices. This implies that experimental video data drawn from these devices is our best research probe. Our new database is intended to help accelerate the development and evaluation of such methodologies.

We are aware of only one other recently designed database [9] that models in-capture video distortions. The videos in CVD2014 [9] were captured using 78 different cameras, all used in automatic mode. The quality of the cameras used varied from low-quality mobile phone cameras designed for video capture and high-quality digital single-lens reflex (DSLR) cameras. Videos from different devices were captured one at a time and later edited by the database creators to be as similar as possible with respect to the video content. While the CVD2014 database has great potential, it has one important difference from our collection: the videos were captured in completely uncontrolled settings with no preset goals regarding the types of significant distortions to be studied, limiting researchers’ ability to dissect the effects of specific distortion types and different mobile devices on an end user’s QoE. Some distortions, like exposure and color-related distortions, for instance, occur more frequently than others, hence designing reliable predictive models and algorithms for ameliorating these distortions is of great interest.

We sought to strike a balance between a completely uncontrolled collection of in-the-wild videos (as done in [9]) and a database of systematically generated single-distortion videos. Towards this end, we used a predetermined set of mobile camera devices, and captured videos with intended “dominant” authentic distortion types in mind (e.g., focus, exposure, and stabilization issues). We assigned each video to a category 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 purposes with regards to our subjective study in Sec. III. Table I, by way of comparison, shows the various ways by which the proposed database differs from the CVD2014 Database.

Further, we will describe the set-up we used to capture videos with significantly overlapping content, allowing for objective comparisons of scenes across mobile devices, which in turn enables direct comparisons between mobile devices. A current standard for comparing the performance of cameras and lenses is DxOMark [25], which implements a system called Dxo Analyzer [26] to test and evaluate camera software and hardware and to measure the quality of captured images and videos. Specifically, a Dxo Analyzer system has hardware, software, and test protocol components. The hardware module is used to capture pictures under a variety of different test conditions (such as varied apertures, focal lengths, ISO settings, and so on) [27]. These captured images are then loaded into the software and individual tests are run by the Dxo Analyzer. These tests measure various aspects of camera sensors such as ISO sensitivity, noise, and color sensitivity, and compare lenses with respect to resolution, distortion, vignetting, light transmission, and chromatic aberrations. Upon extensive testing, a DxOMark score is reported for each of the following objective categories: (a) exposure and contrast, (b) color, (c) autofocus, (d) texture, (e), noise, (f) artifacts, and (g) stabilization. An overall DxOMark score is also reported by the Dxo Analyzer for the capturing device being tested [25].

While objective measurements such as these are valuable indicators of camera performance, perceived video quality depends on an abstruse nonlinear combination of many parameters, and the interactions of multiple different distortions can impact the viewer’s QoE in complex ways. In Section VI-E, we compare the subjective video quality scores of the phones studied here with published phone camera rankings from DxOMark [25]. Based on this, we propose a MOS-based alternative for comparing cameras purely in terms of perceptual quality.

III. THE LIVE-QUALCOMM MOBILE IN-CAPTURE VIDEO QUALITY DATABASE

Figure 1 shows a sample of scenes from the LIVE-Qualcomm Mobile In-Capture Video Quality Database. The database consists of a total of 208 videos. Our goal was to capture videos noticeably afflicted by any one of following six distortions:

- **Artifacts** - Noise and blockiness distortions not part of the video content.
- **Color** - Videos with incorrect or insufficient color representation.
- **Exposure** - Over/under-exposure, making it difficult to see parts or the entirety of the scene.
- **Focus** - Autofocus related distortions, i.e., videos that are intermittently sharp or blurry over time.
- **Sharpness** - General unsharpness, i.e., lack of detail, texture, or sharpness. This distortion differs from out-of-focus distortion in that with sharpness distortion, objects are in focus but do not appear ‘crisp’ or detailed.
- **Stabilization** - Camera shake that overwhelms content.

Based on our understanding of the image processing engines in the most popular mobile phones, we believe that these six distortions are the most significant distortions that frequently occur in mobile videos. Note that our characterization of (un)sharpness refers to generic distracting unsharpness of a video as may arise from motion blur, whereas focus distortions specifically refer to failures of the autofocus capability of the camera to focus properly on the object(s) of interest, or of unstable vacillating focusing. While there may be overlap between the subjects’ perceptions of these distortions, these descriptions were presented to the subjects to help give an intuitive, inexpert understanding of their task. No further explanation was given, to avoid biasing the subject.

We chose the following eight mobile devices to capture video content for our database: Samsung Galaxy S5, Samsung Galaxy S6, HTC One VX, Apple iPhone 5S, Nokia Lumia 1020, LG G2, Samsung Galaxy Note4, and Oppo Find 7. These eight devices were among the most widely-used mobile devices during the time of video content acquisition [25], [28]–[30]. To be able to evaluate different camera devices across
TABLE I: Comparison of the proposed database with CVD2014 database [9].

<table>
<thead>
<tr>
<th>Database features</th>
<th>CVD2014 Database [9]</th>
<th>LIVE-Qualcomm Database</th>
<th>Additional comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of capture devices</td>
<td>78</td>
<td>8</td>
<td>By having enough data samples per device for each distortions, we were able to evaluate the subjective performances of each of the devices we used.</td>
</tr>
<tr>
<td>Average number of contents per phone</td>
<td>3</td>
<td>26</td>
<td>We obtained enough datapoints to draw comparisons between phones.</td>
</tr>
<tr>
<td>Video capturing process</td>
<td>Each content was sequentially captured one after the other using a different device.</td>
<td>We used a rig that allowed the simultaneous capture of videos from four mobile devices at once.</td>
<td>By capturing with four devices simultaneously, the ambient settings and external sources of distortions remained constant across each content, so comparisons could be drawn between capture devices.</td>
</tr>
<tr>
<td>Number of unique scenes</td>
<td>5</td>
<td>54 (9 videos per distortion category)</td>
<td>More variety in content reduces content-dependency concerns when developing VQA algorithms. It also yields a more challenging dataset for evaluating existing VQA algorithms.</td>
</tr>
<tr>
<td>Realism in the content</td>
<td>The 5 contents were chosen to have varied spatial and temporal activity</td>
<td>54 different contents were chosen to be broadly typical of those captured by people on their mobile devices at different times of day.</td>
<td></td>
</tr>
<tr>
<td>Source video formats</td>
<td>9 different video formats, and several different frame rates and video codecs were used.</td>
<td>Each of these parameters was held constant.</td>
<td>Our approach reduces the number of factors influencing video quality making it possible to compare the devices purely based on distortions.</td>
</tr>
<tr>
<td>Video lengths</td>
<td>10-25 sec.</td>
<td>15 sec.</td>
<td></td>
</tr>
<tr>
<td>Choice of distortions</td>
<td>The database was not constructed around a fixed set of distortions</td>
<td>We designed our database around the six major distortions identified by DxOMark [25].</td>
<td></td>
</tr>
<tr>
<td>Subjective Study</td>
<td>• Videos were shown with audio. • Subjects could re-watch videos to answer distortion-specific questions. • Subjects in some sessions answered qualitative questions particular to each video.</td>
<td>• Videos were shown without audio. • Subjects watched the videos only once and provided overall quality opinion scores. • The same questionnaire was administered to all subjects after they completed the study.</td>
<td>Each of these study set-ups were tailored to be able to draw specific conclusions from the data.</td>
</tr>
<tr>
<td>Subjective Data Analysis</td>
<td>Conclusions about how in-capture distortions affect videos on a broader scale were drawn, since their video collection was captured using a wider variety of capture devices.</td>
<td>We drew conclusions regarding phone cameras and the subjective perception of the six distortion categories we modeled.</td>
<td></td>
</tr>
</tbody>
</table>

in-capture distortions, the new database was designed to meet the following requirements:

1) The cameras used to capture the video content should reasonably span the quality spectrum (from low-quality to high-quality).
2) The distorted video contents should model adequate perceptual separability, i.e., they should also reasonably span the quality spectrum, from low quality to high quality. Otherwise, a large number of videos may cluster too tightly at high and/or low quality, making it difficult to distinguish the performances of different quality assessment models.
3) The video content should be interesting and representative of real-life captures.
4) The mobile camera devices used for capture should be widely adopted by consumer users.
5) There should be overlap of video content across different devices, and clips from these devices for a given content should be spatially and temporally aligned.
6) Since we are focused on in-capture distortions, the videos should not contain any post-capture distortions (such as transmission artifacts).

A. Video Capture Process

The video contents in our database were captured with care and effort to satisfy the above criteria. The eight phones were grouped into two groups of four each – the two groups were dynamically changed so that each phone captured an almost equal number of contents. The phones were grouped such that each group had a wide range of expected quality from the “best” to the “worst” phone (based on data available through independent testing). The grouping of phones also took into account similarity of field-of-view (FoV), and thus, phones with similar FoVs were grouped together.

Our goal was to simultaneously capture each video content with a group of four phones. In order for each phone to capture
Fig. 1: Sample frames of the unique video contents contained in our database. Nine unique contents were captured per distortion category (illustrated along each column).

essentially the same content at the same time from similar viewing angles, a rig consisting of four phone holders and a metal rod was set up. Since each phone’s field-of-view (FoV) was different, the FoVs were manually normalized across devices by ensuring that the ordering of the phones and their relative distances from one another on the rig were accounted for. Hence the four devices were constrained to capture near-identical scenes. Of course, perfect alignment was not possible, and hence, while the scenes obtained are highly similar, they are not perfectly identical across phones.

The videos were captured using the default settings on each phone (for exposure, color, white-balance, focus etc.) and the touch-to-focus feature was never used. All videos were captured at a resolution of $1920 \times 1080$ progressive (1080p). While audio was captured, all audio was removed before the final set of videos was created, and no audio was played during the subjective study.

Long segments of videos were captured from which segments of fifteen seconds were selected based on distortion relevance, content variability, and content interest while maintaining as much temporal alignment as possible. Since we did not wish to re-compress the videos, the relevant segment of each video was simply copied over. This operation was achieved by seeking the closest I-frame in each segment that we wished to capture, then chopping 15-seconds of video from that time stamp onward. Since I-frame locations across phones may not be the same (due to different sampling rates or recording start times etc.), there remain small temporal misalignment in the video content. However, most of the phones have one-second groups of pictures (GOP) [31], so the temporal differences are limited to one second.

<table>
<thead>
<tr>
<th>Artifacts</th>
<th>Color</th>
<th>Exposure</th>
<th>Focus</th>
<th>Sharpness</th>
<th>Stabilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>35</td>
<td>34</td>
<td>34</td>
<td>34</td>
<td>35</td>
</tr>
</tbody>
</table>

B. Categorizing Videos by Dominant Distortion

Each video content was captured with an intended distortion in mind (from the six distortions under consideration). For instance, videos intended to test the quality of stabilization were captured by someone walking with the rig to simulate video shake similar to what might be experienced in regular recording. To capture exposure-related distortions, the camera rig was panned across a scene, moving from a very brightly-lit region to a poorly-lit region or vice-versa. To capture color-related distortions, scenes with bright and varied colors were chosen. Most video capture was performed without the aid of a tripod, as the rig provided enough stability. Every content in the database was categorized into one of the six ‘expertly intended’ distortion classes. Finally, a set of contents deemed to provide adequate perceptual separability was selected. The breakdown of capture devices and videos belonging to the different distortion categories for the collection of videos is listed in Table III. It is important to note that while each video is associated with a dominant distortion label, other distortions may be present in it, viz., the distortions are not disjoint. For instance, exposure-related distortions can introduce color distortions; poor video stabilization can impact focus; incorrect auto-focus can impact video sharpness, and so on.

Figure 2 shows side-by-side frame comparisons between Samsung Galaxy S6 videos and HTC One VX videos of the same content, for each distortion. As discussed later (in Section VI-D), these two phones were respectively assigned the highest and lowest overall subjective scores by the human subjects, among the eight phones we considered. Thus we
Fig. 2: Side-by-side comparisons between Samsung Galaxy S6 (first column) and HTC One VX (second column) for each distortion type.
chose to illustrate the differences in the quality of content captured by them. Note that distortions such as focus, which depends on the auto-focus capability of each phone, and stabilization, which depends on processing during video capture, are more apparent in temporal sequences, so their presence (or lack thereof) in Fig. 2 may not be obvious.

IV. SUBJECTIVE STUDY

A. Unbiased and Biased Study Groups

Upon finalizing the video content and the dominant distortions afflicting them, our next task was to conduct a subjective study and gather human opinion scores on these videos. We aimed to gather subjective scores under two different study protocols that we describe next. The participating subjects were also randomly assigned to these study setups.

1) A free viewing, unbiased study setup to obtain overall subjective quality scores recorded by subjects while freely watching videos.

2) A distortion-guided study setup to gather opinion scores assessing the degree to which each of the six distortions affected a subject’s perceived quality.

Motivation for a free-viewing study setup: Under this study setup, a subject was presented with videos in no particular order and was asked to provide an overall quality opinion score at the end of each video’s playback. The presented videos were randomized across contents and dominant distortion categories. However, in order to understand which aspect of the viewed video had influenced the subject’s opinion score, we gathered additional feedback from each subject. Specifically, once the subjects provided their opinion score on a video, another screen appeared asking the subject to select the distortion that most influenced their judgment from a list of seven options (the six distortions mentioned earlier along with a seventh option “no apparent visual distortions”). Gathering subjective dominant distortion labels is useful for discerning the distortions that the observers tended to focus on. More importantly, the overall subjective quality scores obtained under this setup could help in the development of automatic, distortion-agnostic video quality predictors. We call this group of subjects the unbiased subject group, since subjects belonging to this group were not asked to attend to any particular distortion, i.e., no prior information about the afflicting dominant distortion was provided to them, which would have biased their quality opinion.

Motivation for a distortion-guided study setup: One of our key ambitions was to design a database that would also allow us to better understand how each of the considered dominant distortions influences human judgments. The availability of subjective scores that represent the affect of each individual distortion on perception could be useful for designing distortion-specific video quality metrics. Such metrics, particularly for these six distortions, would help us understand how different camera devices handle each of these major distortions. These distortion-specific metrics and the corresponding subjective data could also lead to insights regarding how perceptually significant the performance differences of various mobile devices are. These insights could in turn guide camera designers, since, the causes or ameliorating factors of these six distortions during video capture may be closely tied to certain modules in the camera video processing pipeline. For example, subjective opinion scores indicating that a particular phone handles noise distortions better than other phones could help smartphone camera developers assess and improve their internal denoising algorithms.

However, since these six distortions can be subtle, measuring perceptual quality arising from the underlying distortions becomes a challenging task. Under free-viewing conditions, it is likely that perceived quality might be influenced by factors other than the expertly determined dominant distortion. In other words, when videos are presented at random and subjects are asked to provide their perceived quality scores, subjects may not necessarily look for specific distortions in isolation.

In order to overcome this challenge, we conducted a guided study. Under the guided study protocol, we informed the subjects a priori which type of distortion to be on the alert for while viewing the videos. These biased subjects were each 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 attending to, was displayed on the screen before each distortion-specific test session. We refer to these subjects as the biased subject group, since knowing an expected distortion would ostensibly bias a subject’s quality opinion score. We believe that this way of guiding a subject towards perceiving a given dominant distortion was an effective way to gather human opinion scores that accurately reflect how each of the selected distortion influences human judgments of quality.

Since these subjects were guided towards expected distortions, there was no way to pool their scores across distortions towards understanding the ‘overall performance’ of a device (or an objective VQA model) across distortions. For those tasks, we relied on the subjective data obtained from the unbiased study setup.

B. Subjective Testing Display

We designed a user interface for the subjective study using the XGL toolbox [32] with MATLAB 2015b on a Windows PC. The XGL toolbox was designed for the purpose of presenting psychophysical stimuli to human observers, and we encountered no display issues while using it. Each video sequence was stored in raw YUV 4:2:0 format and was loaded in its entirety into memory before being displayed to subjects in order to avoid any potential playback latencies. The graphics card that was used was an ATI Radeon X600, and the video sequences were displayed on an ASUS VG248QE monitor at their native 1920 × 1080 resolution. Prior to the subjective tests, the display monitor was calibrated using the Datacolor Spyder5PRO Display Calibration System [33], which measures ambient lighting conditions and provides on-screen guidance on changing the monitor settings so that colors
TABLE III: Number of videos per phone and distortion. The rows are indexed by the eight mobile phones used to capture video content. The columns are indexed by the six dominant distortion categories.

<table>
<thead>
<tr>
<th>Phone</th>
<th>Artifacts</th>
<th>Color</th>
<th>Exposure</th>
<th>Focus</th>
<th>Sharpness</th>
<th>Stabilization</th>
<th>Total number of videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samsung Galaxy GS5</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>Samsung Galaxy GS6</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>33</td>
</tr>
<tr>
<td>HTC One VX</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>32</td>
</tr>
<tr>
<td>Apple iPhone5S</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>21</td>
</tr>
<tr>
<td>LG G2</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>Nokia Lumia 1020</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Samsung Galaxy Note4</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>Oppo Find 7</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
<td>35</td>
<td>34</td>
<td>34</td>
<td>34</td>
<td>35</td>
<td>208</td>
</tr>
</tbody>
</table>

Each subject completed three sessions of approximately 30 minutes each. Each biased session consisted of two 15-minute halves, each sub-session 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. Each subject was shown each test video in the database exactly one time. The video playlist order for each session was randomized for each subject, and care was taken such that consecutive videos would always contain unique content.

D. Training Phase

Subjects belonging to both study groups participated in a short training session before the actual study began. Biased subjects watched three or four training videos preceding each of their 15-minute distortion-focused sessions. The set of training videos for each distortion was of a single content, but the videos were captured from different mobile devices. These videos contained sufficient perceptual separability to enable the subjects to understand how each distortion might manifest in a video and to give them practice judging each distortion. Unbiased subjects watched six or seven training videos preceding each of their three 30-minute sessions. The videos were of random content and contained at least one video from each of the expertly categorized distortion categories. These training videos allowed the unbiased subjects to practice viewing videos while focusing on overall video quality and determining which distortion, if any, impacted their judgment the most. The training videos were not included in the set of 208 videos contained in the database, and the training sessions were primarily intended to help subjects become familiar with the testing procedure, the interface, the distortions, and the ratings scale. The training videos were of the same length (15 seconds) as all the videos in the database and were captured in the same manner as the test videos. They were also categorized by the expert-deemed dominant distortions.

E. Test Methodology

Next, we briefly describe the study pipeline for both the biased and unbiased groups.

1) Biased study pipeline: Before the beginning of a distortion-focused sub-session, the subjects were informed via a message on the screen regarding which distortion to attend to when viewing the videos that would be presented during...
Fig. 3: An example of the distortion-specific instructions displayed to the subjects in the biased group prior to each of the training phases of the sub-sessions.

Fig. 4: The continuous rating bar that is displayed to the subject after each video (for both biased and the unbiased groups).

Fig. 5: The screen that was displayed to the unbiased subjects to select a single, dominating distortion that impacted their perception of quality.

the next fifteen minutes (Fig. 3). The subjects were allowed to refer to the description of these distortions throughout the study (a hard copy of the instructions was made accessible to each subject for reference). Once a subject understood and agreed to proceed, videos afflicted with a particular dominant distortion were displayed one after the other. Following the presentation of each video, a screen with a message was displayed to select and submit a quality score based on their perceived quality while focusing on the expertly labeled distortion for that video (Fig. 4). Upon submitting this rating, the next video was displayed and the process repeated. Figure 6 illustrates the pipeline of the biased study.

2) Unbiased study pipeline: As mentioned, each unbiased study session lasted for about 30 minutes, during which videos afflicted by any of the six dominant distortions were presented in a random order. Following the presentation of each video, the unbiased subjects were instructed to select their ratings based on their overall perceived quality. Upon submitting this quality score, the subjects were presented with a screen with instructions to select the distortion which dominated their judgment (Fig. 5). In addition to the six distortions under consideration, a seventh option “no apparent visual distortions” was also listed on this page, allowing the subject to indicate that none of the six distortions bothered them for the video they just viewed. A subject was allowed to choose only one of the seven options presented on this screen. Figure 7 illustrates the pipeline of the unbiased-study.

The quality scale that was presented at the end of the video playback for both the groups was a continuous ratings bar (Fig. 4) with the cursor initially set at the center, qualitatively Fair, so as to not bias the subject. The left end of the bar was labeled Bad, while the right end was labeled Excellent. Labels Poor and Good were equally spaced between Bad, Fair, and Excellent reflecting the ITU-R Absolute Category Rating (ACR) scale [34]. Subjects used a mouse to move the slider on the quality scale to their desired rating. 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, while unbiased subjects were instructed to take approximately 15 seconds to submit both their perceived quality rating and their ‘dominant distortion’ selection. Time guidelines were given so that the subjects would know to provide their feedback immediately following their viewing period, so that their recorded judgments would still be fresh in their minds. The subjects were not allowed to go back and change their score after submission. Every subject (from either study group) viewed and rated around 70 videos during a session, and all of the videos in the database over the span of three sessions. Therefore, each video was viewed and rated by each subject. The quality scores delivered by a subject were labeled as either ‘biased’ or ‘unbiased’ depending on the study group they were assigned to. All of the subjects (from both biased and unbiased groups) participated in a short survey at the end of the study (more details in Sec. VI-C).

V. PROCESSING OF SUBJECTIVE SCORES

As mentioned, the LIVE-Qualcomm Mobile In-Capture Video Quality Database does not contain pristine, high-quality, “reference” videos, thus distinguishing itself from most existing VQA databases. This also means that it was not possible to compute differences between the ratings obtained on distorted test videos and corresponding high-quality reference videos to obtain difference mean opinion scores (DMOS).
Fig. 6: Pipeline of the biased study session, which comprises two sub-sessions. Note that in the training phase, a subject followed the same steps as those followed in the test phase (i.e., watch a video and rate it.)

![Diagram of the biased study session](image)

Thus, we simply computed mean opinion scores (MOS) for each video. If we let $s_{ijk}$ denote the raw score assigned by subject $i$ to video $j$ during session $k$, we compute Z-scores $z_{ijk}$ [35] as follows:

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

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

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

to account for each subject’s variability in their use of the quality scale during each session $k$, 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, because, as mentioned earlier, biased subjects complete two 15-minute sub-sessions back-to-back in one session. Thus, we considered their use of the quality scale to be similar in those two sub-sessions. Moving forward, we followed similar data processing steps as done in [5].

Because each subject saw each test video exactly once, we combined the Z-scores across all sessions to create a matrix $\{z_{ij}\}$ corresponding to the Z-score assigned by subject $i$ to video $j$, where $j = 1, 2, ..., N$ indexes the $N = 208$ test videos in the LIVE-Qualcomm Mobile In-Capture Video Quality Database.

Though the task of assessing perceptual video quality is highly subjective, we wanted to retain only those subjects who were able to rate the videos consistently. Towards this end, we followed the subject rejection procedure detailed in the ITU-R BT.500-11 recommendation [34] using the Z-scores, which accounts for differences in the use of the quality scale between subjects as well as differences in the use of the quality scale by each subject between sessions. With the scores collected in our study, 3 subjects—one biased and 2 unbiased—out of 39 were rejected, leaving the scores of 18 biased and 18 unbiased subjects for analysis.

If the Z-scores are normally distributed, then $> 99\%$ of the scores will lie in $[-3,3]$, as was indeed the case with our data. Thus, as was done in [5], the Z-scores were linearly rescaled so that the scores lay in the range $[0,100]$: $z'_{ij} = \frac{100(z_{ij} + 3)}{6}.$

Finally, we computed the Mean Opinion Scores (MOS) of each video as the mean of the rescaled Z-scores:

$$\text{MOS}_j = \frac{1}{M} \sum_{i=1}^{M} z'_{ij},$$

where $M = 18$ subjects after subject rejection in both the biased and the unbiased groups. Thus, throughout our analysis, we used the average Z-scores as the MOS. We conducted the above analysis on the biased and unbiased groups separately. The MOS scores obtained from the unbiased group were found to lie in the range $[16.56, 73.64]$, while for the biased group,
they fell in the range [13.37, 70.98]. The mean of the standard deviations of the Z-scores obtained from all subjects across all the videos was 11.95 for the biased group and 11.55 for the unbiased group.

The distribution of the MOS scores from the biased and the unbiased groups is depicted in Fig. 8. These plots indicate that there is a good perceptual separation of distortions in the database and that the videos broadly span the entire quality range.

VI. SUBJECTIVE SCORE ANALYSIS AND FINDINGS

A. Subjective Scores of the Biased and Unbiased Groups

We first compared the subjective scores obtained from the biased and unbiased groups. Since the subject pool was divided into biased and unbiased cohorts, we were interested in understanding the impact on the reported subjective quality of biasing the viewers to attend to particular distortions. Therefore, for each of the six distortions, we plotted the unbiased MOS scores (i.e., the MOS values obtained from the unbiased group) against the biased MOS scores, as illustrated in Fig. 9. We also computed the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank Ordered Correlation Coefficient (SROCC) between the biased and the unbiased MOS for each distortion category and report them in Fig. 9. The overall SROCC and PLCC across all 208 videos were found to be 0.76 and 0.80 respectively. These reasonably high correlation scores indicate that the biased and the unbiased groups generally agreed in regards to the perception of video quality. Figure 9 also suggests that color and sharpness distortions were difficult to perceive by both biased and unbiased subjects, leading to relatively low correlation scores. This observation is also supported by the responses to an end-of-study questionnaire that we administered to each subject upon completion of their third session. In the questionnaire, one question asked which distortions were the hardest to detect, and about 53% and 29% of both biased and unbiased subjects selected color and sharpness distortions as the hardest to detect, respectively.

B. Distortion Labeling by the Unbiased Groups

As mentioned in Sec. IV-A, upon viewing and rating every video, the unbiased subjects were also asked to select the distortion that dominated their perception of the quality of the video just watched. We adopted a majority voting policy to aggregate the distortion category labels obtained on every video from several subjects. For all the 208 videos in our database, the degree of agreement between the distortion labels obtained from the unbiased subjects and the expert labels assigned to each video (Sec. III-B) was visualized in a confusion matrix, as shown in Fig. 11. The percentage of the ‘correctly’ labeled videos is located along the diagonal of this matrix. This visualization is useful for understanding which distortions were the most difficult to detect and distinguish from one another.

From the confusion matrix, we can draw a few insights:

- Exposure and color-related distortions seemed to be the hardest distortions to detect and distinguish, given that about 47% of the expertly labeled videos afflicted with exposure-related distortions were labeled by the unbiased subjects as suffering from color artifacts. This is probably because variations of the exposure levels can affect the representation of colors in the captured videos. This likely produced the observed high degree of confusion between the two distortion categories.

- Subjects were able to identify the videos that were impaired by stabilization distortion reasonably well.

- A large fraction of the videos (38.24% and 34.29% respectively) suffering from sharpness and color artifacts were classified as not having any dominant distortion present, corroborating the subtlety of these distortions. This observation also sheds light on the degree of sensitivity to certain kinds of distortions experienced by the human subjects.

- Overall, unbiased viewers did not agree with the expertly labeled distortions, despite the detailed instructions and descriptions of each of the six distortions. The overall distortion classification accuracy over all 6 distortions was found to be 26.44%.

C. Analyzing the Survey Responses

As mentioned earlier, all of the subjects participated in a survey at the end of the study. We summarize all of their responses in the following sections.

1) Level of concentration of the subjects: We asked both the biased and the unbiased group subjects to rate their level of concentration throughout the study on a scale of 1-5, with 1 indicating that the subject was very distracted and 5 indicating that the subject was very concentrated throughout the study. We found that the average concentration level of the biased subjects was 4.03, and that of the unbiased subjects was 4.15. The average standard deviations of the reported concentration levels for biased and unbiased groups were 0.69 and 0.67, respectively.

2) Level of interest in the test video contents: We also asked all of the subjects to rate their interest level in the test video content on a scale of 1-5, with 1 indicating very boring content and 5 indicating very interesting content. Biased and unbiased subjects reported an average interest level of 2.89 and 3.02, respectively.

3) Most annoying distortions: One of the questions posed to the subjects in the survey was to select the distortion that most annoyed them, from a list of 7 options (6 distortions and a seventh option of ‘None’). We plotted the histograms of subject responses in Fig. 10 (a). This plot suggests that unbiased subjects were mostly annoyed by focus-related distortions, while the biased subjects were mostly annoyed by stabilization-related distortions. This possibly suggests that awareness of imminent shake distortion heightens the effect, but conversely for autofocus distortions.

4) Distortions that are hardest to detect: We also asked the subjects to report the distortion that was the most difficult to detect in the test videos using the same list of 7 options
(6 distortions and a seventh option of ‘None’). We plotted the distribution of subject responses to this question in Fig. 10 (b). This plot clearly indicates that subjects from both the groups found color and sharpness distortions to be the hardest to detect.

We believe that other factors, such as content, could play a significant role in making one distortion more difficult or annoying to perceive or detect. A systematic study focused exclusively on the interplay of video content and distortion severity and annoyance might help us better understand the cause for the behavior observed in Fig. 10.

D. Subjective Performance of Cameras

1) Per-Distortion Ranking: As mentioned in Sec. III-A, every video content was simultaneously captured by a group of four phone cameras. For each distortion category, we computed an average of the MOS values over all the video contents captured by each mobile camera. This average MOS was used as an indicator of the performance of each phone in regards to each of the six in-capture distortions under consideration. If a phone had a very high average MOS over all the video contents, then it is reasonable to believe that either the presence of visual distortions in the videos was minimal, or they did not have a significant impact on an end user’s perception of quality. For the purpose of this analysis, we combined the Z-scores obtained from both the biased and the unbiased study groups.

Figure 12 illustrates the average MOS and the standard deviations of the eight phones per distortion category. In each of these plots, the phones are ranked by highest to lowest average MOS values (i.e., best to worst performing). Note that not all eight phones are ranked for each distortion category, because phones that contributed fewer than two videos to a distortion category were not considered (see Table III for the distribution of the number of videos captured per distortion per phone). These rankings suggest that while different phones

Fig. 8: Scatter plot of MOS values obtained from the biased group (left) and the unbiased group (right) on all the videos in the database.

Fig. 9: The biased and unbiased scores plotted against each other along with a logistic fit, computed using [36].
might perform better in regards to each of the six individual distortions, overall, the cameras in Samsung’s Galaxy GS6 and Apple’s iPhone 5S performed better than the other camera devices.

2) Overall Subjective Ranking of the Phones: We further extended our analysis and computed the average MOS values across all the videos in the database (and over all the distortions). This was done to reveal the overall performances of the camera devices across distortions. Figure 13 presents the average MOS values per phone over all of the videos in the database (ranked from highest to lowest)\(^3\). As may be observed from this plot, the Samsung Galaxy GS6 had the highest overall MOS values, followed by the Apple iPhone 5S and the Nokia Lumia 1020. However, it is interesting to note that the differences between the overall MOS values was small for the top-performing phones.

We acknowledge that comparing the average MOS across different contents is only one simple way to understand a camera’s performance when capturing and processing authentic distortions. A more rigorous analysis technique could be designed for each particular distortion, which could help draw more meaningful and deeper insights on individual camera capabilities.

E. Subjective Phone Rankings vs. DxO Rankings

As mentioned in Sec. II, DxOMark [25] is a frequently used and referenced Internet site that uses a suite of software tools, collectively called the DxO Analyzer [26], to report the results of measurements of the quality of commercially available cameras and lenses. The method that DxO Analyzer uses to compute DxOMark scores, and the kinds of images and videos used to test the camera devices, are not made publicly available. Further, the methods employed by the DxO Analyzer tool [26], [27] to aggregate measurements derived in software and hardware to yield a final DxOMark score of a given camera device are also unknown. Despite these limitations, we sought to study the degree of agreement between the published DxOMark rankings of the mobile phones with the subjective phone rankings given in Sec. VI-D. Towards this end, we computed the Spearman Rank Ordered Correlation Coefficient (SROCC) and Pearson’s Linear Correlation Coefficient (PLCC) between:

(a) overall subjective rankings (computed in Sec. VI-D2) and the overall DxO rankings
(b) subjective distortion-specific rankings (computed in Sec. VI-D1) and distortion-specific DxO rankings.

The overall DxO rankings were computed from the DxO scores of the seven\(^4\) phones under consideration. The correlation scores between the distortion-specific and overall rankings are presented in Table IV. From the correlation scores in Table IV, it may be seen that there are strong positive correlations for the artifacts and color distortion categories; however, for the other distortion classes, the subjective rankings of the phones included in our study did not correlate very well with the corresponding DxO distortion-specific rankings. One inference that might be drawn from this is that objective technical

\(^3\)The distribution of the number of videos captured by each phone is already presented in Table II

\(^4\)We excluded the HTC One VX from our analysis because it has not been analyzed and rated by the DxOMark tools.
Fig. 12: Aggregate MOS values (and standard errors presented as error bars) computed on videos captured from each of the eight phones and for each distortions category. Phones that captured less than two videos for any particular distortion were eliminated from the analysis.

Fig. 13: Overall MOS values (and standard errors presented as error bars) over all the videos captured from each of the eight phones considered in the subjective study. Phones with higher overall MOS values were considered to be better performing than phones with low MOS values.

camera measurements are not necessarily very predictive of ultimate perceptual picture quality. This suggests that deeper studies of this question may be warranted, including much larger sets of data specific to these questions.

**Analysis of comparative phone ranking:** There are caveats regarding our comparison of phone rankings derived from the subjective study and the DxOMark rankings. First, our evaluation protocol for measuring phone camera performance is quite different from that used by DxOMark. We used aggregated MOS from nine distinct video contents to evaluate the performance of each phone (per distortion), while DxOMark’s ranking mechanism is unknown. Further, the impact of complex interactions between multiple coexisting distortions on the overall perception of quality is difficult to quantify. In any case, since humans are the ultimate consumers (and arbiters) of digital media, we believe that any rankings of camera quality should also take subjective opinion scores into account.

We could also have computed distortion-specific rankings by selecting the most challenging content for each distortion, i.e., the content with the lowest overall MOS, and ranked phones based on the MOS scores for that content. In this way, phones would be ranked according to how well they captured the content that most severely tested each camera’s capabilities with respect to each distortion. However, each video content in our database was captured by a maximum of four phone cameras (as described in Sec. III-A). Using just one content per distortion would only allow for ranking of the 3 or 4 phones which were used to capture that particular content. For this reason, we chose to rely on aggregate MOS scores across content types as a better and more stable measure of performance.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SROCC</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>PLCC</td>
<td>0.94</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
<td>0.80</td>
</tr>
</tbody>
</table>
VII. PERFORMANCE OF OBJECTIVE QUALITY METRICS

We evaluated the performance of some publicly available no-reference generic IQA and VQA algorithms on the LIVE-Qualcomm Mobile In-Capture Video Quality Database and reported the performance of each in Tables V and VI. We evaluated each algorithm over each distortion category and also over the entire dataset. The data collection for every experiment was divided into non-overlapping training and testing 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 [15], FRIQUEE [13], and BRISQUE [37] are learning-based models, in each iteration, a model was trained from scratch on 80% of the data and evaluated on 20% of the data. We used their publicly-available implementations to extract quality-relevant features from every alternate video frame and average-pooled the features to derive a single feature vector per video. An SVR with a radial basis function was used to train and test V-BLIINDS, FRIQUEE, and BRISQUE. The kernel parameters for every experiment setup were chosen via cross-validation. VIIDEO [38] and NIQE [39] are training-free algorithms, hence for a fair comparison with the learning-based models, we report their performance on the test data alone. NIQE and VIIDEO scores were passed through a logistic non-linearity [40] to map to MOS before computing LCC.

We also applied the temporal technique described in [41] on the NIQE frame scores as an alternative to average pooling method. With this method, per-frame scores were first pooled into low- and high-quality groups using k-means clustering (with \( k = 2 \)). Then these scores were combined using a linear weighting function that gave lower-quality frames more influence on the overall quality score.

When evaluating the overall performance of an algorithm over the entire video collection, we used the MOS scores obtained from the unbiased study group as ground truth scores. To gain insights on the distortion-specific performance of the various IQA/VQA algorithms, we considered only those videos belonging to each distortion category (using the same 80/20 test setup). To evaluate the objective quality prediction models on specific distortions, we used the MOS values from the biased group. Tables V and VI present the median Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Ordered Correlation Coefficient (SROCC) scores computed between the predicted and ground truth quality scores over the 100 iterations.

There are some notable distortion categories where the blind models under consideration performed particularly poorly or well. For instance, BRISQUE [37] correlated poorly with videos categorized as having color distortions, which is likely due to the fact that its NSS-based features are only extracted from luminance values. FRIQUEE [13] on the other hand, includes color-based features and thus performed better than BRISQUE in this category. Also observe that temporal pooling of the NIQE scores correlated well with the subjective opinions as compared to conventional average pooling. Many of the natural scene statistics-based models performed quite well on videos in the artifacts category, which agrees with previous work that showed that NSS-based features effectively predict the impacts of noise, blockiness, and blur on perceived quality [15], [37], [39]. V-BLIINDS performed well on most of the distortion categories, but FRIQUEE achieved the overall top performance on the entire video collection.

There are also dynamic changes of distortion over time in many of the videos, which a single overall quality opinion score gathered per video is often unable to effectively capture. Gathering continuous-time subjective scores to better understand the effects of complex factors such as recency or hysteresis [42] on the perception of quality is an interesting direction to pursue in the future.

In summary, the natural scene statistics-based blind picture and video quality assessment models performed reasonably well, with the exception of VIIDEO [38]. This recent model relies only on temporal scene statistics to make quality predictions, hence it performs well on the motion-rich LIVE VQA database [5]. The videos in the current database contain rich spatial content and authentic mixtures of spatial distortions, in addition to motion and ego-motion. The interactions between these distortion mixtures presented difficulties to the VIIDEO algorithm. Overall, we believe that existing blind IQA/VQA algorithms have room for improvement regarding their ability to accurately predict the quality of videos suffering from naturally-occurring, in-capture video distortions.

VIII. CONCLUSION AND FUTURE WORK

We designed a new database of videos captured using modern mobile camera devices exhibiting authentic in-capture distortions. This new database of videos and associated subjective scores provides a potentially valuable tool that may be used to address some of the limitations of current VQA databases [5–7], [9] 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 and generic (from our unbiased study) blind VQA models that perform well on videos affected by complex in-capture distortions. We are also interested in adapting such models to perceptually optimize mobile cameras and lenses.

REFERENCES

<table>
<thead>
<tr>
<th>Artifacts</th>
<th>Focus</th>
<th>Sharpness</th>
<th>Stabilization</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7638</td>
<td>0.6900</td>
<td>0.5795</td>
<td>0.7034</td>
<td>0.7349</td>
</tr>
<tr>
<td>0.6402</td>
<td>0.5371</td>
<td>0.4821</td>
<td>0.6429</td>
<td>0.5585</td>
</tr>
<tr>
<td>0.6078</td>
<td>0.5371</td>
<td>0.4821</td>
<td>0.6429</td>
<td>0.5585</td>
</tr>
<tr>
<td>0.6766</td>
<td>0.5371</td>
<td>0.4821</td>
<td>0.6429</td>
<td>0.5585</td>
</tr>
<tr>
<td>0.8386</td>
<td>0.6455</td>
<td>0.8077</td>
<td>0.7138</td>
<td>0.6653</td>
</tr>
<tr>
<td>0.2888</td>
<td>0.3312</td>
<td>0.2073</td>
<td>0.3697</td>
<td>0.0982</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Artifacts</th>
<th>Focus</th>
<th>Sharpness</th>
<th>Stabilization</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7500</td>
<td>0.6701</td>
<td>0.5536</td>
<td>0.6607</td>
<td>0.6795</td>
</tr>
<tr>
<td>0.6071</td>
<td>0.5371</td>
<td>0.4821</td>
<td>0.6429</td>
<td>0.5585</td>
</tr>
<tr>
<td>0.5000</td>
<td>0.3314</td>
<td>0.2073</td>
<td>0.3697</td>
<td>0.0982</td>
</tr>
<tr>
<td>0.5357</td>
<td>0.3314</td>
<td>0.2073</td>
<td>0.3697</td>
<td>0.0982</td>
</tr>
<tr>
<td>0.7321</td>
<td>0.6071</td>
<td>0.5371</td>
<td>0.6607</td>
<td>0.6795</td>
</tr>
<tr>
<td>0.1786</td>
<td>0.1429</td>
<td>0.0714</td>
<td>0.1071</td>
<td>0.1414</td>
</tr>
</tbody>
</table>