Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies

Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, *Fellow, IEEE*, and Gustavo de Veciana, *Fellow, IEEE*

***Abstract*—We introduce a new video quality database that models video distortions in heavily-trafficked wireless networks and that contains measurements of human subjective impressions of the quality of videos. The new LIVE Mobile Video Quality Assessment (VQA) database consists of 200 distorted videos created from 10 RAW HD reference videos, obtained using a RED ONE digital cinematographic camera. While the LIVE Mobile VQA database includes distortions that have been previously studied such as compression and wireless packet-loss, it also incorporates dynamically varying distortions that change as a function of time, such as frame-freezes and temporally varying compression rates. In this article, we describe the construction of the database and detail the human study that was performed on mobile phones and tablets in order to gauge the human perception of quality on mobile devices. The subjective study portion of the database includes both the differential mean opinion scores (DMOS) computed from the ratings that the subjects provided at the end of each video clip, as well as the continuous temporal scores that the subjects recorded as they viewed the video. The study involved over 50 subjects and resulted in 5,300 summary subjective scores and time-sampled subjective traces of quality. In the behavioral portion of the article we analyze human opinion using statistical techniques, and also study a variety of models of temporal pooling that may reflect strategies that the subjects used to make the final decision on video quality. Further, we compare the quality ratings obtained from the tablet and the mobile phone studies in order to study the impact of these different display modes on quality. We also evaluate several objective image and video quality assessment (IQA/VQA) algorithms with regards to their efficacy in predicting visual quality. A detailed correlation analysis and statistical hypothesis testing is carried out. Our general conclusion is that existing VQA algorithms are not well-equipped to handle distortions that vary over time. The LIVE Mobile VQA database, along with the subject DMOS and the continuous temporal scores is being made available to researchers in the field of VQA at no cost in order to further research in the area of video quality assessment.**

Manuscript received November 05, 2011; revised April 21, 2012; accepted August 04, 2012. Date of publication August 08, 2012; date of current version September 12, 2012. This work was supported in part by the National Science Foundation under Grant CCF-0728748 and in part by Intel and Cisco corporation under the VAWN program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Touradj Ebrahimi.

A. K. Moorthy was with the Laboratory for Image and Video Engineering (LIVE), and the Wireless Networking and Communications Group (WNCG), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 1084 USA. He is now with Texas Instruments, Inc., Dallas, TX 75243 USA (e-mail: anushmoorthy@gmail.com).

L. K. Choi and A. C. Bovik are with the Laboratory for Image and Video Engineering (LIVE), and the Wireless Networking and Communications Group (WNCG), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA. (e-mail: larkkwonchoi@gmail.com; bovik@ece.utexas.edu).

G. de Veciana is with the Wireless Networking and Communications Group (WNCG), Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-1084 USA (e-mail: gustavo@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2012.2212417

***Index Terms*—Mobile video quality, objective algorithm evaluations, subjective quality, video quality assessment, video quality database.**

I. Introduction

GLOBAL mobile data traffic nearly tripled in 2010 for the third consecutive year, exceeding three times the data volume of the entire global Internet traffic just 10 years ago [1]. According to the Cisco Visual Networking Index (VNI) global mobile data traffic forecast, mobile video traffic accounts for nearly 50% of mobile traffic, and it is predicted that this percentage will steadily increase to more than 75% by 2015. As smartphone usage explodes along with mobile enabled video streaming websites such as Amazon Video on Demand, Hulu, Itunes, Netflix and YouTube[1], it is clear that video traffic on mobile devices will continue to account for an increasingly significant portion of mobile data traffic. While this bodes well for end-users able to watch HD quality video clips at the touch of a button, the picture is not completely rosy for those who provide the spectrum.

In early 2010 U.S. Federal Communications Commission (FCC) Chairman Julius Genchowski summarized the problem succinctly – "The record is pretty clear that we need to find more spectrum" [3]. According to Peter Rysavy, a wireless analyst, mobile broadband will surpass the spectrum available in mid-2013 [4]. The paucity of bandwidth is evident from the bandwidth caps that most of the wireless providers in the U.S. have recently imposed on data-hungry users.

Given that video traffic accounts for a significant portion of this mobile data traffic, the development of frameworks for wireless networks is a topic of intense study. One particularly promising direction of research is *perceptual optimization* of wireless video networks, wherein network resource allocation protocols are designed to provide video experiences that are measurably improved under perceptual models.

The final receivers of most videos transported over wireless networks are humans and therefore visual perception is the ultimate arbiter of the received visual experience. The human visual system (HVS) is complex and highly non-linear, so treating video data as any other data in solving the resource allocation problem can lead to suboptimal end-user perceptual experiences. The study of models for resource optimization that model video traffic using perceptually relevant features is easily motivated. A key ingredient in developing these tools is understanding and predicting user perception of video quality on mobile devices by conducting large scale human/subjective studies.

[1]Netflix usage accounts for almost 30% of all downstream traffic during peak hours; YouTube accounts for just over 11% (as of May 2011) [2].

Here, we describe an extensive study that we have recently conducted in order to gauge subjective opinion on HD videos when displayed on mobile devices.

Several researchers have conducted subjective video quality studies with various aims [5]–[10]. Significant effort has also been applied to designing objective algorithms that are capable of predicting visual quality with high correlation against subjective perception [11]–[14]. Previous subjective studies on VQA have been performed on large format displays such as CRT/LCD monitors, while typically distorted videos have included compressed videos (H.264/MPEG), videos transmitted over wireless/IP channels [5], [6] and jittered and delayed videos [15], [16]. While video quality on mobile devices has not been extensively researched, there have been a few studies on the quality assessment of videos on mobile devices.

Eichhorn and Ni performed a human study to evaluate the quality of H.264 scalable video codec (SVC) encoded video streams at QVGA and QQVGA resolutions on a 2.5-inch screen [17]. Each of the six 8-second clips were encoded at two spatial resolutions using 3 temporal layers and 4 quality layers. Thirty subjects rated the visual quality of the videos yielding a differential mean opinion (DMOS) score for each of the videos in the database. Based on the DMOS obtained, the authors analyzed the effect of reduced spatial resolution as well as reduced temporal sampling and quality. While the analysis presented is interesting, the low-resolution of the videos (QVGA/QQVGA) relative to those displayed by current mobile devices, the fact that some of the videos in the database were un-natural (eg., animations) and the unavailability of the database limit its current utility.

Knoche and colleagues evaluated image resolution requirements for MobileTV by conducting a large-scale human study where over 120 subjects participated (although each video only received 32 ratings) [18]. The subjects were asked to rate the quality of videos which had gracefully decreasing encoding bitrates (using Microsoft Windows Video V8 codec) and varying resolutions on a display of resolution 240×320. The results presented are quite valuable, especially since the authors also varied audio quality in the study. However, from an algorithm design-perspective, the lack of pristine reference videos as well as the manner in which some of the videos were artificially modified (eg., feeds from News which included text scrolls, picture-in-picture etc.), coupled with its unavailability again limits the usefulness of the database.

Jumisko-Pyykko and Hakkinen performed a subjective study where reference clips from video tapes were converted to digital video, then compressed using a variety of video codecs (H.263, H.264 etc.) [19]. The authors evaluated video-only as well as audio-video quality on the Nokia 6600 and the S-E P800. As with other studies of this nature, the very low frame-rates and bit-rates relative to current technology and the lack of public availability reduce the currency of the work.

Ries *et al.* evaluated the quality of five reference videos of 10-seconds each when compressed at varying frame-rates and bit-rates using the H.264/AVC baseline encoder [20]. The authors also detailed an algorithm that would evaluate the quality of these videos so that the objective scores produced would correlate well with the obtained human opinion scores. All of the limitations of the above databases apply to this one as well. Other studies on mobile devices include an investigation on context and its effect on quality [21], and a study of the effect of extremely low bit-rates on perceived quality [22].

Almost all of the above studies suffer from several of the following problems: (1) the dataset is of insignificant size, (2) the distortions and their severities considered are insufficient to make judgments on perception of quality, (3) the videos were obtained from unknown sources and contain unknown corruptions, (4) the video resolutions are too small to be relevant in today's world, (5) the human studies were conducted on a single device with a fixed display resolution and (6) the database is not publicly available. Realizing the need for an adequate and more modern resource, we have endeavored to create a database of broad utility for modeling and analyzing contemporary wireless video networks.

The LIVE Mobile VQA database consists of 200 distorted videos evaluated by over 30 human subjects on a small mobile screen, as well as 100 distorted videos evaluated by 17 subjects on a larger tablet display. The source videos were shot using a RED ONE digital cinematographic camera and the RAW data so obtained was used in the study. The database consists of videos at HD resolution (720p), distorted by a variety of distortions including compression and wireless channel transmission losses. More importantly, the LIVE mobile VQA database also includes dynamically changing distortions resulting in perceptible quality fluctuations in the videos over time.

A brief summary of the distortions follows. (1) Compression, using the H.264 scalable video codec (SVC) [23] to compress the video at four different compression rates. (2) Wireless channel packet-loss, where the H.264 compressed streams were passed through a simulated wireless channel. (3) Frame-freezes, including both live video freezes – loss of temporal continuity after freeze, and stored video freezes – no loss of temporal continuity after freeze. (4) Rate adaptation, where the compression rate is dynamically varied within a video stream between two compression rates. And finally, (5) Temporal dynamics, where the compression rate is varied between multiple compression rates with different rate-switching structures within a single video stream.

We collected and analyzed "summary" scores provided by the subject at the end of the presentation, and also continuously recorded scores that the subjects provided, thereby allowing researchers to understand how temporal quality scores are collapsed by the human into a final opinion score of the video. The database enables a new avenue of research – behavioral modeling of visual quality perception. Finally, the database and the subjective opinion scores (including the temporal scores) are being made available online in order to help further research in the area of visual quality assessment.

While this database and the associated human opinions scores and the analysis carried out below have tremendous value to the video quality assessment community, other fields of inquiry such as human behavior modeling; application and content driven analysis of human behavior; device and context-specific design of objective algorithms; video network resource allocation and so on may also seek benefit from the publicly available data.

Through the rest of this paper, we describe the construction of the database and perform an exhaustive analysis of the subjective opinion scores obtained from the study – for both the mobile and the tablet databases; a comparison between the two databases is also performed. We also analyze the temporal scores and evaluate various possible measures that explain how temporal subjective scores are pooled by humans to produce a final estimate of quality. Finally, we evaluate the performance of a wide range of image and video quality assessment (IQA/VQA) algorithms in terms of correlation with human perception for each distortion and across distortion categories; the analysis includes hypothesis testing and statistical significance evaluation[2].

II. Subjective Assessment of Mobile Video Quality

A. Source Videos

The source videos were obtained using a RED ONE digital cinematographic camera. The sequences of REDCODE (.r3d) images received from the MYSTERIUM sensor, using the RED 50 – 150 mm and 18 – 50 mm T3 zoom lens were stored as 12-bit REDCODE RAW data, at a resolution of $2K$ (2048×1152) at frame rates of 30 fps and 60 fps using the REDCODE 42 MB/s option to ensure the best possible acquisition quality. A tripod was used in most scenes and the ISO was set in the range 100 to 360 according to the weather – ISOs of 100 or 200 were used for outdoor scenes and 200 or 360 were used for indoor scenes; the shutter speed varied between 1/48 to 1/60 s. The automatic white balance mode was used. The RED drive was used to record the videos.

The source videos were then downsampled to resolution 720p (1280×720) and frame-rate of 30 fps, and the .r3d videos were converted into uncompressed .yuv files using a combination of the imresize (option: bicubic) function in MATLAB and VirtualDub. All of the source videos in the database are of duration 15 seconds. A total of 12 videos were selected for this study from a larger subset. These were chosen to be representative of a wide variety of content types that the user might experience. Two of these videos were used to train the subjects (see below) while the rest of the videos were used to perform the actual study. The list below describes each of the videos used in the study.

1) Friend Drinking Coke (fc): Shot at studio with tungsten light and gel. It shows different light ratios on the face with detailed muscle changes occurring under dim lighting. The camera was fixed.
2) Two Swan Dunking (sd): Shot at Lady Bird Lake, Austin Texas on a sunny morning. There are bright twinkles on the waves, and swans are seen dunking into the water. The camera tracked two of the swans.
3) Runners Skinny Guy (rb): Shot at a marathon race early in the morning. Many runners show diverse contrasts and colors and complex motions. The fixed camera zooms in and out.
4) Students Looming Across Street (ss): Shot on the campus of The University of Texas at Austin on a windy morning. Walking students loom towards the camera.
5) Bulldozer With Fence (bf): Shot at a construction area on a sunny afternoon. Different exposures of light, shadowing of trees, motion of bulldozer and complex textures produce a variegated scene. The camera pans across the screen from left to right.
6) Panning Under Oak (po): Shot under a large oak tree under a blue sky on a sunny afternoon. Many small leaves are visible moving slowly.
7) Landing Airplane (la): Shot at Austin-Bergstrom International Airport on a cloudy afternoon. The landing airplane exhibits fast motion, and the background changes rapidly. The camera tracked the airplane from upper right to lower left.
8) Barton Springs Pool Diving (dv): Shot at Austin's Barton Springs Pool on a sunny afternoon. There are sparsely moving people, and one diver who creates a splash. The camera was fixed.
9) Trail Pink Kid (tk): Shot at a Lady Bird Lake trail on a sunny morning. People walk or jog at various speeds in different directions. The camera was fixed.
10) Harmonicat (hc): Shot at Zilker Park in Austin on a sunny afternoon. A musician plays guitar and harmonica in front of a tree. The camera zooms in and out.
11) Fountain Vertical (fv): Shot at LBJ Library fountain on the campus of The University of Texas at Austin on a sunny morning. The fountain jets water into the air in front of a campus skyline. The camera was fixed.
12) Hyein BSP (hy): Shot at Austin's Barton Springs Pool on a sunny afternoon. A child with a colorful dress walks next to the water. The camera pans the scene from right to left.

Fig. 1 shows sample frames from the various video sequences.

B. Distortion Simulation

Each of the reference videos were subjected to a variety of distortions including: (a) compression, (b) wireless channel packet-loss, (c) frame-freezes, (d) rate adaptation and (e) temporal dynamics. In this section we detail how these distorted videos were created.

1) Compression: We used the JM reference implementation of the H.264 scalable video codec (SVC) to compress the 720p HD reference videos [23], [25], [26]. Since the SVC implementation does not allow rate control for layers above the base layer, we use fixed QP encoding. The QP was varied across videos and layers in order to produce the target bit-rates for each layer of every video. The videos were compressed using 6 SNR layers (temporal and spatial scalability were not evaluated in this study), and 4 of these layers (R_1, R_2, R_3, R_4; $R_1 < R_2 < R_3 < R_4$) were manually chosen for each video based on their perceptual separation. As other authors have argued, ensuring perceptual seperation between the videos in QA studies makes it possible for humans (and algorithms alike) to produce consistent judgements of visual quality [5], [6].

Since the video content is quite varied, the bit-rates for each of these layers varies across videos; all videos were compressed with rates between 0.7 Mbps and 6 Mbps. The choices of rates were based on commonly-used parameters for transmission of HD videos over networks as well as rates that are generally seen

[2]A highly condensed summary of the database appears in [24].

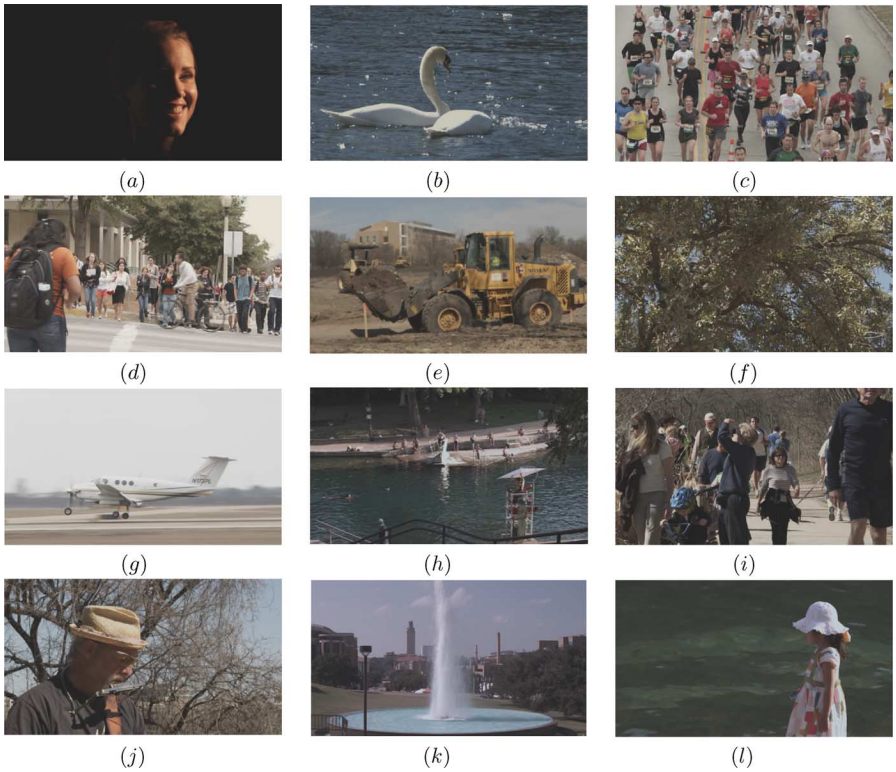


Fig. 1. Example frames of the videos used in the study. fv and hy were used for training the subjects while the rest of the videos were used in the actual study. (a) fc. (b) sd. (c) rb. (d) ss. (e) bf. (f) po. (g) la. (h) dv. (i) tk. (j) hc. (k) fv. (l) hy.

on wifi networks. The videos were encoded with an intra period of 16 and loss aware distortion optimization (LARDO) was enabled with packet-loss rates set to 3%. Instead of fixing the number of macroblocks per slice, the number of bytes per packet was fixed at 200 bytes – as recommended for wireless transmission of H.264 coded video [27].

Thus, for each video, four compressed SVC streams were created, yielding a total of 40 compressed videos.

2) Wireless Channel Packet-Loss: H.264 SVC compressed videos were transmitted over a simulated wireless channel in order to induce loss, thereby affecting perceptual quality. The simulated channel was modeled using an IEEE 802.11-based wireless channel simulator implemented in LabVIEW. The system comprised of a single link channel with coding, interleaving, QAM modulation, and OFDM modulation. A bit stream containing 2,000,000 bits was sent through a frequency selective channel with 5 taps at an SNR of 15 dB; 4QAM and a 1/2 rate convolutional code were used. These kinds of a bit-streams were sent 100 times, and for each transmission an error trace was created by XORing the transmitted bit-stream with the received bit-stream, which recorded the erroneous bit-locations. These error traces were used to induce errors in the compressed video streams. For each video, a random error-trace from the set of 100 traces was picked and applied, where a video packet was considered to be lost if one of the bits of the packet was erroneous [27]. Since the SVC decoder imposes certain requirements on decoding the video due to the layered architecture, care was taken to ensure that the loss of packets would not result in an error at the decoder.

Each of the compressed videos was transmitted over the wireless channel, resulting in a total of 40 wireless channel distorted videos.

3) Frame-Freezes: Two kinds of frame-freeze models were used to create distorted videos: frame freezes for (1) stored video delivery and (2) live video delivery. In the case of stored videos, frame-freezes do not result in the loss of a video segment from the video, i.e., the videos maintain temporal continuity after the freeze. On the other hand, frame-freezes in live video delivery result in a loss of video segments, i.e., a lack of temporal continuity.

For both of the above cases, the model for frame-freeze is as follows. For every x seconds of freeze (where the last frame in the buffer is displayed on the screen until the next frame arrives), the post-freeze video playback is of duration bx seconds ($b > 1$), i.e., the longer the user waits, the longer the post-freeze playback. In our simulations we chose $b = 1.5$.

Three stored video freeze lengths were modeled: (i) 1 second (short bursts of video playback with 8 freezes), (ii) 2 seconds (longer video playback, with 4 freezes) and (iii) 4 seconds (2 freezes, longest continuous video playback); the live video freeze length was set to be 4 seconds. In all cases, there was a lead-in time of 3 seconds, i.e., the first 3 seconds of the video playback did not incorporate a freeze. All frame-freezes were simulated on uncompressed reference videos.

A total of 40 frame-freeze distorted videos (4 for each reference video) were thus obtained.

4) Rate Adaptation: Psychovisual studies have demonstrated that humans are more sensitive to changes in a visual stimulus than to the magnitude of the stimulus [28]. In order to investi-

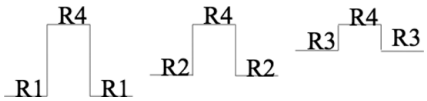


Fig. 2. Rate Adaptation: Schematic diagram of the three different rate-switches in a video stream simulated in this study.

gate whether such behavior translates to judgments of temporal quality, we simulated rate-changes as a function of time as the subject views a particular video. Specifically, the subject starts viewing the video at a rate R_X, then after n seconds switches to a higher rate R_Y, then again after n seconds switches back to the original rate R_X. Comparing such a rate-adapted stream with the appropriate compressed stream may provide important information regarding human behavioral responses to time-varying video data rates.

Such a scheme may also reveal whether humans prefer shorter durations of high quality content in the midst of a low quality stream, or if they prefer to view the low quality stream without any fluctuation in quality. Thus we may find answers to questions like: Does exposing the viewer to better quality increase his expectations, thereby reducing his quality rating for the lower quality segment of the stream? From a resource allocation perspective this condition will provide data that will allow for better allocation of resources, where 'better' is a function of the quality perceived by the end user. This condition may provide answers to questions like: Given that the channel is going to allow a rate higher than the current one for only n seconds before one is forced to revert back to the current rate, should one switch to a higher rate for n seconds, given that you are currently at rate R_X?

It should be clear from the above discussion that such behavioral aspects of quality perception may be a function of the difference between the initial rate and the final rate, as well as of the initial rate itself. Hence, we simulate three different rate switches, where $R_X = R_1, R_2$ and R_3 and $R_Y = R_4$. Although the duration n is another potential influence on human behavior, because of on the length of the subject's sessions, we fixed $n = 5$.

The three rate-adaptations which are illustrated in Fig. 2 yielded to a total of 30 rate-adapted distorted videos.

5) Temporal Dynamics: In the previous section, we simulated conditions that evaluated the effect that a single rate switch has on perceived quality. One would imagine that the subjective perception of quality is also a function of the *number* and *lengths* of the rate-switches that occur in a stream. In order to evaluate this, we simulated a multiple rate-switch condition, where the rate was varied between R_1 to R_4 multiple times (3). This is illustrated in Fig. 3. To ensure an objective comparison between the multiple and single rate-change scenarios, the two conditions are simulated such that the average bit-rate was the same in both cases.

Apart from multiple switches, one may intuit that subjective quality is also influenced by the *abruptness* of the switch, i.e., instead of switching directly between R_1 and R_4, it may be useful to evaluate conditions where the rate is first switched to an intermediate level Rz from the current level and then to the other extreme. Studying responses to this condition may reveal whether easing a user into a higher/lower quality regime is better

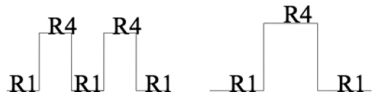


Fig. 3. Temporal Dynamics: Schematic illustration of two rate changes across the video; the average rate remains the same in both cases. Left: Multiple changes and Right: Single rate change. Note that we have already simulated the single rate-change condition as illustrated in Fig. 2, hence we ensure that the average bit-rate is the same for these two cases.

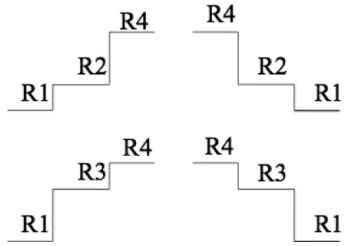


Fig. 4. Temporal Dynamics: Schematic illustration of rate-changes scenarios. The average rate remains the same in all cases and is the same as in Fig. 3. The first row steps to rate R_2 and then steps to a higher/lower rate, while the second row steps to R_3 and then back up/down again.

than abruptly switching between these two regimes. It should be clear that the intermediate rate Rz may have an impact on the perception of quality as well. Hence, we simulated the following rate-switches: (1) $R_1 - R_2 - R_4$, (2) $R_1 - R_3 - R_4$, (3) $R_4 - R_2 - R_1$ and (4) $R_4 - R_3 - R_1$, as illustrated in Fig. 4. Again, the average bit-rate remains the same across these conditions as well as over the conditions in Fig. 3.

Notice that the rate-changes illustrated in Fig. 4 form dual structures – including such models may also reveal whether the user is influenced by the quality observed towards the end of the video. Specifically, we seek to answer the question: Which of the following scenarios is preferable: ending the video with a high quality segment, or ending the video with a low-quality segment? Again, in addition to supplying data on human behavioral responses to time-varying video quality, answering these kinds of questions may also facilitate making better resource allocation decisions. A total of 50 distorted videos with varying temporal dynamics were thus created.

While it is impossible to plot all of the various temporal distortions simulated here, Figs. 5 and 6, show two examples of distorted frames from the distorted videos, along with the reference frames for comparison. The reader is invited to download the freely available database, in order to better visualize the distortions.

In summary, the LIVE Mobile VQA database consists of 10 reference videos and 200 distorted videos (4 compression + 4 wireless packet-loss + 4 frame-freezes + 3 rate-adapted + 5 temporal dynamics per reference), each of resolution 1280×720 at a frame rate of 30 fps, and of duration 15 seconds each.

C. Test Methodology

1) Design: A single-stimulus continuous quality evaluation (SSCQE) study [29] with hidden [5], [6], [30] was conducted over a period of three weeks at The University of Texas at Austin, LIVE subjective testing lab. Each subject was asked to view and rate the videos one video at a time. Each original, uncompressed reference video was randomly placed amongst the set of videos shown to each user in each session, although

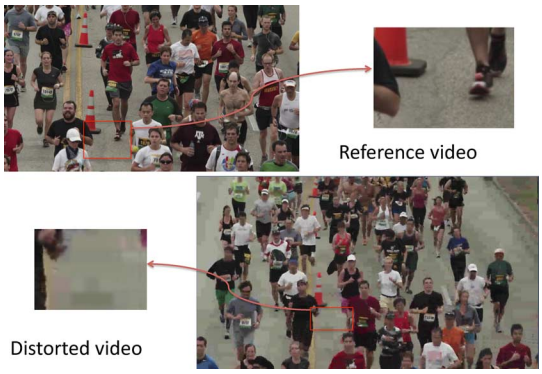


Fig. 5. Figure illustrating the spatial effect of the distortions simulated in this study for a frame from video 'rb'. Also plotted are the reference frame and a zoomed area for comparison purposes.

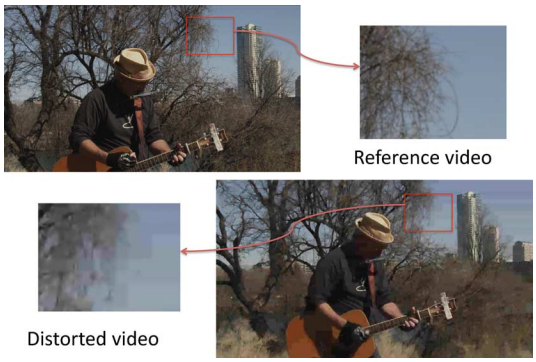


Fig. 6. Figure illustrating the spatial effect of the distortions simulated in this study for a frame from video 'hc'. Also plotted are the reference frame and a zoomed area for comparison purposes.

the subjects were unaware of their presence. The score that the subjects gave these 'hidden' references is representative of the bias that the subject carries. By subtracting the reference video scores from those for the distorted videos, the biases are compensated for yielding differential scores for each distorted video. We believe that SS with hidden reference studies are preferable to longer double-stimulus (DS) studies [5], [6]. Shorter studies make the study duration less likely to fatigue the subjects, while allowing the subjects to evaluate a larger set of conditions, for a given study duration. Perhaps most importantly, a SS study design better models real video experiences; typical users deploying mobile video devices in their daily activities are unlikely to ever encounter side-by-side or sequential back-to-back video comparisons. Moreover, unlike a TV showroom, the visual distortions we are interested in are display-device independent and occur in isolation. The choice of a continuous scale as opposed to a discrete 5-point ITU-R Absolute Category Scale (ACR) has advantages: expanded range, finer distinctions between ratings, and demonstrated prior efficacy [5], [6].

2) Display: The user interface was developed on Eclipse[3] using the Android SDK, since the target platforms for the human study were Android-based devices. Although the platform did not allow for explicit control over the video buffer as is allowed by the XGL toolbox [31] which we have previously used [5], [6], no errors such as latencies were encountered while displaying the videos. Since the Android platform does not allow for RAW video playback, the RAW videos were embedded in a 3gp container and compressed using the MPEG-4 codec via ffmpeg. While this additional compression was undesirable, the choice of the platform made this unavoidable. However, the bit-rate for compression was > 18 Mbps with the QP set at 0 on ffmpeg, and we were unable to detect any differences between the embedded 3gp streams and the original YUV videos.

The videos were displayed on two devices – the Motorola Atrix smartphone and the Motorola Xoom tablet. The Atrix consists of a dual-core 1 Ghz ARM Cortex-A9 processor, with 1 GB RAM, ULP GeForce GPU and the Tegra 2 chipset. Videos were displayed on the Atrix 4-inch Gorilla glass display with a screen resolution of 960×540; the Atrix is capable of playing out videos at 1080p and the processor was powerful enough to avoid any buffering or playback issues when playing the high-resolution content. The Xoom uses a 1 Ghz NVIDIA Tegra 2 AP20H dual-core processor with 1 GB RAM. Videos were displayed on the 10.1-inch TFT display with a screen resolution of 1280×800. As with the Atrix, the Xoom had no problems playing out 720p videos. The devices do not allow for calibration; however, the same devices (with brightness set at max) were used throughout the course of the study.

3) Subjects, Training and Testing: The subjective study was conducted at The University of Texas at Austin (UT) and involved mostly undergraduate students, with a male majority. The study was voluntary and no monetary compensation was provided to the participants. The average subject age was between 22–28 years and the subjects were inexperienced with video quality assessment, types of video distortion and concepts underlying the perception of quality. Though no vision test was performed, a verbal confirmation of soundness of (corrected) vision was obtained from the subject. At this juncture, it may be prudent to explain our choice.

We decided to forego formal screening for visual acuity (e.g., Snellen test) and color vision (Ishihara), instead using informal confirmation of normal corrected acuity directly from each subject. This approach follows our continuing philosophy towards conducting large-scale image and video quality subjective studies: rigorous visual screening of subjects, such as we routinely do in our other vision science work, may bias results as compared to a 'typical user'.

Regarding chromatic perception, we are not (yet) conducting color quality studies nor is there any evidence that that any of the distortions that we are studying are correlated in any manner with color deficiency. One could disregard this, and assume that chromatic perception has an effect on the quality rating provided in the current setup. In this case, a very conservative high-end estimate of chromatic disability is that as many as 8% of the population has some, even very minor color deficiency. Even then, for those videos that were viewed the least (17 times) the chances are less than 1% that as many as 4 subjects might be color affected in the smallest way. However, taking into account that nearly all color blind persons are deuteranomalous or "green weak" which causes at most small differences in the perception of hues, these figures become even more remote. There is a less than 20% chance that any other form of color blindness

[3]Eclipse is an integrated development environment (IDE) for JAVA, C, C++, Perl amongst other languages, and is freely available: http://www.eclipse.org/. It is also the recommended

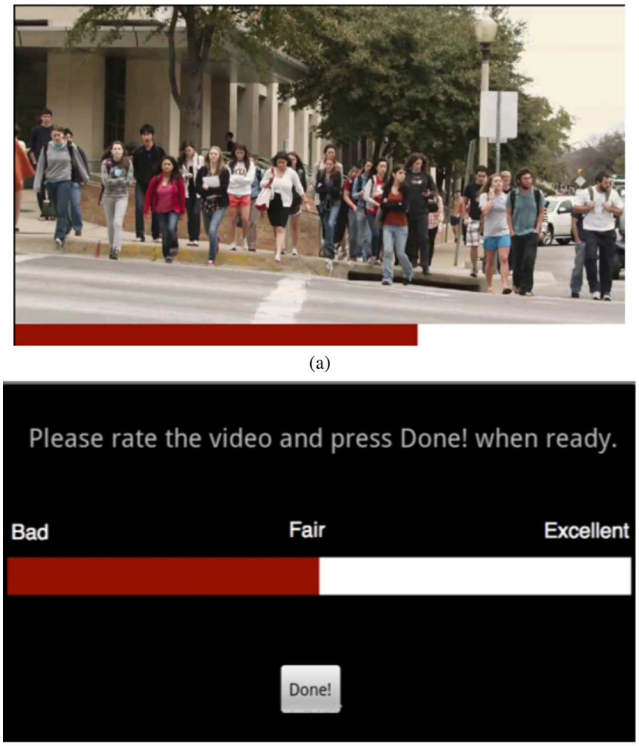
might appear in even one subject in the course of the study, and in line with our philosophy of a 'typical viewer', we would naturally welcome such persons.

Since creating a truly unrestricted "mobile" setting is near impossible, using mobile devices that are "real world" and by not forcing any specific viewing distance, nor demanding perfect acuity, but rather a reasonably representative slice, we believe that the database is far more realistic than one created in a controlled setting. Of course, studies of visual quality in different environments such as that in [22], remain valuable. While looser restrictions, e.g., exiting the laboratory entirely, might alleviate biases that might be introduced due to a rigid lab setup, and while a completely non-rigid setup might simulate real-life better, it also introduces a number of variables that cannot be controlled. Human studies are highly subjective in nature, and human viewing and rating experience is a function not only of the stimulus seen but also of the mental and physical state of the subject. Consider, for example, mobile viewing in the hot sun (discomfort, hard to see the screen, etc.) by someone feeling impatient against a lazy office executive viewing content in air conditioning. Ratings can vary drastically for the same content based on the state and environment of the subject. In other words, with such lack of control the results could quickly become meaningless. In our opinion, our setup supplies a happy median, while still obtaining statistically meaningful results. By ensuring a semblance of uniformity across subjects, the ratings provided are more or less related to the stimulus. Given our incomplete understanding of how the human rates visual stimuli (as our objective QA analysis will demonstrate), attempting to understand and model human behavior in random scenarios may be best tackled at a later date.

While our philosophy in this regard does not necessarily accord with published (and largely outdated) industry standards, we have discussed our view with other vision scientists and received general accord. Aside from the fact that most of the published standards are severely dated, and even setting aside the exceedingly important point that they bear little relevance to a study of this type of videos with temporal distortion variations and using digital mobile monitors (e.g., BT. 500-11 [29] is all about studio quality videos, viewing on CRT screens, etc., which are clearly not relevant in a mobile context), it is important that academic researchers and vision scientists, like ourselves, not feel bound by industry-mandated standards of conduct regarding any kind of studies. Notwithstanding that such recommendations have definite value for standardization within certain realms, for advancing science it is not a good thing: rather, they are limiting and could impede timely advances.

We believe that this approach allows for greater freedom and realism in designing large scale studies such as the one described here, using mobile devices likely to be used in highly diverse conditions and for which there exist no guidelines.

Each subject attended two separate sessions as part of the study such that each session lasted less than 30 minutes, and the sessions were separated by at least 24 hours, in order to minimize fatigue [29]. Informal after-study feedback indicated that the subjects did not experience any uneasiness or fatigue during the course of the sessions. Each session consisted of the subject viewing 55 videos (50 distorted + 5 reference), and a short



(a)

(b)

Fig. 7. Study Setup: (a) The video is shown at the center of the screen and an (uncalibrated) bar at the bottom is provided to rate the videos as a function of time. The rating is controlled using the touchscreen. (b) At the end of the presentation, a similar calibrated bar is shown on the screen so that the subject may rate the overall quality of the video.

training set (6 videos) preceded the actual study. The videos in the training session spanned the entire range of video quality that the user was bound to see during the course of the study; the distortions were a subset of the distortions used in the actual study. The videos were shown in random order across subjects as well as within a single session for a subject. Care was taken to ensure that two consecutive sequences did not belong to the same reference content, to minimize memory effects [29].

The videos were displayed on the center of the screen with an un-calibrated continuous bar at the bottom, which was controlled using the touchscreen. The subjects were briefed about the bar during the training session. Before the video was played, a screen indicating that the video was ready for playback was displayed. Once the subject hit 'play' the video played on the screen. The subjects were asked to rate the videos as a function of time i.e., provide instantaneous ratings of the videos, as well as to provide an overall rating at the end of each video. We sampled the scores at the rate at which the video was played out, so that a single score was available per frame, i.e., at 30 fps. Ideally, the sampling rate should be at least as fast as the amount of time it takes a human to react (approx. 220 ms) and a sampling rate of 30 fps, which is much higher than this reaction time, ensures that the data captured does not miss out on human opinion owing to poor sampling.

At the end of each video a similar continuous bar was displayed on the screen, although it was calibrated as "Bad", "Fair", and "Excellent" by markings, equally spaced across the bar. Although the bar was continuous, the calibrations served

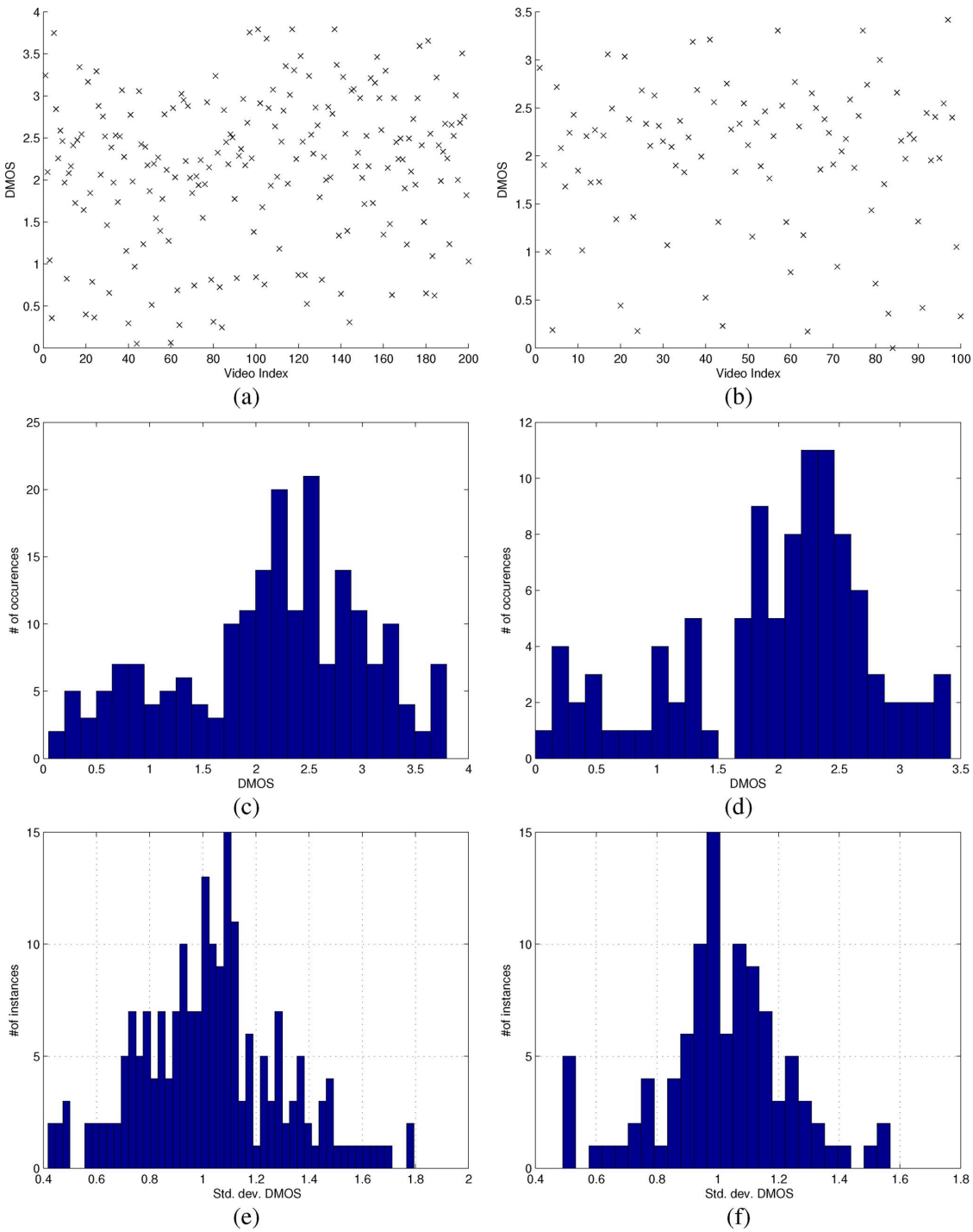


Fig. 8. DMOS scores for all video sequences: (a) Mobile Study, (b) Tablet Study; the associated histograms of scores for (c) the Mobile Study and (d) the Tablet Study; DMOS standard deviation histograms for (e) the Mobile Study and (f) the Tablet Study.

to guide the subject. Once the quality was entered, the subject was not allowed to change the score. The quality ratings were in the range 0–5. The instructions to the subject are reproduced in the Appendix.

Fig. 7 shows the various stages of the study.

D. Processing of the Scores

A total of thirty-six subjects participated in the mobile study and seventeen subjects participated in the tablet study. The mobile study was designed so that 18 subjective ratings were obtained for each of the 200 videos in the study. 100 distorted videos from this set of 200 distorted videos were used for the tablet study, and thus each of the 100 videos in the tablet study received ratings from 17 subjects. The subject rejection procedure in [29] was used to reject two subjects from the mobile study, while no subjects were rejected from the tablet study. The scores from the remaining subjects were then averaged to form a Differential Mean Opinion Scores (DMOS) for each video. The DMOS is representative of the perceived quality of the video. Specifically, let s_{ijk} denote the score assigned by subject i to

TABLE I
MOBILE STUDY: RESULTS OF t-TEST BETWEEN THE VARIOUS COMPRESSION-RATES SIMULATED IN THE STUDY. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE ROW IS STATISTICALLY WORSE (LOWER VISUAL QUALITY) THAN THE COLUMN; A VALUE OF '−' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY EQUIVALENT. EACH SUB-ENTRY IN EACH ROW/COLUMN CORRESPONDS TO THE 10 REFERENCE VIDEOS IN THE STUDY

	R_1	R_2	R_3	R_4
R_1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
R_2	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
R_3	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0
R_4	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -

TABLE II
MOBILE STUDY: RESULTS OF t-TEST BETWEEN THE FRAME-FREEZES SIMULATED IN THE STUDY. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE ROW IS STATISTICALLY WORSE (LOWER VISUAL QUALITY) THAN THE COLUMN; A VALUE OF '−' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY EQUIVALENT. EACH SUB-ENTRY IN EACH ROW/COLUMN CORRESPONDS TO THE 10 REFERENCE VIDEOS IN THE STUDY

	$F1$	$F2$	$F3$	FR_4
$F1$	- - - - - - - - - -	0 0 0 - 0 0 0 - 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$F2$	1 1 1 - 1 1 1 - 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 - 1 - 1 0 - 0
$F3$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	1 1 1 1 1 1 1 1 1 1
FR_4	1 1 1 1 1 1 1 1 1 1	1 1 1 - 0 - 0 1 - 1	0 0 0 0 0 0 0 0 0 0	- - - - - - - - - -

the distorted video j in session k, s_{ijref_k} the score assigned by subject i to the *reference* video associated with the distorted video j in session k, M_j the total number of rating received for video j and let N_{ik} be the number of test videos seen by subject i in session k. The difference scores d_{ijk} are computed as

$$d_{ijk} = s_{ijk} - s_{ijref_k}.$$

The DMOS (after subject rejection) is then

$$DMOS_j = \frac{1}{M_j} \sum_i \sum_k d_{ijk}.$$

DMOS values ideally range continuously from 0 (excellent quality) to 5 (worst quality); however small negative values as possible due to the nature of DMOS computation.

DMOS was computed only for the overall scores that the subject assigned to the videos. Fig. 8 plots the DMOS scores across distorted videos for the mobile and tablet studies, and shows the corresponding histograms for the DMOS and the associated standard deviation in order to demonstrate that the distorted videos span the entire quality range. The average standard error in the DMOS score was 0.2577 across the 200 distorted videos for the mobile study and 0.2461 across the 100 distorted videos for the tablet study.

At this juncture, it may be prudent to comment on the distribution of the subjective ratings. Note that in the current study, the scale ranges from 0–5, with the maximum DMOS rating for the mobile study component being 3.8, and that for the tablet study is 3.5 both of which are at least 75% of the available scale. By comparison, the widely used LIVE VQA database [5] uses 50% of the scale, and the VQEG Phase I database [8] uses 50% (525) and 70% (625) of the entire scale. In our view, in studies of video quality that varies, use of the entire scale would cast the study into question. A set of ratings that use the entire scale would imply that *every single subject* thought that at least one (if not more) of the distorted videos were the *worst* videos they have seen (since these are DMOS scores the scale is reversed 5 is a horrible video), which is almost always impossible.

We believe that studies of this nature are designed not to enforce the designer's view of the scale and the distortions on the subject but rather extract the opinion of the subject himself. It is hence that we do not compel the subject to utilize the entire scale in the ratings, instead using the training session as a 'normalizer' for the range of quality the subject is likely to see in the study. A DMOS of 3.8 for a distorted video implies that the video can get worse and still be within the limits of the subject.

For all further analysis, we assume that the DMOS scores sample a Gaussian distribution centered around the DMOS having a standard deviation computed from the differential opinion scores across subjects.

E. Evaluation of Subjective Opinion

We analyzed the distorted videos with respect to the subjective DMOS for each of the videos and the associated standard deviations of DMOS across the subjects on the mobile and the tablet studies. For each of the subsections below, we conduct a t-test between the Gaussian distributions centered at the DMOS values (and having a associated, known standard deviation) of the conditions we are interested in comparing at the 95% confidence level. Since the conditions being compared are functions of content, we compared each of the 10 reference contents separately for each pair of conditions. In the tables that follow, a value of '1' indicates that the row-condition is statistically superior to the column-condition, while a '0' indicates that the row is worse than a column; a value of '−' indicates that the row and column are statistically indistinguishable from each other. For example, in Table I, for all the 10 contents, videos compressed at rate R_2 have statistically better visual quality than those compressed at rate R_1, while they are statistically worse than those compressed at a rate R_3. Further, for the tablet study, we compared the results obtained from the tablet study to those obtained from the mobile study across all distortions as well as for each distortion subsection.

1) Mobile Study: The results from the statistical analysis are tabulated in Tables I–VII. Due to the dense nature of the content, we summarize the results in the following paragraphs. Note that

TABLE III

MOBILE STUDY: RESULTS OF t-TEST BETWEEN THE VARIOUS RATE-ADAPTED DISTORTED VIDEOS SIMULATED IN THE STUDY. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE ROW IS STATISTICALLY WORSE (LOWER VISUAL QUALITY) THAN THE COLUMN; A VALUE OF '$-$' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY EQUIVALENT. EACH SUB-ENTRY IN EACH ROW/COLUMN CORRESPONDS TO THE 10 REFERENCE VIDEOS IN THE STUDY

	$R_1 - R_4 - R_1$	$R_2 - R_4 - R_2$	$R_3 - R_4 - R_3$
$R_1 - R_4 - R_1$	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$R_2 - R_4 - R_2$	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0
$R_3 - R_4 - R_3$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -

TABLE IV

MOBILE STUDY: RESULTS OF t-TEST BETWEEN THE VARIOUS COMPRESSION-RATES AND THE RATE-ADAPTED VIDEOS SIMULATED IN THE STUDY. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE ROW IS STATISTICALLY WORSE (LOWER VISUAL QUALITY) THAN THE COLUMN; A VALUE OF '$-$' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY EQUIVALENT. EACH SUB-ENTRY IN EACH ROW/COLUMN CORRESPONDS TO THE 10 REFERENCE VIDEOS IN THE STUDY

	R_1	R_2	R_3	R_4
$R_1 - R_4 - R_1$	1 1 1 1 1 1 1 1 1 1	0 0 0 - 0 1 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$R_2 - R_4 - R_2$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
$R_3 - R_4 - R_3$	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 - 0 1 - 0 1 0	0 0 0 0 0 0 0 0 0 0

TABLE V

MOBILE STUDY: RESULTS OF t-TEST BETWEEN MULTIPLE RATE SWITCHES AND A SINGLE RATE SWITCH. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE ROW IS STATISTICALLY WORSE (LOWER VISUAL QUALITY) THAN THE COLUMN; A VALUE OF '$-$' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY EQUIVALENT. EACH SUB-ENTRY IN EACH ROW/COLUMN CORRESPONDS TO THE 10 REFERENCE VIDEOS IN THE STUDY

	$R_1 - R_4 - R_1$	$R_1 - R_4 - R_1 - R_4 - R_1$
$R_1 - R_4 - R_1$	- - - - - - - - - -	0 - - - 0 0 0 0 1 -
$R_1 - R_4 - R_1 - R_4 - R_1$	1 - - - 1 1 1 1 0 -	- - - - - - - - - -

TABLE VI

MOBILE STUDY: RESULTS OF t-TEST BETWEEN THE VARIOUS TEMPORAL-DYNAMICS DISTORTED VIDEOS SIMULATED IN THE STUDY. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE ROW IS STATISTICALLY WORSE (LOWER VISUAL QUALITY) THAN THE COLUMN; A VALUE OF '$-$' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY EQUIVALENT. EACH SUB-ENTRY IN EACH ROW/COLUMN CORRESPONDS TO THE 10 REFERENCE VIDEOS IN THE STUDY

	$R_1 - R_4 - R_1 - R_4 - R_1$	$R_1 - R_2 - R_4$	$R_4 - R_2 - R_1$	$R_1 - R_3 - R_4$	$R_4 - R_3 - R_1$
$R_1 - R_4 - R_1 - R_4 - R_1$	- - - - - - - - - -	- 0 0 0 1 1 0 0 0 0	1 1 - 1 1 1 1 1 1 1	0 0 0 0 - 0 0 0 0 0	1 1 0 0 1 1 1 1 1 1
$R_1 - R_2 - R_4$	- 1 1 1 0 0 1 1 1 1	- - - - - - - - - -	1 1 1 1 1 1 1 1 1 1	0 0 - 0 0 0 - 0 0 0	1 1 1 - 1 1 1 1 1 1
$R_4 - R_3 - R_1$	0 0 - 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	1 - 0 0 0 0 - 0 1 0
$R_1 - R_3 - R_4$	1 1 1 1 - 1 1 1 1 1	1 1 - 1 1 1 - 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	1 1 1 1 1 1 1 1 1 1
$R_4 - R_3 - R_1$	0 0 1 1 0 0 0 0 0 0	0 0 0 - 0 0 0 0 0 0	0 - 1 1 1 1 - 1 0 1	0 0 0 0 0 0 0 0 0 0	- - - - - - - - - -

TABLE VII

MOBILE STUDY: RESULTS OF t-TEST BETWEEN THE VARIOUS WIRELESS PACKET-LOSSES SIMULATED IN THE STUDY. A VALUE OF '1' INDICATES THAT THE ROW IS STATISTICALLY SUPERIOR (BETTER VISUAL QUALITY) THAN THE COLUMN, WHILE A VALUE OF '0' INDICATES THAT THE ROW IS STATISTICALLY WORSE (LOWER VISUAL QUALITY) THAN THE COLUMN; A VALUE OF '$-$' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY EQUIVALENT. EACH SUB-ENTRY IN EACH ROW/COLUMN CORRESPONDS TO THE 10 REFERENCE VIDEOS IN THE STUDY

	WR_1	WR_2	WR_3	WR_4
WR_1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
WR_2	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0
WR_3	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -	0 0 0 0 0 0 0 0 0 0
WR_4	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1	- - - - - - - - - -

the text only provides a high level description of the results in the table, the reader is advised to thoroughly study the table in order to better understand the results.

Compression (Table I): This table confirms that the distorted videos were perceptually separable. Notice that each compression rate is statistically better (perceptually) than the next lower rate over all content used in the study.

Frame-Freeze (Table II): For frame-freezes, the following trend is seen across most of the contents: longer freezes are preferred to shorter freezes, which lead to choppy playback, implying playback immediately after the buffer receives data is less desirable than waiting before playback. We also observe that pauses of 4 seconds are seemingly tolerable. For the frame-freezes with lost segments (real-time freezes), one would conjecture that lost segments are important and became evident when the segments are about 4 seconds long or larger. Further, it seems that shorter freezes (choppy playback) are regarded as worse than lost frames.

Rate Adaptation (Tables III and IV): While conventional wisdom might dictate that people do not prefer fluctuations in video quality, our study seems to indicate that it is preferable to switch to a higher rate if possible, especially if the duration of the higher rate is at least half the duration of the lower rates. Further, if one is capable of maintaining a continuous rate at a

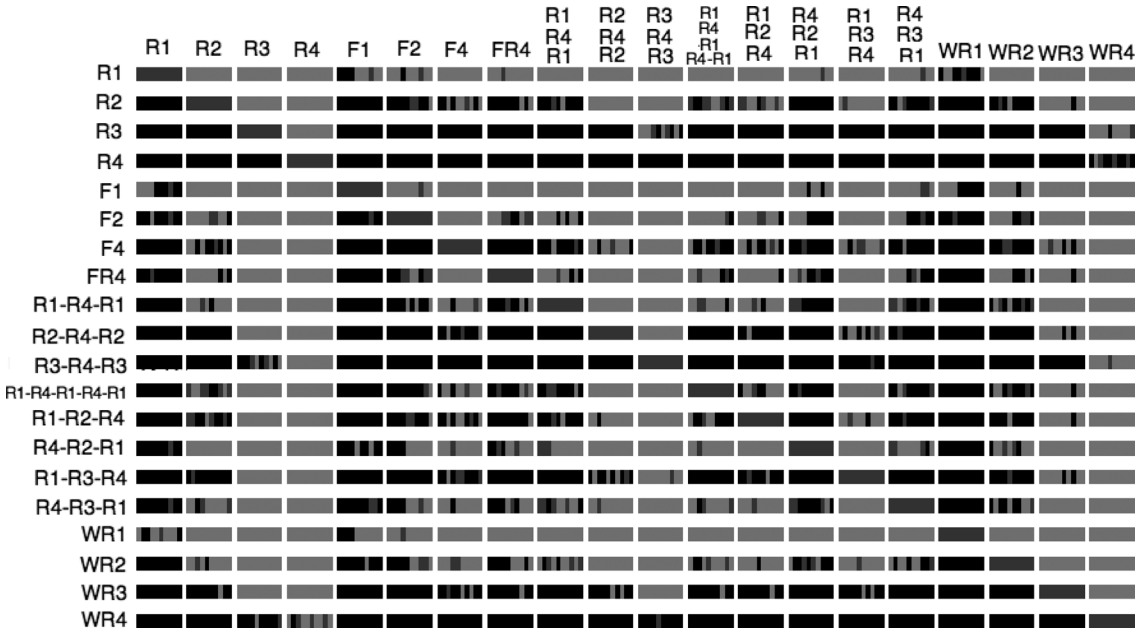


Fig. 9. Mobile Study: Statistical significance analysis of the distortion categories in this studies. Refer text for label explanation. Each (non-white) block represents the results of hypothesis testing for the 10 different video contents in this study. Dark (black) gray levels indicate that the row distortion is statistically superior (better visual quality) than the column; the lightest gray-levels indicates that the row is statistically worse than the column, while mid-grey levels represents that the row and column are statistically equivalent. For example, R_4 is always statistically superior to R_2 and R_3, while for R_3 versus WR_4, one of the R_3 compressed videos is superior to WR_4 and one of them is equivalent, while the rest are statistically worse.

TABLE VIII
CORRELATION AND RESULTS OF THE WILCOXON SUM-RANK TEST FOR EQUAL MEDIANS (IN PARENTHESIS – HYPOTHESIS/p-VALUE) BETWEEN DMOS SCORES FROM THE MOBILE AND TABLET STUDIES. A VALUE OF '1' IN THE BRACKETS INDICATES THAT THE DMOS SCORES FROM THE TWO STUDIES HAVE DIFFERENT MEDIANS, WHILE A VALUE OF '0' INDICATES THAT THE MEDIANS ARE STATISTICALLY INDISTINGUISHABLE AT THE 95% CONFIDENCE LEVEL

Compression	Frame-freezes	Rate Adaptation	Temporal Dynamics	Wireless	All
0.9493 (0/0.93)	0.7981 (1/0.01)	0.8701 (0/0.56)	0.6298 (0/0.92)	0.9359 (0/0.56)	0.9047 (0/0.89)

value higher than the base rate of the switch (eg., $R_2 - R_4 - R_2$ versus R_3), the continuous higher rate is preferred.

Temporal Dynamics (Tables V and VI): Our analysis indicates that multiple rate switches are preferred over fewer switches, if the subject is able to view the high quality video for longer duration. There is a plausible explanation for this behavior. Our hypothesis is that when shown high quality video for a long time, the bar of expectation is raised, and when the viewer is exposed to low quality segments of the video, s/he assigns a high penalty than on videos containing high quality segments of shorter duration. The subject might view the short high quality segments as attempts to improve the viewing experience, thereby boosting overall perception of quality. An even more likely explanation is that long low-quality video segments preceded by much higher quality segments evoke a strong negative response. Of course, our results are conditioned on the degree of quality separation between the low and high quality segments and may not generalize to switches between quality levels exhibiting a lesser degree of quality separation.

Our results also indicate that switching to an intermediate rate before switching to a higher rate is preferred over multiple large-magnitude rate switches, and that the end quality of the video makes a definite impact on perceived quality (see for example, $R_4 - R_3 - R_1$ versus $R_1 - R_3 - R_4$ in Table VI).

Wireless (Table VII): The wireless results mirror the compression results, demonstrating the perceptual separability of the videos in the study.

Finally, in order to visualize how different distortions affect visual quality, Fig. 9 plots a visual map of the statistical significance values for all possible pairs of distortions. The map is comprehensive in that it encompasses all of the videos from the study; the caption explains the figure's interpretation in detail.

2) Tablet Study: We compare the results from the tablet study to those from the mobile study for each distortion category and across all the distortions considered here, and tabulate the (linear) correlation coefficient between these two studies in Table VIII. In the table, we also report the results from a Wilcoxon sum-rank test for equal medians – a value of '1' in the brackets indicates that the DMOS scores from the two studies have different medians, while a value of '0' indicates that the medians are statistically indistinguishable at the 95% confidence level. Also reported are the p-values. The results indicate that while the data is correlated and that the medians are statistically indistinguishable, the degree of correlation is a function of the distortion category. Specifically, for the frame-freeze case, the perception of visual quality varies significantly as a function of the display resolution.

We performed an analysis similar to that for the mobile database and since our results are similar to those for the mobile

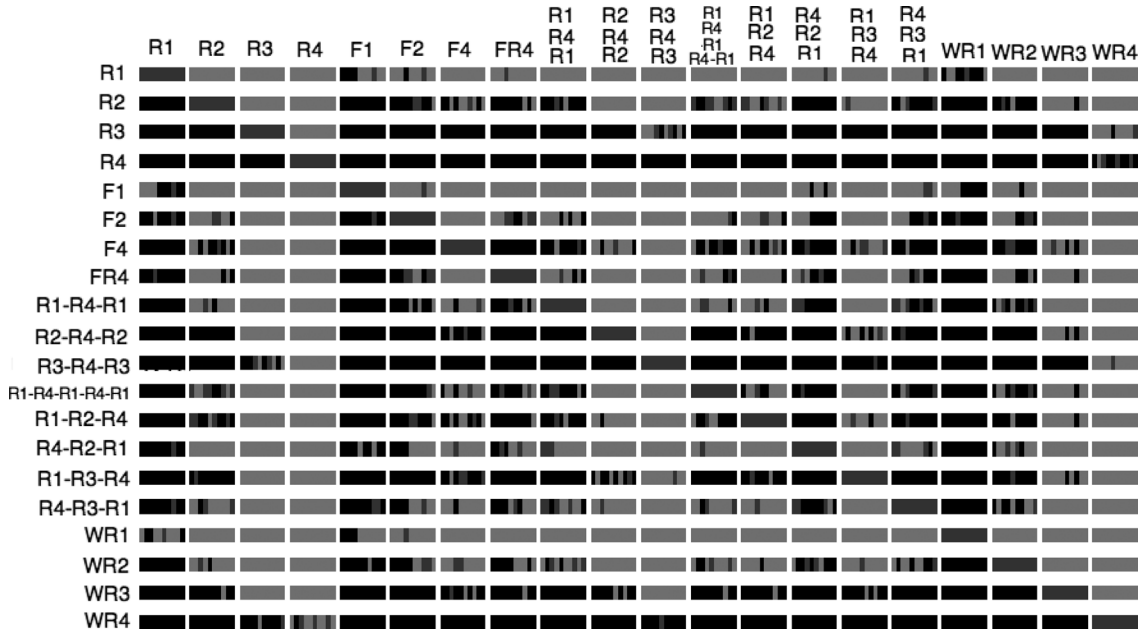


Fig. 10. Tablet Study: Statistical significance analysis of the distortion categories in this studies. Refer text for label explanation. Each (non-white) block represents the results of the hypothesis testing for the 10 different contents in this study. Dark (black) gray levels indicate that the row distortion is statistically superior (better visual quality) than the column; the lightest gray-levels indicates that the row is statistically worse than the column, while mid-grey levels represents that the row and column are statistically equivalent. For example, $R_2 - R_4 - R_2$ rate change is statistically better than R_1 and R_2 across all videos, however, for the $R_2 - R_4 - R_2$ versus F4, while most videos are statistically better than F4, 2 of the videos are worse, while one of them is equivalent to F4.

case, we refrain from reporting those tables here[4]; instead, as in the mobile case, we report a visual map of the distortions and their statistical significance in Fig. 10.

F. Evaluation of Temporal Quality Scores

Recall that we collected subjective opinion scores on time-varying video quality by asking the subject to rate the quality of the video as a function of time. These temporal opinion scores were obtained at a sampling rate equal to that of the frame-rate of the video (i.e., 1/30 fps) for all distortions, except for the frame-freezes where the scores were collected at a rate such that the temporal scores spanned the same support as those for other distortions. Thus a total of 450 temporal scores were collected for each 15 second video. The temporal scores so obtained were then processed as in [32], in order to produce a temporal MOS (z-score) for each video. Specifically, let $f_{ijk}(t)$ be the score assigned to the video j by subject i in session k, where each video is of length T_j. We computed:

$$m_{ik} = \frac{1}{\sum_{j=1}^{N_{ik}} T_j} \sum_{j=1}^{N_{ik}} \sum_{t=1}^{T_j} f_{ijk}(t) \tag{1}$$

$$s_{ik}^2 = \frac{1}{\sum_{j=1}^{N_{ik}} T_j - 1} \sum_{j=1}^{N_{ik}} \sum_{t=1}^{T_j} (f_{ijk}(t) - m_{ik})^2 \tag{2}$$

$$z_{ij}(t) = \frac{f_{ijk}(t) - m_{ik}}{s_{ik}} \tag{3}$$

[4]The interested reader is directed to the supplementary material for the associated tables from the tablet study.

and finally,

$$MOS_j^f(t) = \frac{1}{M} \sum_{i=1}^{M} z_{ij}(t) \tag{4}$$

where $MOS_j^f(t)$ is the mean opinion score recorded over time for video j and M is the number of subjects in the study (after subject rejection, as described earlier).

We analyzed how these temporal scores contribute to the overall perception of visual quality, i.e., how temporal scores might be pooled to reproduce the DMOS that the subject assigned the video at the end of the presentation. The analysis below is simplistic, but much work remains on developing good behavioral models of temporal quality judgements of dynamically changing video distortions. Our first attempt at understanding this new problem is detailed in [32].

We evaluate three different methods of temporal pooling: (1) Mean, (2) Percentile pooling [11], [33], [34], and (3) Memory-effect based pooling.

The temporal mean serves as the baseline and is simply the time-average of $MOS_j^f(t)$. Percentile pooling was proposed in [11], [33], [34] as a method of spatially collapsing image quality scores while emphasizing severe errors. There is some evidence that this type of pooling may relate to the visual quality of videos as well [35]. Here, we sorted the temporal scores in ascending order and averaged the lowest 5% of the sorted scores to produce a single quality score for each video.

One may conjecture that human quality decisions are heavily influenced by the visual quality perceived in the last segment prior to rating. To investigate this claim, we averaged quality scores from a time-window spanning the last n frames of the

TABLE IX
MOBILE STUDY: CORRELATION COEFFICIENT BETWEEN THE TEMPORALLY POOLED SUBJECTIVE SCORES AND THE DMOS FOR VARIOUS POOLING STRATEGIES

	Compression	Frame-freezes	Rate Adaptation	Temporal Dynamics	Wireless	All
Mean	-0.9724	-0.2488	-0.9001	0.3374	-0.9729	-0.7008
Percentile Pooling	-0.8970	0.0501	-0.7991	0.0767	-0.9247	-0.7092
Memory Effect ($t = 1s$)	-0.9788	-0.6251	-0.8054	-0.7399	-0.9805	-0.8337
Memory Effect ($t = 2s$)	-0.9777	-0.6309	-0.7861	-0.7082	-0.9794	-0.8360
Memory Effect ($t = 3s$)	-0.9778	-0.6389	-0.7799	-0.6193	-0.9797	-0.8340

TABLE X
TABLET STUDY: CORRELATION COEFFICIENT BETWEEN THE TEMPORALLY POOLED SUBJECTIVE SCORES AND THE DMOS FOR VARIOUS POOLING STRATEGIES

	Compression	Frame-freezes	Rate Adaptation	Temporal Dynamics	Wireless	All
Mean	-0.9720	-0.1557	-0.9248	0.6757	-0.9847	-0.7031
Percentile Pooling	-0.8543	0.3040	-0.8108	0.4945	-0.9263	-0.5781
Memory Effect ($t = 1s$)	-0.9826	-0.4825	-0.8718	-0.3492	-0.9882	-0.8134
Memory Effect ($t = 2s$)	-0.9850	-0.5565	-0.8343	-0.1702	-0.9899	-0.8092
Memory Effect ($t = 3s$)	-0.9850	-0.5864	-0.8142	0.0794	-0.9900	-0.8116

video, where n is varied between 1–3 seconds in steps of 1 second.

In Tables IX and X, we tabulate the correlation coefficient between the DMOS (as obtained previously) and each of the four pooling strategies, for each distortion as well as across all distortions, for the mobile study and for the tablet study respectively.

Note that the correlations should ideally be negative, since we are comparing the MOS with DMOS; the small positive correlations in the tables are meaningless, and imply that the pooling strategy does not correlate well for those distortion categories.

Tables IX and X indicate that while the temporal and percentile pooling strategies are poor approaches to collapsing temporal scores (especially for the frame-freezes and the temporal dynamics case), the memory-effect pooling seems to function better, lending credo to the observation that humans are influenced by the last few seconds of viewing when assessing overall quality. We note that this effect was not observed in the study of [32], but this may have been due to the shorter durations of those videos. We also note that while the Memory-effect does help, the overall improvement achieved is not great, which may be due to the short durations of the clips used in this study. While the videos in this study were at least 50% longer than those in [5], [32], they are still short relative to the kind of memory effects that can occur.

The tables also indicate that, while most pooling strategies work for videos exhibiting uniform visual quality over time video (for example, compression), almost all pooling strategies performed poorly when the quality changes dynamically – either when the compression rate is varied (eg., temporal dynamics) or if the video freezes. One could conjecture that a good behavioral model of temporal quality pooling should improve correlation with DMOS, and that such temporal pooling models could profitably be incorporated into existing VQA algorithms to provide better predictions of overall visual quality. Finally, we note that temporal pooling had a greater impact in the tablet study than the mobile study. It is possible that the resolution of the display makes dynamically varying distortions even more perceptible on a device with a larger form factor (notice that for compression and wireless distortions the correlations are similar to those for the mobile study). The results seem to indicate that temporal pooling strategy should account for display resolution as well.

TABLE XI
LIST OF FR 2D IQA ALGORITHMS EVALUATED IN THIS STUDY

No.	Algorithm
1.	Peak Signal-to-Noise ratio (PSNR)
2.	Structural Similarity Index (SS-SSIM) [37]
3.	Multi-scale Structural Similarity Index (MS-SSIM) [38]
4.	Visual Signal-to-Noise ratio (VSNR) [39]
5.	Visual Information Fidelity (VIF) [40]
6.	Universal Quality Index (UQI) [41]
7.	Noise Quality Measure (NQM) [42]
8.	Signal-to-Noise ratio (SNR)
9.	Weighted Signal-to-Noise ratio (WSNR) [43]

III. EVALUATION OF ALGORITHM PERFORMANCE

We evaluated a wide variety of full-reference (FR) IQA algorithms against the human subjective scores collected. Table XI lists these algorithms, all of which are available as part of the Metrix Mux toolbox [36]. The reader is referred to the citations for details on these approaches.

The FR IQA algorithms were applied on a frame-by-frame basis and the average score across time used as a final measure of quality. Since it is unclear how FR QA algorithms may be used for frame-freezes (an interesting and important problem for the future), we did not include this case in our evaluation below.

We also evaluated two FR VQA algorithms – Visual Quality Metric (VQM) [11] and the MOtion-based Video Integrity Evaluation (MOVIE) index [12]. VQM was obtained from [44] while MOVIE is freely available at [45]. The version of VQM that we used (CVQM v13) requires input videos in YUV422p format encased in an avi container. The YUV420p videos were converted to YUV422p using ffmpeg, then placed in an avi container (no compression was used). These algorithms were also not evaluated for their performance on frame-freezes.

Algorithm Correlations Against Subjective Opinion

Tables XII and XIII, tabulate the Spearman's rank ordered correlation cofficient (SROCC) between the algorithm scores

TABLE XII
MOBILE STUDY: SPEARMAN'S RANK ORDERED CORRELATION COEFFICIENT (SROCC) BETWEEN THE ALGORITHM SCORES AND THE DMOS FOR VARIOUS IQA/VQA ALGORITHMS

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.8185	0.5981	0.3717	0.7925	0.6780
SS-SSIM	0.7092	0.6303	0.3429	0.7246	0.6498
MS-SSIM	0.8044	**0.7378**	**0.3974**	0.8128	0.7425
VSNR	**0.8739**	0.6735	0.3170	0.8559	**0.7517**
VIF	0.8613	0.6388	0.1242	0.8739	0.7439
UQI	0.5621	0.4299	0.0296	0.5756	0.4894
NQM	0.8499	0.6775	0.2383	**0.8985**	0.7493
WSNR	0.7817	0.5598	0.0942	0.7510	0.6267
SNR	0.7073	0.5565	0.2029	0.6959	0.5836
VQM	0.7717	0.6475	0.3860	0.7758	0.6945
MOVIE	0.7738	0.7198	0.1578	0.6508	0.6420

TABLE XIII
TABLET STUDY: SPEARMAN'S RANK ORDERED CORRELATION COEFFICIENT (SROCC) BETWEEN THE ALGORITHM SCORES AND THE DMOS FOR VARIOUS IQA/VQA ALGORITHMS

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.7910	0.4464	0.0981	0.7564	0.5886
SS-SSIM	0.4947	0.3679	0.0773	0.5609	0.4300
MS-SSIM	0.6602	0.4821	0.1400	0.6451	0.5678
VSNR	0.7714	0.4429	0.0469	0.7053	0.5929
VIF	**0.8917**	0.6714	0.0700	**0.8617**	**0.7261**
UQI	0.5053	0.3500	0.0481	0.4226	0.3642
NQM	0.8406	0.4643	0.0792	0.8075	0.6614
WSNR	0.8361	0.6214	0.1462	0.7353	0.6255
SNR	0.7098	0.6321	**0.2354**	0.6602	0.5474
VQM	0.6316	0.4357	0.0515	0.6692	0.5552
MOVIE	0.7744	**0.7714**	0.0658	0.8451	0.6792

TABLE XIV
MOBILE STUDY: LINEAR (PEARSON'S) CORRELATION COEFFICIENT (LCC) BETWEEN THE ALGORITHM SCORES AND THE DMOS FOR VARIOUS IQA/VQA ALGORITHMS

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.7841	0.5364	0.4166	0.7617	0.6909
SS-SSIM	0.7475	0.6120	0.3924	0.7307	0.6637
MS-SSIM	0.7664	0.7089	0.4068	0.7706	0.7077
VSNR	0.8489	0.6581	**0.4269**	0.8493	0.7592
VIF	**0.8826**	0.6643	0.1046	**0.8979**	**0.7870**
UQI	0.5794	0.2929	0.2546	0.7412	0.6619
NQM	0.8318	0.6772	0.3646	0.8738	0.7622
WSNR	0.7558	0.5365	0.0451	0.7276	0.6320
SNR	0.6501	0.3988	0.0839	0.6052	0.5189
VQM	0.7816	0.5910	0.4066	0.7909	0.7023
MOVIE	0.8103	**0.6811**	0.2436	0.7266	0.7157

and DMOS for the mobile and tablet studies, Tables XIV and XV tabulate the Pearson's (linear) correlation coefficient (LCC) and Tables XVI and XVII, tabulate the root mean-squared-error (RMSE) between the algorithm scores (after non-linear regression, as prescribed in [46][5]) and DMOS.

There are two immediate takeaways from the combined tables. First, that multiscale matters as the display size is reduced. Indeed, the two true wavelet decomposition based algorithms – VSNR and VIF – yielded the best overall performance, exceeding that of true video QA algorithms – the single-scale VQM and the MOVIE index, which is partially-multiscale but omits high frequencies. Multiscale SSIM also does quite well, although it overweights mid-band frequencies. A lesson here is that true multiscale is advisable to achieve scalability against variations in display size, resolution and viewing distance, suggesting future refinements of VQA algorithms.

Secondly as Table XII shows, almost all algorithms fail to reliably predict overall subjective judgements of dynamic distortions – on the set of "temporal-dynamics" distorted videos and to some extent, the set of "rate-adaptation" videos. Some algorithms such as VQM, NQM and VIF perform reasonably well on the wireless distorted videos. For the rate-adaptation case, MS-SSIM and MOVIE were the top performers; however, there clearly remains significant room for improvement. Overall, VSNR, VIF, MS-SSIM and NQM are seemingly well correlated with human perception, while the single-scale UQI is the weakest of the lot probably since it captures the narrowest range of frequencies. The widely criticized PSNR holds its own against compression and wireless distortions, since, while it is not multiscale, it captures high frequency distortions.

[5]Except for MOVIE, where the fitting failed; instead the logistic specified in [8] was used.

TABLE XV
TABLET STUDY: LINEAR (PEARSON'S) CORRELATION COEFFICIENT (LCC) BETWEEN THE ALGORITHM SCORES AND THE DMOS FOR VARIOUS IQA/VQA ALGORITHMS

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.7712	0.4368	0.2520	0.7320	0.6348
SS-SSIM	0.5857	0.4222	0.0814	0.5900	0.4893
MS-SSIM	0.7018	0.5644	0.2134	0.7060	0.6213
VSNR	0.7751	0.5083	0.2202	0.7310	0.6444
VIF	**0.8511**	0.5942	0.0484	0.8541	0.7635
UQI	0.4160	0.2454	0.3043	0.5708	0.3256
NQM	0.8115	0.4124	0.1199	0.8298	0.7178
WSNR	0.8150	0.6704	0.2154	0.7252	0.6665
SNR	0.7158	0.6006	**0.3501**	0.6137	0.5544
VQM	0.6430	0.4897	0.2738	0.7349	0.6150
MOVIE	0.8275	**0.8023**	0.0711	**0.8767**	**0.7828**

TABLE XVI
MOBILE STUDY: ROOT MEAN-SQUARED-ERROR (RMSE) BETWEEN THE ALGORITHM SCORES AND THE DMOS FOR VARIOUS IQA/VQA ALGORITHMS

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.7069	0.5733	0.4179	0.7279	0.6670
SS-SSIM	0.7566	0.6023	0.4228	0.7670	0.6901
MS-SSIM	0.7316	0.4792	0.4199	0.7160	0.6518
VSNR	0.6021	0.5115	0.4157	0.5932	0.6005
VIF	0.5354	0.5078	0.4572	0.4945	0.5692
UQI	0.9283	0.6496	0.4445	0.7542	0.6916
NQM	0.6374	0.4999	0.4280	0.5463	0.5972
WSNR	0.7458	0.5733	0.4592	0.7707	0.7150
SNR	0.8654	0.6230	0.4580	0.8944	0.7887
VQM	0.7312	0.4840	0.4141	0.7279	0.6663
MOVIE	0.6674	0.4974	0.4458	0.7719	0.6444

TABLE XVII
TABLET STUDY: ROOT MEAN-SQUARED-ERROR (RMSE) BETWEEN THE ALGORITHM SCORES AND THE DMOS FOR VARIOUS IQA/VQA ALGORITHMS

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.7057	0.5810	0.2510	0.7205	0.6630
SS-SSIM	0.8985	0.5855	0.2585	0.8538	0.7483
MS-SSIM	0.7896	0.5332	0.2533	0.7489	0.6724
VSNR	0.7004	0.5562	0.2530	0.7216	0.6562
VIF	0.5820	0.5195	0.2590	0.5500	0.5541
UQI	1.0080	0.6261	0.2470	0.8683	0.8113
NQM	0.6477	0.5884	0.2575	0.5902	0.5974
WSNR	0.6424	0.4792	0.2532	0.7281	0.6397
SNR	0.7741	0.5164	0.2429	0.8349	0.7141
VQM	0.8047	0.5922	0.2593	0.7594	0.6980
MOVIE	0.6224	0.3855	0.2593	0.5087	0.5342

The results of algorithms against subjective judgments of videos viewed on the tablet show some interesting contrasts (Table XIII). Whilst VSNR was the top performer for compression in the mobile case, it does not do as well for the tablet case, where multiscale is less of a factor (at finer scales), with MOVIE and NQM eclipsing it and VIF the clear top performer. Since VSNR is a human visual system (HVS)-based measure which takes the number of pixels per visual degree into account, one could conjecture that a recalibration of VSNR based on the viewing distance and form factor of the tablet might boost performance. While all the algorithms still have trouble predicting judgments of dynamic distortions, MOVIE successfully predicts judgements of rate-adaptation. On wireless distortions, VIF again does well, as does MOVIE, while VSNR again sees a drop in performance. The performance increase of MOVIE in the tablet wireless case over the mobile case is instructive. Since MOVIE is only partially multiscale and has only been tested against human judgments of videos viewed on larger screens than mobile phones, it is not surprising that its performance improves on videos displayed on screens with a larger form factor. As in the case of VSNR, a recalibration of MOVIE as a function of the form factor, or by making it fully multiscale, would likely improve its performance on smaller screen sizes. PSNR is again close to the end of the pack, with the single-scale UQI being the worst performer.

A. Hypothesis Testing and Statistical Analysis

1) Inter-Algorithm Comparisons: We performed a statistical analysis of the algorithm scores in order to gauge if the correlations tabulated above were significantly different from each other. In order to evaluate this, we use the method of [5], [46], where the F-statistic is used to evaluate the difference between the variances of the residuals produced after a non-linear mapping between the two algorithms being compared. We perform a similar statistical analysis and report the results in Tables XVIII and XIX for the mobile and the tablet studies respectively. A

TABLE XVIII

MOBILE STUDY: STATISTICAL ANALYSIS OF ALGORITHM PERFORMANCE. A VALUE OF '1' IN THE TABLES INDICATES THAT THE ROW (ALGORITHM) IS STATISTICALLY BETTER THAN THE COLUMN (ALGORITHM), WHILE A VALUE OF '0' INDICATES THAT THE ROW IS WORSE THAN THE COLUMN; A VALUE OF '–' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY IDENTICAL. WITHIN EACH ENTRY OF THE MATRIX, THE FIRST FOUR SYMBOLS CORRESPOND TO THE FOUR DISTORTIONS (ORDERED AS IN THE TEXT), AND THE LAST SYMBOL REPRESENTS SIGNIFICANCE ACROSS THE ENTIRE DATABASE

	PSNR	SS-SSIM	MS-SSIM	VSNR	VIF	UQI	NQM	WSNR	SNR	VQM	MOVIE
PSNR	- - - - -	1 - - 1 1	- 0 - - -	- - - - -	- 1 1 - -	1 1 1 1 1	- - - 0 0	1 1 1 1 1	1 1 1 1 1	1 0 - - -	1 - 1 1 -
SS-SSIM	0 - - 0 0	- - - - -	0 0 - 0 0	0 - - 0 0	0 1 1 0 0	- 1 1 - 1	0 - - 0 0	- 1 1 - -	- - - - 1	- 0 - 0 0	- - 1 - -
MS-SSIM	- 1 - - -	1 1 - 1 1	- - - - -	- 1 - - -	- 1 - - -	1 1 - 1 1	0 - - 0 0	- 1 - - 1	1 1 - 1 1	- - - - -	- - - - 1
VSNR	- - - - -	1 - - 1 1	- 0 - - -	- - - - -	- 1 1 - -	1 1 1 1 1	- - - 0 -	1 1 1 1 1	1 1 1 1 1	1 0 - - -	1 - 1 1 1
VIF	- 0 0 - -	1 0 0 1 1	- 0 - - -	- 0 0 - -	- - - - -	1 - - 1 1	- 0 0 0 0	1 - - 1 1	1 - - 1 1	1 0 0 1 -	1 0 - 1 -
UQI	0 0 0 0 0	- 0 0 - 0	0 0 - 0 0	0 0 0 0 0	0 - - 0 0	- - - - -	0 0 - 0 0	0 - - - 0	- - - - -	0 0 - 0 0	0 0 - - 0
NQM	- - - 1 1	1 - - 1 1	1 - - 1 1	- - - 1 -	- 1 1 1 1	1 1 - 1 1	- - - - -	1 1 - 1 1	1 1 - 1 1	1 0 - 1 1	1 - 1 1 1
WSNR	0 0 0 0 0	- 0 0 - -	- 0 - - 0	0 0 0 0 0	0 - - 0 0	1 - - - 1	0 0 - 0 0	- - - - -	- - - - -	- 0 0 - 0	- 0 - - 0
SNR	0 0 0 0 0	- - - - 0	0 0 - 0 0	0 0 0 0 0	0 - - 0 0	- - - - -	0 0 - 0 0	- - - - -	- - - - -	- 0 - 0 0	- 0 - - 0
VQM	0 1 - - -	- 1 - 1 1	- - - - -	0 1 - - -	0 1 1 0 -	1 1 - 1 1	0 1 - 0 0	- 1 1 - 1	- 1 - 1 1	- - - - -	- - 1 - -
MOVIE	0 - 0 0 -	- - 0 - -	- - - - 0	0 - 0 0 0	0 1 - 0 -	1 1 - - 1	0 - 0 0 0	- 1 - - 1	- 1 - - 1	- - 0 - -	- - - - -

TABLE XIX

TABLET STUDY: STATISTICAL ANALYSIS OF ALGORITHM PERFORMANCE. A VALUE OF '1' IN THE TABLES INDICATES THAT THE ROW (ALGORITHM) IS STATISTICALLY BETTER THAN THE COLUMN (ALGORITHM), WHILE A VALUE OF '0' INDICATES THAT THE ROW IS WORSE THAN THE COLUMN; A VALUE OF '–' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY IDENTICAL. WITHIN EACH ENTRY OF THE MATRIX, THE FIRST FOUR SYMBOLS CORRESPOND TO THE FOUR DISTORTIONS (ORDERED AS IN THE TEXT), AND THE LAST SYMBOL REPRESENTS SIGNIFICANCE ACROSS THE ENTIRE DATABASE

	PSNR	SS-SSIM	MS-SSIM	VSNR	VIF	UQI	NQM	WSNR	SNR	VQM	MOVIE
PSNR	- - - - -	1 - - - 1	1 - - - -	- - - - -	- - - - 0	- - - - 1	- - - - -	- - - - -	- - - - -	1 - - 1 1	- - - 0 -
SS-SSIM	0 - - - 0	- - - - -	- - - - -	0 - - - 0	0 - - 0 0	- - - - -	0 - - 0 0	0 - - - 0	0 0 0 - -	- - - - -	- - - 0 0
MS-SSIM	0 - - - -	- - - - -	- - - - -	- - - - -	0 - - 0 0	- - - - 1	0 - - 0 0	0 - - - 0	- 0 0 - -	- - - - -	- 0 - 0 0
VSNR	- - - - -	1 - - - 1	- - - - -	- - - - -	0 - - 0 0	- - - - 1	- - - 0 -	- - - - -	- 0 - - -	1 - - - 1	- 0 - 0 -
VIF	- - - - 1	1 - - 1 1	1 - - 1 1	1 - - 1 1	- - - - -	1 - - 1 1	- - - - 1	- - - 1 1	1 - 0 1 1	1 - - 1 1	1 - - - -
UQI	- - - - 0	- - - - -	- - - - 0	- - - - 0	0 - - 0 0	- - - - -	0 - - 0 0	0 - - - 0	- 0 - - 0	- - - - -	- 0 - 0 0
NQM	- - - - -	1 - - 1 1	1 - - 1 1	- - - 1 -	- - - - 0	1 - - 1 1	- - - - -	- 0 - 1 -	- 0 - 1 -	1 - - 1 1	- 0 - - -
WSNR	- - - - -	1 - - - 1	1 - - - 1	- - - - -	- - - 0 0	1 - - - 1	- 1 - 0 -	- - - - -	- - - - -	1 - - - 1	- - - 0 -
SNR	- - - - -	1 1 1 - -	- 1 1 - -	- 1 - - -	0 - 1 0 0	- 1 - - 1	- 1 - 0 -	- - - - -	- - - - -	1 1 1 - -	- - 1 0 0
VQM	0 - - 0 0	- - - - -	- - - - -	0 - - - 0	0 - - 0 0	- - - - -	0 - - 0 0	0 - - - 0	0 0 0 - -	- - - - -	- 0 - 0 0
MOVIE	- - - 1 -	- - - 1 1	- 1 - 1 1	- 1 - 1 -	0 - - - -	- 1 - 1 1	- 1 - - -	- - - 1 -	- - 0 1 1	- 1 - 1 1	- - - - -

value of '1' in the tables indicates that the row (algorithm) is statistically better than the column (algorithm), while a value of '0' indicates that the row is worse than the column; a value of '–' indicates that the row and column are statistically identical. In Tables XVIII and XIX, we evaluate this hypothesis for each distortion category as well as for all distortions considered together.

Tables XVIII and XIX validate our observations from the correlations – NQM, VIF, VQM perform well, although interestingly, NQM is the only algorithm that is statistically superior to PSNR overall for the mobile study, while VIF is superior to PSNR in the tablet study, where MOVIE also performed well.

2) Comparison With the Theoretical Null Model: We also performed an analysis to evaluate whether algorithm performances were different from the theoretical null model [5], [46]. Given that we have performed all analysis up to this point using DMOS scores from the database, and given that humans exhibit inter-subject variability , it is important not to penalize an algorithm if the differences between the algorithm scores and DMOS can be explained by the differences between the individual subjective scores and the DMOS. This variance between the differential opinion scores (DOS) and the DMOS is used as a measure of the inherent variance of subjective opinion, and we analyze whether the variances of differences between the algorithm scores and DOS are statistically equivalent to that of DOS and DMOS. Our analysis unfolds as in [5]. Specifically, we compute the ratio between (a) the variances ($\sigma^2_{algorithm}$) of residuals between the differential opinion scores (DOS) and algorithm scores (after non-linear regression) and (b) the variances (σ^2_{null}) of residuals between the differential opinion scores (DOS) and DMOS for each distortion as well as across all distortions. The ratio of two variances $\sigma^2_{algorithm}/\sigma^2_{null}$ is the F-statistic and at the 95% confidence level, for the degrees of freedom exhibited by the numerator and denominator, one can compute the threshold F-ratio. If the computed F-statistic exceeds the threshold F-ratio, then one accepts the null hypothesis – i.e., the algorithm performance is equivalent to the theoretical null model – else, one rejects the null hypothesis. In Tables XX and XXI we report the F-statistic for each distortion and for all distortions for each of the algorithms considered here, as well as the threshold F-ratio for the mobile and tablet study respectively. Fields marked in bold indicate acceptance of the null hypothesis. The tables indicate that across distortions, there does not exist a single algorithm that is equivalent to the theoretical null model, except VIF on the wireless distorted videos. Clearly, there remains much work to do on video quality assessment, both on developing fully scalable VQA algorithms and especially towards understanding human reactions to temporal video dynamics and how to model them.

IV. DISCUSSION AND CONCLUSION

We described a human study to assess video quality which was conducted on multiple mobile platforms and encompassed a wide variety of distortions, including dynamically-varying distortions as well as uniform compression and wireless packet-

TABLE XX
MOBILE STUDY: ALGORITHM PERFORMANCE VS. THE THEORETICAL NULL MODEL. LISTED ARE THE F-RATIOS I.E., RATIO OF (A) VARIANCES OF RESIDUALS BETWEEN THE DIFFERENTIAL OPINION SCORES (DOS) AND ALGORITHM SCORES AND (B) VARIANCES OF RESIDUALS BETWEEN THE DIFFERENTIAL OPINION SCORES (DOS) AND DMOS FOR EACH DISTORTION AS WELL AS ACROSS ALL DISTORTIONS. ALSO LISTED IS THE THRESHOLD F-RATIO. THE ALGORITHM IS STATISTICALLY EQUIVALENT TO THE NULL MODEL IF THE F-RATIO IS GREATER THAN THE THRESHOLD F-RATIO. BOLD FONT INDICATES STATISTICAL EQUIVALENCE TO THE THEORETICAL NULL MODEL

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.8331	0.1365	0.0342	0.8391	0.3821
SS-SSIM	0.7570	0.0212	0.0302	0.7722	0.3526
MS-SSIM	0.7959	0.2384	0.0327	0.8589	0.4010
VSNR	0.9764	0.2054	0.0360	1.0432	0.4614
VIF	1.0555	0.2094	0.0022	**1.1661**	0.4959
UQI	0.4549	0.0407	0.0128	0.7934	0.3507
NQM	0.7845	0.2172	0.0262	1.1043	0.4651
WSNR	0.7739	0.1365	0.0004	0.7658	0.3197
SNR	0.5727	0.0755	0.0014	0.5297	0.2156
VQM	0.7966	0.2337	0.0370	0.8392	0.3830
MOVIE	0.8897	0.2201	0.0117	0.7635	0.4100
Threshold F-ratio	1.1390	1.1622	1.1234	1.1390	1.0672

TABLE XXI
TABLET STUDY: ALGORITHM PERFORMANCE VS. THE THEORETICAL NULL MODEL. LISTED ARE THE F-RATIOS I.E., RATIO OF (A) VARIANCES OF RESIDUALS BETWEEN THE DIFFERENTIAL OPINION SCORES (DOS) AND ALGORITHM SCORES AND (B) VARIANCES OF RESIDUALS BETWEEN THE DIFFERENTIAL OPINION SCORES (DOS) AND DMOS FOR EACH DISTORTION AS WELL AS ACROSS ALL DISTORTIONS. ALSO LISTED IS THE THRESHOLD F-RATIO. THE ALGORITHM IS STATISTICALLY EQUIVALENT TO THE NULL MODEL IF THE F-RATIO IS GREATER THAN THE THRESHOLD F-RATIO. BOLD FONT INDICATES STATISTICAL EQUIVALENCE TO THE THEORETICAL NULL MODEL

	Compression	Rate Adaptation	Temporal Dynamics	Wireless	All
PSNR	0.9773	0.0859	0.0043	0.6947	0.2932
SS-SSIM	0.5638	0.0802	0.0005	0.4514	0.1743
MS-SSIM	0.8095	0.1434	0.0031	0.6463	0.2809
VSNR	0.9873	0.1163	0.0033	0.6930	0.3022
VIF	1.1904	0.1589	0.0002	0.9459	0.4242
UQI	0.2844	0.0271	0.0063	0.4224	0.0771
NQM	1.0823	0.0766	0.0009	0.8928	0.3749
WSNR	1.0915	0.2023	0.0032	0.6820	0.3233
SNR	0.8421	0.1623	0.0084	0.4884	0.2237
VQM	0.7773	0.0717	0.0000	0.6280	0.2462
MOVIE	1.1253	0.2897	0.0000	0.9966	0.4260
Threshold F-ratio	1.1956	1.2292	1.1732	1.1956	1.0831

loss. The large size of the study and the variety that it offers allows one to study and analyze human reactions to temporally varying distortions as well as to varying form factors from a wide variety of perspectives. We make a number of further observations that may prove useful—from the perspective of understanding human reactions to complex, time varying distortions and from the algorithm design perspective.

An obvious conclusion from our analysis is that time-varying quality has a definite impact on human subjective judgments of quality, and this impact is a function of the frequency of occurrence of significant distortion changes and of the differences in quality between segments. Humans seemingly prefer longer freezes over shorter ones – this is not terribly surprising since choppy video playback is not pleasing at all. However, what is surprising about the frame-freeze distortion is that humans appear to be far more forgiving of lost segments than they are of choppy quality. This has interesting implications for those supplying real-time video delivery. It is also prudent to note that while choppy playback is the worst offender, lost segments start to matter relative to small reductions in choppiness. Further, this preference is dependent upon the content being displayed. It would be interesting to study whether the same results hold true when viewing sports – a viewer may prefer choppy playback in this case as opposed to him missing out on the footage of that all important goal being scored. On the flip side, in applications such as video chatting it is possible that our results will be further validated. The data in this study seems to indicate that designers should use algorithms for resource allocation that penalize semi-filled buffers over those that penalize completely empty buffers.

The data from the rate adaptation and temporal dynamics distortions, while somewhat contrary to popularly held notions on human perception of quality are intuitive and interesting. The first observation is that humans are not as unforgiving as one would imagine them to be. In fact they seem to reward attempts to improve quality. As we summarized in the temporal dynamics discussion, when the user is subjected to a long spell of good quality video, s/he has seemingly taken that level of quality for granted, and when the provider switches to a much lower quality level, he is severe with his rating of quality. On the contrary, faster rate changes seemingly push the user to believe that the provider is attempting to maximize his quality of experience and hence these videos are given higher quality scores. Another explanation is that less rapid rate changes can produce long periods of low-quality video bracketed by segments of high-quality videos. In this case, the low quality may be regarded as more enduring, and hence, more annoying. Due to the limitations of study sessions we were unable to include the

other condition – $R_4 - R_1 - R4$ – here, not only is there high quality at the end, but there is also a segment of poor quality in the middle. From the current data it is difficult to predict how the user may react to this situation. Of course, variations on the rate of fluctuation in quality is another area to explore.

The field of analyzing continuous-time human opinion scores of quality is one that is still nascent. We explored a small set of preliminary temporal pooling ideas drawn from the literature or from conventional wisdom. Our results, while encouraging, still do not completely explain human responses to temporally varying distortions. For compression and wireless distortions, the mean of human opinion across time is a good indicator of the final quality – possibly owing to the fact that with stagnant quality, the human simply picks the mean when providing continuous quality scores. What is surprising is the performance of percentile pooling. This strategy works well for larger screen displays (albeit using an indirect method to assess its performance – pooling of objective scores [11], [33], [34]), but humans are seemingly more forgiving of poorer quality when viewing videos on smaller form factors. The observations from the memory-effect pooling are intriguing. While the mean of continuous quality scores is poor indicator of the final quality for videos with dynamically varying distortions, memory-effect based pooling seems to better capture human responses. With a change in the device form factor however, even this pooling strategy begins to fail. This implies that there is a lot more work to be done in understanding how humans integrate continuous quality scores and produce the final summarized score that they give each video. This is even more true for the frame-freeze distortions. It is unclear at this point how humans rate the effects of frame-freeze distortions on the temporal perception of video quality.

While a lot more can be said with regards to the human data, in the interest of space we now move our discussion to the objective algorithms. To us, the main takeaway from the analysis is that scalability, which requires multiscale processing, is a desirable property to assess the quality of videos of diverse sizes, resolutions and display forms. Single-scale algorithms such as VQM and SS-SSIM, which do well on videos shown on larger screens, may not accurately predict the quality of videos displayed on smaller screens.

Results from the temporally varying distortions are both disappointing and encouraging at the same time. It seems that for smaller rate variations, the algorithms manage to do resonably well in predicting quality, however with increased variation in the temporal distortion patterns, the algorithms fail. While this may be due to a multitude of factors, one possible reason could be the temporal pooling strategy applied. For the IQA algorithms, our strategy was simply to use the temporal mean of the frame-level scores, while the VQA algorithms pooled the predicted temporal scores as they were designed to do (eg., MOVIE uses the mean). In light of the results from our temporal pooling analysis of human scores and recent research in temporal pooling strategies for objective algorithms [32], [35], it seems very likely that algorithm performance can be improved by employing more appropriate strategies for integrating quality scores over time. Incorporating knowledge of the device and human responses to temporal quality as a function of the form factor should lead to additional benefits. Clearly, there remains ample room for developing better VQA algorithms – since none of the algorithms are equivalent (or even close) to the theoretical null model.

We hope that the new LIVE mobile VQA database of 200 distorted videos and associated human opinion scores from over 50 subjects will provide fertile ground for years of future research. Given the sheer quantity of data, we believe that our foregoing analysis is the tip of the ice-berg of discovery. We invite further analysis of the data towards understanding and producing better models of human behavior when viewing videos on mobile platforms. Other fields of inquiry that may benefit from this database include human behavior modeling; application and content driven analysis of human behavior; device and context-specific design of objective algorithms; video network resource allocation over time and many others. Given the explosion of mobile devices, and associated load on bandwidth, we believe that the work presented here and the observations made with regards to human behavior will serve as essential tools in modeling video delivery over wireless networks. The database is available at no charge to researchers at: http://live.ece.utexas.edu/research/quality/live_mobile_video.html.

APPENDIX
INSTRUCTIONS TO THE SUBJECT

You are taking part in a study to assess the quality of videos. You will be shown a video at the center of your screen and there will be a rating bar at the bottom, which can be controlled by using your fingers on the touchscreen. You are to provide the quality as function of time – i.e., move the rating bar in real-time based on your instantaneous perception of quality. The extreme left on the bar is bad quality and the extreme right is excellent quality. At the end of the video you will be presented with a similar bar, this time calibrated as 'Bad', 'Poor' and 'Excellent', from left-to-right. Using this bar, provide us with your opinion on the overall quality of the video. There is no right or wrong answer, we simply wish to gauge your opinion on the quality of the video that is shown to you.

REFERENCES

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010–2015, CISCO Corp., 2011, [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11–520862.html

[2] Global Internet Phenomena Spotlight, Sandvine, 2011, [Online]. Available: http://www.sandvine.com/downloads/documents/05–17-2011_phenomena/Sandvine%20Global%20Internet%20Phenomena%20Spotlight%20-%20Netflix%20Rising.pdf

[3] FCC Warns of Impending Wireless Spectrum Shortage, PCWorld, 2010, [Online]. Available: http://www.pcworld.com/article/186434/fcc _warns_of_impending_wireless_spectrum_shortage.html

[4] S. Higginbotham, Spectrum Shortage Will Strike in 2013, 2010, [Online]. Available: http://gigaom.com/2010/02/17/analyst-spectrum-shortage-will-strike-in-2013/

[5] 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. 2, pp. 1427–1441, Feb. 2010.

[6] A. K. Moorthy, K. Seshadrinathan, R. Soundararajan, and A. C. Bovik, "Wireless video quality assessment: A study of subjective scores and objective algorithms'," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 4, pp. 513–516, Apr. 2010.

[7] Final Report from the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment Phase II, Video Quality Experts Group (VQEG), 2003, [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII

[8] Final Report from the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment Phase I, Video Quality Experts Group (VQEG), 2000, [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI
[9] Final Report of Video Quality Experts Group Multimedia Phase I Validation Test, TD 923, ITU Study Group 9, Video Quality Experts Group (VQEG), 2008.
[10] F. D. Simone, M. Naccari, M. Tagliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective assessment of h.264/avc video sequences transmitted over a noisy channel," in *Proc. 1st Int. Workshop Quality of Multimedia Experience (QoMEX)*, Jul. 2009.
[11] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.
[12] 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.
[13] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Amer.*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.
[14] A. Watson, J. Hu, and J. McGowan, III, "Digital video quality metric based on human vision," *J. Electron. Imag.*, vol. 10, p. 20, 2001.
[15] S. R. Gulliver and G. Ghinea, "The perceptual and attentive impact of delay and jitter in multimedia delivery," *IEEE Trans. Broadcasting*, vol. 53, no. 2, pp. 449–458, Jun. 2007.
[16] Q. Huynh-Thu and M. Ghanbari, "Impact of jitter and jerkiness on perceived video quality," in *Proc. Workshop Video Process. Quality Metrics*, 2006.
[17] A. Eichhorn and P. Ni, "Pick your layers wisely-a quality assessment of h. 264 scalable video coding for mobile devices," in *Proc. 2009 IEEE Int. Conf. Commun.*, 2009, pp. 5446–5451.
[18] H. Knoche, J. McCarthy, and M. Sasse, "Can small be beautiful?: Assessing image resolution requirements for mobile tv," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 829–838.
[19] S. Jumisko-Pyykko and J. Hakkinen, "Evaluation of subjective video quality of mobile devices," in *Proc. 13th Annual ACM Int. Conf. Multimedia*, 2005, pp. 535–538.
[20] M. Ries, O. Nemethova, and M. Rupp, "Performance evaluation of mobile video quality estimators," in *Proc. Eur. Signal Process. Conf.*, Poznan, Poland, 2007.
[21] S. Jumisko-Pyykko and M. Hannuksela, "Does context matter in quality evaluation of mobile television?," in *Proc. 10th Int. Conf. Human Comput. Interact. Mobile Devices and Services*, 2008, pp. 63–72.
[22] S. Winkler and F. Dufaux, "Video quality evaluation for mobile applications," in *Proc. SPIE Conf. Vis. Commun. Image Process.*, Lugano, Switzerland, 2003, vol. 5150, pp. 593–603.
[23] Joint Draft ITU-T Rec. H.264 — ISO/IEC 14496-10/Amd.3 Scalable Video Coding, Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, [Online]. Available: http://www.hhi.fraunhofer.de/fileadmin/hhi/downloads/IP/ip_ic_H.264-MPEG4-AVC-Version8-FinalDraft.pdf
[24] A. K. Moorthy, L. K. Choi, G. deVeciana, and A. C. Bovik, "Subjective analysis of video quality on mobile devices," in *Proc. 6th Int. Workshop Video Process. Quality Metrics (VPQM) (Invited Article)*, Jan. 2012.
[25] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h. 264/avc standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.
[26] SVC Reference Software (JSVM Software), Joint Video Team (JVT), [Online]. Available: http://ip.hhi.de/imagecom_G1/savce/downloads/SVC-Reference-Software.htm
[27] T. Stockhammer, M. Hannuksela, and T. Wiegand, "H.264/avc in wireless environments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 657–673, Jul. 2003.
[28] B. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer, 1995.
[29] BT-500-11: Methodology for the Subjective Assessment of the Quality of Television Pictures, Int. Telecommuncation Union Std..
[30] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. SPIE Vis. Commun. Image Process.*, 2003, vol. 5150.
[31] The XGL Toolbox, 2008 [Online]. Available: http://128.83.207.86/jsp/software/xgltoolbox-1.0.5.zip
[32] K. Seshadrinathan and A. Bovik, "Temporal hysteresis model of time varying subjective video quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2011, pp. 1153–1156.
[33] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process., Iss. on Visual Media Quality Assess.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.
[34] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2006, pp. 2945–2948.
[35] J. Park, K. Seshadrinathan, S. Lee, and A. Bovik, "Spatio-temporal quality pooling accounting for transient severe impairments and egomotion," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2010, pp. 2509–2512.
[36] M. Gaubatz, Metrix Mux Visual Quality Assessment Package, [Online]. Available: http://foulard.ece.cornell.edu/gaubatz/metrix_mux/
[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," *IEEE Signal Process. Lett.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[38] Z. Wang, L. Lu, and A. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.
[39] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.
[40] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[41] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
[42] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Sep. 2002.
[43] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 4, pp. 525–535, Jul. 1974.
[44] Video Quality Metric, [Online]. Available: http://www.its.bldrdoc.gov/n3/video/VQM software.php
[45] K. Seshadrinathan, Movie Software Release, 2010 [Online]. Available: http://live.ece.utexas.edu/research/Quality/movie.html
[46] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

Anush Krishna Moorthy received the B.E. degree in electronics and telecommunication with a Silver Medal from the University of Pune, Pune, India, in 2007, the M.S. degree in electrical engineering from The University of Texas at Austin, in 2009, and the Ph.D. degree from The University of Texas at Austin, in 2012.

He joined the Laboratory for Image and Video Engineering (LIVE), The University of Texas, Austin, in January 2008 and was the Assistant Director of LIVE from 2008 to 2012. Anush Moorthy is the recipient of the Continuing Graduate Fellowship for 2010–2011, the Professional Development Award, Fall 2009, Fall 2010 and the Center for Perceptual Systems Travel Grant, Spring 2010, from The University of Texas at Austin and the TATA scholarship for higher education abroad. He currently works as an advanced imaging engineer at Texas Instruments, Dallas, Texas.

His research interests include image and video quality assessment, image and video compression, and computational vision.

Lark Kwon Choi received the B.S. degree in Electrical Engineering from Korea University, Seoul, Korea, in 2002, and the M.S. degree in Electrical Engineering and Computer Science from Seoul National University, Seoul, Korea, in 2004, respectively. He worked at KT (formerly Korea Telecom) as a senior engineer from 2004 to 2009 on IPTV platform research and development. He participated in IPTV standardization in International Telecommunication Union (ITU-T) and Telecommunications Technology Association (TTA).

He is currently pursuing his Ph.D. degree as a member of the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin under Dr. Alan C. Bovik's supervision. His research interests include image and video quality assessment, spatial and temporal visual masking, and motion perception.

Alan Conrad Bovik (F'96) is the Curry/Cullen Trust Endowed Chair Professor at The University of Texas at Austin, where he is Director of the Laboratory for Image and Video Engineering (LIVE). He is a faculty member in the Department of Electrical and Computer Engineering and the Center for Perceptual Systems in the Institute for Neuroscience. His research interests include image and video processing, computational vision, and visual perception. He has published more than 650 technical articles in these areas and holds two U.S. patents. His several books include the recent companion volumes The Essential Guides to Image and Video Processing (Academic Press, 2009). Al was named the SPIE/IS&T Imaging Scientist of the Year for 2011. He has also received a number of major awards from the IEEE Signal Processing Society, including: the Best Paper Award (2009); the Education Award (2007); the Technical Achievement Award (2005), and the Meritorious Service Award (1998). He received the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin, the Distinguished Alumni Award from the University of Illinois at Champaign-Urbana (2008), the IEEE Third Millennium Medal (2000) and two journal paper awards from the international Pattern Recognition Society (1988 and 1993). He is a Fellow of the IEEE, a Fellow of the Optical Society of America (OSA), a Fellow of the Society of Photo-Optical and Instrumentation Engineers (SPIE), and a Fellow of the American Institute of Medical and Biomedical Engineering (AIMBE). He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996–1998; co-founder and Editor-in-Chief, IEEE TRANSACTIONS ON IMAGE PROCESSING, 1996–2002; Editorial Board, The Proceedings of the IEEE, 1998–2004; Series Editor for Image, Video, and Multimedia Processing, Morgan and Claypool Publishing Company, 2003-present; and Founding General Chairman, First IEEE International Conference on Image Processing, held in Austin, Texas, in November, 1994. Dr. Bovik is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.

Gustavo de Veciana (S'88–M'94–SM'01–F'09) received his B.S., M.S., and Ph.D. in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993 respectively. He is currently a Professor at the Department of Electrical and Computer Engineering and recipient of the Temple Foundation Centennial Fellowship. He served as the Director and Associate Director of the Wireless Networking and Communications Group (WNCG) at the University of Texas at Austin, from 2003–2007.

His research focuses on the analysis and design of wireless and wireline telecommunication networks; architectures and protocols to support sensing and pervasive computing; applied probability and queueing theory. Dr. de Veciana has served as editor for the IEEE/ACM TRANSACTIONS ON NETWORKING. He was the recipient of a National Science Foundation CAREER Award 1996, co-recipient of the IEEE William McCalla Best ICCAD Paper Award for 2000, co-recipient of the Best Paper in ACM Transactions on Design Automation of Electronic Systems, Jan 2002–2004, co-recipient of the Best Paper in the International Teletraffic Congress (ITC-22) 2010, and of the Best Paper in ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems 2010. In 2009 he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He is on the technical advisory board of IMDEA Networks .