Video Quality Assessment for Mobile Devices on Mobile Devices

Milan Mirkovic(✉), Dubravko Culibrk, Srdjan Sladojevic, and Andras Anderla

Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
mirkovic.milan@gmail.com

Abstract. Pervasiveness of mobile devices and ubiquitous broadband Internet access have laid foundations for video content to be consumed increasingly on smart phones or tablets. As over 85% of the global consumer traffic by 2016 is estimated to be generated by streaming video content, video quality as perceived by end-users of such devices is becoming an important issue. Most of the studies concerned with Video Quality Assessment (VQA) for mobile devices have been carried out in a carefully controlled environment, thus potentially failing to take into account variables or effects present in real-world conditions. In this paper, we compare the results of traditional approach to VQA for mobile devices to those obtained in real-world conditions by using a physical mobile device, for the same video test-set. Results indicate that a difference in perceived video quality between the two settings exists, thus laying foundations for further research to explain the reasons behind it.

Keywords: Video quality assessment · Subjective · Mobile devices

1 Introduction

Recent years have witnessed a tremendous increase in usage of mobile devices to access the Internet and its services. Fierce competition on the end-user electronics market has caused smart phones to become accessible to everyone, and similar situation in the telecommunications department has made broadband Internet access cheap and ubiquitous. As a consequence, vast majority of population in developed countries now owns a cell phone [1], while one quarter of the smart phone owners also possess a tablet [2]. These devices are becoming more potent and versatile by the day and the result of this evolution is not only a change in people's habits when day-to-day tasks are in question, but a shift in the way some traditional services are perceived (such as TV, mail or telephony). This shift is especially noticeable when video content is concerned, as more and more of it gets consumed "on the move" [3]. In fact, some estimates have it that 86% of the global consumer traffic by 2016 will be generated by streaming video content [4]. Such high percentage inevitably raises the issue of quality of the delivered content, as perceived by the end-users.

V. Murino et al. (Eds.): ICIAP 2015 Workshops, LNCS 9281, pp. 555–562, 2015.
DOI: 10.1007/978-3-319-23222-5_67

Successful design and validation of different quality assessment approaches first requires the ground truth data, which in the domain of Video Quality Assessment (VQA) takes the form of degraded sequences and the Mean Opinion Scores (MOS) gathered mostly in laboratory tests on human observers. The sequences are degraded through multimedia coding and decoding processes and the effects of transmission are simulated by eliminating a certain proportion of packets from the encoded data, before passing the data to the decoder [5]. Important constraint when VQA for mobile devices is concerned however, is that close-to-ideal laboratory conditions can rarely be encountered in the real world.

In this paper, we consider network-induced video impairments and their effect on video quality as perceived by the consumers when observed in controlled laboratory conditions on a desktop monitor, versus when observed on a physical mobile device (i.e. a tablet) in a real-world scenario. Our main research hypothesis is that there are differences in the way the same video material is perceived (quality-wise) depending on the screen/device and setting (environment, context) of its presentation. As artefacts introduced to streaming videos are to a large degree dependant on network conditions and people are familiar with them appearing when watching content on mobile devices, we chose to focus on these to derive a set of videos for our experiments.

The rest of the paper is organized as follows: Section 2 provides an overview of the related work in the field, Section 3 describes the proposed approach in more detail, Section 4 presents the results of experiments conducted and Section 5 concludes the paper with a brief discussion and pointers for future research.

2 Background and Related Work

A recent survey by Winkler provides a fairly exhaustive list of publicly available video quality databases [5]. Winkler lists a total of 11 different databases, covering MPEG-2, Dirac wavelet, and H.264 codecs and simulated packet losses for wireless and wired IP transmission. When focusing on mobile devices, the only concession made in published databases is the reduction of the spatial resolution of sequences, due to the expected smaller screen sizes. The subjective quality assessment experiments are, however, carried out in the laboratory environment and conditions recommended for general multimedia were laid out at the time when majority of it was consumed over classic television and computer screens.

Most of the research in the field of VQA for mobile devices is aimed towards the evaluation of effect of different codecs, bitrates and content on perceived video quality. Winkler et. al. compare two coding standards used for mobile applications (MPEG-4 and Motion JPEG2000) by simulating the transmission errors of a WCDMA channel using representative bit error patterns provided by ITU-T [6]. They analyse the subjective scores obtained from assessors and use them to compare codec performance as well as the effects of transmission errors on visual quality, in a controlled environment using desktop computers for assessment. In subsequent work, they use similar experimental setting to explore the interactions between audio and video on perceived audiovisual quality, and

confirm that both the product and linear combination of the two components are an effective model of the audiovisual quality [7]. Similarly to Winkler, Jumisko-Pyykko et. al. compare the performance of different codecs and audio/video bitrates but use physical mobile devices (smart phones) for obtaining subjective quality scores, and conclude that presenting different content clearly requires different audio-video bitrate ratios at relatively low bitrates levels [8]. In addition, work of Jumisko et. al. [9] as well as that of Mirkovic et. al. [10] shows that the personal interest in content presented to the assessors might be an important factor for video quality evaluation, and they recommend measuring the evaluator's interest in content in subjective assessment studies. Although in the work of Jumisko et. al. actual mobile devices have been used to obtain MOS, the experiments were executed in a controlled environment.

At the same time, there is an ongoing debate spanning different research areas on whether and how do results obtained in a controlled environment (i.e. in a laboratory) correlate with those observed in field experiments. Opinions – and indeed results – vary across domains, but a consensus on the matter has still not be reached. While it is generally accepted that laboratory studies are good at telling whether or not some manipulation of an independent variable causes changes in the dependent variable, many scholars assume that these results do not generalize to the "real world" [11]. Even though some studies have shown otherwise [12][13], they acknowledge that the failure to find high correspondence between lab and field studies in a given domain or with a specific phenomenon should not be seen as a failure of the researchers in either setting, but should be seen as an indicator that further conceptual analysis and additional empirical tests are needed to discover the source of the discrepancy.

As we are aware of no mobile-device-targeted VQA study that attempts analyse the effect of network impairments as they occur in real-world scenarios and compare them to quality results obtained through traditional approach in a controlled environment, we aim to establish a connection (or show the lack of) in this research.

3 Methodology

To create the test set, we adopted the following approach: a video was streamed by the server and delivered to an Android mobile phone, which logged the packets received as well as the times they were received at, starting from the first packet received. An application that relies on the FFMpeg library [14] was developed for the Android platform and used to this extent. At the same time, we have developed a client (PC) application that is able to receive a video stream and uses a log file generated on the mobile device as input, to impair the stream it is receiving.

The client application drops packets not present in the log file, and decodes video frames using the rest. To ensure successful decoding, first 10 packets containing coding parameters are not dropped. Timestamps from the log file are used to withhold the received packets until the reception time-stamp time has

passed as measured from the first packet received. The result is a set of decoded frames. The decoded frames are generated using default error concealment of the FFMpeg. Those frames missing completely are created by copying them from the last decoded frame, to create the video of the original frame rate.

Using this approach, we created a test set consisting of 50 impaired videos and 10 reference videos, that we used to obtain MOS from human assessors. To do this, we employ a standard methodology (as recommended in [15]), but first divide assessors into two groups: (i) "laboratory" – where the experiment is conducted in a controlled environment as proposed by the ITU-T and (ii) "field" – where the experiment is conducted in real-world conditions.

We also ask assessors to provide us with more details about their opinion on the perceived quality of the test-set as a whole. With the "laboratory" group we do this via a post-experiment questionnaire where the participants are asked to evaluate the effect of 4 distinct factors (namely "freeze", "jerkiness", "blockiness/distortion" and "large grey areas") on a grading scale, and leave them with the option to provide additional qualitative comments. With the "field" group we conduct brief post-experiment interview where we ask assessors for their opinion.

3.1 Test Set

To create the test set, we used reference videos available within the LIVE Video Quality Database. Videos used were: Pedestrian Area, River Bed, Rush Hour, Tractor, Station, Sunflower, Blue Sky, Shield, Park Run and Mobile & Calendar.

The spatial resolution of all videos was 768x432 pixels, but they were first resized to a resolution of 384x216, as this is a resolution more appropriate for mobile device applications. In addition, all videos were converted to 25 fps frame rate. Finally, they were compressed using x264vfw, which is the the VfW (Video for Windows) version of a well known x264 encoder and ffh264 decoder (from FFmpeg/Libav project) [14]. Since the sequences are short, the logs in various scenarios were acquired by concatenating all ten videos and streaming this content in a single session. The merged video is 493s long, and has 12,325 frames. The video is split into 29,615 RTP packets for streaming. In this aspect, we evaluated three common scenarios: (i) static device over GSM, where the stream was received by a static mobile device, (ii) driving scenario, where the stream was received over GSM, while driving on the highway at the average speed of around 65 mph, and (iii) WiFi scenario, where content was streamed over WiFi network to a static mobile device.

The ratio of received packets and decoded frames was as follows: (i) static GSM scenario – 9,210 packets received and 4,377 frames decoded, (ii) highway GSM scenario – 1,908 packets received and 657 frames decoded, and (iii) WiFi scenario – 11,837 packets received and 5,239 frames decoded. For each scenario a log file was created, for the whole duration of streaming. As our logs were much longer than the average length of the test sequence, we first chose interesting transmission segments based on the running average of the number of packets received over an interval twice the maximum sequence length (20 second) interval. Based on this, our three multimedia use scenarios yielded 5 network

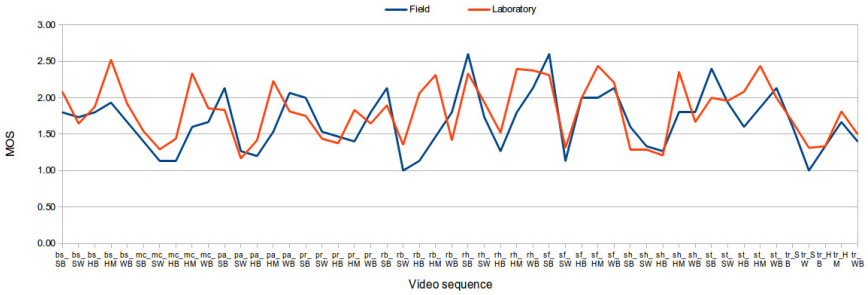

Fig. 1. Comparison of MOS obtained for each of the videos

impairment scenarios: static GSM best segment (SB), static GSM worst (SW), highway GSM best (HB), highway GSM average (HM) and WiFi best (WB) segment. The performance in the worst segments of highway scenario yielded too few packets to create viable impaired test sequences, so we opted for average segments, while the performance in the WiFi scenario was selected to represent ideal streaming conditions, so we opted for the best segment.

Once the segments were chosen, we reproduced those reception scenarios using our impairment application, which was run on the same machine as the server, in order to avoid any additional packet loss. The decoded frames were saved to disk as bitmaps, along with the decoding times from the log. To create impaired videos, a Matlab script was created to concatenate the frames and fill in the missing frames by copying the last decoded frame. Thus, the freeze effects were created when no frames were available from the decoder. For each original sequence, we created 5 impaired videos. E.g. for video named bs, we created bs1, bs2, bs3, bs4 and bs5 which represent network impairments scenarios logged in scenarios SB, SW, HB, HM and WB.

4 Experiment and Results

Subjective assessment was conducted for the 50 videos created. DSIS Variant I method was used [15], where the assessor is first presented with an unimpaired, reference sequence, and then with the same sequence impaired. He/she is then asked to vote on the second sequence, keeping in mind the first. Voting is done on a 1 to 5 scale, 1 being the lowest score where perceived impairments are very annoying and 5 being the highest, where impairments cannot be perceived. Sequences with different levels of impairments are randomly presented to assessors to avoid ordering effects.

The final MOS value for a sequence is the average score over all assessors for that sequence. Average MOS for a Scenario is calculated as the average of all MOS assigned to videos impaired using the same Scenario settings.

For the "laboratory" group, the environment where the tests were conducted was set up in a dedicated computer laboratory as proposed in [16] [17]. Sequences were presented to assessors on a 18.5" LG Flatron monitor (E1941), which was

operated at its native resolution of 1366x768 pixels. For the "field" group, the tests were conducted on a Asus Transformer Pad TF300T tablet device (10.1" touch-screen operated at a native resolution of 1280x800 pixels) in an office, and a custom-developed software was used for voting. Assessors in this group were allowed to freely position the tablet device in terms of viewing distance and angle, in order to make themselves comfortable and to reduce the potential glare. In addition, each of them was allowed to choose the position they found the most comfortable for performing the assessment (sitting at the table, in the easy chair or being laid-back in the lazy-bag), to account for the fact that different people have different habits when it comes to watching videos on mobile devices.

The actual test in case of each group consisted of one session of about 25 minutes, including training. Before the test, written instructions were given to subjects, and a test session was run that consisted of videos demonstrating the extremes of expected video quality ranges. It was also specifically pointed out that the assessment should be focused on the quality and not the content of the video. In the "laboratory" group, 48 subjects – 41 male and 7 female participated in the test, their age ranging from 18 to 54. In the "field" group, 15 subjects – 11 male and 4 female participated, their age ranging from 24 to 64. Even though there is a considerable difference in the size of groups, number of assessors in the "field" group conforms with the recommendations for the minimum number of participants given in [15]. None of the assessors in either group were familiar with video processing, nor had previously participated in similar tests. All of the subjects reported normal or corrected vision prior to testing.

Obtained MOS results for each video separately and over all scenarios are presented in Figure 1 and Table 1 respectively. Due to the severe impairments introduced to original videos that were easily identifiable by simple overview of the resulting test-set, we have expected MOS to gravitate towards the low end of the scale. To identify differences between different scenarios, we performed Analysis of Variance (ANOVA) for scores obtained within each group. When "laboratory" group is concerned, ANOVA showed that for all but one video (video labelled "st") statistically significant differences exist between scenarios ($p<0.05$), meaning that assessors found a difference in quality of videos belonging to different Scenarios. Post-hoc analysis was run (Tukey's test) to identify where exactly these differences laid, and it revealed that for majority of videos this difference is due to scenario HM, mean of which significantly differs from mean of at least one other scenario (and often differs from 3 or all 4 of scenarios).

For the "field" group, assessors found differences in video quality (which are statistically significant) between different scenarios for only 4 out of 10 videos ($p<0.05$ for those labeled "pa", "rb", "rh" and "sf"). Post-hoc analysis showed that these differences were largely due to scenario SB, which scored higher than at least one other scenario (usually scenario SW).

Analysis of additional questionnaire and interview materials revealed that the two groups identified the same impairment ("video freeze") as the most influential on their opinions. What is interesting though, is that in the case of "laboratory" group this impairment contributes positively to the score given (i.e.

Table 1. MOS and % of decoded frames for different scenarios

Scenario	Average % of decoded frames	MOS "Laboratory"	MOS "Field"
SB	61.20	1.87	2.03
SW	17.30	1.47	1.38
HB	15.60	1.63	1.42
HM	6.60	2.27	1.71
WB	60.40	1.84	1.86

to a degree, the more prominent the effect, the higher score is given), while in the case of "field" group it affects the score in the opposite manner (more "freezing" will annoy the assessors and cause video to receive lower score).

5 Conclusions and Future Work

Results suggest that people tend to be more tolerant to impairments when observing videos on mobile devices in a relaxed environment than when watching the same sequences in controlled, laboratory conditions. Part of explanation for this might lay in the fact that people behave differently (e.g. they are more concentrated on the task or pay attention to details more closely) when they are asked to do something in a laboratory as opposed when being asked to do the same thing in a more informal setting. This has to do with human psychology and the wish to perform well, but also might be a good cue for researchers in the domain of VQA to relax their criteria a little when it comes to real-world applications, or to find means to exploit these phenomenons to devise more efficient ways for video streaming. Also, another important contributor to the results obtained might be the tacit expectation when it comes to observing videos on mobile devices; people might simply have a lower criteria for the quality they expect to observe because of effects they are used to experiencing when consuming videos on phones or tablets – such as "jerkiness" (which occurs commonly due to video buffering) or "freezing" (loss of connection).

We presented results of an initial study that was not designed to take into account or control all the variables that might bear some weight on final conclusion whether people really have a different perception of video quality when observed in different settings and on different devices. We discovered that some discrepancy exists between the MOS obtained through traditional methods and those acquired "in the field", so future research aimed at discovering what are the reasons for this disagreement is needed. In particular, we intend to design experiments which should take into account both psychological (i.e. behaviour related) and technological (i.e. intrinsic to devices) factors that might influence the VQA process in order to improve the current methods for obtaining subjective quality scores.

Acknowledgments. This research was supported by the FP7 Marie-Curie project QoSTREAM (Grant Agreement 295220).

References

1. Lenhart, A., Purcell, K., Smith, A., Zickuhr, K.: Social media & mobile internet use among teens and young adults. Technical report, Pew Internet & American Life Project, Washington, DC (2010)
2. ComScore: Today's U.S. Tablet Owner Revealed (2012)
3. O'Hara, K., Mitchell, A., Vorbau, A.: Consuming video on mobile devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 857–866. ACM (2007)
4. CISCO: Cisco Visual Networking Index : Forecast and Methodology, 2011–2016. Technical report, CISCO (2012)
5. Winkler, S.: Analysis of Public Image and Video Databases for Quality Assessment. IEEE Journal of Selected Topics in Signal Processing **6**, 616–625 (2012)
6. Winkler, S., Dufaux, F.: Video quality evaluation for mobile applications. In: Proceedings of SPIE Conference on Visual Communications and Image Processing, Lugano, Switzerland, pp. 593–603 (2003)
7. Winkler, S., Faller, C.: Audiovisual quality evaluation of low-bitrate video. In: SPIE/IS&T Human Vision and Electronic Imaging, pp. 139–148. Citeseer (2005)
8. Jumisko-Pyykko, S., Hakkinen, J.: Evaluation of subjective video quality of mobile devices. In: Proceedings of the 13th Annual ACM International Conference on Multimedia - MULTIMEDIA 2005, pp. 535–538 (2005)
9. Jumisko, S., Ilvonen, V.: Vaananen-vainio mattila, K.: Effect of TV content in subjective assessment of video quality on mobile devices. In: Proceedings of SPIE, vol. 5684, pp. 243–254 (2005)
10. Mirkovic, M., Vrgovic, P., Culibrk, D., Stefanovic, D., Anderla, A.: Evaluating the role of content in subjective video quality assessment. The Scientific World Journal **2014** (2014)
11. Campbell, D.: Factors relevant to the validity of experiments in social settings. Psychological Bulletin **54**, 297 (1957)
12. Anderson, C., Lindsay, J., Bushman, B.: Research in the psychological laboratory truth or triviality? Current Directions in Psychological Science **8**, 3–9 (1999)
13. Wolfe, J., Roberts, C.: A further study of the external validity of business games: five-year peer group indicators. Simulation & Gaming **24**, 21–33 (1993)
14. Bellard, F., Niedermayer, M.: FFMpeg (2007)
15. ITU-T: Subjective video quality assessment methods for multimedia applications (1999)
16. Winkler, S., Campos, R.: Video quality evaluation for internet streaming applications. In: Proceedings of SPIE Human Vision and Electronic Imaging, pp. 104–115 (2003)
17. ITUT: Methodology for the subjective assessment of the quality of television pictures (2002)