SUBJECTIVE QUALITY OF VIDEO BIT-RATE REDUCTION BY DISTANCE ADAPTATION

Qing Song[⋆]*, Pamela Cosman*[⋆]*, Morgan He*[⋆]*, Rahul Vanam*[†]*, Louis J. Kerofsky*[†]*, Yuriy A. Reznik*[†]

[⋆]UC San Diego, Dept. of Electrical and Comp. Engr., 9500 Gilman Dr, La Jolla, CA 92093-0407 USA
[†] InterDigital Communications, Inc., 9710 Scranton Road, San Diego, CA 92121 USA

ABSTRACT

We investigate the potential to reduce video bit-rate by adapting to specifics of a viewer's display device and viewing conditions. We conducted a subjective test to demonstrate the performance of a pre-processing filter for video compression which adapts to the viewing conditions of the user, specifically the viewing distance. We studied three viewing distances, corresponding to holding a tablet in the hand, on the lap, or on a stand. The visual quality of the compressed videos with and without the pre-filtering was compared, and we found that the pre-filtering can save on the average 30% and 3% approximately of the bit-rate for the on-lap and in-hand viewing modes, without degrading perceptual quality. Adapting to conditions of an individual viewer provides a promising area to reduce bit-rate without sacrificing video quality.

1. INTRODUCTION

In video transmission, reducing bit-rate is desirable as long as the video quality is preserved. Factors such as viewing conditions and visual attention have been found to affect the visibility of information displayed on a screen. The capability of the display device is implicitly a factor due to display device aspects such as resolution, physical size, and ambient reflectivity. Traditional approaches assume conservative viewing conditions and have explored methods to exploit perceptual viewing phenomena under assumed fixed viewing conditions. For example, some methods are proposed to select regions of interest (ROI) in the video and to preserve their high quality. The non-ROI regions could be blurred by a smoothing filter, allowing lower bit-rate, based on the assumption that the user's attention is not on those areas [1–3]. Besides visual attention, viewing conditions such as display size, brightness, pixel density, viewing distance, and ambient illumination also play a role in the visibility of information. For example, a device held farther away may have fewer details visible compared to a device held closer. Similarly, a device under sunlight may have fewer visible details compared to one seen in the dark. Transmission of such invisible details is wasteful.

The same video content may be viewed on any of a variety of devices under dynamically varying viewing conditions. The work of [4] examined typical usage of tablet devices and determined common usage clustered into modes such as On-Lap and On-Stand. These modes correspond to different viewing distances. The physical size of the display can also result in the device occupying different portions of a viewer's visual field. Consider a small mobile phone held in the hand on a sunny day, a mini tablet held at arm's length, or a tablet placed on a stand to watch long format content. The viewing conditions vary due to the display device but also due to the dynamic use of the device. A user may hold a tablet near while watching a short video clip but discomfort will prevent the user from holding a tablet at a close distance for the duration of long format content. Relevant viewing parameters of a mobile device will vary based on usage mode, display device, as well as ambient environment conditions.

Compressed bit-rate and video quality are inversely related with the relation depending upon content and viewing conditions. We are interested in exploiting the variation in viewing conditions to achieve rate reduction without sacrificing perceived video quality. Xue et al. [5] proposed a strategy to select quantization parameters based on an environment-aware quality assessment model which uses viewing distance, display size, ambient luminance and body movement. Another perceptually motivated technique is to filter the video prior to encoding based on the anticipated viewing conditions. A perceptual pre-filter in [6] removes the spatial oscillations in a video that are invisible under given viewing conditions, resulting in lower complexity images which can be compressed at a lower bit-rate without loss of subjective quality. Bit-rate savings can be documented but potential impact on subjective video quality requires visual testing. That is the goal of this work. To evaluate the perceptual quality performance of the pre-filter and the whole user-adaptive video delivery system, we conducted a subjective test based on the pair comparison (stimulus-comparison) method [7, 8]. Observers compared the quality of compressed videos shown on a tablet with and without pre-filtering, and graded each pair's difference. We examined three common viewing distances corresponding to using a tablet on a stand, on the lap, and in the hand.

The paper is organized as follows. In Section 2 we review the design of a viewing condition adaptive system. In Section 3 we describe the subjective testing. Results are in Section 4, and Section 5 concludes.

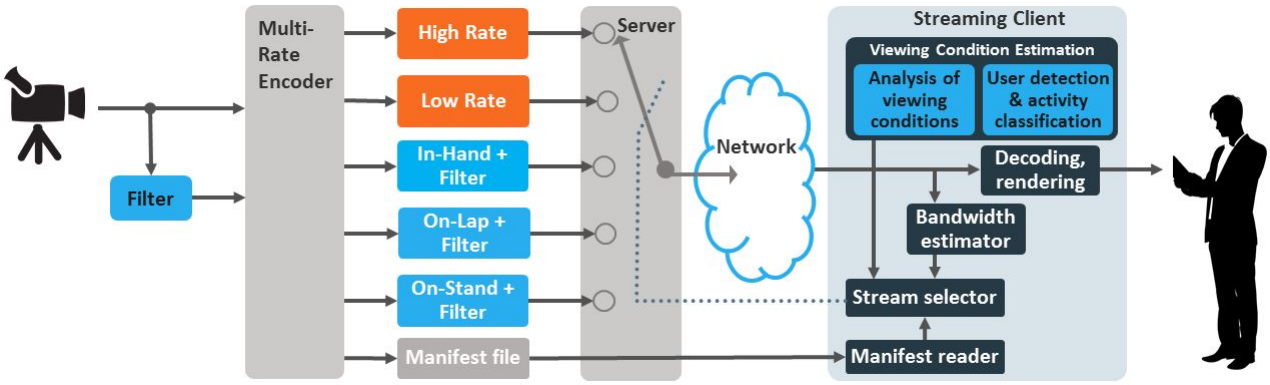

Fig. 1. Architecture of user-adaptive video delivery system

2. VIEWER ADAPTIVE SYSTEM

In conventional video coding and delivery systems, viewing condition parameters are not known and are assumed to be within typical ranges (e.g., viewing distance equal to 3 to 4 times screen height). However, as exemplified in Fig. 1, one can design an adaptive system that classifies user state and viewing conditions and then uses them to select one of the available encoded versions of the content (representations) on the HTTP server. The representations may include versions with different pre-filtering applied prior to encoding, as well as traditional encodings performed using different target bit-rates. A special manifest file is also placed on the HTTP server to describe properties of all available representations. In performing stream selection, the client software (media player) may find the best matching encoded video representation given a combination of current viewing conditions and network bandwidth limits. The design of such a user-adaptive video delivery system was first proposed in [9]. The implementation of user-adaptive streaming utilizing an MPEG-DASH streaming standard was described in [10].

As mentioned above, the representations of content may differ in the pre-filtering applied in addition to traditional factors. Given viewing conditions, the pre-filter may be used to remove details from the content which would be invisible but still require bits to transmit to the device. The perceptual pre-filter described in [6] exploits three basic phenomena of human vision: (1) Contrast sensitivity function (CSF): relationship between frequency and contrast sensitivity thresholds of human vision, (2) Eccentricity: rapid decay of contrast sensitivity as angular distance from gaze point increases, and (3) Oblique effect: lower visual sensitivity to diagonally oriented spatial oscillations as opposed to horizontal and vertical ones.

Fig. 2 shows examples of encodings produced with and without perceptual filtering. The encodings in sub-figures (c) and (d) use the same rate, however, the filtered version looks softer with fewer coding artifacts. When viewed from a certain distance, the softness introduced by the pre-filter becomes invisible, but bit-rate savings remain.

Fig. 2. Examples of different encodings (1st frame from Old town sequence [6]): (a) Original uncompressed frame, (b) Compressed at *High* rate, (c) Compressed at *Low* rate, (d) Filtered and Compressed at "On Stand" rate.

3. SUBJECTIVE TEST

We conducted a subjective test of the performance of the pre-filter using the pair comparison method [7, 8]. HD video source sequences were obtained from [11]. Video clips compressed with and without the pre-filtering are shown sequentially in some randomized order to the subjects who provide a comparative preference score. The videos were displayed on a tablet (Nexus 7). To begin, we defined three viewing modes: In-Hand, On-Lap, and On-Stand. The three viewing modes correspond to three viewing distances, i.e., three sets of filter parameters. For "In-Hand" mode, the device is held in both hands. Subjects sat in an armless chair, so their hands were not steadied against anything. For "On-Lap" mode, the device rests on the lap. Subjects could tilt the device to make a good viewing angle but the device remains on the lap. For "On-Stand" mode, the device is on a stand on a table, and the

subject does not touch it after the initial comfortable positioning. We assume the viewing distances of In-Hand, On-Lap and On-Stand modes are 12", 20" and 24" respectively [4].

3.1. Video versions

For each viewing mode, we apply the pre-filter to the original uncompressed video. Longer viewing distance results in stronger filtering so that more details are removed. Then the filtered videos are compressed by the x264 encoder [12], configured to produce High-Profile H.264/AVC-compliant bitstreams. We denote the compressed filtered videos as *user adaptive videos (UAV)*.

For comparison, we also compress the original video by the same encoder without pre-filtering. The video is compressed at two bit-rates: one bit-rate (called *High*) is higher than the highest *UAV* bit-rate, and the other (called *Low*) is approximately equal to the lowest *UAV* bit-rate. *High* and *Low* versions serve as negative and positive controls. The goal is that *UAV* should have quality equivalent to *High*, given the corresponding viewing conditions. However, if only *UAV* and *High* are compared and no difference is found, it is possible that this outcome arose because the observers are sleepy, distracted, or in some way unreliable, or because both data rates are so low (or so absurdly high) that no difference between them can be discerned. So we also compare *Low* with *UAV*, to be able to exclude these possibilities. If the pre-filter works for all modes, the outcome would support that all *UAV* versions have quality equal to that of the unfiltered version *High*, and the *UAV* versions have better quality than the unfiltered version *Low*.

The filtering parameters are based on the viewing modes. The three viewing modes (In-Hand, On-Lap and On-Stand) result in three filtered versions, which are compressed at different bit-rates. Together with the *High* and *Low* bit-rates, each video sequence is compressed at five bit-rates using the following steps:

1. Compress the unfiltered sequence with a high bit-rate such that there is no visual artifact. The full encoding capability of H.264 high profile and 1-pass rate control are used to encode the sequence. The output bitstream is the *High* version.
2. For each viewing mode, compress the filtered sequence with multiple bit-rates. The one that has the lowest bit-rate and is visually very close to *High* under the given viewing conditions is selected. The output bitstreams are the *UAV*: In-Hand, On-Lap, and On-Stand versions.
3. Encode the original unfiltered sequence at a bit-rate which is close to but slightly higher than the rate of On-Stand. It gives the bitstream *Low* version. The encoder settings except for the bit-rate are the same as the settings in step 1 for all versions.

The five bit-rates are selected manually for each sequence by experts. The relationship of the bit-rates of the five test versions are $High > Hand > Lap > Stand \approx Low$. The rates of each version of each test sequence are in Table 1.

Table 1. Bit-rate of each test sequence. All sequences are at 25fps with the exception of Kimono which is at 24fps. The bit-rate of *High* is in kb/s, while others are represented as the percentage compared to *High*.

Sequence	Bit-rate				
	High	UAV			Low
		Hand	Lap	Stand	
Basketball	4008	98.7%	76.6%	65.5%	66.2%
Into trees	8414	99.0%	72.8%	62.4%	62.5%
Old town	3420	97.1%	67.7%	54.8%	60.1%
Sunflower	2290	88.0%	66.0%	45.1%	45.8%
Pedestrian	13058	99.7%	80.0%	56.5%	58.1%
Station	3494	99.0%	76.2%	64.1%	66.6%
Tractor	6512	97.8%	67.3%	54.7%	55.9%
Rush hour	6689	97.9%	66.3%	52.2%	55.6%
Kimono	4980	99.2%	63.2%	61.9%	65.7%
Average	-	97.4%	70.7%	57.5%	59.6%

3.2. Comparison method

We used the pair comparison (stimulus-comparison) method [7, 8] to compare video quality. The subject was presented with a series of sequence pairs, each from the same source, but the rate and/or the compression (with or without filtering) are different. Videos were presented sequentially on the same device. The subject provides a score of the second sequence (test) relative to the first one (reference) of -1 = worse, 0 = same, 1 = better. We did not follow the 7-point grading in [7] as the differences were very subtle. For each mode, the three versions (*UAV, High, Low*) were shown as reference/test in pseudo-random fashion. The comparisons of each viewing mode included, in randomized order, *UAV* vs. *High*, *UAV* vs. *Low*, *High* vs. *Low*, and *High* vs. *High*. The first two comparisons are the main purpose of our test. *High* vs. *Low* provides a sanity check of the results. *High* vs. *High* is a null test to check for subject accuracy.

We used the pair comparison method because our experiment deals with very small differences in quality. The pair comparison method is more sensitive than the double stimulus continuous quality scale (DSCQS) method used in [2]. DSCQS requires subjects to mark both videos, then DMOS is calculated to do the comparison. Pair comparison, however, asks subjects to mark the difference between two videos directly. It is known to work better for very subtle differences. Since the rating includes the option of "the same," it requires fewer subjects than forced choice when the purpose is to show that two videos are subjectively the same. The rating scale does not bias subjects as does degradation category rating [8], which assumes that the test video has lower quality than the reference.

In our experiment, each video clip was 10 seconds long. Long sequences can produce a "forgiveness" effect, in which users forget and forgive quality lapses which occurred early on. One second of gray screen was shown between the videos in each paired comparison. Our videos all have spatial resolution of 1920×1080. The video clips used had a range of content: high motion and low motion, as well as content which is spatially simple and spatially complex.

3.3. Subjective test

The test was held in a room with typical office lighting conditions. We included 10 test sequences. There are 3 viewing modes and 3 pairs to be compared in each mode. Therefore, we had 90 pairs to be shown in total, excluding null tests. Each pair was compared by 15 observers. Thirty subjects (20 male, 10 female, average age 25.2 years) participated in the test. Each subject compared 45 pairs of test videos and 6 null tests. After the experiment, a playback problem was found with one sequence (the playback of the *High* version was jerky, leading it to be liked less than *Low*) so this sequence (not included in Table 1) was excluded from our data analysis. An experimental session was divided into six parts, where the modes were In-Hand, On-Lap, On-Stand, In-Hand, On-Lap and On-Stand. In each of the first 3 parts, subjects compared 8 pairs, and in the last 3 parts, they compared 9 pairs. There was one null test randomly placed in each part. After the 2nd and 4th parts, subjects were asked to take a break.

Written user instructions were provided at the beginning to each subject. The instructions described the three viewing modes, the experiment procedure, the grading scale and the interface. The three viewing modes were demonstrated by the experimenter. The subject then did a practice run (using unrelated sequences) to become familiar with the experiment procedure. The whole experiment took about 40 minutes.

4. RESULTS AND DISCUSSION

From the scores provided by the subjects, we use a one-sided test because in each case if the difference is not zero, there is a clear direction in which we would expect the difference to lie. The null hypothesis is that the mean score μ is equal to 0, i.e., the compared pair has the same subjective quality. For different comparisons, our alternative hypotheses are selected as: (1) *UAV-High*: $\mu < 0$, (2) *UAV-Low*: $\mu > 0$, (3) *High-Low*: $\mu > 0$.

The ideal result for this experiment would be that: for *UAV-High*, we cannot reject the null hypothesis that the tested pair has the same subjective quality; and for *UAV-Low* and *High-Low*, we can reject the null hypothesis. We use a one-sided test because it would be significant for us if *UAV* has lower quality than *High*, and if *Low* has lower quality than *UAV* and *High*.

Table 2. Results of t-test for data from all the subjects

Mode	*UAV-High*	*UAV-Low*	*High-Low*
Hand	fail to reject	reject	reject
Lap	$p = 0.06$	fail to reject	reject
Stand	reject	fail to reject	fail to reject

The results of t-tests for each comparison in each viewing mode are in Table 2. The table has "fail to reject" when $p > 0.1$ and "reject" when $p < 0.01$. We give the p-value in Table 2 if $0.01 < p < 0.1$. We also plot the means and 95% confidence intervals (CIs) in Fig. 3.

Table 2 shows that all comparisons of *UAV, High, Low* in In-Hand mode correspond to the ideal result. The null hypothesis of (*UAV-High*) cannot be rejected, and the null hypothesis of (*UAV-Low*) and (*High-Low*) can be rejected.

On-Lap mode: Table 2 shows that the null hypothesis of both (*UAV-High*) and (*UAV-Low*) cannot be rejected (though the p-value of *UAV-High* is marginal), which may indicate that no difference was observed among the three. However, when *High* was compared with *Low*, subjects seemed to notice the difference between them as the null hypothesis is rejected. So there is an inconsistency here.

On-Stand mode: the null hypothesis of (*UAV-High*) can be rejected, whereas the null hypothesis of (*UAV-Low*) and (*High-Low*) cannot. Again there is an inconsistency.

When we check the CIs of the null tests, we find that the CI of the null test in In-Hand mode unexpectedly does not include 0. There are relatively fewer of the null tests than there are of the other comparisons. Some subjects reported anecdotally after the experiment that a large number of sequences were very similar, and that it was hard to find differences. This difficulty is to be expected, since the test was designed to see whether video versions which were designed to be visually equivalent were in fact visually equivalent. It may be that the paucity of clear differences led viewers to sometimes find differences when there were none.

Given these observation, we examine subject reliability in more detail.

4.1. Analysis of null tests

The histogram of the number of subjects who reported a difference when none existed is shown in Fig. 4. It shows, for example, that only six subjects out of 30 did not report any difference on any of their null tests. Ten out of 30 subjects reported differences on two or more null tests, and six out of 30 subjects reported differences on three or more null tests. Their data may be less reliable.

To check for fatigue, we looked at whether subjects are more likely to report difference in the null tests as they watch more videos. Table 3 shows the fraction of subjects who reported no difference in the jth null test. As mentioned before, the first and fourth parts are In-Hand, the second and fifth

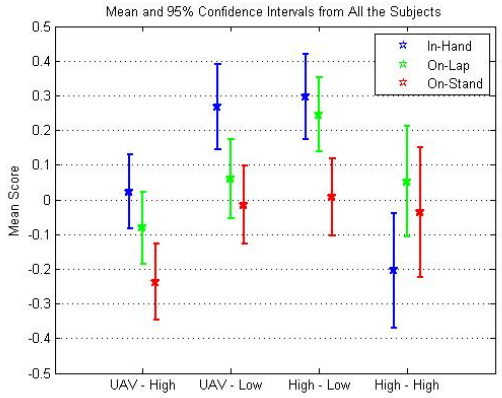

Fig. 3. Mean scores and CIs from all the subjects

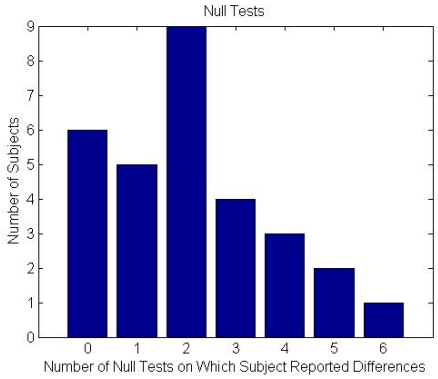

Fig. 4. Histogram of numbers of subjects who reported difference on null tests

parts are On-Lap, and the third and sixth are On-Stand. After the second and fourth parts, the subjects were notified to take a break. Table 3 shows that the subjects are slightly more likely to give accurate scores at the beginning of the experiment and after breaks. For example, 77.3% of the subjects reported no difference in the first null test, while only 51.9% reported no difference in the fourth null test (the second In-Hand part). In the On-Lap parts, more subjects reported no difference in the fifth part which followed a break, than in the second part. On-Stand is similar, with slightly higher correctness in the third part than in the sixth part.

Table 3. Fraction of subjects who did not report difference in each null test.

Mode	First Null Test		Second Null Test	
	Part No.	Correct%	Part No.	Correct%
Hand	1	77.3%	4	51.9%
Lap	2	62.1%	5	65.5%
Stand	3	58.6%	6	50.0%

Table 4. p-values of t-test for data from reliable parts and subjects

Mode	*UAV-High*	*UAV-Low*	*High-Low*
Hand	fail to reject	reject	reject
Lap	fail to reject	fail to reject	reject
Stand	$p = 0.03$	fail to reject	fail to reject

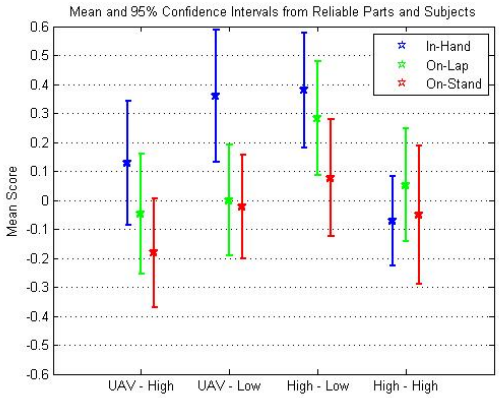

Fig. 5. Mean scores and CIs from reliable parts and subjects

4.2. Results from reliable subjects and reliable parts

As the null tests show that some subjects are more reliable than others, and some parts may have more of a fatigue effect, we re-analyze the data from reliable subjects (reported difference in at most two null tests) and from the more reliable parts of the experiment (first part of the experiment for In-Hand mode, fifth part for On-Lap, third part for On-Stand). The fraction of subjects who reported no difference in null tests in those 3 parts is 95%, 90% and 75%.

Table 4 shows the results of t-test of the reliable data. We plot the means and 95% CIs in Fig. 5. The results change slightly from the previous results which used all data.

In-Hand mode: as before, the null hypothesis of (*UAV-High*) cannot be rejected, and the null hypothesis of (*UAV-Low*) and (*High-Low*) can be rejected with strong evidence, corresponding to the ideal result.

On-Lap mode: the data shows that we cannot reject the null hypothesis of both (*UAV-High*) and (*UAV-Low*), but we can reject the null hypothesis of (*High-Low*). The p-value of *UAV-High* is no longer marginal. So there is more of an inconsistency than before.

On-Stand mode: the null hypothesis of (*UAV-Low*) and (*High-Low*) cannot be rejected, while the null hypothesis of (*UAV-High*) is on the margin. If we take 0.01 as the significance level, the null hypotheses of the three comparisons cannot be rejected, which means we cannot exclude that the three versions have the same subjective quality. If we take 0.05 as the significance level, the result shows inconsistency.

4.3. Discussion

The subjective visual quality of a high rate encoding of original content was compared with an encoding at a lower rate and with encoding content pre-filtered for the anticipated viewing conditions.

For In-Hand mode corresponding to the shortest viewing distance (most demanding viewing conditions), the visual quality of the *Low* version was worse than both the *High* version and the *UAV* version. For this mode, the 3% bit-rate savings of *UAV* did not degrade perceptual quality, but the attempt to realize 40% rate savings with *Low* results in visibly reduced quality.

For the intermediate case of On-Lap, the results are inconclusive but suggest that the pre-filter may be able to save on average 29% of the bit rate without degrading perceptual quality. The *Low* version is also not equivalent to the *High* version for this mode.

At the longest viewing distance (least demanding viewing conditions) of On-Stand, the results are inconsistent when using all data. When using the data from reliable subjects and parts, the data suggest that all three versions (*High*, *Low*, and *UAV*) may be perceptually equivalent. It would be important to ascertain whether the distance people use for the On-Stand mode is actually the distance for which the filtering was intended.

The videos in the experiment had subtle differences. Some subjects reported that the test was frustrating because so many videos looked equal. Many subjects could not reliably identify identical videos as being identical (nonzero scores in the null tests). We suspect that this fact and the previous one are related, in that some subjects did poorly in the null tests because the experiment overall aimed at barely visible differences, and so the subjects were scrutinizing for any possible difference.

5. CONCLUSION

Bit-rate reduction can be implemented by merely lowering encoding rate based on viewing conditions at the expense of increasing compression artifacts. Alternatively user-adaptivity may be implemented more gracefully by using a pre-filter in combination with the reduction of coded bit-rate. The benefits of adapting to the viewing conditions are expected to be enjoyed by a range of video encoding technologies. We presented a subjective test which confirms the ability to reduce encoded bit-rate without impacting the visual quality by adapting the representation and encoded bit-rate to the variable viewing conditions. We tested three viewing modes which correspond to three viewing distances. A very substantial bit rate savings can be realized if the tablet device can determine its viewing conditions and the content delivered to the device is adapted to these conditions. Average rate savings of 3% in critical In-Hand viewing and 30% approximately in an intermediate On-Lap usage modes without degradations in subjective quality were supported. Specifically, for the In-Hand and On-Lap versions, the video with pre-filtering is statistically equivalent to the video without pre-filtering *High*, but the pre-filtered video has lower bit-rate. Since the bit-rates were selected manually, it is possible that the actual bit-rate savings could be larger than what we showed in the paper. The particular tests used H.264 as the video encoder but this method of reducing video bit-rate based on adapting to viewing conditions is independent of the codec technology.

6. REFERENCES

[1] L.S. Karlsson and M. Sjostrom, "Improved ROI video coding using variable gaussian pre-filters and variance in intensity," in *Image Processing, IEEE Intl. Conference on*, Sept 2005, vol. 2, pp. II–313–16.

[2] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Video coding based on audio-visual attention," in *Multimedia and Expo, IEEE Intl. Conference on*, June 2009, pp. 57–60.

[3] N. Tsapatsoulis, K. Rapantzikos, and C.S. Pattichis, "An embedded saliency map estimator scheme: Application to video encoding," *Int. J. Neural Syst.*, vol. 17, no. 4, pp. 289–304, 2007.

[4] J.G. Young, M. Trudeau, D. Odell, K. Marinelli, and J.T. Dennerlein, "Touch-screen tablet user configurations and case-supported tilt affect head and neck flexion angles," *Work: A Journal of Prevention, Assessment and Rehabilitation*, vol. 41, no. 1, pp. 81–91, 2012.

[5] J. Xue and C.W. Chen, "Mobile video perception: New insights and adaptation strategies," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 3, pp. 390–401, June 2014.

[6] R. Vanam and Y.A. Reznik, "Perceptual pre-processing filter for user-adaptive coding and delivery of visual information," in *Picture Coding Symposium (PCS), 2013*, Dec 2013, pp. 426–429.

[7] Recommendation ITU-R BT.500-13, *Methodology for the subjective assessment of the quality of television pictures*, 2012.

[8] Recommendation ITU-T P.910, *Subjective video quality assessment methods for multimedia applications*, 2008.

[9] Y. Reznik, E. Asbun, Z. Chen, Y. Ye, E. Zeira, R. Vanam, Z. Yuan, G. Sternberg, A. Zeira, and N. Soni, "User-adaptive mobile video streaming," in *Visual Communications and Image Processing, 2012 IEEE*, Nov 2012.

[10] Y.A. Reznik, "User-adaptive mobile video streaming using MPEG-DASH," in *SPIE Optical Engineering+ Applications*. International Society for Optics and Photonics, 2013, pp. 88560J–88560J.

[11] "HD sequences," https://media.xiph.org/video/derf/.

[12] "x264," http://www.videolan.org/developers/x264.html.