See erratum at end

Factors Affecting Enhanced Video Quality Preferences

Prem Nandhini Satgunam, Russell L. Woods, P. Matthew Bronstad, and Eli Peli

***Abstract*— The development of video quality metrics requires methods for measuring perceived video quality. Most of these metrics are designed and tested using databases of images degraded by compression and scored using opinion ratings. We studied video quality preferences for enhanced images of normally-sighted participants using the method of paired comparisons with a thorough statistical analysis. Participants (n = 40) made pair-wise comparisons of high definition video clips enhanced at four different levels using a commercially available enhancement device. Perceptual scales were computed with binary logistic regression to estimate preferences for each level and to provide statistical inference of the differences among levels and the impact of other variables. While moderate preference for enhanced videos was found, two unexpected effects were also uncovered: 1) participants could be broadly classified into two groups: a) those who preferred enhancement ("Sharp") and b) those who disliked enhancement ("Smooth") and 2) enhancement preferences depended on video content, particularly for human faces to be enhanced less. The results suggest that algorithms to evaluate image quality (at least for enhancement) may need to be adjusted or applied differentially based on video content and viewer preferences. The possible impact of similar effects on image quality of compressed video needs to be evaluated.**

***Index Terms*— Television watching, image enhancement, image quality.**

I. Introduction

MEASURING image quality preferences (*i.e.*, perceived image quality), and proper statistical analyses of such preferences, are essential for developing devices and techniques for image acquisition, image processing, image display and for setting broadcast, storage, and display standards. Computational image quality metrics that attempt to predict the perceived image quality perception are desirable, because direct measurement of perceived image quality tends to be laborious [1], [2] and expensive [3]. Besides using such computational image quality metrics to develop new display systems [4]–[6], such metrics can also be integrated within a system to adjust image appearance in real time (*e.g.*, by controlling compression, bandwidth or adjusting the amount of enhancement applied).

Most widely used image quality metrics and their associated image databases along with human-subject preference measures are centered on images degraded by compression (though earlier metrics addressed low pass filtering and random noise as the degradations of interest). One inherent assumption of current computational image quality metrics is that all human observers respond in a similar manner, and that differences between responses to the same stimulus reflect measurement noise, both within- and between-subjects. In this paper, we present data that shows substantial between-subject differences in the preference responses to a video enhancement. This suggests that there may not be a (single) standard observer, at least for image enhancement.

A second assumption implied in current computations of human-subject image-quality responses is that responses are independent of image or video content. While some effects of image content have been acknowledged within the image-quality-assessment community, we are not aware of any published reports specifically addressing this issue. Small differences in favored video enhancement between content categories [7], and reductions in the impact of image compression with content desirability [8]–[10] have been noted. In our study, we found significant content-dependent effects. If this is the case in general, image enhancement and image quality algorithms may need to accommodate such content differences.

Progress in computational image quality metrics requires a solid understanding and accurate measurement of perceived image quality [11]. Sensory scaling measures can be used to determine the preferences of human observers [12]–[17]. Use of such grading or rating scales is included in the ITU-R BT.500-13 standard recommendation for assessment of perceived television picture quality [18] and have been applied in measuring (mean) opinion scores for various image quality databases. Such ratings can be analyzed to create estimates similar to Thurstone scales [19]–[21] using custom software [22], [23]. Pair-wise comparisons is an alternative measure that generates reliable and informative data about perceived image quality [24]–[27]. Pair-wise comparisons are widely used in applied psychology, marketing, food tasting, and advertising research [28]–[31]. Recently, two statistical methods have been described that produce an outcome very similar to Thurstone scaling while also providing inferential statistics [32], [33]. We [34] compared these two approaches, binary logistic regression [32] and linear regression [33], and found very similar outcomes. The Bradley-Terry-Luce model [35], [36],

Manuscript received December 29, 2012; revised July 15, 2013; accepted August 29, 2013. Date of publication September 20, 2013; date of current version October 4, 2013. This work was supported in part by NIH under Grants EY05957, EY16093, and EY019100, and in part by a grant from Analog Devices Inc. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jesus Malo.

P. N. Satgunam is with Hyderabad Eye Research Foundation, L. V. Prasad Eye Institute, Hyderabad 500034, India (e-mail: premnandhini@lvpei.org).

R. L. Woods, P. M. Bronstad, and E. Peli are with the Harvard Medical School, Schepens Eye Research Institute, Boston, MA 02114 USA (e-mail: russell.woods@schepens.harvard.edu; matthew_bronstad@meei.harvard.edu; eli.peli@schepens.harvard.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/TIP.2013.2282120

an alternative to Thurstone scaling, also has been used for paired comparisons [26], [27], [37]. We prefer the binary logistic regression approach [32], as the statistical significance can be obtained with commonly-used statistical software without the need for additional calculations or custom software. In addition, since logistic regression is a widely used statistical method it is well understood and it facilitates the development and testing of models that account for experimental variables and uncontrolled or confounding variables (as illustrated in the Appendix), a capability that is not currently available with the Bradley-Terry-Luce analyses [26], [27], [37]. In our study, we analyzed the pair-wise comparisons of four video enhancement levels using logistic regression models that included potential confounding factors as covariates. The consequences of these findings for video processing (enhancement and compression) and for computational video quality metrics are discussed.

The original purpose of this project was to measure the effect of image enhancement on perceived video quality. However, we found the results, regarding between-observer variability and image content effects to be of general importance and, therefore, we present these as the main emphasis of this paper.

II. Methods

A. Procedure

Participants viewed two versions of a 30s video clip on two side-by-side HDTVs each connected to a commercially-available video enhancement device that was set to one of four ordinal enhancement levels (Off, Low, Medium or High). Participants indicated their preference for one side over the other (left or right display; two alternative forced choice) using a computer mouse. Participants watched each video clip for the entire duration or stopped the clip as soon as they decided which one they preferred. If a participant had no preference s/he was asked to select left or right at random. Participants practiced the preference task until they understood the procedure. Every subsequent trial proceeded automatically after the participant indicated his or her preference on the previous trial. All 40 participants made 64 comparisons of 64 video clips, which took about an hour (more information available in section 2.5).

Participants were told that the two video clips on the two displays may or may not look different. We did not use the word "enhancement" in our instructions to avoid biasing the participants, as this word has a positive connotation. The participants were unaware that video quality was being assessed; they were told that they were control participants in a study of a rehabilitation device for vision impairment. This created an application-independent environment that is considered desirable for image quality evaluation [38].

During pilot testing, debriefing of participants made us realize that there may be substantial differences in individual preferences. Some participants claimed to like a more "natural" or "smoother" appearance to video (*i.e.*, enhancement Low or Off). We named this group "Smooth". Other participants preferred "brighter" or "sharper" (typically, more enhanced) video. We named this group "Sharp". Individual differences in preferences are potentially important, so combining data from all participants may mask such variability in enhancement preferences, possibly misrepresenting some or all participants. To account for such individual differences in enhancement preferences, upon completion of the 64 video clip comparisons, participants in the main study were debriefed by asking them to describe (in their own words) how they made their preference decisions. Using those descriptions, we classified participants into the two groups identified in the pilot testing. In addition, contrary to our expectations, many participants made comments during the debriefing that indicated that their preferences were affected by video content, so a post-hoc analysis was performed, as described in section 2.4.

B. Hardware Components

Two 42" HDTVs (VIZIO VO42L FHDTV10A, 16:9 aspect ratio) that were manufactured in the same month, had consecutive serial numbers, and were essentially identical in all important respects were used for pair-wise comparisons. There were no appreciable differences in the measured luminance and color properties of these two displays. A video clip was shown on these two HDTVs simultaneously. The source video was duplicated using a HDMI splitter (HSP12 HDMI Splitter-1-in 2-out, ConnectGear, Inc., Fremont, CA) and the video clips were processed independently by two video enhancement devices (PureAV RazorVision, Belkin International, Inc., Los Angeles, CA) that were each connected to one of the HDTVs (Figure 1). Measured luminance variability for the grayscale range 16–235 was within 5% between the two HDTVs with all hardware connected (*i.e.*, PureAV Razor Vision device and HDMI splitter). The Rec.709 specification for digital image data [39] recommends using the range 16 to 235 for HDTVs, rather than 0–255. This places "pure black" at code 16 and "pure white" at code 235. Interface codes 0 and 255 are prohibited from video data. 8-bit codes between 1 and 15 provide "footroom" and 236–254 provide "headroom" to accommodate filter undershoots and overshoots. Although the HDTVs were connected to a computer, which would normally use the full 0-255 range, when video clips were displayed by the computer, the appropriate transformations were made to conform to Rec 709. The video signal came from a Windows computer (EVGA Nforce680i Sli motherboard, Intel Core 2 Quad, 2.5GHz, video card: EVGA NVIDIA GeForce 9800 GX2).

To allow examination of potential display differences, the two HDTVs were physically swapped after testing the first 20 of the 40 participants. Also, to minimize the influence of side biases on the final outcome measure, the presentation side and the level of enhancement were counterbalanced for each participant (*i.e.* each participant saw each of the 16 possible permutations of enhancement level pairs crossed with presentation side an equal number of times).

C. Viewing Distance

Viewing distance recommendations for a 42" monitor from various sources that we located range from 5.3' to 10.5' and may not have any scientific basis. The scan-line pitch subtends an angle of about one minute of arc for these 42" HDTVs

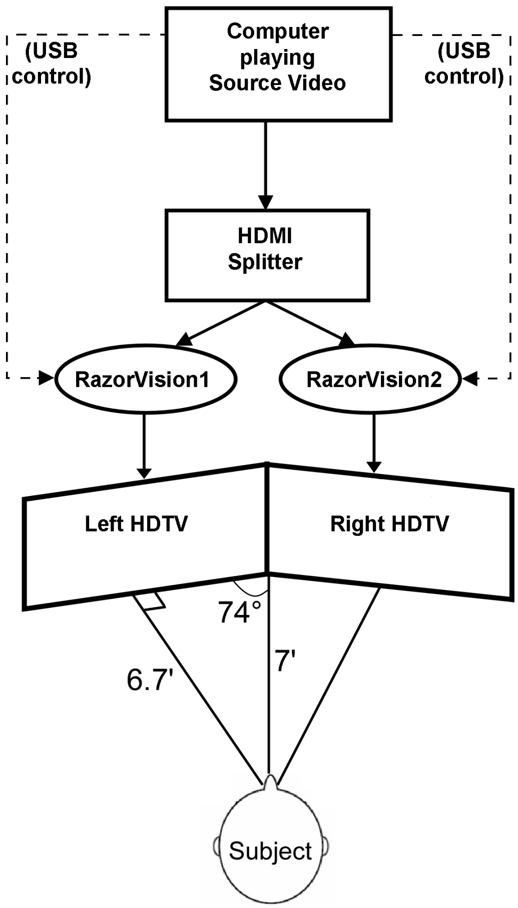

Fig. 1. Hardware components and their connections used in the experimental setup.

from about 5.3'. In a small pilot study, it was determined that a viewing distance of about 7' was more comfortable and this distance was used for the main study. The two HDTVs were angled towards each other (148°) so that the center of each display was perpendicular to the participant (Figure 1).

D. Stimuli

High definition 1080p (scaled to 1920 × 1080) movie trailers and documentary videos were downloaded (dates: October 2008 to February 2009) from http://www.apple.com/trailers/ and http://www.apple.com/quicktime/guide/hd/. The calculated video compression was estimated to be about 100 to 1. The downloaded videos were edited using QuickTime 7 Pro (Apple Inc., Cupertino, CA) into 30-s video clips. Seventy six video clips were selected for use in the main study. Additional video clips were used for the practice trials.

Fullerton *et al.* [7] found only small differences in desired enhancement level between some of their four video categories (low motion, high motion, cartoon and dark). Therefore, we did not select the 76 video clips based on content. The movie trailers tended to have substantial human face content and the documentaries mainly had nature scenes. Despite our expectation that video content would not affect the responses to video enhancement, during the debriefing some of our participants reported making preference decisions based on content, especially when the video clips had human faces. Therefore, we conducted a *post-hoc* rating of video content for four categories: presence of human faces, human figures, nature, and man-made objects ("human figures" indicated that a person was present, but his or her face was not visible or not important). Four naïve participants who did not participate in the main study of video preferences were asked to rate the presence based on the importance of each category, as though they would have to describe the video clip to another person who had not seen it. Each rating scale ranged from 0-5, with 0 being absent and 5 being always present. The responses of the four participants were averaged for each rating category and for each video clip. To include video content in the binary logistic regression analysis, video clips were considered to have high content (e.g. presence of faces) if the average rating was 3 or greater, and to have low content if the average rating was 2 or less. The number of video clips that were found to have high and low ratings for each video-content category is shown in Table 1.

TABLE I

THE NUMBER OF VIDEO CLIPS (OUT OF 76) THAT WERE RATED AS HAVING HIGH AND LOW LEVELS OF THAT CATEGORY OF VIDEO CONTENT. THE NUMBERS OF VIDEO CLIPS FOR EACH VIDEO CATEGORY DO NOT SUM TO 76 BECAUSE SOME VIDEOS HAD AN AVERAGE RATING BETWEEN 2 AND 3

Video-Content Category	High Content (rating ≥3)	Low Content (rating ≤2)
Faces	36	27
Nature	23	49
Human Figures	25	36
Man-made Objects	18	49

The remaining video clips (e.g. the 13 in the Face video content category) were not included in the video content analysis on preference (but were used in all other analyses). The Face scale was negatively related to the Nature (Fisher exact test, $p < 0.001$) and Man-made Objects ($p = 0.004$) scales. For example, few video clips had both high Face and high Nature content ($n = 1$) or both low Face and low Nature content ($n = 7$). The Human-figure and Man-made Objects categories were slightly positively related ($p = 0.06$) with 29 video clips having a low rating on both, 9 clips having high rating on both, and 14 clips that were rated differently on the two scales. There were no significant relationships between other video-content category pairs. Because we had not planned an analysis of video content, video content was not balanced between participants and with respect to enhancement comparison pairs (e.g. low-content with Low versus High). There were some participants who did not experience particular enhancement level combinations (e.g. low Man-made Objects with High versus Medium). However, mixed-effects regression analyses are robust to such "missing data", particularly since, across all participants, there were many comparisons of each enhancement comparison pair for each

Fig. 2. Illustration of the Original image (upper left) and the three enhancement levels: Low (upper right), Medium (lower left) and High (lower right).

video content category (the least was 67 for comparisons for high-Man-made, Original versus Medium).

E. Image Enhancement

Video clips were enhanced in real time using the commercially available PureAV RazorVision device (Belkin International, Inc., Los Angeles, CA) that performed an adaptive local contrast enhancement [40], [41]. Briefly, the enhancement algorithm calculated the mean local luminance, which is then subtracted from the original image to produce a high-pass filtered version. The high-pass filtered version is amplified by an enhancement gain control that may be determined by multiple variables, including local mean luminance. To avoid image distortions ($e.g.$, saturation), higher enhancement gain is adaptively applied to pixels in areas with moderate local mean luminance more than to pixels that have extreme luminance (bright or dark). The size of the Gaussian kernel used for local averaging was 9.1% of the screen width and thus spatial frequencies above 0.5 cycles/degree were enhanced in our viewing distance. To illustrate the available enhancement levels, an example still image photographed from one of the video clips at the four enhancement levels is shown in Figure 2.

Each HDTV was connected to a separate PureAV RazorVision device to process the video clip independently (Figure 1). RazorVision can be set to one of four enhancement levels (Off, Low, Medium or High). All 16 (4 × 4) possible combinations of enhancement levels were compared (Table 2). Four of these 16 comparisons were between identical enhancement levels (shaded cells in Table 2), which were included to test for response (side) and display biases. Each of the 16 comparisons was made 4 times for a total of 64 comparisons per participant. A MATLAB program controlled the presentation order of the 64 comparisons, randomly selecting 64 video clips from the pool of 76 video clips. No video clip was seen twice by any participant. Control of the PureAV RazorVision device was automated to present the desired enhancement level using a Visual C++ program, triggered by the MATLAB program. We used VLC media player (0.9.6) (www.videolan.org) to play the video clips.

F. Data Analysis

Thurstone perceptual scales derived from pair-wise comparisons represent the relative preference for one stimulus (e.g. enhancement level) over each other [19], [20], [31], [34]. By convention, the lowest relative preference score is set to zero, and the remaining relative preference scores are scaled accordingly to preserve the relative distances among the rated items. Thus, this perceptual scale orders the items from least to most preferred. Binary logistic regression has been used to obtain a Thurstone-like perceptual scale along with the statistical significance for the differences between stimuli [32], [34]. Since logistic regression does not account for the correlation between responses ($i.e.$ assumes independence of data), we used crossed-random, mixed-effects logistic regression, as described in detail in the Appendix. In steps described in the Appendix, the final model was constructed:

$$a_{ij} = \left(\beta_1 + \beta_{g1} + \beta_{c1}\right) X_{ij1} + \left(\beta_2 + \beta_{g2} + \beta_{c2}\right) X_{ij2}, + \left(\beta_3 + \beta_{g3} + \beta_{c3}\right) X_{ij3} + \beta_S X_S + \beta_d X_d + \varphi_{\mathrm{i}} + \vartheta_{\mathrm{j}} + \varepsilon_{\mathrm{ij}} \quad (1)$$

where a_{ij} was the exponent of the logistic function (Eq. A3), β_k were coefficients for each stimulus, X_{ijk}, at enhancement level k, β_{gk} were coefficients of indicator variables for the reported-preference Group (*Sharp* or *Smooth*) at each enhancement level, β_{ck} were coefficients of indicator variables for the video-Content (low or high), β_s was a coefficient for the Side (left or right), β_d, was a coefficient for the Display (a or b), ϕ_i were coefficients for each participant, i, and θ_j were coefficients for each video, j.

To construct a Thurstone-like perceptual scale, coefficients for each of the enhancement levels obtained were normalized to have a range of 1 unit. Unlike the traditional approach of setting the least preferred level to zero, we anchored this perceptual scale by fixing the preference level for the original video clips (Off condition) to zero, while maintaining the range of 1. Thus, an enhancement level that was preferred less than the original video clip received a negative value. The corresponding significance levels (p-values) of the relative preferences were calculated for the Wald χ^2 statistic[1] for the coefficients, providing comparisons between enhancement levels.

G. Participants

Forty normally-sighted participants (ages: 20–83y, median age: 32y) consented to participate. The study was approved by the Institutional Review Board of the Schepens Eye Research Institute. Preliminary screening of the participants included self-report of ocular health, measures of their visual acuity and contrast sensitivity for a 2.5 degree-high letter target and evaluation of central retinal health using retinal photography (Nidek MP-1, Nidek Technologies, Vigonza, Italy). All the participants had visual acuity of 20/25 or better, letter contrast sensitivity of 1.675 log units or better and steady central fixation with no evidence of retinal defects.

[1]Caution needs to be applied when proportions approach 100% or 0% as it can result in the Hauck-Donner effect. [34], [42].

TABLE II

EXAMPLE PREFERENCE MATRICES FOR TWO PARTICIPANTS, REPORTED FOR THE 16 POSSIBLE ENHANCEMENT-LEVEL COMPARISONS, EACH MADE 4 TIMES. PARTICIPANT 1 (LEFT MATRIX) IS REPRESENTATIVE OF THE SHARP GROUP AND PARTICIPANT 2 (RIGHT MATRIX) IS REPRESENTATIVE OF THE SMOOTH GROUP. THE SHADED CELLS ON THE DIAGONAL REPRESENT COMPARISONS OF THE SAME ENHANCEMENT LEVELS APPLIED TO THE LEFT AND RIGHT HDTVS. EACH CELL IS READ AS LEFT HDTV PREFERRED OVER THE RIGHT HDTV. HENCE, FOR THE FIRST ROW, FOR PARTICIPANT 2, LEFT OFF WAS PREFERRED 3 TIMES (OUT OF 4) OVER RIGHT OFF, RIGHT LOW AND RIGHT MEDIUM, AND WAS PREFERRED 4 TIMES OVER OR RIGHT HIGH. FROM THE FIRST COLUMN FOR PARTICIPANT 2, LEFT LOW WAS PREFERRED ONCE OVER RIGHT OFF. THUS, FOR PARTICIPANT 2, IN TOTAL LOW WAS PREFERRED 4 TIMES, OUT OF 8, OVER OFF

Participant 1	Right Off	Right Low	Right Medium	Right High
Left Off	1	0	0	0
Left Low	4	1	0	0
Left Medium	4	4	1	1
Left High	4	4	2	0

Participant 2	Right Off	Right Low	Right Medium	Right High
Left Off	3	3	3	4
Left Low	1	0	0	3
Left Medium	1	1	2	1
Left High	0	4	2	3

III. RESULTS

A. Overall Preference Results

All 40 participants completed the 64 planned trials. Crossed-random, mixed-effects binary logistic regression for all participants' preferences combined (Equation A5,[2] $\chi^2 = 26.7$, df = 3, p < 0.001) was performed with all 2560 trials from the 40 participants. When including Display and Side (Equation A6,[3] $\chi^2 = 31.0$, df = 5, p < 0.001), no bias was found for Display (HDTVa or HDTVb; p = 0.46), that is, both Displays were equally preferred, but there was a preference for the right side (p = 0.04). The side bias was mainly associated with the condition in which the two stimuli had equal levels of enhancement (n = 640 trials, $p = 0.06$), and not when the stimuli had different enhancement levels (n = 1920 trials, $p = 0.18$). The same-stimulus comparisons were included to evaluate side bias and were not used to analyze the overall preference data. Our results support prior reports that participants (or at least some of them) were poor in equally dispersing their 'guess' responses in a two alternate forced choice task [43], [44]. Overall, Low and Medium were preferred over Off and High ($p \leq 0.024$), and Low and Medium ($p = 0.35$) and Off and High ($p = 0.14$) were not significantly different from one another (Equation A6). Figure 3 shows that, overall, there was a non-monotonic response (inverted U-shape) to the video enhancement. As discussed below, such non-monotonic behavior is not expected when compression level is increased. The inverted U-shape obtained here is comparable to studies involving quality judgment for brightness [45], color [46] and stereoscopic depth [47].

B. Two Types of Preferences

Participants described their preference criteria for video quality in their own words at a debriefing following the completion of the 64 trials. After reviewing these criteria 39

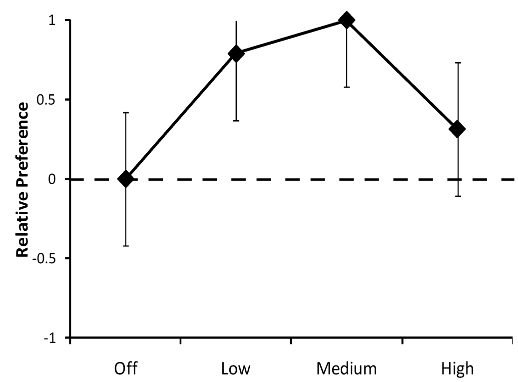

Fig. 3. Video-enhancement preferences, for all 40 participants, were non-monotonic. Error bars are 95% confidence intervals of the relative preferences derived from the standard errors of the logistic regression coefficients.

of the 40, participants could be assigned to two groups. One group of participants (n = 12) preferred more clarity and scrutinized the clarity of small details in the image ("Sharp" Group). The other group of participants (n = 27) preferred smoother human faces yet they too preferred more clarity for nature scenes ("Smooth" Group). One participant preferred enhancement for neither human faces nor for nature scenes. This participant was not assigned to either group but was retained for the overall data analysis (section 3.1). Binary logistic regression was performed with the participant group (Sharp or Smooth) included (Equation A7,[4] $\chi^2 = 83.3$, df = 8, $p < 0.001$) and provided an improvement ($\chi^2 = 61.3$, df = 3, $p < 0.001$) over the all-participants model (Equation A6).

Figure 4 shows that the Sharp group most preferred Medium enhancement, which was significantly preferred to Low ($p = 0.008$) and Off ($p < 0.001$) but was not significantly different from High ($p = 0.51$). Off was least preferred, significantly less than Low ($p < 0.001$). The Smooth group most preferred Low enhancement, but it was not significantly

[2]Equivalent to equation A8 with $\beta_{gi} = \beta_{ci} = \beta_s = \beta_d = 0$.

[3]Equivalent to equation A8 with $\beta_{gi} = \beta_{ci} = 0$.

[4]Equivalent to equation A8 with $\beta_{ci} = 0$

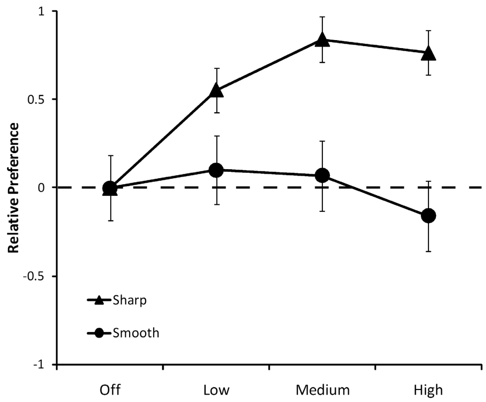

Fig. 4. Enhancement preferences of the Sharp (n = 12 triangles) and Smooth (n = 27, circles) groups. Error bars show 95% confidence intervals derived from the standard errors of the logistic regression coefficients.

different from Medium ($p = 0.64$) or Off ($p = 0.15$). High was significantly least preferred ($p \leq 0.025$) when compared with the other 3 levels (analysis applied for all images).

C. Preferences Depend on Video Content

Many of the Smooth participants reported preferring enhancement of nature scenes but not of faces. If so, relative preferences should differ based on the video content rating levels (high or low) for the face and nature categories. For example, more enhancement should be preferred for Non-Face (low face content) video clips than for Face (high face content) video clips. Conversely, more enhancement may be preferred for Nature (high nature content) than for Non-Nature (low Nature content) video clips. Post-hoc analyses for each video classification were performed (Equation A8): Face ($\chi^2 = 135.0$, df = 11, p < 0.001); Nature ($\chi^2 = 127.9$, df = 11, p < 0.001); Human-figure ($\chi^2 = 86.8$, df = 11, p < 0.001); and Man-made-Objects ($\chi^2 = 83.4$, df = 11, p < 0.001) categories. Compared to Equation A7, the inclusion of both Face and Nature video categories (Equation A8) substantially improved the model ($\chi^2 > 53$, df = 3, p < 0.001), while the inclusion of the Human-figure ($\chi^2 =$ 15.2, df = 3, p = 0.002) and Man-made ($\chi^2 = 10.1$, df = 3, p = 0.02) categories made smaller improvements in the fit.

Both the Smooth and Sharp groups had a lower preference for enhancement of Face than Non-Face video clips ($p < 0.001$; Figure 5a). Conversely, both groups had a higher preference for enhancement of Nature than Non-Nature video clips ($p \leq 0.001$; Figure 5b). Similarly, for both the Human-figure ($p \leq 0.02$; Figure 5c) and the Man-made-Objects (p ≤ 0.08; Figure 5d) categories, there was a higher preference for enhancement of video clips with a higher rating than for those with a low rating (“Non”). That all the other three video-content categories had a different enhancement preference pattern to the Face video content category supports the conclusion that human face content was a major factor in the enhancement preferences of participants in our study.

D. Multimodal Preference Distribution of Objective Scores

Since the two groups, Sharp and Smooth, were determined based on the responses to questions about preference decisions, we sought additional evidence of discrete preference patterns using an objective method. A scalar was calculated for each participant to represent his or her overall enhancement preferences. The three enhancement levels, Low, Medium and High, were arbitrarily assigned ranks of 1, 2 and 3, respectively (Off was normalized to zero on the relative preference scale). The normalized coefficients obtained for each participant were weighted by the ranks and then summed. Examples for the two participants from Table 2 are shown in Figure 6. The resulting weighted sum was the participant’s Enhancement Preference (EP) score. A higher EP score indicated a preference for higher enhancement levels (e.g. Participant 1 in Figure 6). The possible range for the EP scores was from −6 to +6, the obtained EP scores ranged from −5.29 to +6 (mean +1.23). Figure 7 shows the distribution of EP scores for all participants.

The Smooth group had a wider range of EP scores than the Sharp group (Figure 7), as expected since their preference greatly varied with image type. The Sharp group had significantly higher EP scores than the Smooth group ($t = 5.21$, $p < 0.001$). This is consistent with the logistic regression analysis (Figure 4).

The EP score distribution appeared to be tri-modal (Figure 7). Mixture modeling using Mixmod 2.1.1 [48] indicated a significantly improved fit ($\chi^2 = 16.75$, df = 3, $p < 0.001$) if the data were modeled with three Gaussian distributions rather than two Gaussian distributions [49], and two Gaussian distributions were significantly better than one ($p = 0.05$). No significant improvement was observed if more than three Gaussian distributions were used. The adjusted Bayesian information criterion (BIC) improved from two to three Gaussian distributions ($p < 0.001$), with no improvement noted for models using more than three Gaussian distributions ($p = 0.26$ for four Gaussian distributions). There was fair agreement between the subjectively-defined groups and the groups identified by the mixture model (Kendall’s *Tau-b* = 0.465, $p = 0.002$). It seems that the Smooth group included the participants identified as having EP scores in the low EP score and middle EP score distributions, while the Sharp group were all within the high-EP score distribution.

E. Effect of Incomplete Comparisons

We applied a complete testing design comparing all enhancement levels to each other. The preference analysis method used here also enables determination of the relative preference (*i.e.*, the coefficients) and statistical significance with an incomplete set of comparisons. Reducing the number of comparisons made can reduce the burden and cost of such studies and in some situations comparisons may be limited by other factors. For example, Fullerton and Peli [50] used a very similar device that implemented the same adaptive enhancement algorithm, but only one enhancement level could be displayed at a time. Therefore, only partial comparisons were conducted; Off (no enhancement) was compared to each of the other three enhancement levels but no comparisons were made amongst the three enhancement levels. To determine the effect of using partial comparisons like those used by

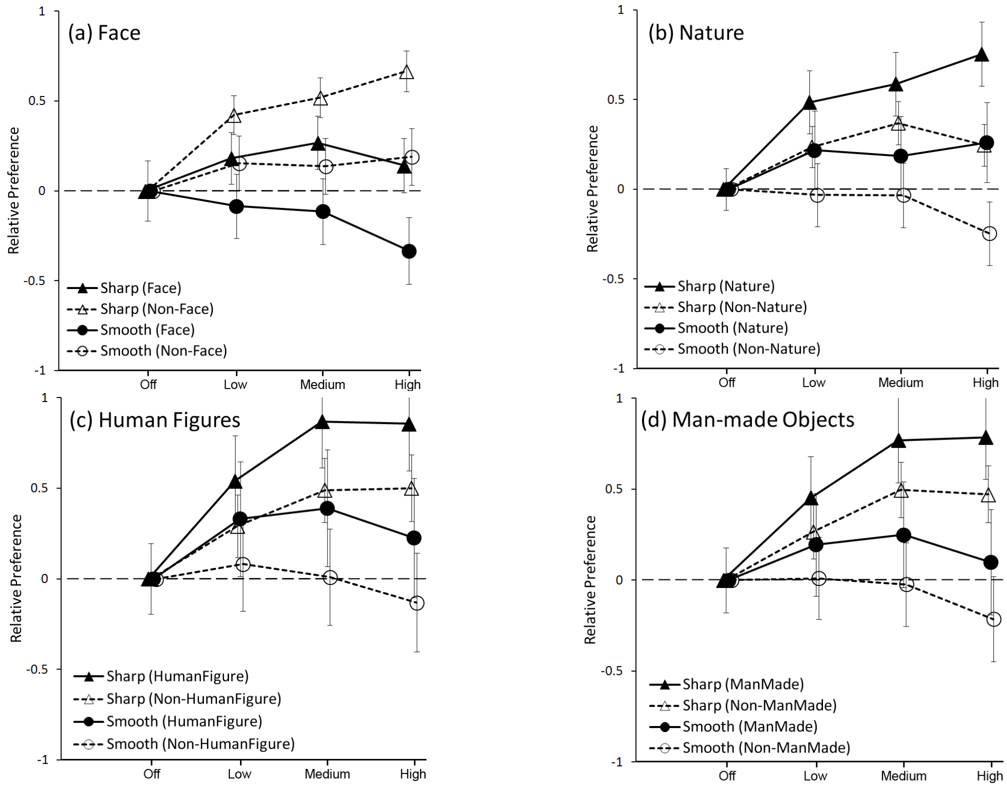

Fig. 5. Relative preferences by group (Sharp and Smooth) for high and low ratings of video-content categories (a) Faces, (b) Human figures, (c) Nature, and (d) Man-made Objects. The Face video-content category showed a response pattern that differed from the other three video-content categories, with enhancement being less preferred for high-rating (Face) video clips than for low rating (Non-Face) video clips (note the reversal of order of filled and open symbols from top to bottom). Error bars show 95% confidence intervals. To improve clarity, the group symbols are plotted with small offsets.

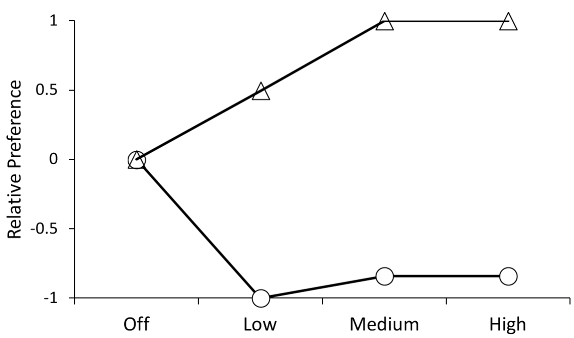

Fig. 6. Relative preferences of Participant 1 (triangles) and Participant 2 (circles) reported in Table 2. The calculated EP scores for Participant 1 (Sharp) was +5.49 and that of Participant 2 (Smooth) was −5.21.

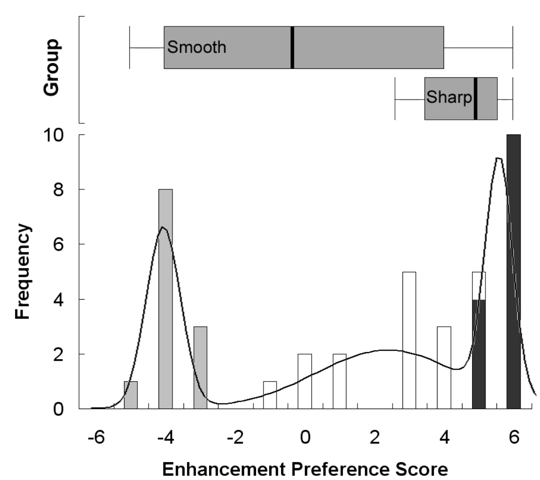

Fig. 7. The box plots for enhancement preference (EP) score for the two subjectively-defined preference groups shows that the Sharp group had higher EP scores. The median EP score is marked by the line within each box. The horizontal extent of each box represents the interquartile range (IQR) and the whiskers represent values within 1.5 times the IQR. The EP score distribution was tri-modal as shown in the frequency plot. The three shades correspond to the three groups identified by the multimodal fit. One column has members of two different groups due to overlap of the fitted distributions.

Fullerton and Peli [50], a subset of data from the present study, that only included the comparisons made between the Off level to the other three levels, was analyzed for all the participants. The logistic-regression results obtained from this partial comparison [34] were different from the results obtained when all comparisons were made (Figure 8) and were similar to the results obtained by Fullerton and Peli [50] with the same constrained comparisons, particularly at the High level of enhancement. Also, the results for this subset of data obtained with logistic regression were different from those obtained using Thurstone scaling, but were consistent with the raw proportion of responses [34].

F. Other Individual Differences

We found no evidence that image quality preferences are related to conventional vision measures; there was no significant correlation between EP scores and either visual acuity (Spearman $\rho_{39} = 0.15$, p $= 0.37$) or letter contrast

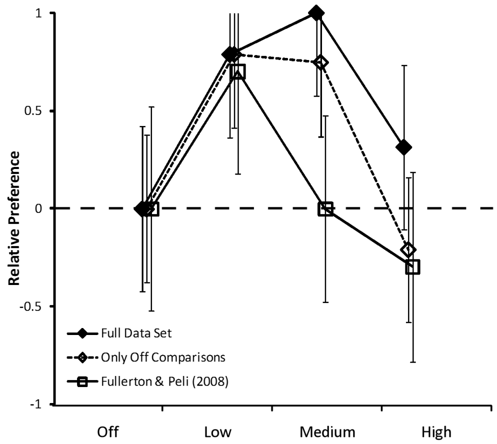

Fig. 8. Enhancement preferences shown for full data (same as Figure 3) and when analyzed using only a partial set of the available comparisons in the current study, those that included the Off level. The results of that comparison are more similar to an earlier study [50] in which the 3 enhancement levels were compared to Off. Error bars are 95% confidence intervals. To improve clarity, the group symbols are plotted with small offsets.

sensitivity (ρ_{39} = −0.13, p = 0.42). In a normally-sighted population the range of visual acuities and contrast sensitivities are not large, thus reducing the probability of finding a significant correlation. Even so, with increasing age, visual acuity (ρ_{39} = 0.37, p = 0.02) and letter contrast sensitivity (ρ_{39} = −0.45, p = 0.004) became worse, while EP score did not change (ρ_{39} = 0.07, p = 0.69). Participants in the Sharp group tended to be older than those in the Smooth group (z = 1.87, p = 0.06) and were more likely to be male (z = 1.88, p = 0.06), but age and gender were confounded. So, when corrected for age there was no gender difference (z = 0.21, p = 0.83), and when corrected for gender, there was no relationship between age and group (z = 0.80, p = 0.42). There were no differences between the groups in visual acuity (z = 0.40, p = 0.69) or letter contrast sensitivity (z = 0.41, p = 0.68). We thought that video enhancement preferences might relate to personality, however, a personality questionnaire related to tolerance of blur [51] did not reveal any relations with preferences in a subset of 25 participants to whom the questionnaire was administered.

IV. Discussion

Our study was designed to measure the subjective preference of normally-sighted observers to motion videos enhanced by a commercially-available device. That objective was achieved. In general, the two lower enhancement levels were preferred, but the highest available level was not liked (Figure 3). Such a non-monotonic response to increasing enhancement is common and expected with image enhancements. However, we also uncovered two important effects: one related to the observers, and the other related to the video content.

We found that observers could be divided into at least two groups; one that liked enhancement over all (Sharp) and one that did not (Smooth) (Figure 4). That grouping of participants based on a debriefing interview was confirmed by an objective analysis (EP Score), which suggested a possible third group with weaker or intermediate preferences (Figure 7). That third group seemed to have been included in the Smooth group by the debriefing classification. As it is possible that the Smooth-Sharp group differences were an artifact or were peculiar to our sample, we reanalyzed the data from two published studies [50], [52] that measured preferences from pair-wise comparisons. In the Fullerton and Peli study [50] that investigated a very similar video-enhancement device, the EP scores ranged from −6 to 6 for their 11 normally-sighted participants and from −3 to 6 for their 20 low-vision participants. The distribution of EP scores was slightly, but not significantly, better fit with two Gaussian distributions than with a single Gaussian (Mixmod 2.1.1: $\chi^2 = 6.57$, df = 3, $p = 0.17$). In the Satgunam *et al.* study [52] that investigated static-image enhancement with a different enhancement algorithm, the EP scores ranged from −5 to 4.3 for their 24 low-vision participants, and it was best fit with a bi-modal distribution ($\chi^2 = 12.12$, df = 3, $p = 0.014$, over uni-modal, and tri-modal). Thus, our finding that observers can vary widely in their preferences for enhancement was confirmed; the between-observer effect was found to occur with two image enhancement algorithms and in different populations. Whether there are two or three separate groups of preference types is not certain.

It has been suggested that differences between laboratories in video quality measures can be attributed to differences in the experience (with image quality) of the observers in those study samples [18]. All observers were naïve in our study and the other two studies [50], [52]. Our finding that enhanced video quality preferences were not uniform across observers (*i.e.* that between-observer variance reflects real differences in response) may have important consequences for the evaluation of computational image (video) quality metrics, particularly if there are similar between-observer differences in preference responses to image degradation (e.g. due to compression). Most image degradation studies differ from our image enhancement studies in two important ways: the observer's task (rating versus paired comparison) and the effect of the intervention on perceived image quality (monotonic versus non-monotonic). The non-monotonic preference that we found for the enhancement (Figures 3 to 6 and 8) is not expected for image compression or other degrading effects. Most such studies and metrics presume that perceived image quality data is normally distributed (e.g [12]–[17]), effectively that the between-participant variance is due to measurement noise. To evaluate this, we examined data from three studies of image compression in which the observers reported perceived image quality ("opinion score"). For two studies by one group [13]–[17] the distributions of relative (difference) opinion scores of the study participants (N = 13 to 29) were uni-modal Gaussian distributions, and for a third study [12], between-observer differences were apparent, but there was no obvious shape to the distribution in the responses of the 16 participants.

In a study where noise was added to natural images of tree bark, two participant preference patterns were noted [53]. One group found the noise-added image to be sharp while another found it blurred. The authors attributed this difference to the observers' specific attention to different image details, as noted in the observers' introspection. Using synthetic simple images they found that added noise sharpened low spatial frequency content and blurred high spatial frequency content. Two preference patterns were also found in another study [54] where one group of observers found rough textures to be pleasant while another group did not. The authors in that study however were unsure of the presence of two groups, and raised concern about their experimental paradigm.

Future video- and image-quality observer studies may have to consider that there may be at least two different groups of observers. To classify an individual (to Sharp or Smooth), representative sample videos, enhanced by at least two levels, could be shown to observers who indicate their preferences over the original videos. Such a method should easily identify their preference pattern. Alternately, the enhancement could be provided with two settings, one expected to be preferred by each group, and the user can determine their preference by switching between the two settings while watching a sample of videos. If such differences are found with compressed images, the computational metrics will have to be able to address such a dichotomy in the population.

The second effect that we found was that video content affected preferences for video enhancement. When human faces were an important aspect of the content, our participants, independent of their overall preference group (smooth or sharp), preferred less enhancement than when face content was not important (Figure 5). For the other three video-content categories, Nature, Human Figures and Man-made Objects, video enhancement preference patterns were the opposite of that for Faces (Figure 5). This could not be explained by the categorizations for Faces being the inverse of these three categories (it was for Nature, but not for the other two). The reason for this video-content effect is not clear. Discussions with our study participants suggest that the enhancement increased the visibility of facial features (e.g. skin blemishes) in a way that was not "natural" and thus not preferred. This is in agreement with an earlier study [55] that noted the appearance of human skin to be a critical component in the judgment of 'naturalness' of color reproduction. A similar preference pattern was reported in a study that examined enhancement of static images for mobile displays [56]. The authors in that study recommended using one preset image enhancement parameter for images human figures (defined by detection of human skin coloration). However, our results indicate different responses between faces and human figures (Figure 5) suggesting that a face detection algorithm would be more appropriate. To further examine whether such content effects were found in image degradation studies, we analyzed the data from such studies [13]–[15]. There were no images with face content in one image-degradation database [12] and the videos were not available for another [16], [17]. In the LIVE image database [13]–[15], participants gave higher ratings to images with face content independent of bitrate ($p < 0.001$) and images with man-made-object content received lower ratings when the bitrate was lower ($p < 0.001$). Thus, it seems that image content can affect subjective ratings of image quality when quality is degraded.

In a similar setting, studies of image compression, the "desirability" (how much the participant liked the content of the video) had an impact on video quality responses, with a more desirable (liked) video clip being given a higher rating [8]–[10]. Kortum and Sullivan [10] suggested that the desirability effect may be related to engagement with the content, in which case attention may be directed to regions such as faces. We did not measure the desirability of our video clips. We do not know how desirability can be determined computationally.

This video-content effect suggests that preferences could be predicted from current computational image quality metrics with some video contents, but not with others. For example, the visibility of minor facial skin blemishes may indicate superior image quality from computational measurements but may not be preferred by human observers, while the increased visibility of details of room furnishings by the same algorithm may be both superior for the computational metric and be preferred by observers. For assessment of video quality, it may be possible to include face detection within a computational metric, then using separate algorithms or parameters for video content with and without predominantly human faces. When using image enhancement, to overcome this preference difference, it may be necessary to apply less enhancement when faces are present or to regions that contain faces. Computational measurements to evaluate enhanced video quality should capture the non-monotonic preferences of human observers.

Most current computational image quality metrics are likely to find an enhanced image to be of lower quality than the original, even though our participants found the low and medium levels of enhancement to be preferred (Figure 3). Image quality metrics should be able to capture such effects as image enhancement could be used to reduce bandwidth or improve quality when bandwidth reduces image quality.

V. Conclusion

Video image quality can be effectively tested using side-by-side pair-wise comparisons and scored using logistic regression that enables examination of possibly confounding factors. While an incomplete test design that reduces the test burden may be used and analyzed, care should be taken in such a design, as incomplete comparison sets may adversely affect the results. To avoid the effect of bias it may be important to permit a "no difference" or equal quality response [43], [44]. When choosing videos for image quality studies and analyzing video quality, video content should be considered, especially the presence or absence of human faces. Videos used in observer study or computational study should match the anticipated images to be used in an application. Between-participant differences may reflect real individual differences in perceived image quality. If between-observer effects or content effects are found they need to be addressed in computational image quality metrics. While we studied preference for video enhancement, similar considerations and tools may need to be applied to image quality evaluation in the field of image compression. It is possible that not all our findings will

TABLE A1

ILLUSTRATION OF THE LOGISTIC REGRESSION MATRIX FOR ALL 64 TRIALS FOR ALL 40 PARTICIPANTS AND USED FOR THE BINARY LOGISTIC REGRESSION ANALYSES. STIMULI (ENHANCEMENT LEVELS) NOT PRESENTED DURING THE TRIAL ARE INDICATED BY 0. PRESENTED STIMULI WERE ALLOCATED 1 OR −1 DEPENDING ON THE REPORTED PREFERENCE AND THE IDENTITY VECTOR (E), AS DESCRIBED IN THE TEXT. THE PREFERRED STIMULUS IN EACH TRIAL IS SHOWN IN BOLD IN EACH ROW

Participant ID	Trial	Off	Low	Medium	High	Side	Display	Group	*e*
1	1	1	0	**-1**	0	1	-1	0	0
1	2	-1	0	**1**	0	0	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	64	0	0	**1**	-1	0	-1	0	1
2	1	0	1	0	**-1**	1	-1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
40	63	-1	**1**	0	0	1	-1	1	1
40	64	1	**-1**		0	1	1	1	0

be replicated in studies of perceived quality or preference for compressed imagery.

APPENDIX

For the binary logistic regression analysis, each paired comparison was entered in a logistic regression matrix (Table A1), where for each participant, i, for each video, j, of stimuli $k = r$ and $k = s$, the preferred stimulus (column), X_{ijr}, was allocated $+1$ and the non-preferred stimulus, X_{ijs}, was allocated -1:

$$\text{if } X_{ijr} \succ X_{ijs} \text{ then } X_{ijr} = +1 \text{ and } X_{ijs} = -1 \qquad \text{(A1)}$$

where $\succ$ is a left preferred indicator. All other row entries are allocated zeros. As described by Lipovetsky and Conklin [32], an identity vector (dependent variable column), e_i, was randomly assigned to a value of 0 or 1. When $e_i = 0$ ("false"), the signs of the responses for that comparison are reversed, such that:

$$\text{if } X_{ijr} \succ X_{ijs} \text{ then } X_{ijr} = -1 \text{ and } X_{ijs} = +1. \qquad \text{(A2)}$$

The outcome of the analysis is independent of the proportion of $e_{ij} = 0$ comparisons, so long as $0 < p(e_{ij} = 0) < 1$, where $p(e_{ij} = 0)$ is the proportion of comparisons with $e_{ij} = 0$. The binary logistic regression involved fitting the equation:

$$P_{ij} = f(a_{ij}) = \frac{e^{a_{ij}}}{e^{a_{ij}} + 1} = \frac{1}{1 + e^{-a_{ij}}} \qquad \text{(A3)}$$

where p_{ij} was the probability of the response (for each combination of participant and video, there was only one comparison of two enhancement levels, *i.e.* each participant only saw one video once). For the model, the responses to one stimulus (*e.g.*, Low enhancement) were fixed by not including it, and the model was forced through the origin (no intercept term). Thus, to compare four stimuli, the term a_{ij} in equation A3 was defined as:

$$a_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \varepsilon_{ij} \qquad \text{(A4)}$$

where β_k, were the coefficients for each enhancement level (stimulus), X_{ijk}, and ε_{ij} was the residual error. Those unknown parameters β_k were estimated by a maximum likelihood procedure, common to generalized linear models, with SPSS 11.5.0 (SPSS, Chicago, IL) and Stata/IC 12.1 for Mac (StataCorp, College Station, TX). The computed coefficients thus represented the relative preferences of the included stimuli with the excluded stimulus having a relative preference of zero. The statistical significance for stimulus k reported in each analysis was for the difference between the excluded stimulus and stimulus k. When n stimuli were compared, to obtain statistical significance for all $n(n-1)/2$ comparisons, the analysis was performed $n-1$ times, with a different stimulus kept constant (excluded) from each analysis. In our case of four stimuli the analysis was conducted three times.

Logistic regression presumes independence between data. However, our data includes repeated measures from each participant and for each video. No participant saw all videos, and the comparison pair of enhancements applied to each video varied between participants. This is known as a crossed-random experimental design. Therefore, we fit a crossed-random, mixed-effects logistic regression:

$$a_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \varphi_i + \vartheta_j + \varepsilon_{ij} \qquad \text{(A5)}$$

where φ_i were coefficients for each participant and θ_j were coefficients for each video. The side on which enhancement was presented could influence responses, as some participants may have an inherent response bias [43], [44] to choose the display on the left or right side and the two HDTVs, while perceived to be virtually identical, cannot be guaranteed to be completely identical in every respect and so could have influenced participants' responses. To measure these effects, *Side* and *Display* were included as covariate factors in the model. With the responses, X_{si}, for Side and, X_{di}, for Display, term a_{ij} became:

$$a_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_s X_s + \beta_d X_d + \varphi_i + \vartheta_j + \varepsilon_{ij}. \qquad \text{(A6)}$$

When the preferred stimulus was on the right monitor, $X_{si} = 1$, and when on the left, $X_{si} = 0$, except for $e_i = 0$ ("false") when the assigned value of X_{si} was reversed. Similarly, when the preferred stimulus was on HDTVa,

$X_{di} = 1$, and when on HDTVb, $X_{di} = 0$, except for $e_i = 0$ ("false") when the assigned value of X_{di} was reversed.

To examine between-group differences (between participants in the *Sharp* and *Smooth* groups), an indicator variable, g, was used, and the term a_{ij} became:

$$a_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_{g1} X_{ij1} + \beta_g X_{ij2} + \beta_{g3} X_{ij3} + \beta_s X_s + \beta_d X_d + \varphi_i + \vartheta_j + \varepsilon_{ij}. \quad \text{(A7)}$$

In this model, for simplicity of interpretation, the original video clip (Off condition) was excluded. For participants in the Smooth group, $g = 0$ and for the Sharp group, $g = 1$. Thus, for the Smooth group, $gX_{ji} = 0$, and the perceptual scale was defined by the first three terms in Equation A7 only (*i.e.* those that did not include g). The statistical significance of differences between the two groups in responses to stimulus j was then found using the coefficient β_{gj}.

Equation A7 can also be written as:

$$a_{ij} = (\beta_1 + \beta_{g1}) X_{ij1} + (\beta_2 + \beta_{g2}) X_{ij2} + (\beta_3 + \beta_{g3}) X_{ij3} + \beta_s X_s + \beta_d X_d + \varphi_i + \vartheta_j + \varepsilon_{ij}. \quad \text{(A7a)}$$

To also examine the effects of video content, video-content category (as described in section 2.4) an indicator variable, c, was used, and the term a_{ij} became:

$$a_{ij} = (\beta_1 + \beta_{g1} + \beta_{c1}) X_{ij1} + (\beta_1 + \beta_{g1} + \beta_{c1}) X_{ij1} + (\beta_1 + \beta_{g1} + \beta_{c1}) X_{ij1} + \beta_s X_s + \beta_d X_d + \varphi_i + \vartheta_j + \varepsilon_{ij}. \quad \text{(A8)}$$

For each of the four video categories, when the video clip had a high rating on that scale (*e.g.*, Face ≥ 3) video clips were coded with $c = 1$ and when the video clip had a low rating (*e.g.*, Face ≤ 2), the video clip was coded with $c = 0$. Statistical significance for the differences between the high and low content levels in responses to stimulus j were obtained from the coefficient β_{cj}. A more complex model that included an interaction between subjective group and video content was also investigated, but it was found to provide no improvement in the model (Wald $\chi^2 \leq 5.7$, df = 3, p $\geq$ 0.13).

References

[1] P. G. J. Barten, *Contrast Sensitivity of the Human Eye and its Effects on Image Quality*. Bellingham, WA, USA: SPIE Press, 1999.

[2] M. Miyahara, K. Kotani, and V. R. Algazi, "Objective picture quality scale (PQS) for image coding," *IEEE Trans. Commun.*, vol. 46, no. 9, pp. 1215–1226, Sep. 1998.

[3] A. C. Bovik, *The Essential Guide to Video Processing*. Burlington, MA, USA: Academic, 2009.

[4] J. A. J. Roufs, "Brightness contrast and sharpness, interactive factors in perceptual image quality," in *Human Vision, Visual Processing, and Digital Display*, vol. 1077, B. E. Rogowitz, Ed. Los Angeles, CA, USA: SPIE Press, 1989, pp. 66–72.

[5] P. G. J. Barten, "Evaluation of subjective image quality with the square-root integral method," *J. Opt. Soc. Amer. A, Opt., Image Sci. Vis.*, vol. 7, no. 10, pp. 2024–2031, 1990.

[6] J. Lubin, "A visual discrimination model for imaging system design and evaluation," in *Vision Models for Target Detection*, E. Peli, Ed. Singapore: World Scientific, 1995, pp. 245–283.

[7] M. Fullerton, R. L. Woods, F. A. Vera-Diaz, and E. Peli, "Measuring perceived video quality of MPEG enhancement by people with impaired vision," *J. Opt. Soc. Amer. A, Opt. Image Sci. Vis.*, vol. 24, pp. B174–B187, Dec. 2007.

[8] P. Kortum and M. Sullivan, "Content is king: The effect of content on the perception of video quality," in *Proc. 48th Annu. Meeting Human Factors Ergonom. Soc.*, 2004, pp. 1910–1914.

[9] M. Sullivan, J. Pratt, and P. Kortum, "Practical issues in subjective video quality evaluation: Human factors vs. psychophysical image quality evaluation," in *Proc. 1st Int. Conf. Design. Interact. Exper. TV Video*, Silicon Valley, CA, USA, 2008, pp. 1–4.

[10] P. Kortum and M. Sullivan, "The effect of content desirability on subjective video quality ratings," *Human Factors*, vol. 52, pp. 105–118, Feb. 2010.

[11] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Austin, TX, USA: Morgan & Claypool, 2006.

[12] Z. M. Parvez Sazzad, Y. Kawayoke, and Y. Horita. (Accessed 2012, Dec. 28). *Image Quality Evaluation Database* [Online]. Available: http://mict.eng.u-toyama.ac.jp/database_toyama/

[13] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. (Accessed 2012, Dec. 28). *LIVE Image Quality Assessment Database Release 2* [Online]. Available: http://live.ece.utexas.edu/research/quality

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[15] 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.

[16] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "A subjective study to evaluate video quality assessment algorithms," *Proc. SPIE*, vol. 7527, pp. 75270H-1–75270H-10, Feb. 2010.

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

[18] "Recommendation ITU-R BT.500-13: Methodology for the Subjective Assessment of the Quality of Television Pictures," International Telecommunications Union, Geneva, Switzerland, Jan. 2012.

[19] L. L. Thurstone, "A law of comparative judgment," *Psychol. Rev.*, vol. 34, no. 4, pp. 273–286, 1927.

[20] L. L. Thurstone, "Psychophysical analysis," *Amer. J. Psychol.*, vol. 38, nos. 3–4, pp. 368–389, 1927.

[21] M. C. Gacula, *Design and Analysis of Sensory Optimization*. Trumbull, CT, USA: Food & Nutrition Press, 1993.

[22] M. C. Boschman, "ThurCatD: A tool for analyzing ratings on an ordinal category scale," *Behavior Res. Methods Instrum. Comput.*, vol. 32, pp. 379–388, Aug. 2000.

[23] M. C. Boschman, "DifScal: A tool for analyzing difference ratings on an ordinal category scale," *Behavior Res. Methods Instrum. Comput.*, vol. 33, pp. 10–20, Feb. 2001.

[24] J. E. Farrell, "Image quality evaluation," in *Colour Imaging: Vision and Technology*, L. W. MacDonald and M. R. Luo, Eds. New York, NY, USA: Wiley, 1999, pp. 285–313.

[25] D. A. Silverstein and J. E. Farrell, "Efficient method for paired comparison," *J. Electron. Imag.*, vol. 10, pp. 394–398, Apr. 2001.

[26] F. Wickelmaier and S. Choisel, "Modeling within-pair order effects in paired-comparison judgments," in *Proc. 22nd Annu. Meeting Int. Soc. Psychophys.*, St. Albans, U.K., 2006, pp. 89–94.

[27] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 882–893, Oct. 2011.

[28] C. H. Coombs, "Thurstone's measurement of social values revisited forty years later," *J. Pers. Soc. Psychol.*, vol. 6, no. 1, pp. 85–91, 1967.

[29] Y. K. Kwan, W. C. Ip, and P. Kwan, "A crime index with Thurstone's scaling of crime severity," *J. Crim. Justice*, vol. 28, no. 3, pp. 237–244, 2000.

[30] P. E. Green, D. S. Tull, and G. S. Albaum, *Research for Marketing Decisions*, 5th ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[31] M. C. Gacula and J. Singh, *Statistical Methods in Food and Consumer Research*. Orlando, FL, USA: Academic, 1984.

[32] S. Lipovetsky and M. W. Conklin, "Thurstone scaling via binary response regression," *Stat. Methodol.*, vol. 1, nos. 1–2, pp. 93–104, 2004.

[33] R. Rajae-Joordens and J. Engel, "Paired comparisons in visual perception studies using small sample sizes," *Displays*, vol. 26, pp. 1–7, Jan. 2005.

[34] R. L. Woods, P. Satgunam, P. M. Bronstad, and E. Peli, "Statistical analysis of subjective preferences for video enhancement," *Proc. SPIE*, vol. 7527, pp. 75270E1–75270E10, Feb. 2010.

[35] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: The method of paired comparisons," *Biometrika*, vol. 39, nos. 3–4, pp. 324–345, 1952.

[36] R. D. Luce, *Individual Choice Behavior—A Theoretical Analysis*. New York, NY, USA: Wiley, 1959.

[37] F. Wickelmaier and C. Schmid, "A Matlab function to estimate choice model parameters from paired-comparison data," *Behavior Res. Methods Instrum. Comput.*, vol. 36, pp. 29–40, Feb. 2004.

[38] P. G. Engeldrum, "A theory of image quality: The image of quality circle," *J. Imag. Sci. Technol.*, vol. 48, pp. 446–456, Sep./Oct. 2004.

[39] C. A. Poynton, *Digital Video and HDTV Algorithms and Interfaces*. San Francisco, CA, USA: Morgan Kaufmann, 2003.

[40] E. Peli and T. Peli, "Image enhancement for the visually impaired," *Opt. Eng.*, vol. 23, pp. 47–51, Jan./Feb. 1984.

[41] R. G. Hier, G. W. Schmidt, R. S. Miller, and S. E. DeForest, "Real-time locally adaptive contrast enhancement: A practical key to overcoming display and human-visual-system limitations," in *SID Int. Symp., Dig. Tech. Papers*, Seattle, WA, USA, 1993, pp. 491–494.

[42] W. W. Hauck and A. Donner, "Wald's test as applied to hypotheses in logit analysis," *J. Amer. Stat. Assoc.*, vol. 72, pp. 851–853, Dec. 1977.

[43] M. A. Garcia-Perez, "Denoising forced-choice detection data," *Brit. J. Math. Stat. Psychol.*, vol. 63, pp. 75–100, Dec. 2010.

[44] M. A. García-Pérez and R. Alcalá-Quintana, "The difference model with guessing explains interval bias in two-alternative forced-choice detection procedures," *J. Sensory Stud.*, vol. 25, no. 6, pp. 876–898, 2010.

[45] J. A. J. Roufs, V. J. Koselka, and A. A. van Tongeren, "Global brightness contrast and the effect on perceptual image quality," *Proc. SPIE*, vol. 2179, pp. 80–89, May 1994.

[46] T. J. W. M. Janssen and F. J. J. Blommaert, "Image quality semantics," *J. Imag. Sci. Technol.*, vol. 41, no. 5, pp. 555–560, 1997.

[47] W. A. Ijsselsteijn, H. De Ridder, and J. Vliegen, "Subjective evaluation of stereoscopic images: Effects of camera parameters and display duration," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 10, no. 2, pp. 225–233, Mar. 2000.

[48] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet, "Model-based cluster and discriminant analysis with the MIXMOD software," *Comput. Stat. Data Anal.*, vol. 51, no. 2, pp. 587–600, 2006.

[49] A. Agresti, *An Introduction to Categorical Data Analysis*. New York, NY, USA: Wiley, 1996.

[50] M. Fullerton and E. Peli, "Digital enhancement of television signals for people with visual impairments: Evaluation of a consumer product," *J. Soc. Inf. Display*, vol. 16, no. 3, pp. 493–500, 2008.

[51] R. L. Woods, C. R. Colvin, F. A. Vera-Diaz, and E. Peli, "A relationship between tolerance of blur and personality," *Invest. Ophthalmol. Vis. Sci.*, vol. 51, pp. 6077–6082, May 2010.

[52] P. Satgunam, R. L. Woods, G. Luo, P. M. Bronstad, Z. Reynolds, C. Ramachandra, B. W. Mel, and E. Peli, "Effects of contour enhancement on low-vision preference and visual search," *Optometry Vis. Sci.*, vol. 89, no. 9, pp. E1364–E1373, 2012.

[53] T. Kurihara, N. Aoki, and H. Kobayashi, "Analysis of sharpness increase by image noise," *Proc. SPIE*, vol. 7240, pp. 7240141–7240149, Feb. 2009.

[54] R. van Egmond, P. Lemmens, T. N. Pappas, and H. de Ridder, "Roughness in sound and vision," *Proc. SPIE*, vol. 7240, pp. 72400B-1–72400B-12, Feb. 2009.

[55] S. N. Yendrikhovskij, F. J. J. Blommaert, and H. de Ridder, "Color reproduction and the naturalness constraint," *Color Res. Appl.*, vol. 24, no. 1, pp. 52–67, 1999.

[56] Y.-G. Lee, Y.-J. Kang, H.-E. Kim, K.-H. Kim, and C.-W. Kim, "Content dependent selection of image enhancement parameters for mobile displays," *Proc. SPIE*, vol 7866, pp. 786608-1–786608-11, Jan. 2011.

Prem Nandhini Satgunam received the B.S. degree in optometry from the Elite School of Optometry, BITS, Pilani, India and the M.S. and Ph.D. degrees in vision science from Ohio State University, Columbus, OH, USA.

She is currently a Research Faculty with the L. V. Prasad Eye Institute, Hyderabad, India. Her research interests include pediatric vision rehabilitation, visual development in children with visual impairments, and binocular vision.

Russell Woods received the Ph.D. degree from the City University, London, U.K.

He is an Assistant Scientist with the Schepens Eye Research Institute and an Assistant Professor of ophthalmology with Harvard Medical School, Boston, MA, USA. He was trained as an Optometrist with the University of New South Wales, Sydney, Australia. He conducts research primarily about low vision (vision impairment) and its rehabilitation. His major interest is the development and evaluation of novel aids for people with low vision, including devices to aid with mobility and to aid viewing television, movies, and other electronic media. He has published research on the optics of bifocal contact lenses, ocular aberrations, retinal imaging, screening for eye disease, clinical measurement systems, and visual perception.

P. Matthew Bronstad received the Ph.D. degree in cognitive psychology from the University of Texas at Austin, Austin, TX, USA.

He is currently an Instructor of ophthalmology with Harvard Medical School, Boston, MA, USA and a Research Associate with the Schepens Eye Research Institute, Massachusetts Eye and Ear, Boston. His research interests include low vision, mobility, and individual differences in preferences.

Eli Peli received the B.S.E.E. and M.S.E.E. degrees from the Technion-Israel Institute of Technology, Haifa, Israel and the O.D. degree from the New England College of Optometry, Boston, MA, USA.

He is currently a Moakley Scholar in aging eye research, a Senior Scientist with the Schepens Eye Research Institute, and a Professor of ophthalmology with Harvard Medical School, Boston. He is a Faculty Member with the New England College of Optometry and Tufts University School of Medicine, Boston. His research interests include image processing in relation to visual function and clinical psychophysics in low-vision rehabilitation, image understanding, evaluation of display-vision interaction, and oculomotor control and binocular vision.

Dr. Peli is a fellow of the American Academy of Optometry, the Optical Society of America, the International Society of Optical Engineering, and the Society for Information Display. He is currently an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING.

Erratum to "Factors Affecting Enhanced Video Quality Preferences"

PremNandhini Satgunam, Russell L. Woods, P. Matthew Bronstad, and Eli Peli

In [1], a delay in communicating proof errors occurred. The paper was published without corrections. The following corrections are important.

The first name of the first author is PremNandhini.

The dates in the first footnote should read as follows.
Manuscript received March 14, 2012
Final revision received July 15, 2013
Accepted for publication August 29, 2013

Mentions in the text to section 2.4, section 2.5, and section 3.1 should be read as Sections II.D, II.E, and III.A, respectively.

Mentions in the text to Table 1 and Table 2 should be read as Table I and Table II, respectively.

Table II legend should read as follows:
"…thus for participant 2, in total low was preferred **2** times, out of 8, over off." (not 4 times, out of 8, over off).

Page 5149: Equation (1) should read:

$$\begin{aligned} a_{ij} = &\left(\beta_1 + \beta_{g1} + \beta_{c1}\right) X_{ij1} \\ &+ \left(\beta_2 + \beta_{g2} + \beta_{c2}\right) X_{ij2} \\ &+ \left(\beta_3 + \beta_{g3} + \beta_{c3}\right) X_{ij3} \\ &+ \beta_s X_{si} + \beta_d X_{di} + \phi_i + \theta_j + \varepsilon_{ij}. \end{aligned} \tag{1}$$

Equation (1) is the same as Equation (A8) in the appendix.

Manuscript received October 28, 2013; accepted October 28, 2013. Date of current version January 9, 2013.

P. Satgunam is with Hyderabad Eye Research Foundation, L. V. Prasad Eye Institute, Hyderabad 500034, India (e-mail: premnandhini@lvpei.org).

R. L. Woods, P. M. Bronstad, and E. Peli are with the Harvard Medical School, Schepens Eye Research Institute, Massachusetts Eye and Ear, Boston, MA 02114 USA (e-mail: russell_woods@meei.harvard.edu; matthew_bronstad@meei.harvard.edu; eli_peli@meei.harvard.edu).

Digital Object Identifier 10.1109/TIP.2013.2288513

Page 5150: Footnotes should read as:
[2]Equivalent to equation 1 with $\beta_{gk} = \beta_{ck} = \beta_s = \beta_d = 0$.
[3]Equivalent to equation 1 with $\beta_{gk} = \beta_{ck} = 0$.
[4]Equivalent to equation 1 with $\beta_{ck} = 0$.

On page 5155: Equation (A5) should read as:

$$a_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \phi_i + \theta_j + \varepsilon_{ij}. \tag{A5}$$

On page 5155: Equation (A6) should read as:

$$\begin{aligned} a_{ij} = &\beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_s X_{si} \\ &+ \beta_d X_{di} + \phi_i + \theta_j + \varepsilon_{ij}. \end{aligned} \tag{A6}$$

On page 5156: Equation (A7) should read as:

$$\begin{aligned} a_{ij} = &\beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_{g1} X_{ij1} + \beta_{g2} X_{ij2} \\ &+ \beta_{g3} X_{ij3} + \beta_s X_{si} + \beta_d X_{di} + \phi_i + \theta_j + \varepsilon_{ij}. \end{aligned} \tag{A7}$$

On page 5156: Equation (A7a) should read as:

$$\begin{aligned} a_{ij} = &\left(\beta_1 + \beta_{g1}\right) X_{ij1} + \left(\beta_2 + \beta_{g2}\right) X_{ij2} + \left(\beta_3 + \beta_{g3}\right) X_{ij3} \\ &+ \beta_s X_{si} + \beta_d X_{di} + \phi_i + \theta_j + \varepsilon_{ij}. \end{aligned} \tag{A7a}$$

On page 5156: Equation (A8) should read as:

$$\begin{aligned} a_{ij} = &\left(\beta_1 + \beta_{g1} + \beta_{c1}\right) X_{ij1} \\ &+ \left(\beta_2 + \beta_{g2} + \beta_{c2}\right) X_{ij2} \\ &+ \left(\beta_3 + \beta_{g3} + \beta_{c3}\right) X_{ij3} \\ &+ \beta_s X_{si} + \beta_d X_{di} + \phi_i + \theta_j + \varepsilon_{ij}. \end{aligned} \tag{A8}$$

References

[1] P. Satgunam, R. L. Woods, P. M. Bronstad, and E. Peli, "Factors affecting enhanced video quality preferences," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5146–5157, Dec. 2013.