https://doi.org/10.2352/ISSN.2470-1173.2017.12.IQSP-244

A Framework for Auto-exposure Subjective Comparison

Seungseok Oh[1], Clayton Passmore[12], Bobby Gold[1], Taylor Skilling[13], Sean Pieper[1], Taek Kim[1], Margaret Belska[1]
1: NVIDIA, Santa Clara, CA, 2: University of Waterloo, Waterloo, Canada, 3: Northeastern University, Boston, MA

Abstract

Subjective testing has long been used to quantify user preference in the field of imaging. The majority of subjective testing is done to analyze still images, leaving the ever-growing field of video overlooked. With little work put into this area of study, not much is known about the preferential behavior of dynamic auto control functions such as automatic exposure (AE). In this study, we focus on subjective preferences for two aspects of video auto exposure convergence: convergence time and convergence curve type, with each tested individually. This experiment utilizes a novel framework for subjective testing, where a collection of videos are captured with simulated changes in light. This method allows for much more precise control of the capture device and constitutes better repeatability of experiments, as opposed to recording real changes. A paired comparison model is employed to conduct the subjective analysis of the videos. In a web application, two videos are played side by side with a slight delay and the user is asked to pick which video they prefer. Results from the experiments show that users prefer monotonic, gradual transition in AE, with no sharp or abrupt changes. Users also preferred transition times of 266-500 milliseconds.

Introduction

Many of the current efforts in understanding image quality have focused on still captures [1,2,3]. While consumers have more access than ever to many different video-recording devices, little is known about their preferences. Even for auto-control functions such as auto-exposure (AE), which is inherently dynamic, image quality assessments are usually performed on the final, converged image, not for an entire video [4].

Subjective tests fill the void where objective metrics alone often fail to capture important insights and findings [2]. Dynamic auto-control behavior is of particular interest as video recordings are getting more and more popular and many new challenging use cases such as those in drones and automotive markets are emerging.

Video subjective evaluation is more complicated than its still counterpart. It is especially difficult to capture test videos with the environment controlled dynamically in a consistent way and to present videos in a way that allows subjects to evaluate two videos simultaneously [6].

This paper presents a framework for evaluating subjective preference, which we have found useful for video AE evaluation. We employ paired-comparison approaches [9] over using the Mean Opinion Score because scale definition in Mean Opinion Score experiments can be ambiguous in that their interpretation could be dissimilar among subjects [5]. In the proposed paired-comparison, two videos are compared at a time in a web application, which permits simultaneous video viewing.

The proposed framework also includes a method for test video generation, which uses simulated changes in light to provide more accurate and repeatable light level transitions. This can prove to be useful in the process of designing and developing AE algorithms or quality metrics. Developers, designers and engineers alike can benefit from the knowledge of how a certain metric fits in with human perception. For example, we would like to know whether people care about convergence speed, which speed is preferred, and whether the preferred speed depends on the magnitude of the illumination change. This method can capture test videos to exhibit the intended AE dynamic behaviors with a simulated approach.

This paper also presents experimental results with the proposed framework to gain insights on user preference for dynamic convergence behavior of auto-exposure algorithms, especially on convergence curve type and convergence time.

Procedure

Test planning

We first define research questions and identify test scenarios.

Sometimes we need to answer one research question for multiple setups, just in case people's preference could be different for different types of exposure changes. For example, preferred convergence time may be different depending on the magnitude or the direction of lighting change [6]. In this work, we propose four setups: Big Up, where the lux level starts at 7,000 Lux and increases exposure by 0.35 stops, Big Down, where exposure starts at 100 Lux and decreases by 4 stops, Small Up, where exposure starts at 600 Lux and increases by 0.15 stops and Small Down, where the exposure starts at 600 Lux and drops by 1 stop. Here, Big or Small denotes the magnitude of lux change and Up or Down signifies the direction of the lux change. Note that we had to use smaller lux changes in upward transitions to avoid saturation of the camera's sensor after the lighting change.

Test video generation

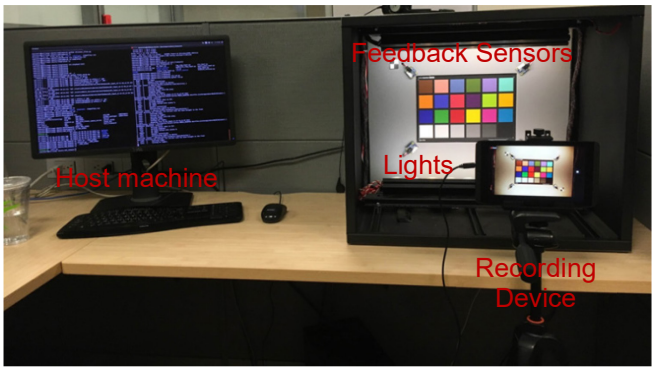

Figure 1: Example video capture setup. The camera, feedback sensors and lights are all controlled by the host machine. The sensors are used to calculate exposure of the scene and to ensure exposure and uniformity remains constant across the X-Rite ColorChecker.

Before subjective analysis can begin, test videos are generated for comparison.

Each video represents a certain AE dynamic behavior, such as a type of convergence curve or convergence time. More specifically, the AE behavior is in response to a step function in lighting. A step response is more convenient for algorithm development and evaluation because it has been well studied in control theory, and it is easy to translate it to engineering parameters such as convergence speed, acceleration, and jerk, which are related to the first, second, and third derivatives, respectively [11].

In this experiment, we use simulated lighting changes instead of real lighting changes for three reasons. First, it is difficult to find LED lights that can produce ideal, sharp step increases or decreases in light level. In our experiments, the transition took up to 100ms depending on the degree of lux change. Second, this simulation approach makes video generation more consistent by eliminating differences between test runs and lighting setups, which would happen if we use real lighting changes. Third, this methodology allows us to evaluate an arbitrary transition pattern, even if we have no idea how to build the control law that would produce such a response to a natural change in lighting. For example, oscillatory and sawtooth convergence, shown in Fig. 4, can be experimented with the use of simple override logics.

An example of the physical setup used to record videos is shown in Fig. 1. The scene simply consists of an X-Rite ColorChecker in the center of the neutral background. The lights are positioned to uniformly illuminate the color checker. Lighting level can be controlled by a host machine. After the video is captured by a recording device, the video file, and the lux information recorded by the light sensors are transferred to the host machine. They are used to calculate exposure and exposure error as explained below in the video analysis section.

Figure 2 illustrates how we simulated these lighting changes in more detail. In the real world, the lights are on and unchanged. Instead of a real change in lighting, we change exposure to force an error of known magnitude and direction on a single frame in the video stream through a modification in the camera driver. In this example, the change in exposure simulates an increase of light and produces instant overexposure. Then, the convergence curve which was programmed in advance will take over and converge to the proper exposure level. We tested several curve types (Linear, Accelerating, Decelerating, S-Shaped, Oscillatory, and Sawtooth) and several convergence times, as shown in Fig.4.

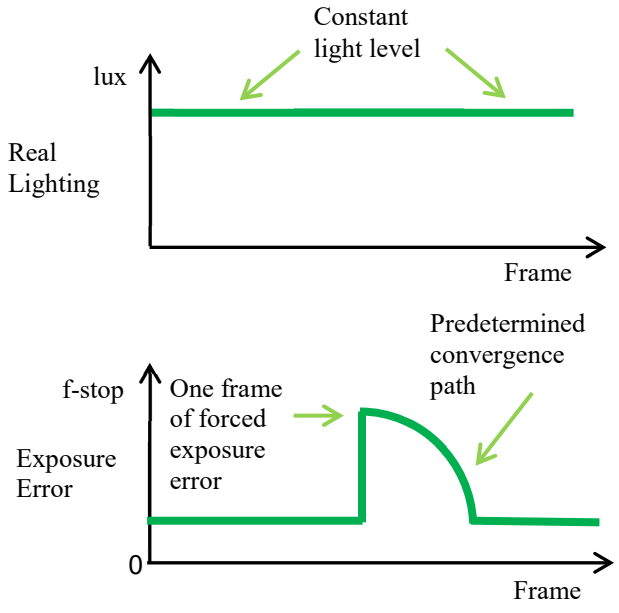

Figure 2: An example plot of the real and simulated scene over time. The forced exposure error allows for simulated changes in light rather than having to modify the lights to the desired transition. This figure illustrates an accelerated convergence curve.

Video Analysis

To ensure the validity and accuracy of the simulated scene it is strongly recommended to double-check whether the exposure over time in the recorded videos are captured as intended. The video analysis tool takes in an input video and a text file that contains light sensor data and produces a plot of exposure and exposure error.

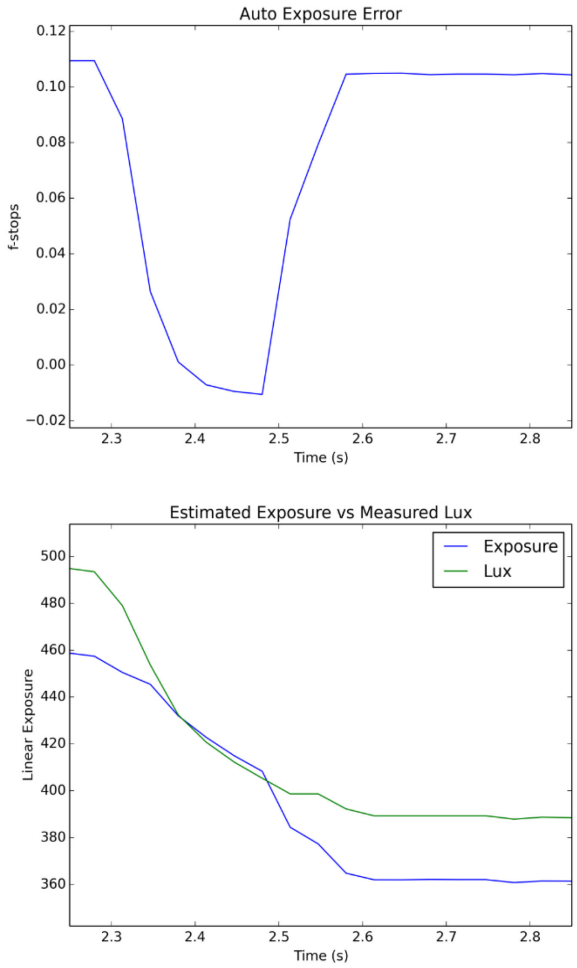

Figure 3a & b: A (above) shows exposure over time calculated from the video analysis script. B (below) is an example of how exposure can be back calculated. Actual scene illuminance is plotted for reference.

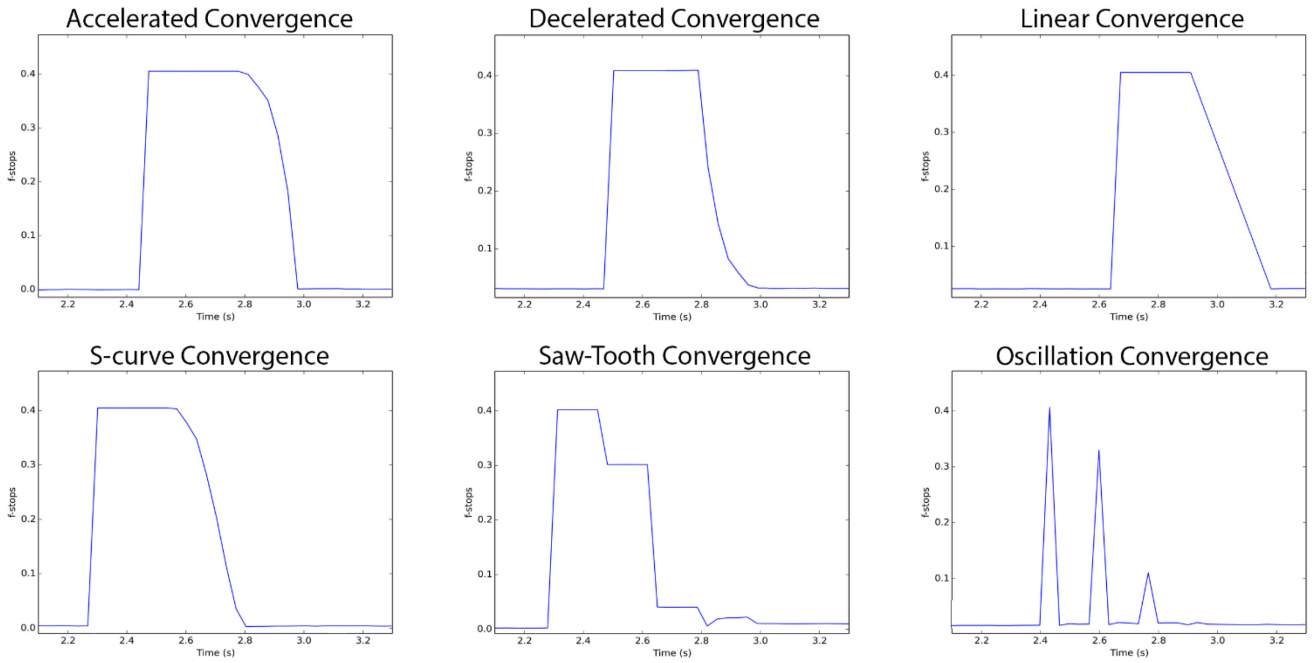

Figure 4: Exposure error plots of all convergence curves used in the Big-up lighting scenario, where the camera is adapting to a light source that gets significantly brighter than its current state. From top left to bottom right: Accelerated Convergence, convergence starts slow and quickens as the camera approaches its normal state; Decelerated Convergence, convergence starts with rapid change and slows as the camera approaches its proper exposure; Linear Convergence, rate of change remains constant over the entire convergence period; S- curve Convergence, changes is slower at the beginning and end with more rapid rate of change in the middle; Saw-tooth convergence, a stair step function which incrementally, and abruptly changes until convergence is reached; Oscillation Convergence, multiple iterations of decreasing over and under compensation to find the proper exposure.

The analysis tool created with this framework calculates exposure error by comparing the known reflectances of the neutral patches (#20-23) on the X-Rite ColorChecker to the same patches in the captured test video. Patches 19 and 24 are omitted so any clipping present in the highlights or shadows will not affect the measurement. We then take the mean signal strength of the patches captured in the test video as the measured reflectance and compare it against the known reflectance of the same patches on the X-Rite ColorChecker, where the unit of exposure error is f-stop and the constant 2.2 originates from gamma correction, as shown in Fig 3a.

$$\mathrm{E}\boldsymbol{xposureError} = \frac{1}{2.2} \times \log_2 \frac{\boldsymbol{measured\ reflectance}}{\boldsymbol{known\ reflectance}}$$

Then, camera exposure can be back-calculated with:

$$\mathrm{E}\boldsymbol{xposure} = \boldsymbol{MeasuredLux} \times 2^{-\boldsymbol{ExposureError}}$$

In this way, we can plot the estimated exposure against the light intensity, as shown in Fig. 3b.

Subjective comparison test

For subjective comparison tests, we employ the paired comparison model where users are required to pick one video over another or state that they appear the same for every comparison [9].

The comparison tests are performed in a self-administered way through a web application. We do not specify a viewing condition as the comparison is relative to just the two videos being displayed. The only condition required in this experiment is that the web application is used as a full screen, so the 50% gray background of the application fills the screen [10]. Before starting the tests, subjects are given instructions on the app as follows:

"You are about to watch a sequence of super exciting videos! We will play two videos at a time, side by side. The one on the left will start slightly before the one on the right, so look at that one first. Each pair of videos will replay 3 times in a row. Tell us if you could tell the difference between the videos and which one you liked better. If you couldn't tell the difference, pick randomly! Once you make your choice, click submit and a new pair of videos will appear!

In the videos that you are about to see, pay attention to how the lighting changes and how long it changes for. We want to know how to best transition from one lighting level to another.

Ready?"

The comparisons then start, as shown in Fig. 5.

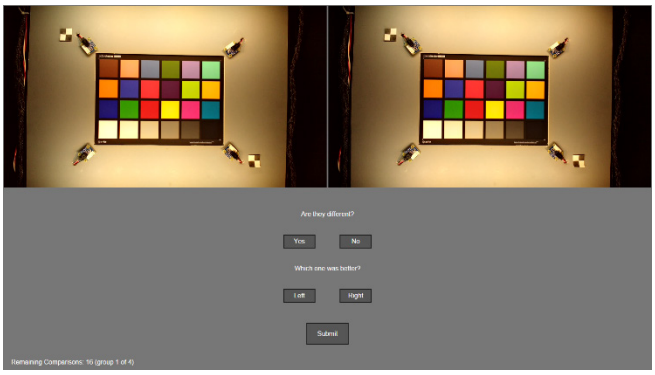

Figure 5: Example pair comparison from web application used in user study.

Each comparison runs videos three times and the user can replay the videos additional times if necessary. There is also a 1.5-second delay between the start of the two videos to allow the user to watch each transition separately, but still, make a direct comparison between the two. The time it takes for a user to choose between videos is recorded. This can be an indicator of how difficult the comparison was and helps identify possible random guesses, where the time to compare the videos is too short.

Redundant comparisons are intentionally added, to determine whether the subject is paying attention or how skillful the subject is at comparisons. For each user, we showed one duplicate comparison in each set of videos [10]. Ideally, the subject should answer consistently to which video they preferred. Additionally, there is one comparison in which the same video is shown side-by-side, where the subject should answer "No" to the question "Are they different?". This indicates that they can successfully identify videos.

Randomization of sequences

Video sequences are randomized at runtime so that every user is presented videos in a different order, removing any potential bias caused by a specific ordering of videos. When multiple groups of comparisons are performed as explained in test planning, video sequences will be grouped and played together, not interleaved. The order in which the groups are displayed are randomized as well. Within every group, every video is compared against every other video, which makes analysis much easier and provides more samples. Subjects are given a break after each group to prevent them from losing their concentration when the full set of tests is expected to take a relatively long time.

Statistical analysis

The first step in analyzing the data is to determine the total wins and losses of each AE behavior. It is a simple task, but there are a couple of details that need to be decided on. First, we have three choices to handle with redundant tests:

1. Redundant test results are not factored into the final count. The drawback is that this would not exploit all the comparison data.
2. Results are calculated with redundant scores and each of the tests carries the same weight. The drawback is that the comparison which has a redundant test has twice the weight as the others.
3. Results are calculated both with and without redundant scores, but the weight for the redundant tests is reduced to half each.

We settled on option 3 because it doesn't penalize two matching redundant trials but can also represent a split in user preference. If a user switches their answer in redundant response it shows that there isn't a clear, definitive preference and the scoring should reflect that.

Secondly, there are two choices on how to handle tests in which the user could not tell the difference between the two videos:

1. They are not counted in the total tally.
2. They are counted in the total tally.

We settled on option 1 so the results would not be skewed by any false positive results. Once the tally is calculated, we apply statistical analysis [7,8]. The final score reported is a z-

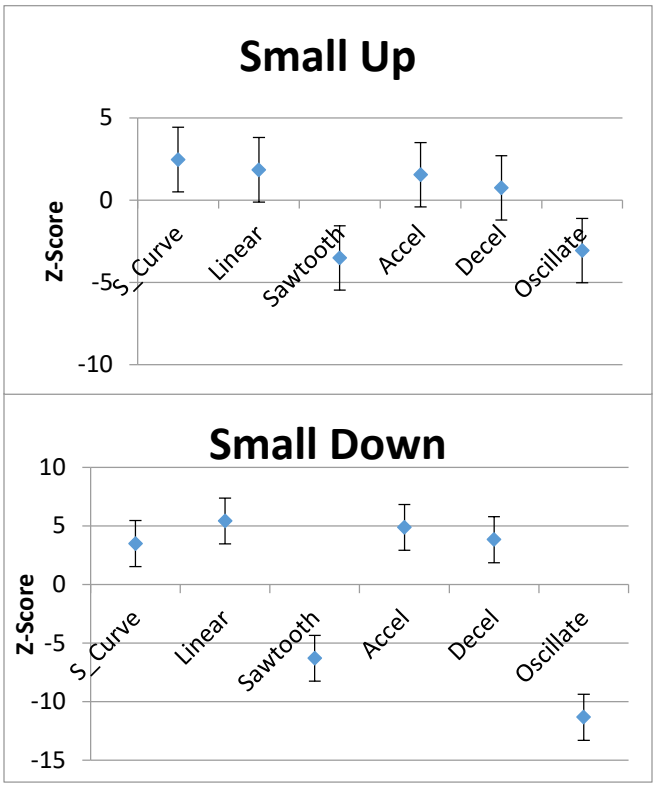

Figure 6: Results plots of Small lighting categories from Round 1 of convergence curve tests.

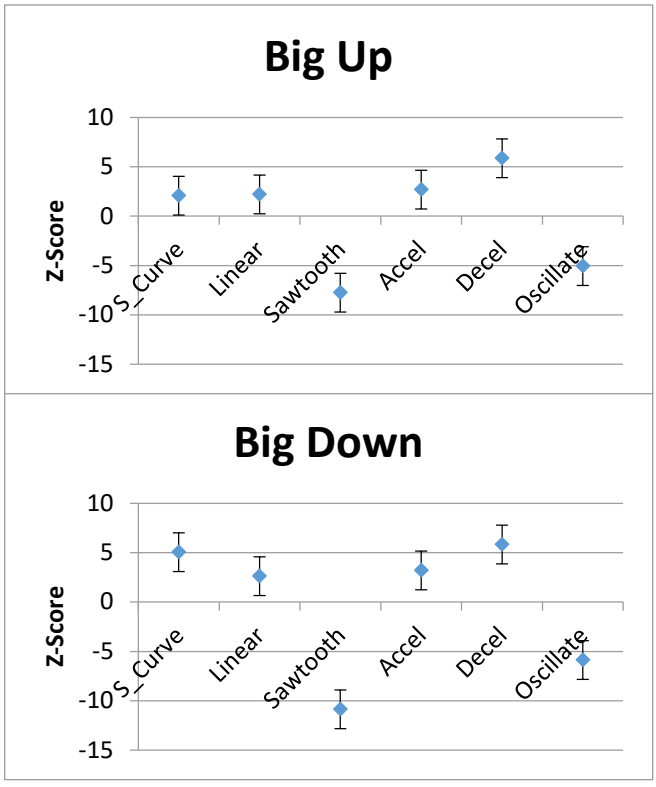

Figure 7: Result plots of Big lighting categories from Round 1 of convergence curve tests.

score, which is a statistic normalized by mean and standard deviation:

$$Z_{Score} = \frac{X - Mean}{Standard\ Deviation}$$

To get a total z-score for each configuration we first calculate a z-score based on the number of wins in each pair and then sum all the z-scores for that configuration. The z-score plus/minus 1.96 are marked in the plots as its confidence interval of 95%.

Results

To get insights on user preferences for auto-exposure dynamic behavior we applied the test framework outlined above. We focused on two aspects of AE behavior: convergence curve type and convergence time.

This user study aimed to use a diverse selection of users to best represent many different types of consumers and markets [10]. A total of 35 users participated in this experiment, some of which were software engineers, image quality and tuning experts, videographers, both professional and amateur, and average consumers with little knowledge of cameras and their workings.

Transition curve type

There are multiple ways to converge from a certain AE state to another and the research question in this experiment was whether curve types are important to human perception and if so, which are preferred.

In the first round, six curve types were tested: accelerating, decelerating, linear, S-curve, oscillatory, and sawtooth, illustrated in Fig 4 for Big-up case. Here we tested with all four types of lighting conditions to see if there are different preferences for different lighting scenarios (Big-up, Big-down, Small-up, Small-down).

The results showed that users did not prefer non-smooth transitions between light levels. Sawtooth and oscillatory transitions scored significantly lower than the rest of the transitions which were all smooth continuous curves. This trend persists in both small and big changes, as seen in Fig. 6 and Fig. 7 respectively.

In an attempt to clarify the results, we performed another round of testing where the less preferential, non-smooth transitions were removed. This aimed to create a larger distinction of preference between the top four transition curves: Accel, Decel, Linear, and S-curve. We also focused on large lux transitions and large convergence time, presuming that it would help convergence curves to be more distinguishable.

As shown in Fig. 8, a decelerated transition curve remained preferential for the Big-down lighting. On the other hand, an accelerated convergence curve was the most preferred for Big-up lighting. However, it would be difficult to draw a definite conclusion for the following three reasons:

1. This result was different from the previous round in Fig. 7, where the decelerated transition curve was preferred by users for both the Big-up and Big-down categories.
2. The rankings were not consistent across Rounds and settings, as shown in Figs 6 and 7.

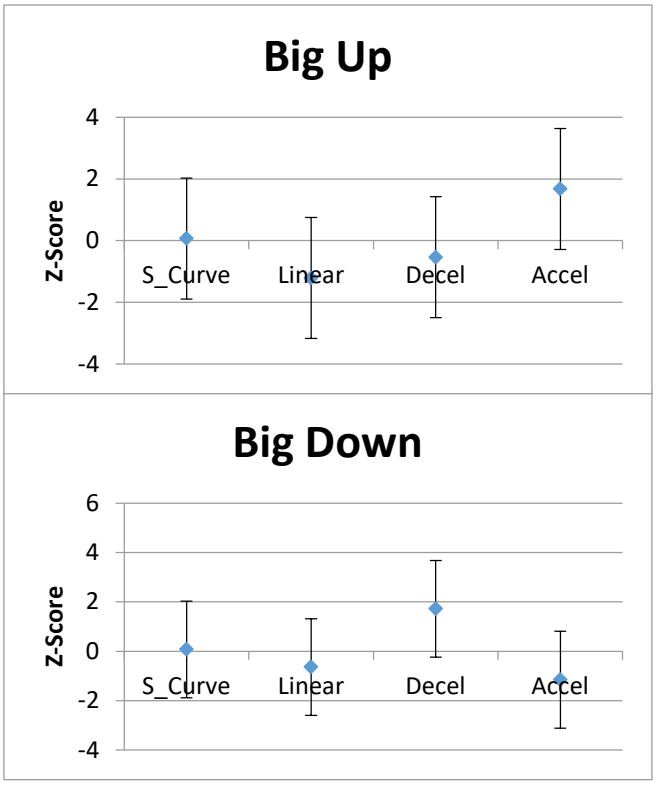

Figure 8: Result plots of from Round 2 of convergence curve tests.

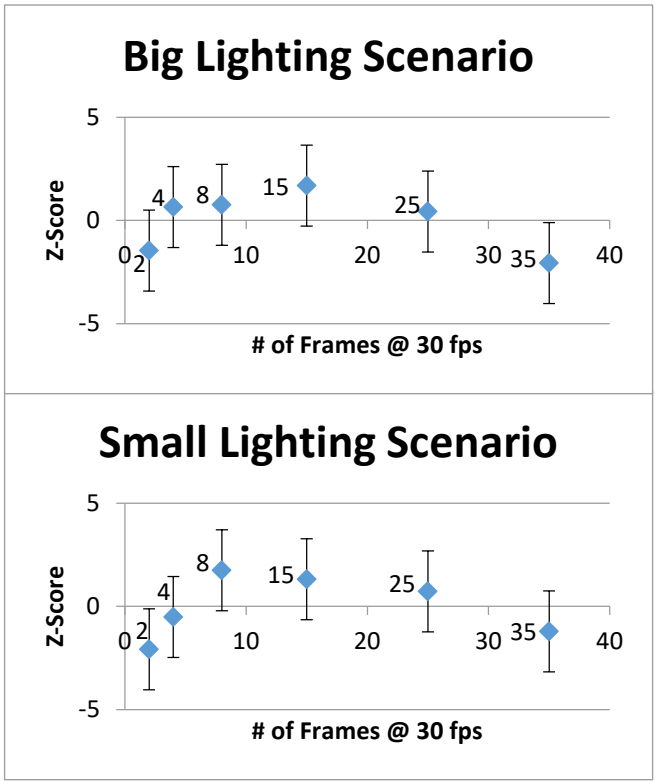

Figure 9: Plots of transition time in Big and Small Lighting Scenarios.

3. None of the results rejected null hypothesis that the z-scores of Accel and Decel are the same.

Transition time

Preferences of transition times were also tested for both upward and downward transitions. Six different convergence times were tested where time is measured in number of frames: 2, 4, 8, 15, 25 and 35 frames with a frame rate of 30 frames per second. The tests were repeated for both Big and Small changes to investigate whether the magnitude of the light changes affects users' preference on transition time.

Figure 9 shows the results for transition time experiments. We averaged the z-scores for each of the Big and Small changes because the direction of lighting conditions appeared not to make meaningful differences. All four lighting scenarios yielded similar trends, which were all centered around the same range of values. Generally, users preferred convergence times of 8-15 frames, which corresponds 266 -500 milliseconds. Users increasingly disliked trends that were longer or shorter than this range.

The magnitude of lighting changes had a slight effect: users preference slightly peaked at an 8 frame (266ms) transition for smaller light level transitions and a slight preference peaked at 15 frames (500ms) in larger transitions. However, when looking at the average number of wins for transition time, user preference was close to 17 frame or 560ms for both Big and Small changes. This was calculated considering total number of wins (before z-normalization) and compensating for the different intervals between frame rates.

Conclusions and Discussions

In order to investigate the user's preference towards the dynamic behavior of a camera's auto-exposure within changing light conditions, this study proposed a novel subjective testing framework. This included the use of simulated light changes, which was used to generate test videos with very precise control and a web viewer that allows for simultaneous viewing and comparison of the two videos. A paired comparison evaluation is used to compare the subjective results of different types of convergent curves and rates.

Experiments with this framework revealed some insights on subjective preferences on auto-exposure dynamic behavior. Users strongly disliked non-smooth, non-monotonic convergence curves; oscillatory curve and sawtooth curves consistently scored very low. Among the monotonic smooth convergence curves, some preference was observed. However, these preferences were all within the statistical variation range.

Users preferred transitions that took between 8 and 15 frames at 30 fps or 266 -500 milliseconds. While the direction of the change did not matter, the magnitude of change had a slight effect. Users prefered the slightly shorter 8 frame transition for smaller changes in light and the slightly longer 15 frame transition for larger changes. The average number of wins for each time, adjusted for the framerate interval, shows that users preferred a transition of about 17 frames for all lighting conditions.

The proposed test framework allowed for easy subjective testing of videos. The auto-exposure experiments in this paper are one of the many uses for this framework. The framework is left open ended so any videos can be placed in it, allowing for testing of other various auto control functions. This can include auto white balance, preferences in local tone mapping of videos, and noise reduction algorithms.

References

[1] Wang Zhou, et al., "Image Quality Assessment: From Error Visibility to Structural Similarity." IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, April 2004

[2] ISO 20462 Photography – Psychophysical experimental methods for estimating image quality

[3] Wang Zhou, "Objective image quality assessment: Facing the real-world challenges," in Image Quality and System Performance XIII, San Francisco, Feb, 2016.

[4] Zhen He, et al., "Development of a perceptually calibrated objective metric for exposure quality," in Image Quality and System Performance XIII, San Francisco, Feb, 2016.

[5] Robert C. Streijl, et al., "Mean opinion score (MOS) revisited: methods and applications, limitations, and alternatives," in Journal of Multimedia Systems, vol. 22, no. 2, pp. 213-227, March 2016.

[6] K. Moorthy, et al., "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," IEEE Journal of Selected Topics in Signal Processing, vol. 6, no. 6, pp. 652–671, Oct. 2012.

[7] K. Tsukida, et al., "How to Analyze Paired Comparison Data," Department of Electrical Engineering: University of Washington, 2011.

[8] R. K. Mantiuk, et al., "Comparison of Four subjective methods for image quality assessment," Computer Graphics Forum, vol. 31, no. 8, pp. 2478–2491, Aug. 2012.

[9] S. Winkler, Digital video quality: Vision models and metrics. Chichester, United Kingdom: Wiley, John & Sons, 2005.

[10] TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU, "P.910: Subjective video quality assessment methods for multimedia applications," 2008. [Online]. Available: http://www.itu.int/rec/T-REC-P.910-200804-I.

[11] K. Ogata, Modern Control Engineering (5th Edition), Pearson, 2009.

Acknowledgments

We thank the NVIDIA camera team, Elaine Jin and her colleagues at Google, and Professor Nitin Sampat and his students at the Rochester Institute of Technology for helping with this user study.

Author Biography

Seungseok Oh *received his B.S. and M.S. degrees in Electrical Engineering from Seoul National University (South Korea) and his Ph.D. degree (2005) in Electrical and Computer Engineering from Purdue University, West Lafayette, Indiana. He worked for Fujifilm Medical Systems USA and Morpho Detection Inc. He is currently with NVIDIA as a Senior Software Engineer. His interest includes algorithms and software in imaging and computer vision.*

Clayton Passmore *is a Canadian undergraduate student at the University of Waterloo. He majors in Computer Science and expects to receive his BCS in 2017. Clayton has worked for five different companies across the United States and Canada while completing his degree. At NVIDIA corporation, his work focused on developing infrastructure for automated image quality testing and subjective image quality comparisons.*

Bobby Gold *received his B.S. in Photographic Imaging and Technology from the Rochester Institute of Technology in 2016. He is*

currently working at NVIDIA as a Software Engineer with a focus on image quality in both still imaging a video. His interest includes the impact of image quality on machine learning, computer vision algorithms, and image quality in video systems.

***Taylor Skilling** is currently pursuing his BS in Electrical and Computer Engineering from Northeastern University (2017). While attending, he has completed six-month internships at Systems and Technology Research, a Department of Defense contractor, and NVIDIA with a focus on design and implementation of data collection systems.*

***Sean Pieper** received his BS and MS in electrical and computer engineering from Carnegie Mellon University (2003,2004) and was a Ph.D. candidate electrical and computer engineering at the University of Wisconsin-Madison from 2004 to 2007. For the last 7 years, he has worked at NVIDIA corporation focusing on image processing architecture and system software.*

***Taek Kim** received his BS in Photographic Imaging and Technology (1986) and MS in Imaging Science (1991) from the Rochester Institute of Technology. Since then he has worked on digital proofing for DuPont and digital camera/photo printers for Hewlett-Packard company as an imaging/color engineer. He is currently with NVIDIA focusing on improving the image quality of ISP and display.*

***Margaret Belska** manages the Customer Camera team at NVIDIA, with an objective of ensuring image quality across all segments employing Tegra Mobile Processors, including mobile, automotive and embedded. Margaret has 20+ years' experience in imaging, ranging from space-based scientific cameras to DSLRs and mobile cameras. Margaret is Secretary of IEEE P1858 Camera Phone Image Quality (CPIQ) and Chair of ICAP CPIQ Conformity Assessment Steering Committee.*