Visual acuity and task-based video quality in public safety applications

Joel Dumke

Institute for Telecommunication Sciences, 325 Broadway, Boulder CO 80305-3337, USA

ABSTRACT

This paper explores the utility of visual acuity as a video quality metric for public safety applications. An experiment has been conducted to track the relationship between visual acuity and the ability to perform a forced-choice object recognition task with digital video of varying quality. Visual acuity is measured according to the smallest letters reliably recognized on a reduced LogMAR chart.

Keywords: Visual acuity, Video quality, Public safety, Task-based, Video quality metrics

1. INTRODUCTION

This paper describes an experiment performed as a part of the Public Safety Communications Research (PSCR) program at the Institute for Telecommunication Sciences. Our investigation of video quality for public safety applications is funded by the Department of Homeland Security's Office for Interoperability and Compatibility (OIC). The goal of this work is to provide guidance for public safety practitioners who are interested in purchasing or designing video systems. This includes specifying the minimum resources, especially bandwidth, that must be applied to produce a sufficient level of quality. To date, our efforts have primarily focused on measuring the minimum video bit rate required for adequate video in public safety applications. The video quality efforts of OIC and PSCR led to the creation of the Video Quality in Public Safety (VQiPS) working group. This group brings researchers and video industry representatives together with end users.

In these research efforts it quickly became clear that subjective video quality, as measured by Mean Opinion Score (MOS), was not particularly meaningful in a public safety context. Public safety practitioners are primarily concerned with whether or not a video system is adequate to perform a particular task. This led us to develop a concept of task-based video quality that is measured by a viewer's ability to perform a task rather than a viewer's subjective quality judgments. These ideas allowed us to develop the methodology adopted in ITU-T Recommendation P.912.[1]

It also became clear that it would be inefficient to conduct separate experiments and publish separate recommendations for different types of agencies (e.g., police, firefighters, emergency medical services, transportation). For this reason, VQiPS developed the concept of the Generalized Use Class (GUC).[2] This assumes that there are five factors of primary importance in describing a particular use case for a video system. These factors are usage time-frame (live or recorded), the required discrimination level, the size of the area of interest (referred to as the target) in the frame, the amount of motion in the video, and the lighting conditions. Use cases can be classified by describing them in terms of these five aspects. By making recommendations for a small number of GUCs, we hope to provide meaningful guidance without making recommendations for a large number of specific use cases.

VQiPS also determined that public safety tasks generally involve recognition. Therefore, experiments were conducted around an object recognition task for both simulated live video[3] and recorded video.[4] Each of these tests included a variety of target sizes, lighting conditions, and amounts of motion. Hence, the only GUC dimension that we had not explored was the discrimination level. To measure the requirements for each of the four discrimination levels defined by VQiPS, we needed to conduct new experiments with at least four different tasks of varying difficulties. Rather than reproduce each large subjective test four times with new tasks, we sought to apply a video quality metric. By measuring the video quality required for each task with one test and measuring the quality metric provided by each video system of interest with another, we can make all of the desired recommendations with few experiments. The experiment in this paper focuses on measuring the quality provided by video systems, while also giving us some indication of the utility of our quality metric.

Image Quality and System Performance X, edited by Peter D. Bums, Sophie Triantaphillidou,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8653, 865306 • © 2013 SPIE-IS&T
CCC code: 0277-786X/13/$18 • doi: 10.1117/12.2004882

The metric we have chosen is visual acuity.[5] As stated above, we would not necessarily expect metrics that correlate highly with MOS to be the most useful quality metrics for public safety applications. Visual acuity is fundamentally a measurement of a viewer's ability to recognize characters on a chart. Measured with human viewers, it incorporates high-level and low-level elements of the human visual system that other metrics do not. Because recognition seems particularly important to public safety applications, we believe visual acuity will be more highly correlated to the performance of public safety tasks than other metrics. It also has the advantage of being easily measured for a particular video system in the field.

Section 2 explains the experiment in detail. Section 3 summarizes the results, and section 4 provides some concluding remarks.

2. EXPERIMENTAL DESIGN

In this experiment, viewers were shown a variety of video sequences and asked questions about each one. For each sequence, the viewer was asked to recognize an object and to read an acuity chart from the video. The general test methodology adhered to ITU-T Recommendations P.910[6] and P.912.[1]

2.1 Acuity Charts

To measure visual acuity, we generated a set of reduced Logarithm of the Minimum Angle of Resolution (LogMAR)[7] charts. These charts are roughly similar to the Snellen eye charts commonly used by physicians, but a few key differences recommend LogMAR charts for any visual acuity research. LogMAR charts are made up of rows of Sloan letters.[8] A subset of the English alphabet consisting of the letters C, D, H, K, N, O, R, S, V, and Z is used. When the charts are generated, each letter is randomly selected with a uniform distribution. Each row is made up of characters of a particular size. At the bottom of the chart, the smallest row is sized so that the height of each character will be exactly five pixels when the chart has been resized to fit within a VGA resolution frame. This is regarded as the minimum number of pixels necessary to reliably recognize characters with good contrast and perfect quality. For this reason, there is no purpose in using any smaller characters. Each successive row uses characters $\sqrt{2}$ times the height of the row directly beneath it so that the size of the characters doubles every two rows. Eight rows were used for each chart. Because we are not interested in measuring an individual viewer's visual acuity, we included only three letters on each line. We can combine results from multiple viewers to measure the acuity of each video sequence, and save time for individual viewers by reducing the number of letters in each row. Figure 1 shows examples of the charts used in this experiment.

2.2 The Object Recognition Task and Video Sequences

ITU-T Recommendation P.912[1] introduces the concept of scenario groups. A scenario group is a set of video sequences that are as nearly identical as possible; only one element varies. For this test, the changing element is the target object. We used the same scenario groups as in our previous experiments.[3,4,9] The scenario groups are designed to capture the three lighting levels, two motion levels, and two target sizes as defined by VQiPS. Although these parameters are not defined in a quantitative way, our scenes were designed to capture a variety

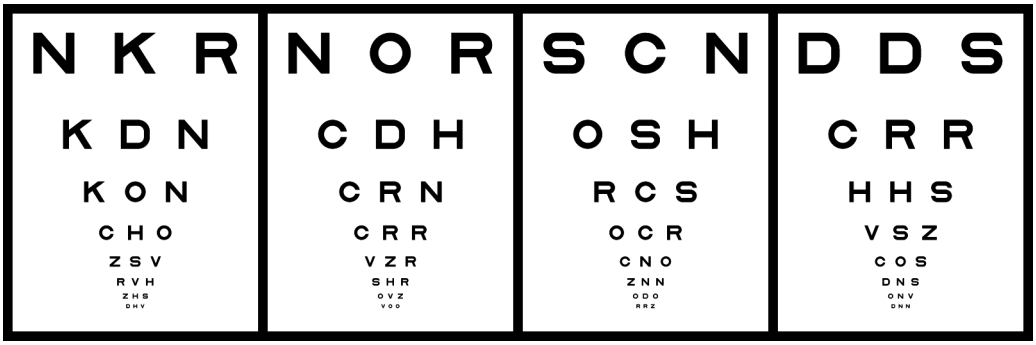

Figure 1. Examples of reduced LogMAR acuity charts used in the experiment.

of conditions. Low motion scenes were produced by placing the objects on a table in front of the camera. High motion scenes were produced by having a police officer walk in front of a camera carrying the target object. The target size was controlled by the lens focal length and the distance between the object and the camera. Video sequences were captured both outdoors and in a variety of indoor lighting conditions.

As in our previous tests,[3,4,9] seven different objects served as targets. They were a gun, an electroshock weapon, a police radio, a cell phone, a flashlight, a mug, and a soda can. For each video sequence, each viewer chose the object that he or she believed was shown. There was always exactly one correct answer. This corresponds to the "multiple choice" method put forward in ITU-T Recommendation P.912.[1] All seven objects were shown in twelve scenario groups and an additional two scenario groups used six of the seven objects. This resulted in a total of 96 source sequences.

The source material for this test was originally captured in high definition. All video sequences were converted to VGA (640×480) and CIF (352×288) resolutions using a Lanczos anti-aliasing filter. The frame rate and color space were not changed and kept at 29.97 fps, 8-bit YCbCr. An acuity chart was then synthetically inserted into each sequence. A different, randomly generated chart was used each time. The chart moved across the screen at the same rate (as measured by pixels per frame) as the target object. The contrast of the chart was also adjusted to reflect the lighting on the object, and this changed over time for video sequences with variable lighting conditions. For the moving charts, a physical camera was simulated by averaging the light that the lens would capture over the entire period of time between two frames. This matches our observations of the workings of actual video cameras. The chart images were generated at very high resolution of 2550×3300 pixels. For each video sequence, great care was taken to ensure that the charts were resized accurately, taking any simulated motion into account.

The clips were then compressed via the MainConcept* H.264 encoder[10] at various bit rates. The H.264 baseline profile was used. Five bit rates were chosen for each resolution. The bit rates were chosen to represent a wide range of resultant video quality and to represent a wide range of bandwidth requirements. Table 1 lists the bit rates used. Each combination of resolution and bit rate is referred to as a Hypothetical Reference Circuit (HRC) and describes the distortion to the video sequence that is being tested. After being processed through an HRC, each sequence was decoded and resized to VGA resolution. This minimized the computational requirements for our test computer to display the sequences, although it required faster storage. By processing all 96 source sequences through all ten HRCs, we generated 960 processed video sequences.

2.3 Viewers and Test Environment

Thirty-nine viewers participated in the test. In accordance with ITU-T Recommendation P.912,[1] expert viewers were recruited. These viewers had experience as practitioners in law enforcement, fire service, or emergency medical services. Viewers were screened for visual acuity and color vision by way of Snellen and Ishihara tests, respectively. Viewers were not automatically excluded from the test if they demonstrated impaired acuity or color vision. In our previous tests,[3,4] an analysis revealed performance of the recognition task was not significantly affected by such visual impairments.

Viewing conditions generally followed the recommendations in ITU-T Recommendation P.910.[6] One exception was that the viewers could choose their viewing distance, and it was not recorded. Viewing distance is

Table 1. H.264 Encoder Bit Rates

Resolution	Bit Rates (kbps)
CIF	64, 128, 256, 512, 1024
VGA	128, 256, 512, 1024, 2048

*Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

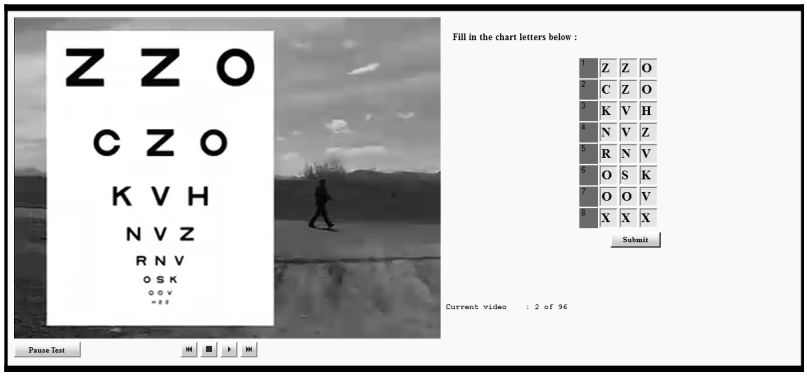

Figure 2. Test software user interface, acuity chart input

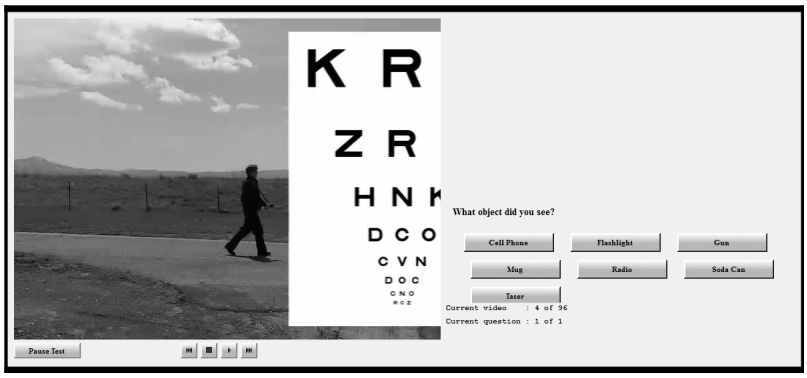

Figure 3. Test software user interface, multiple-choice input

measured relative to the height of the picture being displayed. It is reasonable to assume that the chosen viewing distances most likely fell into the recommended range of 1 to 8 times the picture height, given an approximate picture height of five inches.

Figures 2 and 3 illustrate the user interface for the test. Viewers were shown processed video sequences, and were asked to identify letters on the acuity chart. Viewers entered the letters with a computer keyboard. Viewers then performed the object recognition task by selecting the object they believed was in the sequence. Viewers were allowed to view the clip as many times as they chose, and had the option to pause the video and advance the video frame by frame as much as they wished. This degree of control allowed viewers to find the individual frame that was most useful to the task at hand. The user interface software recorded all such interactions for future analysis.

At the beginning of each viewing session, each viewer was shown a video displaying each target object with a label. Each viewer then performed a practice session consisting of four processed video sequences selected from the 960 sequences used in the test. This familiarized the viewers with the test software. To avoid any memorization effects, each viewer only saw one HRC for a given source sequence. The HRCs were chosen so that the distribution of HRCs among viewers was as uniform as possible. Hence, each viewer saw 96 processed video sequences over the course of the session, generally resulting in a session time between 60 and 90 minutes. Because the viewers controlled their own interaction with the video, the total time for a session could vary widely among viewers. Subjects were free to take breaks as needed.

3. RESULTS

Viewers' responses to each clip were compared against the ground truth for that clip. The total number of characters correctly recognized in each row of the acuity chart (ranging from zero to three) was tallied. We also determined whether each viewer correctly recognized the object they were shown in each clip. We then

treated all viewers as statistically equivalent and combined their results for each clip. We assumed that the object recognition task is essentially the same regardless of which of the seven objects is shown, and combined all the results across objects. For each combination of scenario group and HRC, we were then left with the total number of times viewed, the total number of correct object recognitions, and the total number of characters recognized at each of the eight sizes. By dividing the number of correct object recognitions by the total times a sequence was shown, we calculated the object recognition rates.

The exact definition of visual acuity we have chosen to use is the inverse of the height (measured in pixels) of the smallest reliably-recognizable characters on the acuity charts. Taking the inverse produces a metric that increases as the ability to recognize characters increases and is consistent with previous work on visual acuity.[5] We chose a 90-percent recognition rate as the minimum to be considered reliable. This threshold was heuristically determined to be the one that produced a metric with the best correlation to the object recognition rate. With this approach, we were able to calculate acuity values based on the totals of correctly-recognized characters for each processed video sequence.

3.1 Acuity and Object Recognition

We began investigating visual acuity with the hope that it could serve as a one-dimensional quality metric that would describe the utility of video systems for various public safety tasks. The data gathered in this experiment provide an opportunity to explore how closely visual acuity tracks the performance of the object recognition task. Figure 4 shows a scatter plot of the object recognition rate versus acuity for every combination of scenario group and HRC. Figures 5 and 6 show scatter plots for individual scenario groups that are particularly interesting. Both represent scenario groups involving good lighting and high motion. Figure 5 shows a scenario group involving large targets and Figure 6 shows a scenario group involving small targets.

4. CONCLUSIONS

Unfortunately, Figure 4 makes clear that there is not an extremely high correlation between visual acuity and object recognition rate, but there does appear to be some relationship. Figures 5 and 6 show particularly problematic scenario groups. Figure 5 represents a scenario group with large targets, and shows examples of very high recognition rates for sequences with low acuity. Figure 6 represents a scenario group with small targets, and shows examples of low recognition rates for sequences with high acuity. These outcomes seem natural when we consider that asking viewers to recognize a small target may be a very different task from asking viewers to recognize a large target. In light of this hypothesis, the differences between Figures 5 and 6 are not caused by the acuity metric measuring quality differently, but because we are trying to correlate acuity with two different tasks.

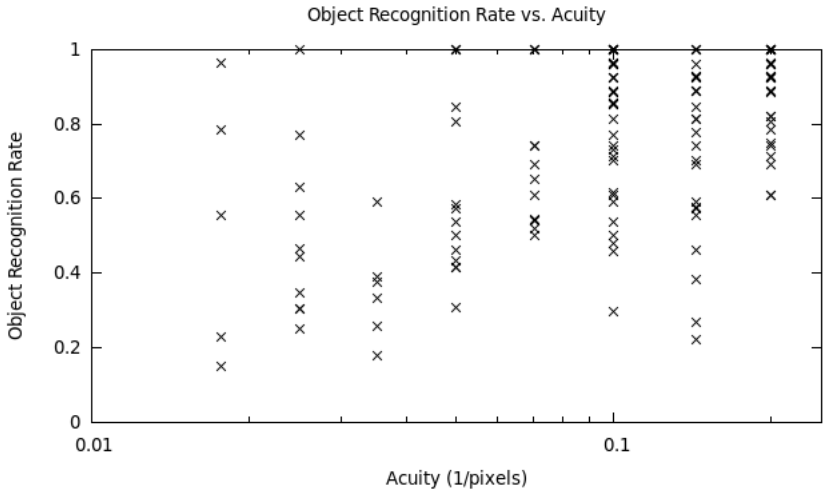

Figure 4. Object Recognition Rate vs. visual acuity for every combination of scenario group and HRC

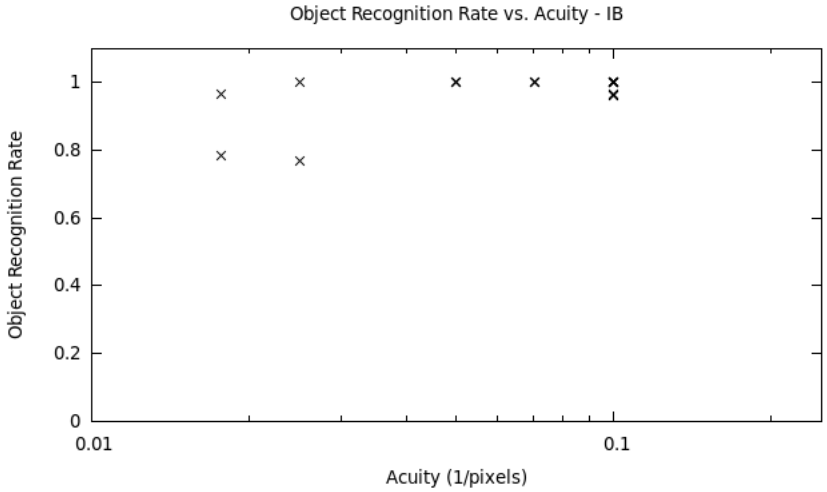

Figure 5. Object Recognition Rate vs. visual acuity for the high-motion, good-lighting, large-target scenario group and every HRC

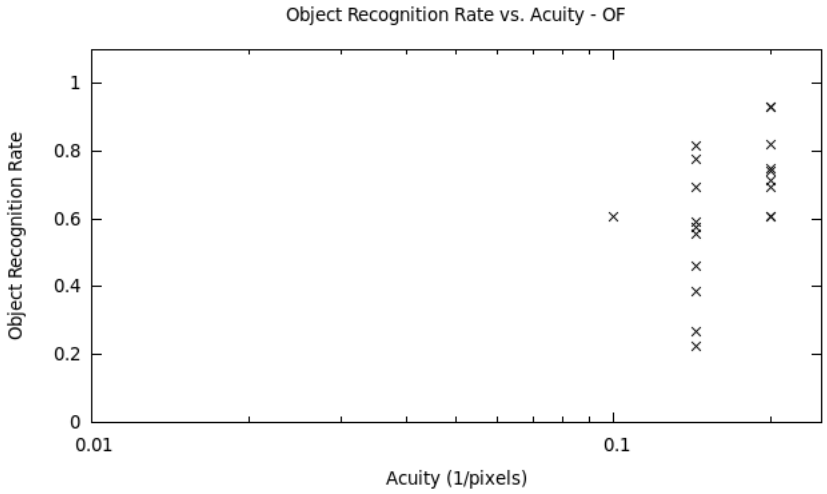

Figure 6. Object Recognition Rate vs. visual acuity for the high-motion, good-lighting, small-target scenario group and every HRC

When the target is small, recognition is difficult even with high acuity. When the target is large, recognition is much easier, even if the overall acuity of the sequence is low. Because acuity is a measurement of the smallest characters that are reliably recognized, it is natural that recognizing small targets would require more acuity than recognizing large targets.

These findings call into question the way that VQiPS has defined its GUC framework. Based on these results, VQiPS is wrong to classify target size as a scene parameter. Target size is not something that contributes to or detracts from the quality of the scene. It describes an aspect of the task the user needs to perform and should be considered a "use characteristic" along with "discrimination level." Furthermore, these dimensions of the GUC framework are clearly not independent, and should be combined or significantly altered. Alternatively, a new acuity metric could be developed for recognition tasks that measures acuity relative to the size of the target. This would probably be a more robust metric with varying target size, but its applications would be limited to somewhat narrowly defined recognition tasks.

Fortunately, in other scenario groups, these issues are not apparent and visual acuity tracks the object recognition rate relatively well. Figures 5 and 6 represent scenario groups with good lighting and high motion. Scenario groups with good lighting and low motion are somewhat degenerate in that the video quality is generally very high. In these cases, object recognition rates and visual acuity measurements are high as well. For scenario groups with low lighting, acuity appears to track object recognition much better. We hypothesize that lighting dominates in these situations because without enough contrast to see the target objects, their size is unimportant. We would also expect that different encoder bit rates will produce very different outputs in low-contrast situations. Again, this would make lighting a dominant parameter over target size.

Figure 4 reveals that visual acuity does not track recognition rate as closely as we would have hoped, but it is still possible that visual acuity is relatively robust compared to other video quality metrics. The next step in our analysis of these results will be to provide context by investigating other quality metrics such as Peak Signal to Noise Ratio (PSNR) and ITS's Video Quality Metric (VQM). We can use the processed video sequences from this test to determine how strongly they correlate to the object recognition rate. We hypothesize that visual acuity will outperform these metrics.

REFERENCES

[1] "Subjective video quality assessment methods for recognition tasks," in [*Recommendation P.912*], ITU-T, Geneva (2008).

[2] "Guide to defining video quality requirements." `http://www.pscr.gov/outreach/vqips/vqips_guide/define_vid_qual_reqs.php` (May 2012).

[3] "Video quality tests for object recognition applications," tech. rep., `http://www.safecomprogram.gov/library/Lists/Library/Attachments/231/Video_Quality_Tests_for_Object_Recognition_Applications.pdf` (Sept 2010).

[4] "Recorded-video quality tests for object recognition tasks," tech. rep., `http://www.pscr.gov/outreach/safecom/vqips_reports/RecVidObjRecogn.pdf` (Sept 2011).

[5] Watson, A., "Video acuity: A metric to quantify the effective performance of video systems," in [*Imaging Systems Applications*], Optical Society of America (2011).

[6] "Subjective video quality assessment methods for multimedia applications," in [*Recommendation P.910*], ITU-T, Geneva (1996).

[7] Grosvenor, T., [*Primary care optometry*], Butterworth-Heinemann, St. Louis, Missouri (2007).

[8] Sloan, L., "New test charts for the measurement of visual acuity at far and near distances," *Am. J. Ophthalmol.* **48**, 807–13 (Dec 1959).

[9] Dumke, J., Ford, C., and Stange, I., "The effects of scene characteristics, resolution, and compression on the ability to recognize objects in video," in [*Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*], **7865**, 23 (2011).

[10] "Advanced video coding for generic audiovisual services," in [*Recommendation H.264*], ITU-T, Geneva (2007).