Video Quality Tests for Object Recognition Applications

Homeland Security

Science and Technology

DHS-TR-PSC-10-09

U.S. Department of Homeland Security

Public Safety Communications

Technical Report

This page intentionally left blank.

Homeland Security
Science and Technology

Support to the Homeland Security Enterprise and First Responders: Office for Interoperability and Compatibility

Defining the Problem

Emergency responders—police officers, fire personnel, emergency medical services—need to share vital voice and data information across disciplines and jurisdictions to successfully respond to day-to-day incidents and large-scale emergencies. Unfortunately, for decades, inadequate and unreliable communications have compromised their ability to perform mission-critical duties. Responders often have difficulty communicating when adjacent agencies are assigned to different radio bands, use incompatible proprietary systems and infrastructure, and lack adequate standard operating procedures and effective multi-jurisdictional, multi-disciplinary governance structures.

OIC Background

The Department of Homeland Security (DHS) established the Office for Interoperability and Compatibility (OIC) in 2004 to strengthen and integrate interoperability and compatibility efforts to improve local, tribal, state, and Federal emergency response and preparedness. Managed by the Science and Technology Directorate within the Support to the Homeland Security Enterprise and First Responders, OIC helps coordinate interoperability efforts across DHS. OIC programs and initiatives address critical interoperability and compatibility issues. Priority areas include communications, equipment, and training.

OIC Programs

OIC programs address voice, data, and video interoperability. OIC is creating the capacity for increased levels of interoperability by developing tools, best practices, technologies, and methodologies that emergency response agencies can immediately put into effect. OIC is also improving incident response and recovery by developing tools, technologies, and messaging standards that help emergency responders manage incidents and exchange information in real time.

Practitioner-Driven Approach

OIC is committed to working in partnership with local, tribal, state, and Federal officials to serve critical emergency response needs. OIC's programs are unique in that they advocate a "bottom-up" approach. OIC's practitioner-driven governance structure gains from the valuable input of the emergency response community and from local, tribal, state, and Federal policy makers and leaders.

Long-Term Goals

Long-term goals for OIC include:

- Strengthen and integrate homeland security activities related to research and development, testing and evaluation, standards, technical assistance, training, and grant funding.
- Provide a single resource for information about and assistance with voice and data interoperability and compatibility issues.
- Reduce unnecessary duplication in emergency response programs and unneeded spending on interoperability issues.
- Identify and promote interoperability and compatibility best practices in the emergency response arena.

This page intentionally left blank.

Public Safety Communications Technical Report

Video Quality Tests for Object Recognition Applications

DHS-TR-PSC-10-09
September 2010

Reported for: The Office for Interoperability and Compatibility by the Public Safety Communications Research program

Homeland Security

This page intentionally left blank.

Publication Notice

Disclaimer

DHS' Science and Technology (S&T) Directorate serves as the primary research and development arm of the Department, using our Nation's scientific and technological resources to provide local, state, and Federal officials with the technology and capabilities to protect the homeland. Managed by S&T, OIC currently assists in the coordination of interoperability efforts across the Nation.

Certain commercial equipment, materials, and software are sometimes identified to specify technical aspects of the reported procedures and results. In no case does such identification imply recommendations or endorsement by the U.S. Government, its departments, or its agencies; nor does it imply that the equipment, materials, and software identified are the best available for this purpose.

Contact Information

Please send comments or questions to: SandT.CCI@hq.dhs.gov

This page intentionally left blank.

Contents

This page intentionally left blank.

Abstract

This report describes a laboratory study to investigate how the interaction of the following scene content parameters affect a viewer's ability to recognize a given target, or object, in the video stream:

- Target size
- Scene motion
- Scene lighting levels

Further, the report describes effects of the preceding scene content parameter combinations on object recognition with the following video processing procedures applied:

- Resolution reduction
- H.264 compression

The task-based subjective tests this report describes follow the test methods described in ITU-T Recommendation P.912 [1].

Key words: object recognition, video quality, subjective test methods

1 Introduction

The Public Safety Communications Research (PSCR) program[1]—in partnership with OIC—is conducting video quality research for public safety applications to determine performance specifications for certain network conditions required to provide minimum viewing levels of quality for video systems based on the specific needs of public safety practitioners and their applications. Collectively, this video quality research is known as the Public Safety Video Quality project.

Each public safety agency may have one or more very specific video applications. However, public safety video applications all share something in common at a higher level: performing a recognition task in which the viewer can recognize a desired target at a particular level of discrimination. Therefore, seemingly different applications may have similar quality requirements for video equipment. Upon closer examination, seemingly disparate video applications may actually have the same minimum requirements to perform their individual desired recognition tasks.

The Video Quality in Public Safety User Guide [2] defines a framework for describing a recognition task in terms of five parameters:

- Timeframe of use
- Discrimination level required
- Target size
- Motion in the scene
- Scene lighting

1. The PSCR program is a joint effort between the National Institute of Standards and Technology/Office of Law Enforcement Standards and the National Telecommunications and Information Administration/Institute for Telecommunication Sciences.

This report describes a laboratory study to investigate how the interaction of the following scene content parameters affect a viewer's ability to recognize a given target, or object, in the video stream:

- Target size
- Scene motion
- Scene lighting levels

Further, the effects of the preceding scene content parameter combinations on object recognition are studied with the following video processing procedures applied:

- Resolution reduction
- H.264 compression

The usage timeframe (i.e., video used for real-time applications versus recorded for later use) for this study was live or real time. Likewise, the discrimination level (i.e., the level of detail necessary to recognize a target or object) was positive recognition (e.g., a face, an object, or alphanumeric characters).

2 Experimental Method

The method used in this study followed [1], and the test conditions followed the recommendations in [3]. Viewers watched video clips at varying quality levels and performed specific recognition tasks, using the multiple-choice method. Next, viewers identified objects given a number of choices.

2.1 Scene Target Objects

The target item in a scene is the subject within the video frame that the viewer must recognize to perform the application task (e.g., face, alphanumeric characters, or object). Target recognition video (TRV) provides the ability to recognize specific targets of interest. The objects that were included in this test were:

- Gun
- Taser
- Radio
- Mug
- Soda
- Flashlight
- Cell phone

2.2 Scenario Groups

The test's video clips contain several scenarios. A scenario provides directions for the actions and contents of a scene (e.g., man walks by camera carrying an object). Because test measurements focus on a viewer's ability to identify objects and actions, the test plan addresses the possibility that a viewer may memorize the scene content and use other visual clues to remember the identity of the target. Therefore, instead of using one scene per test, each test uses a set of scenes (i.e., a scenario group) containing multiple scene versions, with controlled differences between the versions. For example, the scenario could include a

person who walks across the field of view carrying an object. The scenario group would consist of multiple clips using different objects or different people. The number of scenes in a scenario group should be large enough that scene memorization is unlikely.

Because this study focuses on object recognition tasks, the scenario groups consist of each object under test being used in various situations. The scene parameters under study are target size, motion, and lighting. Therefore, the scenario group designs create combinations of the parameters, as shown in Table 1.

Table 1: Summary of test scenario groups

Scenario Group	Motion	Clip Length	Field of View	Distance	Locale	Lighting Condition
stationary object	stationary	5s	23’ 6”	35’ 9”	outdoor	daylight
carried object: right	walking speed	6s	32’ 7”			
carried object: left	walking speed	6s				
stationary object	stationary	5s	48’ 11”	48’		
carried object: right	walking speed	9s	58’ 8”			
carried object: left	walking speed	9s				
carried object: right	walking speed	5s	12’ 8”	17’ 2”	indoor	bright/flash[a]
carried object: left	walking speed	5s				
stationary object	stationary	5s			indoor	dim/flash (lighting: 3.1 lumens)
carried object: right	walking speed	5s				
carried object: left	walking speed	5s				
stationary object	stationary	5s			indoor	dark/flash (lighting: 2.2 lumens)
carried object: right	walking speed	5s				
carried object: left	walking speed	5s				

a. “Flash” refers to the use of a law enforcement light bar to create the lighting condition for the scenario.

Figure 1 shows a still frame from one of the scenario groups. The object in this example is a gun.

Figure 1: Frame from the daylight/stationary object recognition scenario group

2.3 Clip Creation

The test clips were created and impaired using H.264 compression and resolution reduction. Original source sequences were filmed in high-definition (HD) video format with a frame size of 1920x1080 pixels and a frame rate of 29.97 frames per second (fps).

The seven objects listed in the previous section were used as targets. The objects were filmed as they sat on a pedestal, then filmed again as an actor carried the objects at walking speed. Some scenes were filmed with the camera at two different distances from the target object so as to change the object's apparent size.

As detailed in Table 1, there were 14 scenario groups with 7 source sequences in each (1 source sequence per target.) Two of the scenario groups did not use the flashlight as a target. Therefore, there were 96 original source sequences. Source sequences ranged from 5 to 9 seconds long.

The HD source sequences were down-converted to two display resolutions: Video Graphics Array (VGA—640x480 pixels) and Common Intermediate Format (CIF—352x288 pixels). The frame rate was constant at 29.97 fps. The MainConcept H.264 software encoder was used. For each of the two resolutions, the video was impaired by forcing various values for the encoder bit rate. Five bit rates were chosen for each resolution to represent a wide spectrum of resultant video quality.

A Hypothetical Reference Circuit (HRC) is a specific combination of video bit rate and display resolution. Ten total HRCs were tested. The total number of clips generated was 960. Each viewer watched three clips

from each scenario group for each HRC. The total number of clips seen by each viewer was 424, with four exclusively training clips. Table 2 summarizes the number of clips and values of each parameter.

Table 2: Summary of test design

Parameter	Number	Values
Target	7	Listed in Section 2.1
Scenario groups	14	See Table 1
Original clips	96	(7 Targets x 14 Scenario Groups)[a]
HRCs	10	(2 Resolutions x 5 Bit Rates)
Clips for tests	960	(Original Clips x HRCs)
Resolutions	2	VGA, CIF
Bit rates (kbps)	5 per resolution	VGA: 128, 256, 512, 1024, 1536 CIF: 64, 128, 256, 512, 1024

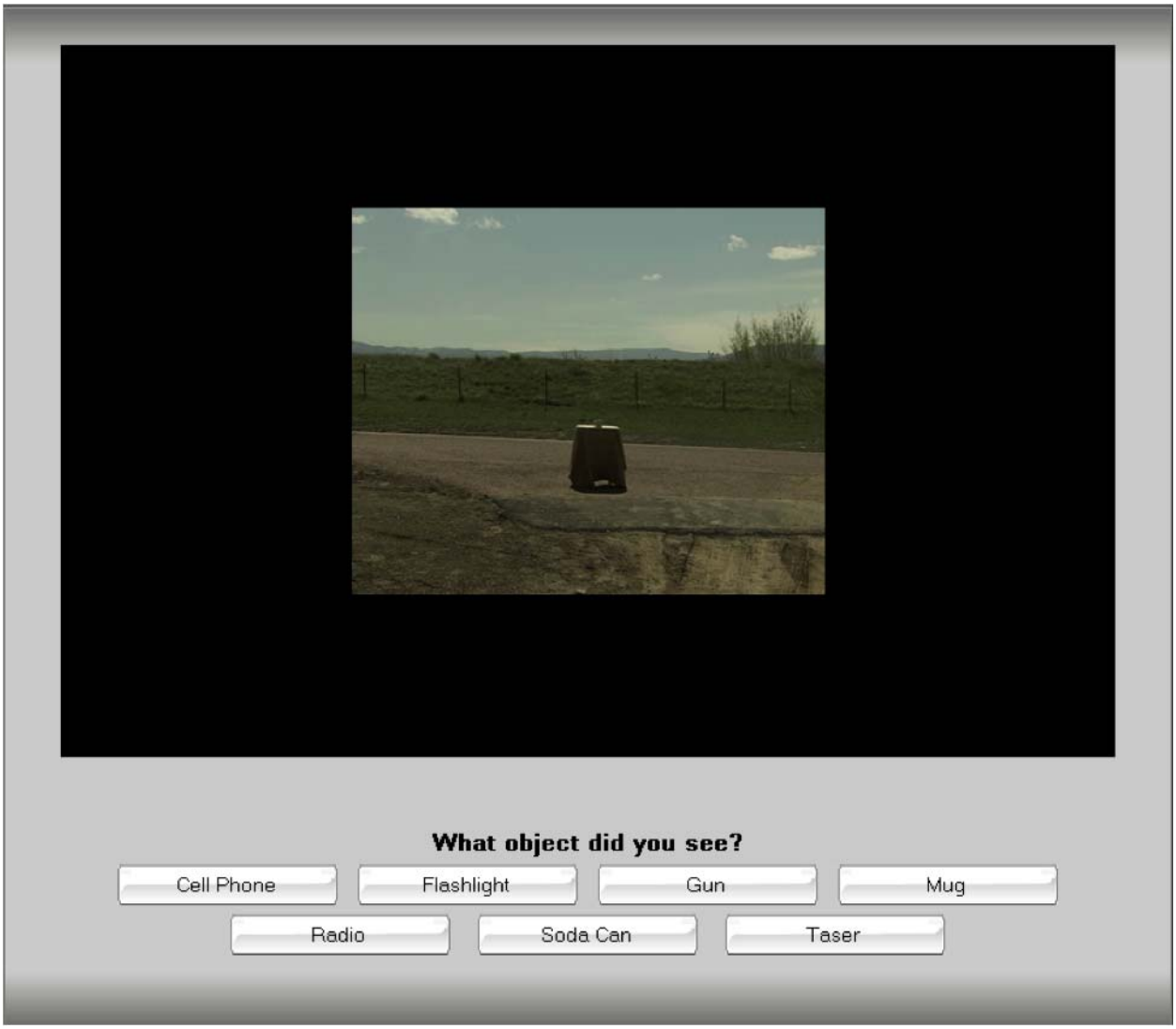
a. For two of the scenario groups, the flashlight is not used as an object, reducing the total number of clips.

Instead of viewing all of the clips, each viewer watched three clips from each scenario group for each HRC. Those items were selected in advance and distributed uniformly among the scenario groups and viewers, and each viewer saw a different randomization of the order of the clips. Every HRC was used for each clip.

2.4 Viewer Response

The viewers watched short video clips and performed a specific recognition task, using a multiple-choice format. For the multiple-choice format, a clip from a particular scenario group was shown above a list of written labels representing the possible answers. After the video was presented, viewers were asked to choose the label closest to what they recognized in the clip. Viewers were offered seven choices. The number of choices offered depended on the number of alternative scenes presented within the scenario group. An optional answer of “Unsure” was not offered [4]. Figure 2 shows the viewer response screen.

Figure 2: Media player with viewer response choices



2.5 Viewers

Thirty-seven viewers participated in this exercise. Each had employment experience in at least one area of public safety. Viewers' visual acuity and color vision were screened prior to inclusion in the test. Three viewers demonstrated impaired color vision; however, according to an analysis of their scores, they did no worse than other viewers of the test. For this reason, the results include their data.

2.6 Instructions for Viewers

Viewers received the following text as instruction:

> *Thank you for coming in today to participate in our study. This study concerns the quality of video images for use in Public Safety applications. As a likely user of next-generation devices for Public Safety applications, we are interested in whether the videos to be*

presented are of sufficient quality to be used by you to perform several different potential tasks.

Today's study examines video used in a live, real-time situation, and the ability to use this video to make real-time decisions on how to respond to an incident. This study does not apply to video which has been recorded for later examination. The application currently being focused on is object recognition. You will be asked to answer specific questions regarding content in the video. The scenes you will be shown, and the response requested, are from the following categories:

Scene Description	Response
Person walking by, holding an object *Lighting scenario* ■ *Indoor flashing lights* ■ *Indoor, dark, flashing lights* ■ *Outdoor, daytime*	*Multiple choice: Identify the object from a list*
Stationary objects *Lighting scenario* ■ *Indoor flashing lights* ■ *Indoor, dark, flashing lights* ■ *Outdoor, daytime*	*Multiple choice: Identify the object from a list*

Each scene will be approximately 7 seconds long. You will be shown the scene, then asked to answer the question relating to the scene as described in the table above. Since this study relates to real-time video applications, you will not be allowed to pause or replay the video.

**** Please wait for the video clip to finish playing before answering the question, and please do not close the media player window at any time during the test. ****

Multiple Choice Instructions

Please choose the answer that most matches what you saw in the video. For this study there is no "other" or "I don't know" option. Therefore, please select the answer you believe to be most likely.

You will be asked to participate in one viewing session which is approximately 90 minutes long. A practice session will be presented to help you get familiar with the scene material and rating process, as well as a clip showing the objects you might see in the videos. You may take a break at any time during the session.

2.7 Data Analysis

Data is reported as percentages of correct answers. For each aggregation of answers, each viewer was a sample.

Because guessing was likely, each score was normalized for the probability of a correct guess based on the following equation:

$$R_A = R - \frac{W}{n-1}$$

Where R_A is the adjusted number of right answers, R represents the number of right answers, W represents the number of wrong answers, and n represents the number of answer choices [5]. Ninety-five percent confidence intervals were calculated using the Clopper-Pearson method [6].

3 Results

The figures in this section represent the percent-correct data, calculated as described in Section 2.7, for each of the scenario groups. A single object that was carried left in a clip and carried right in another was calculated as one scenario group, with its left and right clips analyzed together. Therefore, the data for carried object scenario groups is based on twice as many data points as the stationary scenario groups.

3.1 Best Conditions

Figure 3: Results for stationary objects, in daylight, at nearer distance

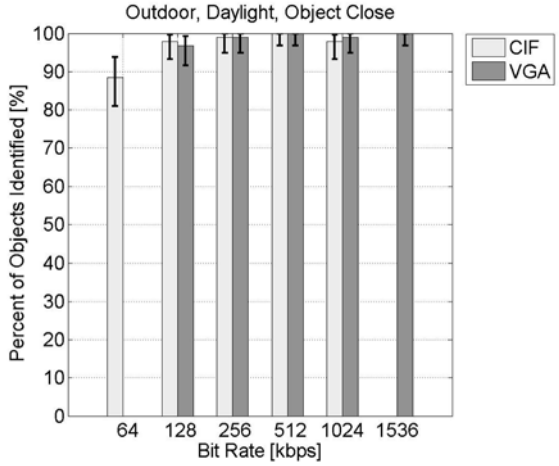


Figure 4: Results for carried objects, indoors under bright light, at nearer distance

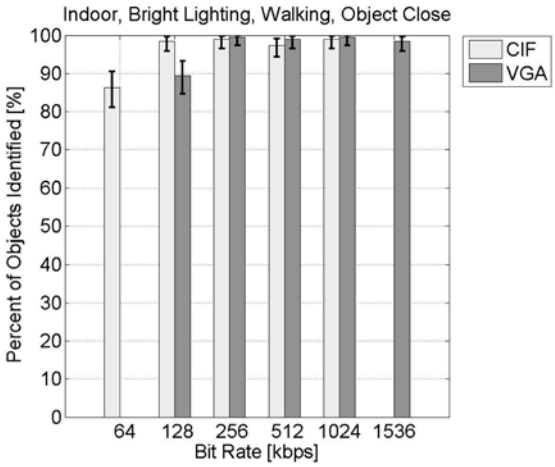
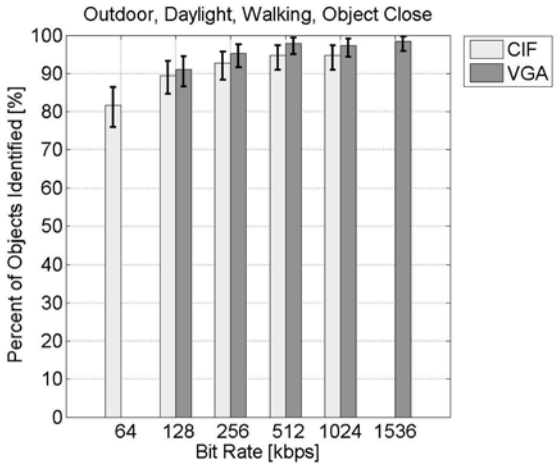


Figure 5: Results for carried objects, in daylight, at nearer distance



Under the best motion and target size conditions (i.e., stationary and nearer distance), viewers achieved nearly 100-percent recognition with outdoor lighting, as Figure 3 shows. Similarly, as Figure 4 shows, viewers achieved nearly 100-percent recognition with bright lighting indoors. In fact, Figures 4 and 5 demonstrate that at 256 kbps, bright indoor lighting outperforms sunlight; however, there is no data for bright indoor lighting with stationary objects. This makes a comparison difficult because one would normally wish to isolate the effect of lighting by comparing data gathered under the best possible conditions in terms of the other factors under test. However, both bright and outdoor lighting conditions show such high levels of recognition that either are considered sufficient for a recognition task.

3.2 Impact of Lighting

In contrast to the results showing the best lighting conditions, Figure 6 shows that dim lighting conditions may be insufficient for a recognition task. Here, the recognition levels never substantially exceed 90 percent—even as the bit rate is increased from 256 kbps to 1536 kbps. This implies that no amount of bandwidth allocated to the video transmission is enough to overcome the fact that the scene was poorly lit. Similarly, Figure 7 shows that in dark conditions with flashing lights, the recognition never substantially exceeds 80 percent; the result was observed even though this rate can be achieved at only 256 kbps for CIF resolution. From this, it could be concluded that improper lighting conditions can create a saturation effect where increasing the bit rate fails to increase recognition beyond a certain level. Generally, the data suggests that a distinction can be made between "enough lighting," such as bright and outdoor conditions, and "not enough lighting," which can cause these saturation problems. This effect, however, should be studied more closely before firmly concluding that such a binary distinction is appropriate.

Figure 6: Results for stationary objects, indoors with dim lighting and flashing lights, at nearer distance

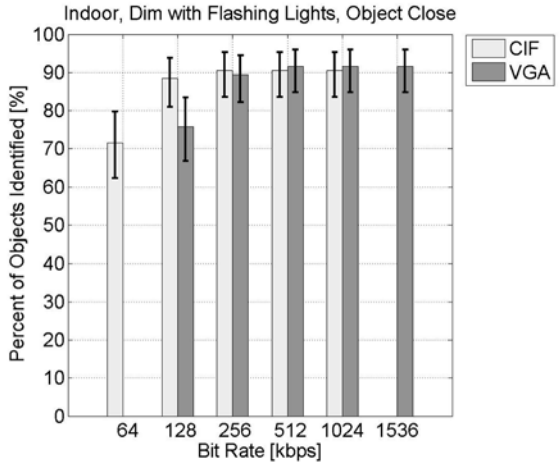
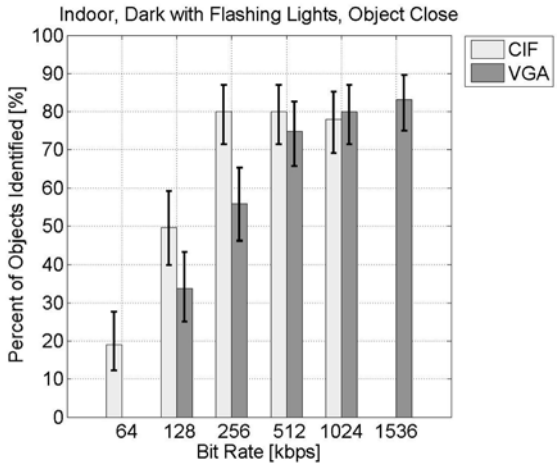


Figure 7: Results for stationary objects, indoors with dark conditions and flashing lights, at nearer distance



Another interesting aspect of the data with regard to lighting conditions is that the lower CIF resolution significantly outperforms the higher VGA resolution in bad lighting conditions. Figures 6, 7, 8, and 9 show the apparent benefit of lower resolution in poor lighting conditions. This observation is somewhat counterintuitive. Generally, logic suggests that higher resolution is better, but that depends very much on

what a particular video coder does with the additional information it is given. At this time, a definitive explanation for the apparent benefit of lower resolution in poor lighting conditions cannot be given.

Figure 8: Results for carried objects, indoors with dim lighting and flashing lights, at nearer distance

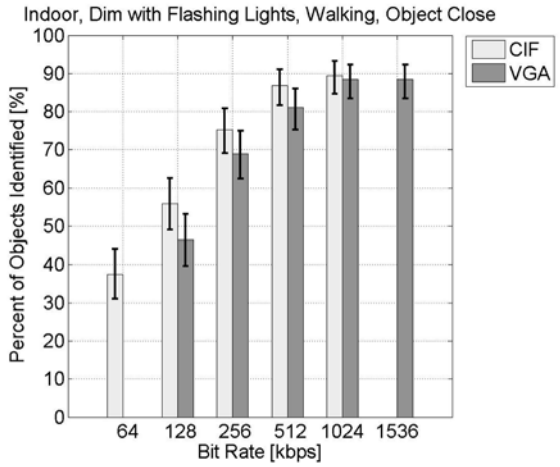
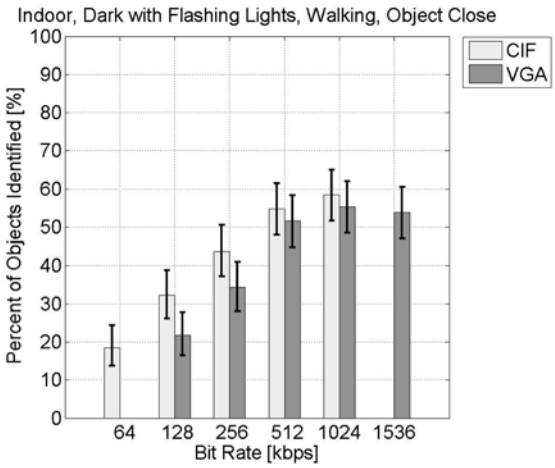


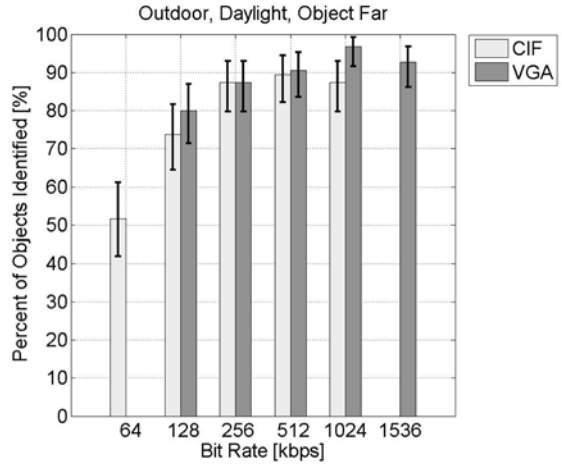
Figure 9: Results for carried objects, indoors with dark conditions and flashing lights, at nearer distance



3.3 Impact of Target Size

The saturation effect seen under certain lighting conditions can also be observed when analyzing the effect of target size. Figure 10 shows that under the best lighting and motion conditions, 512 kbps and even 1536 kbps do not provide significantly higher recognition rates than 256 kbps. There is a significantly higher recognition rate for 1024 kbps at VGA resolution, but not for CIF resolution. Generally, this figure suggests that recognition rates substantially higher than 90 percent cannot be reliably achieved for a small target, regardless of the video stream's bandwidth.

Figure 10: Results for stationary objects, outdoors, at greater distance

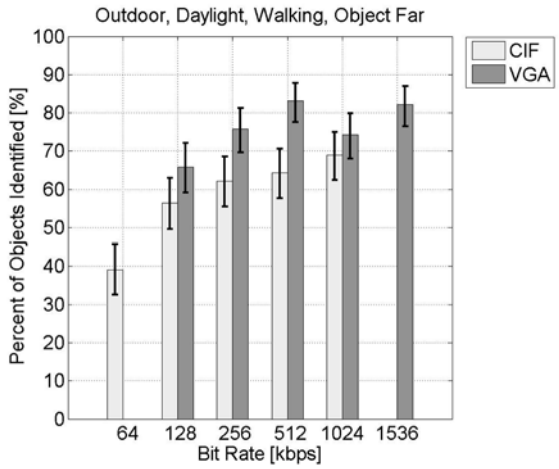


Logic suggests that higher resolutions would provide a substantial recognition advantage in the comparison of data between large and small target sizes. In the data, however, it is not entirely clear that there is any such advantage in the case of stationary objects. It may be the case that the advantage is obscured by the saturation effect Section 3.2 describes. In Section 3.4, Figures 5 and 11 show that when motion is present, higher resolution can act as a significant advantage for a small target size. However, the advantage that resolution offers for a large target is much smaller. This suggests that, given a particular target size, there is a resolution that is "sufficient," beyond which increasing resolution further would have limited benefits.

3.4 Impact of Motion

Generally, the data suggests that for low-motion video, positioning the camera so that it views large targets may be more important than using a high-resolution camera. However, high resolution can mitigate impairments that may result from high motion. The data is not entirely clear on this issue. A more rigorous study of the effects of resolution must be conducted to draw firm conclusions.

Figure 11: Results for carried objects, in daylight, at greater distance



For all lighting conditions and target sizes, the data shows that higher motion makes recognition more difficult than no motion. Figures 7 and 9 show that motion can reduce the recognition performance saturation level from about 80 percent to about 60 percent under flashing lights. However, Figure 5 shows that under the best possible conditions (i.e., outdoor lighting and a large target size), viewers can achieve 100-percent recognition with walking-speed motion. The data suggests that high motion may not place the same kind of limits on recognition that poor lighting and small target size can, but it can significantly worsen the effects of those conditions. Clearly, there is a non-linear relationship between the different scenario group variables, and the effects of those variables on recognition. The effect of motion combined with poor lighting is worse than just adding the penalties of the two separate effects.

3.5 Recommendations

The data suggests that increasing video resolution does not offer a clear-cut advantage. Based on these results, recommendations could be made for higher resolution under high-motion conditions and lower resolution under poor lighting conditions. Further study is required to determine the best possible resolution for each set of conditions. The data also shows that certain impairments can prevent reliable recognition at any bit rate and that even very low bit rates can be useful under proper lighting conditions, with a scene properly framed in the camera. Nevertheless, given a particular scenario and desired reliability, the data allows the formulation of bit rate and resolution recommendations. Table 3 summarizes these recommendations.

Table 3: Recommended bit rates for H.264 encoding

Scenario	Bit rate for 90 percent recognition		Bit rate for 50 percent recognition	
	VGA	CIF	VGA	CIF
Outdoor, stationary, large target	128 kbps	128 kbps	128 kbps	64 kbps
Outdoor, stationary, small target	512 kbps	512 kbps	128 kbps	64 kbps
Outdoor, moving, large target	128 kbps	128 kbps	128 kbps	64 kbps
Outdoor, moving, small target	N/A	N/A	128 kbps	128 kbps
Bright lighting and motion	128 kbps	128 kbps	128 kbps	64 kbps

Table 3: Recommended bit rates for H.264 encoding (Continued)

Scenario	Bit rate for 90 percent recognition		Bit rate for 50 percent recognition	
	VGA	CIF	VGA	CIF
Dim and flashing lighting and stationary	256 kbps	256 kbps	128 kbps	64 kbps
Dim and flashing lighting with motion	N/A	1024 kbps	256 kbps	128 kbps
Dark with flashing lights and stationary	N/A	N/A	256 kbps	128 kbps
Dark with flashing lights with motion	N/A	N/A	512 kbps	512 kbps

4 Limitations

This study did not include a scenario group for bright indoor lighting with stationary objects. As a result, comparison with outdoor recognition tasks will be difficult because the effect of lighting—by comparing data gathered under other test conditions—cannot be isolated. However, both bright and outdoor lighting conditions show high levels of recognition, and these conditions should be considered desirable for a recognition task.

5 Future Work

The research discussed in this report addressed the subjective effects of size, motion, and lighting on the ability to positively recognize targets under various compression rates. The subjective tests were designed to mimic live surveillance applications.

The next steps in this line of research are to:

1. Conduct a study focused on the recorded usage time frame under the same test parameters.
2. Extend the study to include the effects of the same test parameters for less stringent discrimination levels (e.g., recognizing broad target characteristics or elements of the action without requiring positive target recognition).
3. Study objective measurements of the loss of effective resolution under the same test parameters, using standard test charts instead of subjective testing processes.

6 Summary

The data validates, general expectations about which scenario groups would present greater difficulty for the object recognition task. The data bears the assumption that viewers should be able to recognize a close target easier than a target that is far away. Similarly, outdoor daytime lighting would likely provide the needed factors to make the object recognition task easiest, followed by bright lighting, dim lighting with flashing lights, and dark with flashing lights as the most difficult. The data shows that a slight advantage may exist for bright lights indoors versus outdoor sunlight; otherwise, the expected outcomes are confirmed. As for motion, motion blurring might make moving objects more difficult to recognize.

However, viewers may more easily recognize a moving object because they view it at different angles and sampled at different points in the pixel lattice—allowing a moving object to project more independent information than a stationary one. It is not obvious which of these two competing effects would be expected to dominate. The data for this experiment shows that viewers recognize stationary objects more easily than moving ones, indicating that motion blurring represents the dominant effect in this case. This is true for all target sizes and lighting conditions. This study allows the formulation of recommendations regarding bit rates required for various size, lighting and motion conditions of a video scene.

7 References

[1] ITU-T. 2008 August. Recommendation P.912. Subjective video quality assessment methods for recognition tasks. Recommendations of the ITU, Telecommunications Standardization Sector.

[2] *Defining Video Quality Requirements: A Guide for Public Safety,* Volume 1.0, July 2010. http://www.safecomprogram.gov/NR/rdonlyres/5BCA1CBF-1500-4B29-9370-81B823575DE8/0/3aVideoUserRequirementGuidedoc.pdf. Cited September 2010.

[3] ITU-R. 2002. Recommendation BT.500-11. Methodology for Subjective Assessment of the Quality of Television Pictures. Recommendations of the ITU, Radiocommunication Sector.

[4] D. Green, "Application of Detection Theory in Psychophysics," Proceedings of the IEEE, Vol. 58, No. 5, May 1970.

[5] ANSI S3.2, American National Standard Method for Measuring the Intelligibility of Speech Over Communications Systems, 1989.

[6] N. Johnson, S. Kotz, and A. Kemp. *Univariate Discrete Distributions*, p. 129, Wiley, New York, second edition, 1992.

This page intentionally left blank.