Validating Video Analytics in Mission Critical Applications

Stephen J. Mitchell[1], Sukhpreet Gill[1], Steve Loveless[1], and Brent Auernheimer[2]

[1] Pelco, 2500 Pelco Way, Clovis CA 93612-5699 USA
[2] Computer Science, California State University, Fresno CA 93740 USA
{smitchell,sukhpreet.gill,sloveless}@pelco.com,
brent@csufresno.edu

Abstract. Video Analytics (VA) automates and aids intelligent decision making in video surveillance applications. A video security system combines sophisticated video analytics with the oversight and judgment of human operators to save lives and critical resources. The complexity of the user experience and the range of deployments of VA make testing particularly challenging. Trade-offs between functional tests and focused usability tests are used to build a case for trustworthiness. This paper describes a method for validating video analytics for use in mission critical applications.

1 Introduction

Users depend on Video Analytics (VA) functionality in mission and safety critical situations like airports, banks, museums and casinos. Users need assurance that they can interact with VA reliably and efficiently during critical and non-critical times. However, VA functionality and interfaces can be daunting and users overwhelmed with the complexity of configuration, use, and verification. How can developers test a VA system to ensure users will correctly configure the product for their environment and be able to use this product in critical moments?

1.1 Video Security

The video security industry provides products to assure customers that they, and their valuables, are safe. Traditionally this meant that security personnel monitor a number of cameras that, in turn, monitors the valuables and records events. Security personnel observe multiple camera monitors simultaneously, which can lead to fatigue and errors. The hope and goal of VA is to augment the performance of security personnel [4].

1.2 Video Analytics

The purpose of VA is to "describe the real time use of computer vision in a security environment to monitor the CCTV camera feeds and assist the guard in his or her decision making process" [8]. VA provides role-based features to monitor digital video and communicate to security personnel when designated events occur. VA typically includes functions such as Adaptive Motion, Abandoned Object, and Object Counting:

J.A. Jacko (Ed.): Human-Computer Interaction, Part IV, HCII 2011, LNCS 6764, pp. 394–402, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Adaptive Motion detects and tracks objects that enter a scene, triggers an alarm when the objects enter or touch a user-defined zone, and monitors the objects until they exit the scene. Typical installations are banks, casinos, homeland security, the retail industry and public transportation.

Abandoned Object detects objects placed within a predefined zone and triggers an alarm if the object remains in the zone too long . Security personnel use this algorithm for detecting suspicious objects in a scene. Typical installations for this function are airports and public areas.

Object Counting calculates the number of objects that enter a defined zone or cross a virtual trip wire and then triggers an alarm once the count reaches a predefined number. Typical installations for this behavior include counting people (or vehicles) at an exit/entry door (or parking lot).

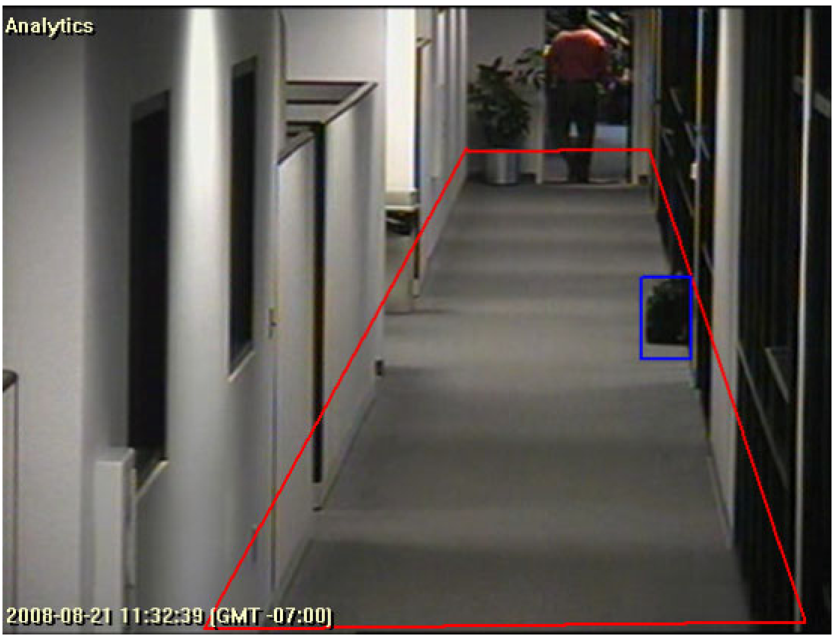

Fig. 1. An abandoned object

2 Validating Video Analytics: Two Techniques

We consider two techniques to build customer trust and assure users of correct VA operation: functional testing (component, then system level) and focused usability testing of the system in context. In order to provide high performance and a rich user experience, we recognize three elements of context: users, equipment, and environment.

After developers implement features they believe users need, functionality is tested at the component and system level. Then, after validating functionality of each component, and then components together as a system, testers need to validate the interaction of users and the system in varied environments. Focused usability testing fulfills this need.

Focused usability testing is inspired by ethnographic approaches to usability testing. Hollan et al. describe a distributed cognition approach to the interaction of users and complex systems saying that "Distributed cognition ... is specifically tailored to understanding interactions among people and technology" [2]. However, Don Norman cautions against unfocused approaches to this type of testing, saying that explorative ethnography helps in interpreting human behavior but cannot be applied to the real-world due to its dispersed nature [10]. We believe that focused usability testing not only helps determine if users can effectively use the system, but also ensures the system responds to diverse needs of users. The goal is to provide a rich user experience.

3 Problems Validating Video Analytics

While VA developers intend for their applications to be comprehensive and useful, we show later in this paper that users can find them complex and difficult to configure. Also, even after users configure the system to suit their environment, they can be misled and their expectations be faulty. Sometimes the VA application does not facilitate users developing their own assurance that they have successfully configured the functions for their purposes. Our experience leads us to believe that a combination of user interface improvements as well as user preparation is missing. Developers and testers must consider this throughout the software development process.

3.1 Functionality Does Not Suffice

Functionality for VAs is, by its nature, tied to a real-world scene. VA cannot operate without a scene. The process of validating functionality consequently leads to examining how customers configure the system for their environment. When VA requires configuring, the software experts (usually the software developers) put the performance of the product in to the hands of the novice (the software user).

3.2 Chaos in Context of Use

Context of use can be defined as "users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used" [3] (see Fig. 2). The context of use for VA systems varies greatly due to the chaotic nature of the environment (scene).

While VA developers can document required general environmental conditions for operation, it is difficult to be precise, concise, and yet capture the range of real-world environments. Users are therefore required to interpret what real-world environmental conditions apply to them, and then ensure that they control their environment as much as possible (such as camera placement and choice of scenes). This minimizes the variability of the context of use. After making proper adjustments for their environment, users then have to configure their VA system to capture the data that they need.

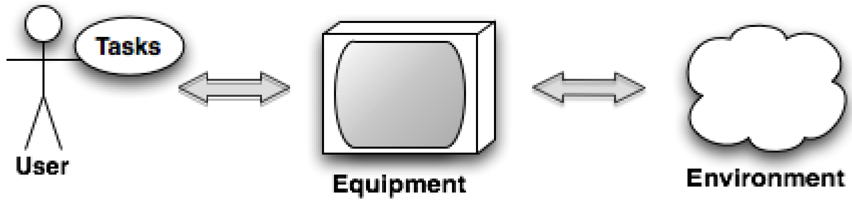

Fig. 2. Context of Use

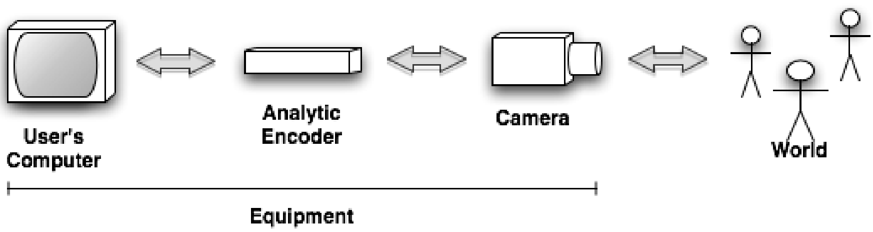

Fig. 3. An Environment for Video Analytics

A characteristic of video systems is they provide users an interface to a collection of equipment, which itself interfaces to the real world (see Fig. 3). Events in the real world may differ day to day, complicating the context of use. Worst case, these real-world scenarios are chaotic.

4 Applying Standards to Context of Use

We recognize three elements of context of use which play important roles in the performance of VA: users, equipment, and environment.

4.1 User: Training and Experience

Don Norman says "People form mental models through experience, training and instruction. The mental model of a device is formed largely by interpreting its perceived actions and its visible structure" [9]. User interfaces, like those for VA, can be complex, creating a potential mismatch between the developers' implementation model and the users' mental models. Training and experience help form a user's initial mental model.

We provided training to users in the form of face-to-face training as well as through written instructions and manuals. In the former method, trainers configure VA on a representative scenario and then give the user control to configure the system. Training ends with an open-ended session where users voice questions and concerns. Manuals are also provided.

We also consider constructivism when assessing the performance of VA users. Martin Dougiamas defines constructivism as "building on knowledge known by the student" [7]. That is, users learn not only by instruction but also by experimentation. Users tweak parameters to make their own discoveries and conclusions.

But because users' environments vary greatly, users cannot solely depend on training to prepare them for achieving their goals for VA. Users must have an interface that directs them and provides feedback so they gain confidence in the system.

4.2 Equipment: Usability Study

When conducting our VA usability study we followed guidelines in NISTIR 7432 [5]. This standard provides information required to conduct usability tests and serves as a test protocol. After recruiting users, and attaining test scripts and a test system, moderators carry out the usability test. Once users have executed the tasks, the moderator gathers data from the users through a questionnaire and one-on-one interviews.

4.3 Environment: i-LIDS

In order to control context-of-use variables, we constrain the test environment. The UK government's Home Office Scientific Development Branch offers standardized sets of videos, "i-LIDS", for this purpose [1]. i-LIDS provides the precise conditions of the environment the video represents as well as how the VA algorithm under test should behave. The description of how the VA should behave is a performance standard for each VA function. Our usability studies used i-LIDS videos representing low chaotic environments so that users in the studies will have a higher potential for positive results.

4.4 Standards

IS0 9241-11 emphasizes that "Quality In Use" (QIU) depends on context of use and that the level of QIU achieved depends on the specific circumstances in which a product is used [3]. Also, ISO 9126 states that in order to ensure customer satisfaction, quality must be established at each phase of the development of a product [3]. "Internal" quality characteristics are tested when defining the product. For example, requirements and specifications fulfill metrics that were a result of considering desired quality characteristics. Testers then apply "external" quality characteristics to the product to ensure that developers implemented requirements and specifications properly [4].

In order to better assess if customers will be satisfied, we applied ISO 9126 QIU quality characteristics. QIU testing is similar to usability testing in that testers observe users with the product to determine if goals for the product are met [3]. When testers measure QIU, they are effectively measuring how well internal and external quality characteristics were applied for the end user [4]. ISO defines four software QIU characteristics within a specified context of use:

- *Effectiveness*: enabling users to achieve specified goals with accuracy and completeness.
- *Productivity*: enabling users to expend appropriate amounts of resources in relation to effectiveness.
- *Safety:* achieving acceptable levels of risk of harm to people, business, software, property or the environment.
- *Satisfaction:* satisfying users.

5 Applying "Quality in Use"

Because the performance of VA depends on the ability of the user to configure the analytics, we emphasize that user effectiveness is essentially tied to overall success.

5.1 Our Quality in Use Test

QIU tests how well users are able to configure the system so that it will detect an event defined by the VA function. We measured the performance of users and VA functions as a single system in terms of the QIU characteristics listed in Section 4. Metrics were used to compare the number of actual events in a video to how well the VA algorithm detected the events given a user's configuration. The VA industry

Table 1. Metrics Used Applying QIU (based on [4])

Characteristic	Metric	Measurement	Interpretation
Effectiveness	1. User/System Effectiveness 2. Error frequency	1. X = A/B 2. Y = C/B A = number of true positives B = number of ground truths C = number of false positives	1. 0 <= X <=1 The higher the better 2. 0 <= Y <= 1 The lower the better.
Productivity	1. Task time 2. Economic productivity	1. X = mean time taken by users to complete assigned tasks 2. Y = A/B A = task effectiveness B = total task cost	1. 0 minutes <= X <= 15 minutes The lower the better 2. 0<=Y The higher the better
Safety	Safety of people affected by the system	X = 1-A / B A = number of false negatives B = number of ground truths	0 <= X <= 1 Closer to 1 is safer.
Satisfaction	1. Satisfaction scale 2. Satisfaction questionnaire	1. X = Rating on Likert scale [4] 2. Y = A/B A = positive comments B = total comments	1. 0 <= X <=5 5 is completely satisfied. 2. 0 <= Y <= 1 The higher the better.

refers to actual events in a video as "ground truths", correctly identified events are referred to as "true positives", incorrectly identified events are referred to as "false positives" (i.e., if reported event was not an actual event), incorrectly unreported events are referred to as "false negatives" [14].

In order to verify the VA interface is simple to use, and to discover errors and areas for improvement, two rounds of usability studies were conducted.

Round One. We tested both expert and novice users. We expected expert users to have higher values of effectiveness and productivity than novice users, but after conducting the first round of usability study this was not the case. Results indicated that although the productivity of expert users was double the productivity of novice users, there was not much difference between the effectiveness of expert and novice users.

We explained the higher productivity of expert users by the novices' learning curve [6]. The possible explanation for lower effectiveness of expert users relative to novice users might be explained by the Hawthorne effect [5]. We observed that novices were more concentrated and performed each task carefully whereas experts overlooked some of the simple configuration parameters that play an important role in VA. Novices may have the perception that their performance was being judged and hence performed conscientiously, whereas the level of attention given by experts was low and therefore their performance was poor. We also speculate a reason for low performance was due to the use of more chaotic video scenes for experts and less chaotic videos for novices.

When analyzing recorded screens and videos of users in round one of usability testing, we discovered a common set of mistakes committed by most users that resulted in the decrease of effectiveness of the system. We concluded that specific training might address these errors and make a difference in performance. We conducted another round of usability testing using 12 novice users, where six were "normally" trained and another six were "specially" trained.

Round Two. Because we found interface and functionality problems in the Object Counting algorithm, we decided to test Adaptive Motion (first) and Abandoned Object (second) for this round of testing. We also decided to more tightly control our test environments (which are part of the context of use for the test), thus widening the gap between a "specially trained" group and the "regularly trained" group. Regularly trained users were provided basic instruction to acquainting them with the look, feel, and operation of the VA system. They were instructed about the parameters in the setup that could affect the functionality of the VA functions. Specially trained users not only received "regular" training, but also were given specific knowledge about tweaking each parameter in the configuration and the effect on VA performance.

5.2 Quality in Use Results

Again, we measured the performance of users and analytics together as a system using metrics of QIU characteristics.

Effectiveness. The results indicated that effectiveness of specially trained users was 40 percent higher than that of trained users for both Adaptive Motion and Abandoned Object algorithms. Error frequency for specially trained users was 2 percent less than that of trained users. This indicates that specially trained users are more apt to properly respond to real events.

Productivity. Productivity of specially trained users was 27 percent greater than that of normally trained users. Also, when we compared productivity for Adaptive Motion and Abandoned Object, it was higher for Abandoned Object for both trained and specially trained users as they had more experience while conducting tasks on the Abandoned Object algorithm (all 12 users performed tests on Adaptive Motion first and then on Abandoned Object). This confirms that productivity increases with experience.

Safety. False negative counts could represent occurrences where something bad might have happened but users weren't alerted. When comparing results from our QIU testing, specially trained users' environments were 51 percent safer than the regularly trained users' configurations.

Satisfaction. When we analyzed satisfaction, we found that trained users were four percent less satisfied than specially trained users. The satisfaction questionnaire indicated that specially trained users were 14 percent more satisfied than trained users. Not surprisingly, satisfaction goes hand in hand with performance: users are more satisfied if they can use the system effectively.

6 Conclusion

As with any software product, testers must thoroughly validate VA before product release. In order to ensure product readiness for a variety of users in diverse environments, QIU testing for VA shows that functional testing is not sufficient to show product readiness. QIU shows that VA developers should train their users in a controlled environment so that users learn how the software interacts with various environments. Unprepared users will not have confidence in a newly acquired VA system. Also, if we only train users on how to use the equipment, they stand a higher risk of using the system poorly. Users must be trained on their entire context of use. If users understand their environments in relation to their equipment, they mitigate variability (similar to our use of i-LIDS in our QIU testing). If customers are deploying VA in a mission or safety critical environment, such as an airport or casino, lack of training could mean missed opportunity for saving lives or resources.

We've shown the importance of end-user training for the successful VA deployment. This research will help guide the development of user training as well as VA with affordances to enable successful operation

Acknowledgements. Pelco is the video security brand of Schneider Electric, the global specialist in energy management. Pelco is fully integrated within the Schneider Electric Buildings Business Unit. The work illustrated in this publication is the result of ongoing product development in the Buildings Business Unit's Video Product Development department and is supported by the Building Business Unit's Solutions and Technology Office. Thanks to Greg Millar, CTO and Sr. VP of Solutions and Technology, and to Ray Still, Director of Systems Software for supporting this work.

References

1. i-LIDS Team: Imagery Library for Intelligent Detection Systems (i-LIDS): A Standard for Testing Video Based Detection Systems. In: Proceedings 2006 40th Annual IEEE International on Carnahan Conferences Security Technology, pp. 75–80. IEEE Press, Los Alamitos (2006), doi:10.1109/CCST.2006.313432
2. Hollan, J., Hutchins, E., Kirsh, D.: Distributed cognition: toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction 7, 174–196 (2000)
3. ISO/IC 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. International Organization for Standardization (1998)
4. ISO/IEC 9126. Software engineering - Product quality. International Organization for Standardization (2002)
5. NISTIR 7432. Common Industry Specification for Usability - Requirements Annex C. National Institute of Standards and Technology. IUSR: CISU-R v0.90, pp. 22–24 (June 2007)
6. Kitchenham, B.A., Sadler, C.: Evaluating Software Engineering Methods and Tools Part 4: The influence of human factors. ACM SIGSOFT Software Engineering Notes 21, 11–13 (1996)
7. Dougiamas, M.: A Journey Into Constructivism (1998), http://dougiamas.com/writing/constructivism.html#intro
8. Video Content Analysis, What Is It and Why Would I Want It? http://smartcctv.blogspot.com/2006/11/video-content-analysis-what-is-it-and.html
9. Norman, D.A.: The Design of Everyday Things. MIT Press, Cambridge (2002)
10. Norman, D.A.: Workarounds and hacks: the leading edge of innovation. Interactions 15, 47–48 (2008)
11. Pelco. Endura Abandoned Object Behavior Operation/Configuration manual C3602M (2007)
12. Pelco. Endura Adaptive Motion Behavior Operation/Configuration manual C3600M (2007)
13. Pelco. Endura Object Counting Behavior Operation/Configuration manual C3603M (2007)
14. Yin, F., Makris, D., Velastin, S.: Performance Evaluation of Object Tracking Algorithms. In: Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. In: PETS 2007 (2007), http://www.cvg.rdg.ac.uk/PETS2007/data.html