TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

On evaluating video surveillance systems

Aleksandra Karimaa

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 71 | TECHNICA – INFORMATICA | TURKU 2025

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

ON EVALUATING VIDEO SURVEILLANCE SYSTEMS

Aleksandra Karimaa

TURUN YLIOPISTON TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. F OSA – TOM. 71 | TECHNICA – INFORMATICA| TURKU 2025

University of Turku

Faculty of Technology
Department of Computing
Information and Communication Technology
The Doctoral Programme in Technology (DPT)

Supervised by

Associate Professor Tuomas Mäkilä
University of Turku

Professor Ville Leppänen
University of Turku

Reviewed by

Associate Professor Mikołaj Leszczuk
AGH University of Science and Technology, Poland

Professor Marcin Grzegorzek
University of Lübeck, Germany

Opponent

Associate Professor, Jussi Kasurinen
LUT University, Finland

ISBN 978-952-02-0480-8 (PRINT)
ISBN 978-952-02-0481-5 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/ Online)
Painosalama, Turku, Finland 2025

To my sister

UNIVERSITY OF TURKU
Faculty of Technology
Department of Computing
Information and Communication Technology
ALEKSANDRA KARIMAA: On Evaluating Video Surveillance Systems
Doctoral Dissertation, 156 p.
The Doctoral Programme in Technology (DPT)
December 2025

ABSTRACT

This dissertation considers the problem of systems evaluation in the context of video surveillance systems. While multiple definitions of evaluation exist, one can agree that the goal of the evaluation is to reflect the "value" of the system. However, in case of surveillance systems, there is a lack of consensus on what objective should be measured to reflect this value, and most of objectives, such as video quality, target detectability or performance are qualitative or complex to measure.

My hypothesis states that every video surveillance system can be evaluated in a quantitative way reflecting evaluation needs of its stakeholders across the lifecycle of the system. Based on presented existing research as well as own publication works, I have proposed a new, holistic framework for video surveillance system evaluation - a metrics-based scoring evaluation system. Finally, the scoring system is presented by two example case studies. To create the most holistic view on system value I have studied different aspects of surveillance systems and researched what evaluation methodologies and metrics can be used for other similar systems and their applicability to video surveillance systems. I have focused on metrics and methodologies that can reflect different contexts of use and solutions, as well as satisfy the needs of various stakeholders. I have demonstrated how such set of metrics can be utilized to build scoring system and demonstrated how example set of metrics can be used to reflect various system properties and illustrate the final performance. Finally, I have demonstrated the proposed scoring system concept by modelling and analysing existing solutions. This work is expected to support innovation development in the industry providing a concept that could be used for evaluating new solutions from both business, technology, and research innovation potential perspectives. It can provide common evaluation vocabulary improving collaboration between research and business and therefore improving the speed of commercialization of innovations. The motivation for this research is based on my experience from my position as Product Manager at Teleste Corporation, where I was responsible for development and commercialization of VMX surveillance system. The work has been supported with literature research, technology analysis, relevant patents analysis, as well as video surveillance market research.

KEYWORDS: Video Management Systems, Video Surveillance, Evaluation, Efficiency, Metrics, Scoring system

TURUN YLIOPISTO
Teknillinen tiedekunta
Tietotekniikan laitos
Tietotekniikka
ALEKSANDRA KARIMAA: Videovalvontajärjestelmien arviointi
Väitöskirja, 156 s.
Teknologian tohtoriohjelma (DPT)
Joulukuu 2025

TIIVISTELMÄ

Tämä väitöskirja käsittelee järjestelmäarvioinnin ongelmaa videovalvontajärjestelmien kontekstissa. Vaikka arvioinnille on olemassa useita määritelmiä, voidaan olla yhtä mieltä siitä, että arvioinnin tavoitteena on heijastaa järjestelmän "arvoa". Valvontajärjestelmien tapauksessa ei kuitenkaan ole yksimielisyyttä siitä, mitä tavoitetta tulisi mitata tämän arvon heijastamiseksi ja useimmat tavoitteista, kuten videon laatu, kohteen havaittavuus tai suorituskyky, ovat kvalitatiivisia tai monimutkaisia mitata.

Hypoteesini on, että jokaista videovalvontajärjestelmää voidaan arvioida kvantitatiivisesti, mikä heijastaa sen sidosryhmien arviointitarpeita järjestelmän koko elinkaaren ajan. Esitettyjen olemassa olevien tutkimusten ja omien julkaisujen perusteella olen ehdottanut uutta, kokonaisvaltaista viitekehystä videovalvontajärjestelmien arvioinnille - metriikkaan perustuvaa pisteytysjärjestelmää. Lopuksi pisteytysjärjestelmä esitetään kahdella esimerkkitapauksella. Luodakseni mahdollisimman kokonaisvaltaisen kuvan järjestelmäarvosta olen tutkinut valvontajärjestelmien eri näkökohtia ja selvittänyt, mitä arviointimenetelmiä ja -mittareita voidaan käyttää muissa vastaavissa järjestelmissä ja niiden sovellettavuutta videovalvontajärjestelmiin. Olen keskittynyt mittareihin ja menetelmiin, jotka voivat heijastaa erilaisia käyttökonteksteja ja ratkaisuja sekä tyydyttää eri sidosryhmien tarpeita. Tämän työn odotetaan tukevan innovaatioiden kehitystä toimialalla tarjoamalla konseptin, jota voitaisiin käyttää uusien ratkaisujen arviointiin sekä liiketoiminnan, teknologian että tutkimuksen innovaatiopotentiaalin näkökulmista. Se voi tarjota yhteistä arviointisanastoa, mikä parantaa tutkimuksen ja liiketoiminnan välistä yhteistyötä ja siten nopeuttaa innovaatioiden kaupallistamista.

Tutkimuksen motivaatio perustuu kokemukseeni tuotepäällikkönä Teleste Corporationilla, jossa vastasin VMX-valvontajärjestelmän kehittämisestä ja kaupallistamisesta. Työtä on tuettu kirjallisuustutkimuksella, teknologia-analyysillä, asiaankuuluvien patenttien analysoinnilla sekä videovalvonnan markkinatutkimuksella.

ASIASANAT: Videonhallintajärjestelmät, videovalvonta, arviointi, tehokkuus, mittarit, pisteytysjärjestelmä

Table of Contents

Abbreviations **8**

List of Original Publications **9**

1 Introduction **10**

2 Research approach **14**

3 Status quo and literature review **19**
3.1 On video surveillance evaluation 19
3.1.1 Video surveillance evaluation standards and programs 21
3.2 On video surveillance evaluation objectives 22
3.2.1 Quality objectives – measuring system attributes. 23
3.2.2 Performance and efficiency objectives- measuring system functions 24
3.2.2.1 Design, architecture and system complexity 24
3.2.2.2 Source data 26
3.2.2.3 Data acquisition and processing 29
3.2.2.4 User interface and workflow support 30
3.2.2.5 System intelligence 32
3.3 On multi-dimensional evaluation 34

4 Evaluation framework proposal – scoring system design **36**
4.1 Rationale for framework 36
4.2 Evaluation framework requirements 38
4.3 Evaluation framework design 40
4.3.1 Scoring specification at a glance 42
4.3.2 Results uncertainty 42
4.3.3 Ethics and compliance consideration 43
4.4 Evaluation framework 43
4.4.1 Quality metrics 43
4.4.1.1 Technology Readiness 44
4.4.1.2 System availability 45
4.4.2 System design performance 47
4.4.2.1 System Software Complexity 47
4.4.2.2 Video Network utilization 48
4.4.2.3 End to End Latency 48
4.4.3 Source data performance 49

4.4.3.1 Object resolution 49
4.4.3.2 Efficient Frame rate 50
4.4.3.3 Encoding/Decoding Delay 52
4.4.4 Data acquisition and processing 53
4.4.4.1 Data acquisition - Precision 53
4.4.4.2 Data acquisition - Recall 54
4.4.4.3 Data acquisition –MODA 55
4.4.4.4 Data acquisition –MOTA 56
4.4.5 User interface efficiency 57
4.4.5.1 User interface friction 58
4.4.5.2 Workflow– Clicks to Achieve 59
4.4.6 System intelligence 59
4.4.6.1 Efficient Operation - False alarms 59
4.4.6.2 Efficient Operation – Probability of Detection 60
4.4.6.3 Learning- speed of learning 61
4.5 Original publications contribution 61

5 Testing hypothesis-case studies 65
5.1 Case study of deployed system 65
5.1.1 Calculations 66
5.1.1.1 System Quality 66
5.1.1.2 System Design performance 67
5.1.1.3 Source Data performance 69
5.1.1.4 Data acquisition performance 70
5.1.1.5 User Interface Efficiency 71
5.1.1.6 System Intelligence 72
5.1.2 Results and analysis 72
5.1.3 Calculations 75
5.1.3.1 System Quality 75
5.1.3.2 System Design performance 77
5.1.3.3 Source data 79
5.1.4 Data acquisition 79
5.1.4.1 User Interface 79
5.1.4.2 System Intelligence 80
5.1.5 Results and summary 80
5.2 Reflection to hypothesis and Research Questions 83

6 Conclusion 85
6.1 Academic conclusions 85
6.1.1 Results uncertainty 86
6.1.2 Ethics and compliance consideration 87
6.2 Practical implications 88
6.3 Future work 90
6.4 Summary 91

List of References 92

Original Publications 95

Abbreviations

AI	Artificial Intelligence
CCTV	Closed Circuit Television
FAR	False Alarm Ratio
FAT	Factory Acceptance Test
HR	Hit Ratio
IP	Internet Protocol
IR	Information Retrieval
IT	Information Technology
MODA	Multiple Object Detection Accuracy
MOTA	Multiple Object Tracking Accuracy
MTBF	Mean Time Between Failures
NIM	Native Intelligence Metric
NIST	National Institute of Standards and Technology
ONVIF	Open Network Interface Forum
PC	Personal Computer
POD	Probability of Detection
PSIA	Physical Security Interoperability Alliance
PTZ	Pan Tilt Zoom
TRL	Technology Readiness Level
UAT	User Acceptance Test
UI	User Interface
UIF	User Interface Friction
VMS	Video Management System

List of Original Publications

This article-based thesis is built upon the following original publications, which are referred to in the text by their Roman numerals:

I Karimaa, Aleksandra. Security Aspects of the Complexity of Modern Surveillance Systems - An Experience Report. *Proceedings of 14th IEEE International Conference on Engineering of Complex Computer Systems*, 2009 (Karimaa, 2009).

II Karimaa, Aleksandra. Video Surveillance in the Cloud: Dependability Analysis. *The Fourth International Conference on Dependability DEPEND2011*, 2011 (Karimaa, 2011a).

III Karimaa, Aleksandra. Efficient Video Surveillance: Performance Evaluation in Distributed Video Surveillance Systems. *Video Surveillance*. 2011 (Karimaa, 2011b).

IV Karimaa, Aleksandra. Mobile and Wireless Access in Video Surveillance System. *International Journal of Digital Information and Wireless Communications*, 2011; Vol. 1(No. 1): 267-272 (Karimaa, 2011c).

V Karimaa, Aleksandra. Rating Functional Design: Evaluation and Scoring for Systems with Video Sources. *International Journal of Soft Computing and Software Engineering*, 2013; Vol. 3. (Karimaa, 2013a)

VI Karimaa, Aleksandra. Value-Aware Approach to Management of Innovative Software Products and Services. *Journal of Economics, Business and Management*, 2013; Vol. 1: pp. 66-71 (Karimaa, 2013b).

VII Karimaa, Aleksandra. A Survey of Hardware Accelerated Methods for Intelligent Object Recognition on Camera. In: Swiątek, J., Grzech, A., Swiątek, P., Tomczak, J. (eds) *Advances in Systems Science. Advances in Intelligent Systems and Computing, Proceedings of the International Conference on Systems Science 2013 (ICSS 2013)*, 2014; Vol. 240: pp. 523-530. Springer International Publishing Switzerland. (Karimaa, 2014)

1 Introduction

The original motivation for this research was based on my professional experience from video surveillance industry. In 2005-2012 I was working as a Product Manager in Teleste Corporation 2005-2012, where I was responsible for the development and commercialization of video surveillance products.

I noticed that (in opposite to other industries at that time) the video surveillance struggled to form successful universal multi-stakeholder multi-dimensional evaluation programs. It was especially difficult due to the qualitative nature of data (video) and functionality (situational awareness).

At the time of preparing this work the industry developed significantly. Across only 20 years, old CCTV (Closed Circuit Television) were replaced by local video surveillance installations and then expanded into wide network based distributed video surveillance solutions. As video surveillance systems developed, the software heart of the system, became the product in itself. In the industry, this software heart of the system is called Video Management System (VMS). A VMS is a software solution providing manage and control video surveillance system. It is responsible for system operations, viewing, record, and storing across the system. VMS enables central view to system capabilities, such as central control, integrations and capabilities such as analytics, motion detection, or facial recognition. As VMS matured, became more independent from underlying hardware and networks, it started to play central role in evaluation of system capabilities. Despite this technical progress and system technology transition, there was very little progress when it comes to availability of evaluation standards that would work for all types of systems, stakeholders and across systems lifecycle.

This work considers the problem of systems evaluation in the context of video systems. My hypothesis states: *Every video surveillance system can be evaluated in quantitative way reflecting evaluation needs of various stakeholders across various stages of system life cycle*.

This hypothesis is investigated by researching the following research questions:

> **RQ1: What aspects of evaluation are important when building a framework for video surveillance evaluation?**

RQ2: What system characteristics should be subject of evaluation, and can they be quantified?

RQ3: How to form the evaluation framework reflecting multiple dimensions of system "value" as relevant for various stakeholders, use cases and valid across various stages of system life cycle?

These questions are explored in frame of research on this work as well as in scope of my earlier publications. While unarguably the key contribution of this thesis is proposal of the **framework for multi-dimensional multi-stakeholder quantitative evaluation of video surveillance systems**, this contribution was built on three phases of earlier research.

During the first phase of the research, I focused on understanding the **background of video surveillance evaluation**, researching academy and industry needs for evaluation programs and the landscape of available evaluation standards and programs. I aimed to understand the requirements for future framework as well as explore and structure the taxonomy of evaluation to be used in context of video surveillance.

Second phase of research was focused on **evaluation objectives** and their measurability and metrics selection. The publications from this period aimed to explore evaluation objectives as relevant for industry and academia stakeholders as well as relevant for various cases. This research was built on background knowledge from previous phase, so the evaluation objectives (system characteristics or attributes) can be understood in the context of stakeholder or use case needs. Special attention has been put into evaluation quantification, so the research focus was on the objective measurability and metrics selection and their relevancy to various stakeholders and use cases. Some examples of publications from this period include:

- Publication I *Security Aspects of the Complexity of Modern Surveillance Systems: An Experience Report* (Karimaa, 2009) evaluating system complexity and its security aspects.
- Publication II *Video Surveillance in the Cloud: Dependability Analysis* (Karimaa, 2011a) exploring topic of system quality changes measuring system dependability when moving a system into a cloud and/or virtual environments.
- Publication III *Efficient Video Surveillance: Performance Evaluation in Distributed Video Surveillance Systems* (Karimaa, 2011b) identified areas critical for system performance where performance is introduced as major characteristics of system efficiency next to system functionality and cost.

This phase was concluded in 2013 with Publication V *Rating Functional Design: Evaluation and Scoring for Systems with Video Sources* (Karimaa, 2013a), which provided the first draft version of **quantitative evaluation framework.**

While second phase of the research contributed to building deeper understanding of selected evaluation dimensions, it concluded in observation that the evaluation of complex and qualitative-data systems (as video systems) is in fact of **multidimensional nature**. Conclusively, the third phase of my research (2011-2015), the publications contributed to building understanding of multi-dimensional nature of evaluation. This phase of research was opened by publication on multi-dimensional aspect of value. In Publication VI *Value-Based Software Product: Value assessment and management for complex software products* (Karimaa, 2011) I explored how multi-dimensional value of system is built combining different aspects of values of software solutions and its specific functional areas. The work laid the fundaments for current multi-dimensional evaluation concept and provided motivation for creating multi-stakeholder evaluation frameworks. Monitoring of system for business and technical development as well as contributes to better understanding of innovation aspects helping multiple stakeholders to cooperate in building better future solutions.

This three-phase research produced a baseline for evaluation framework, which was then tested and presented with two case studies.

In 2015 I left video surveillance industry and research focusing on other data systems. After video systems I have worked with other situational awareness and qualitative data-based systems, including gas spectrum sensory systems or satellite imagery. These experiences made me believe that the challenges described in this work and scoring system framework are more universal and worth expanding to other qualitative systems and the work is worth finalizing.

I resumed this research in 2023, motivated by reflection on universal character of the evaluation framework. I revisited the research, and I have updated the status quo of existing academia research as well as evaluation standardization attempts in the industry. This longer research process gave me possibility to adapt the framework to be more universal, and the dimensions (or evaluation characteristics) were restructured to reflect changes in taxonomy.

The content of this thesis is structured as described below. First, I explained the research approach, describing research concepts and strategies, as well as methodologies used in this work and justification for selection of these. Next, the research status quo is described providing review of knowledge relevant for research questions and connected areas. Following research on questions the hypothesis stated in this work is validated with demonstration of an evaluation concept in form of the scoring system based on set of balanced metrics. Finally, the scoring system is demonstrated by two example case studies.

This thesis is an article-based dissertation. However, it is not presented using typical article-based dissertation style, but it is mostly organized in monograph format to summarize better contributions of individual publications.

No artificial intelligence has been used in this work.

2 Research approach

This chapter describes research concepts and strategies, as well as methodologies used in this work.

The aim of this work is to investigate if it is possible to build an evaluation framework satisfying expectations of industry and academia as defined in the hypothesis:

> **Every video surveillance system can be evaluated in quantitative way reflecting evaluation needs of various stakeholders across different stages of system life cycle.**

The topics relevant for research questions were studied through combining desk and field research. Academic approach to video surveillance evaluation was studied based on desk research and included review of existing research as well as discussions with research groups working with similar topics. The knowledge about the evaluation of video surveillance systems in the industry was gathered mainly by means of combining desk research (such as industry solutions benchmarking) and field research as well as via cooperation in the industry. The research was explored as described in this work, but also in my earlier publications (refer to List of Publications).

The Figure 1 illustrates research approach defining how research was formed and executed. It presents the research phases relating to research questions, as well as visualizes how research phases and their results contributed to hypothesis. Figure 1 also presents the basic structure behind forming of video surveillance evaluation framework.

Figure 1. Research approach.

The first phase of the research aimed to answer Research Question *RQ1: What aspects of evaluation are important when building a framework for video surveillance evaluation?* This phase of the research was impacted by opportunities given by my role in the industry. As a Product Manager I had an opportunity to not only study, use and benchmark existing video surveillance evaluation programs and standards, but also to discuss existing evaluation approaches with professionals from both industry and academia. It gave me the opportunity to have a direct user experience in video surveillance evaluations but also to verify why some evaluation programmes were not usable. Based on this research the evaluation framework requirements were created. In this stage of the research, I have also performed literature research and other background research to better understand evaluation requirements in the context of both research and the industry. Some background publications were created to research different aspects of surveillance systems, including, Publication IV *Mobile and Wireless Access in Video Surveillance* (Karimaa, 2011c) and also later published Publication VII *Survey of Hardware Accelerated Methods for Intelligent Object Recognition of Camera* (Karimaa, 2014).

Second phase of the research was focused on evaluation objectives, i.e. system characteristics, functionalities and other aspects. This phase research resulted in multiple publications summarizing and presenting in context different aspects of evaluation. Publication I *Security Aspects of the Complexity of Modern Surveillance Systems: An Experience Report* (Karimaa, 2009) explored **security** aspects of evaluation, Publication II *Video Surveillance in the Cloud: Dependability Analysis* (Karimaa, 2011a) focused on system's **dependability.** Publication III *Efficient Video Surveillance: Performance Evaluation in Distributed Video Surveillance Systems* (Karimaa, 2011b) researched **performance** evaluation.

Publications as well as their feedback from industry and academia allowed me to identify evaluation focus and objectives as relevant for different stakeholders and different use cases. This research was performed based on desk research, practical experience from different evaluation programs, own research publications and was followed with discussions with experts from both industry and academia. This allowed me to identify which aspects of video surveillance systems should be captured in evaluation framework so they would be satisfying existing evaluation needs of stakeholders. Additionally, I have reviewed these aspects from perspective of their usability in different phases of the life cycle and this way the framework would work similarly independently on the life cycle of the system.

When researching system characteristics and functionalities to be considered as evaluation objectives, I have investigated how these evaluation objectives can be measured and what metrics are typically used. My aim was to focus on metrics, which are general enough to be independent of the details of system design yet reflecting the impact of selected system design on system value. Also, selected

metrics should be easy to gather and analyse with less complex tools, so the system would be easy to use by possibly large fora of stakeholders.

Selection of metrics used in evaluation framework has been developed with the following principles:

- Metric role in evaluation of specific objective: how well it is reflecting specific objective? Will the improvement of the metric result in improvement of the objective value?
- Metrics that can be gathered and analysed using simple acquisition and analysis methodologies and well-available tools. In this study I have presented metrics collected by means of activity logging of VMX system and network traffic analytics (for example for studying delays or network utilization). Additionally, I have used direct observation for example for measuring data acquisition efficiency by comparing number of events detected by system with this detected by observation. Direct observation can provide quantitative data to describe also qualitative values.

Third phase of the research was focused on **capturing multi-dimensional** aspect of system value. Combining the results and experiences from previous phases of the research I aimed to addressing research question RQ3: *How to form the evaluation framework reflecting multiple dimensions of system "value" as relevant for various stakeholders, use cases and valid across various stages of system life cycle?* The research on multi-dimensional nature of evaluation was performed via research on multi-dimensional aspect of existing evaluations standards and programmes and relevant publications. This research phase was opened by Publication V *Rating Functional Design: Evaluation and Scoring for Systems with Video Sources* (Karimaa, 2013a), where I described the first proposal of functional evaluation aiming to combine earlier researched evaluation characteristics the research entered a next phase searching for a way to combine multiple evaluation characteristics.

Another publication, Publication VI *Value-Aware Approach to Management of Innovative Software Products and Services* (Karimaa, 2013b) summarized this phase of research on value-based software concept and multi-stakeholder approach to evaluation of software product. This publication was relevant to validate concept against its usability for reflecting relevant aspect of value for multiple stakeholders.

It is also worth to mention another research exploring management aspects of multi-dimensional software products in scope of my MBA thesis: *Value-Based Software Product: Value assessment and management of complex software products* (Karimaa, 2011). This work, while not directly contributing to this research, allowed me to understand better practical managerial implications of having multi-dimensional evaluation for technical products.

The result of the research on multi-dimensional aspects of evaluation concluded with building the draft of evaluation framework as described in Chapter 4.

Finally, based on the research performed I have validated the hypothesis by presenting evaluation framework tested by two cases studies. The use of framework has been demonstrated on example of **use case studies methodology**. The strategy to use case studies in validation phase of the research was motivated by the need to reflect the implementation context and illustrate evaluation framework. Utilization of case studies research in this work was inspired by Robert Yin (Yin, 1994 (2003)) case study research that where case study was described as "an empirical inquiry that investigates a contemporary phenomenon within its real-life context". According to Robert Yin the case study research, and its capabilities to reflect contextual conditions, has clean advantage over other "contextual" research methods, such as an experiment where the context is strictly controlled (they are executed typically in controlled laboratory environment), or surveys capturing very limited number of variables and responders. Broader definition of case study was described by Robert Yin in (Yin, 1994 (2003)), where the case study inquiry (1) copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result; (2) relies on multiple sources of evidence, with data needing to converge in a triangulating fashion, and as another result (3) benefits from the prior development of theoretical propositions to guide data collection and analysis.

This definition reflects also the concept presented in this work by incorporating different aspects of evaluation creating one universal common result (1), with multiple metrics contribution (2) and benefits surveillance evaluation process participants experience as well as the role of previous evaluations in excelling the development of evaluation toolbox provided in this work (3). The last argument also suggests that the evaluation concept proposed is a skeleton of evaluation framework that will be improved based on learnings from previous evaluations, which is one of motivations in publishing this work. Similar conclusions were published in Publication VI *Value-Aware Approach to Management of Innovative Software Products and Services* (Karimaa, 2013b) creating a baseline for an evaluation framework design.

3 Status quo and literature review

This chapter reviews the existing state of knowledge in the subject of evaluating video surveillance systems and literature review relevant in answering research questions.

First, Section 3.1 describes developments and status quo of evaluation concepts in the industry and academia. The aim is to provide understanding of the challenge of forming the requirements for evaluation framework and the answer to **RQ1: What aspects of evaluation are important when building a framework for video surveillance evaluation?** This overview helps drafting the requirements for forming the evaluation framework as described in Section 4.2 Evaluation framework requirements.

Next, Section 3.2 focuses on research on evaluation objectives providing the overview of knowledge relevant for answering research question **RQ2: What system characteristics should be subject of evaluation, and can they be quantified?** The aim is to capture evaluation objectives relevant for specific stakeholders, use-cases, and life-cycle stages.

Finally, Section 3.3 illustrates industry and academia attempts in forming multi-dimensional aspects of evaluation, aiming to answer the question **RQ3: How to form the evaluation framework reflecting multiple dimensions of system "value" as relevant for various stakeholders, use cases and valid across various stages of system life cycle?**

3.1 On video surveillance evaluation

It is widely understood that system evaluation is beneficial for both business and research domains—but does it serve both effectively? From a business perspective, system evaluation provides input on how to develop a system in terms of functionality, performance, and cost to meet customer requirements and expectations. From a research perspective, system evaluation offers insights into what can be improved to create state-of-the-art solutions.

If the same evaluation framework is used by both industry and academia, it would create an opportunity to better understand the business value of research efforts in potentially promising areas. At the same time, it would clarify where concentrated

research efforts deliver the greatest business impact. Ultimately, such a shared evaluation approach would establish a platform for communication between business and research communities regarding the development and value of future solutions. The more stakeholders an evaluation system can support, the greater its usability. System efficiency analysis should also offer guidelines and verification tools for selecting critical edge components that contribute to managerial qualities of the system (Ulrich, 1995).

Despite these benefits, there are no clear recommendations or standards for evaluating video surveillance systems. There are no established methods for assessing proposed designs, benchmarking multiple solutions, or even defining how system design objectives should be analysed. A good example of this gap is visible in Calls for Tenders. The requirements outlined in these documents often contradict each other, as they are based on "state-of-the-art" features compiled from various commercial solutions. Such a feature set and system requirement description do not guarantee optimal system performance or efficiency.

It is also generally agreed that having a unified evaluation approach across different phases of the system life cycle is beneficial. Early evaluation of system efficiency should help identify significant risks of failing to meet objectives and enable timely corrective actions. However, in the case of surveillance systems, there is no reliable method for modelling overall system "value" during the early development stages. Evaluation in later stages may also fail to reflect the true efficiency, quality, performance, or other intended metrics of the implemented system. Furthermore, the design phase—critical for determining system performance and efficiency—is entirely absent from existing evaluation tools. As a result, there is no universal evaluation toolbox, and in some phases of the product life cycle, evaluation is simply not feasible.

In the later stages of system design, some evaluation is conducted through Factory Acceptance Tests (FAT) or User Acceptance Tests (UAT). However, these tests are performed only after a specific solution has been selected, and their objective is limited to verifying whether predefined use cases are fulfilled (pass/fail).

Ultimately, the more adaptable an evaluation framework is to different solution designs—both current and future—the more useful it becomes for supporting diverse cases and comparing individual surveillance systems. Even when solutions are commercially available and evaluation should be straightforward, it is often difficult to compare offerings from different providers. Comparisons are typically based on individual features claimed by vendors and shaped by the subjective experience and knowledge of individual consultants. Relevant field tests are often not feasible, and available results tend to be unreliable and "over-marketed".

Based on the current state of research and experience with industry evaluation programs and standards, I believe it is possible to develop more precise evaluation

tools that support the needs of multiple stakeholders and are applicable across various phases of the system life cycle.

3.1.1 Video surveillance evaluation standards and programs

The importance of standardized system evaluation has already been recognized in the industry, as most major surveillance systems today form part of national security infrastructures. This is especially relevant for distributed systems that provide surveillance of cities and roads, as well as monitoring systems that secure public spaces, buildings, and institutions. Several organizations aim to provide national guidelines for evaluating such surveillance solutions, with leading efforts coming from the USA, United Kingdom, and Australia.

The Australian Security Industry Association (ASIAL) developed the grading and certification standard AS 2201.2 for assessing monitoring center performance and security. However, this standard focuses on compliance with basic security and performance aspects—such as intruder detection time, physical security of facilities, and recording storage duration—and therefore cannot be used for system benchmarking or development purposes.

In the UK, international standards BS EN 62676 have been developed based on input from several organizations, including the British Security Industry Association (BSI). These standards provide a framework for grading the set of functionalities offered by a system. They are used in the industry to benchmark system functionalities and assess the level of basic performance and security features. However, they are not designed to support development, improvement, or innovation, making them relevant for procurement but less useful to the research community.

In the USA, multiple organizations have developed standards and frameworks for evaluating video surveillance and security systems. NIST has published cybersecurity and physical security standards applicable to video management systems (NIST SP 800-171). The National Fire Protection Association (NFPA) has issued the NFPA 731 standard for the installation and maintenance of security systems. Underwriters Laboratories (UL) developed the UL 2802 standard for the performance and safety of video surveillance systems. Additionally, the Department of Homeland Security (DHS) has released various guidelines and handbooks, such as the CCTV Technology Handbook, which provide comprehensive information on video surveillance technologies.

The NIST standard for security assessment is particularly noteworthy. Although its framework serves a different purpose—evaluating the severity of vulnerabilities rather than system characteristics—it is a strong example of an evaluation framework built around a scoring system. While scoring systems were initially viewed as

unreliable and overly simplistic, the evolution of NIST's scoring system into a sophisticated evaluation program demonstrates their potential. The development of this standard inspired me to explore scoring systems as an evaluation framework and to believe that a simple scoring system can be refined and expanded through practical use.

Between 2005 and 2010, several evaluation programs were created to provide testing setups for improving the performance of specific functions such as retrieval and detection. Nghiem et al. (A. T. Nghiem, 2007) introduced the ETISEO framework for evaluating system performance in terms of video-related metrics and the dependency between algorithms and video characteristics. In the same year, Kasturi et al. developed a benchmark for the VACE (Video Analysis and Content Extraction) program, initiated by the United States Government (Disruptive Technology Office, 2007), to support video retrieval performance testing. The CLEAR program (CLEAR , 2007) provided a foundation for evaluating video data acquisition. In 2010, the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance was organized, offering multiple evaluation frameworks, testing setups, and ground truth data. This phase of research was well summarized by Mabrouk et al. (Mabrouk, 2018), who provided an overview of evaluation measures and compared their effectiveness for specific aspects of system performance in selected video surveillance functionalities. The publication proposed a framework for evaluating and documenting performance, but it focused on specific functions within the system.

Additionally, since video surveillance systems can be considered a specialized type of IT system, IT system evaluation standards may also be applicable. If a video surveillance system is treated as an IT system, its quality can be evaluated using the ISO/IEC TS 25011:2017 quality models. This standard is part of the System and Software Quality Requirements and Evaluation (ISO/IEC, 2017) framework and is specifically designed for multi-stakeholder evaluation. The proposed quality model can be used by providers and developers to identify, evaluate, and improve quality; by customers to specify quality requirements and benchmark services; and by other stakeholders to assess IT service quality objectively. The IT service quality model and its taxonomy are widely used in the industry to evaluate both new and existing services.

3.2 On video surveillance evaluation objectives

In this section, I identify evaluation objectives and relevant research. My aim to provide a description of evaluation approaches expanded with a review of theories, ideas and concepts in topic in surveillance system evaluation.

Most of the available research focuses on the selected system characteristics, such as efficiency, performance, or dependability. It is interesting to point out that while taxonomy used when evaluating video surveillance system in the industry is based on concepts like **functionality, security and retrieval and recording performance**, the taxonomy used in research arena is using **system efficiency, system/component/functionality performance or system dependability or (video) quality**.

These video surveillance evaluations can be classified into two categories:

- Evaluations focused on quality attributes (security, quality, dependability, usability). The goal of this evaluation is to measure system attributes.
- Evaluations focused on performance and efficiency targeting specific functional objectives of the video surveillance system, typically concerning video retrieval and processing. The goal if such evaluation is to measure system functional performance.

3.2.1 Quality objectives – measuring system attributes.

The taxonomy used for system quality evaluation in academic research aligns with the ISO/IEC TS 25011:2017 quality models described in the earlier chapter. These models consist of the following attributes: Suitability, Usability, Security, Reliability, Tangibility, Responsiveness, Adaptability, and Maintainability.

While reliability, availability, safety, integrity, and maintainability are often presented as separate quality attributes, research on video surveillance systems typically refers to **dependability**, which encompasses multiple aspects of these qualities. Other quality attributes are less frequently represented in the research, but they can be illustrated using metrics from IT system evaluations. This may explain why they are not considered particularly relevant in the context of video surveillance. Based on these observations, I have included dependability as a quality characteristic in the evaluation and focused specifically on availability, as security is addressed by the video surveillance standards mentioned in earlier chapters.

Additionally, we identified another quality aspect relevant to the life cycle of video surveillance systems. Intuitively, system value depends on solution maturity, as the introduction of new, untested technologies may reduce the accuracy of efficiency estimations. Although system maturity is not typically considered a quality attribute, it does impact system quality and performance design. Therefore, it should be considered when determining whether the framework supports different stages of the product life cycle. A similar concept was proposed by Meystel et al. (Meystel, 2007), where Technology Readiness Levels (TRL) were used to evaluate intelligent systems. TRL was originally introduced by NASA in 1995 (Mankins,

1995) and has since been adopted by multiple organizations, including the U.S. Department of Defense. The higher the TRL of the technologies used in a solution, the greater the system maturity and the more accurate the evaluation. To reflect the impact of maturity on video surveillance quality, I have included it as a quality attribute.

3.2.2 Performance and efficiency objectives- measuring system functions

Most evaluations of video surveillance systems focus on the performance or efficiency of specific system functionalities. The available research on functional evaluation in video surveillance concentrates on system functionalities that are essential from the perspective of surveillance applications, as described further in this Section.

3.2.2.1 Design, architecture and system complexity

Efficiency of computer systems is one of the most well-researched areas in the field of IT. Improvements in the performance and efficiency of computer networks are naturally motivated by the goal of providing cost-effective solutions for corporate environments. The architecture of computer systems and networks can be evaluated using general metrics such as availability, scalability, channel capacity, utilization (bandwidth and throughput), end-to-end latency, and service or response time. However, the evaluation should also offer insights into challenges such as system expansion or downsizing and their impact on performance.

In modern surveillance systems, considerable attention is given to embedded devices such as cameras, encoders, and mobile sensors. These devices are designed to be mobile, fit into limited spaces, and operate across a wide range of temperatures. This design introduces constraints, such as limited battery life (in mobile devices) and heat distribution issues (in compact devices), which affect their computational capabilities. When the capabilities of embedded devices are limited, core system functionality and intelligence mechanisms rely primarily on the resources of industrial PCs. In such cases, surveillance systems can be considered as groups of networked computers. This is expected to improve in the future, and I anticipate that a significant number of architecture evaluation methods will need to be developed for distributed embedded surveillance systems. Nevertheless, core system functionalities and intelligence mechanisms will likely continue to rely on networks of PCs, making it reasonable to consider surveillance systems as a subset of computer systems from an architectural efficiency perspective.

Software architecture efficiency is critically important when evaluating architectural performance. Modern software-only and service-oriented business models are independent of hardware platforms. The earlier the assessment is performed in the software life cycle, the less expensive and faster it is to identify potential risks and implement solutions to mitigate or eliminate them.

A comprehensive review of evaluation methods for software efficiency is provided by Woodside et al. (Woodside, 2007), Sharma et al. (Sharma, 2005), Hauck et al. (Hauck, et al., 2009), Jun-Tao and Xiao-Yuan (Jun-Tao, 2009) , and Meneely et al. (Meneely, 2013). While most research focuses on early evaluation, Woodside et al. (Woodside, 2007) illustrate two approaches: early-cycle predictive evaluation, which is model-based, and late-cycle evaluation, which is measurement-based. This demonstrates that it is possible to provide evaluation instruments that cover the entire lifecycle of a developed system.

Software architecture evaluation is typically based on architectural modelling, where software complexity is simplified by dividing the platform into layers or functional components. A good example of such methodology is presented by Sharma et al. (Sharma, 2005), who proposed a formal model-based software evaluation approach. This methodology includes defining environmental parameters and a layered model for system architecture. The environment is then modelled (e.g., using network parameters), including its limitations, and finally, a performance model and its outputs (performance factors) are proposed. Jun-Tao and Xiao-Yuan (Jun-Tao, 2009) propose system evaluation models based on UML collaboration and sequence diagrams. Hauck et al. (Hauck, et al., 2009) challenge traditional evaluation methods, arguing that they are unsuitable for complex systems and unreliable for early evaluation. Traditional methods tend to be one-dimensional, primarily aiming to prevent the implementation of low-quality systems. They do not reflect the complexity and multi-dimensionality of modern environments, such as operating systems and virtual machines, which form the foundation of software component applications. As a result, Hauck et al. present an extension to architecture evaluation models that enables accurate performance modelling and prediction for systems in complex environments, including those using disk arrays, virtual machines, and application servers.

However, the above architecture evaluation methods are labour-intensive and complex, making them impractical for commercial evaluations. Inverardi et al. (Inverardi, et al., 2009) propose a method that automatically derives a performance evaluation model from software architecture specifications. While the approach is interesting, it is not applicable in the context of surveillance systems.

System design broadly defines a system's ability to evolve, scale, and support new functionalities. In the case of video systems, this may involve aspects such as data and media stream distribution or internal system communication—both of

which critically impact final system performance. Moreover, as emphasized by Ulrich (Ulrich, 1995), system design significantly influences product performance, adaptability to change, and product development management, all of which are of critical managerial importance. This means that design decisions directly affect system quality attributes, including performance, maintainability, and availability.

3.2.2.2 Source data

The most important components providing source data to video systems are cameras and encoders. The efficiency of these devices in delivering video compression, content processing, or object accuracy defines and limits the system's capabilities to support such functions. Additionally, camera capabilities—such as intelligence and computational power—can influence the flexibility of system design and impact transmission, power consumption, and resource management. Therefore, efficiency evaluation should intuitively reflect, for example, the high performance of modern intelligent heterogeneous sensor networks, where multiple sensing functions are interconnected and deliver only relevant data to the system. Moreover, the efficiency of data source devices should correspond to their functional purpose. For instance:

- An efficient camera used for incident detection should have resolution or frame rate settings optimized to detect objects or events effectively, even under changing environmental conditions. This may require dynamic resolution switching.
- The efficiency of a camera used for identification and classification of license plates depends on its ability to accurately "read" the plate.
- A camera responsible for providing data for evidentiary purposes may require accurate colour reproduction and high resolution.

Most often, the quality of video data is measured by image resolution. Generally, higher resolution implies better image quality. However, in the context of functionality, the best image quality does not necessarily equate to the best efficiency. As resolution increases, demands on computational power, bandwidth, and storage also rise proportionally, which can negatively affect overall system efficiency. Therefore, optimizing efficiency requires careful trade-offs, such as those guided by the Johnson criteria methodology (Sjaardema, et al., 2015). Johnson's criteria define the minimum resolution required to detect, recognize, and identify specific object types with 50 percent probability (refer to **Table 1**).

Table 1. Johnson criteria.

	RESOLUTION PER MINIMUM DIMENSION			
TARGET BROADSIDE VIEW	Detection	Orientation	Recognition	Identification
TRUCK	0.90	1.25	4.5	8.0
TANK	0.75	1.2	1.5	7.0
HALF-TRACK	1.0	1.5	4.0	5.0
JEEP	1.2	1.5	4.0	5.0
COMMAND CAR	1.2	1.5	4.3	5.5
SOLDIER (STANDING)	1.5	1.8	3.8	8.0

The Johnson methodology is based on empirical tests, and similar tests have since been repeated by various research groups, including those in the field of video surveillance. It contributed to the establishment of multiple standards for approximate linear object resolution used in detection, recognition, and identification, based on the size of objects defined in meters of their smallest dimension. Table 2 provides an example of these recommendations. **Figure 2** illustrates the impact on image quality to better demonstrate its effect.

Table 2. Image quality recommendations for face and register plates systems.

IMAGE QUALITY	OBJECT RESOLUTION
IDENTIFICATION FOR FORENSICS	500 pixels/m^2
IDENTIFICATION	250 pixels /m^2
RECOGNITION	100 pixels /m^2
DETECTION	20 pixels /m^2

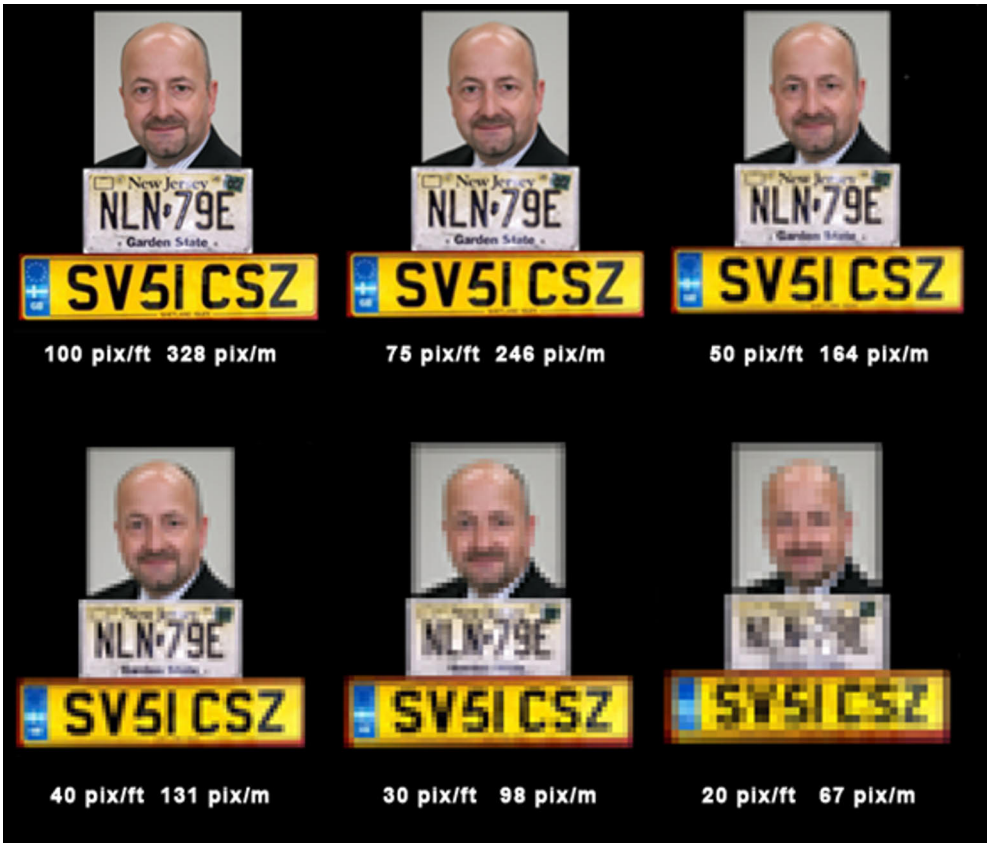


Figure 2. Image quality for different levels of pixel/m.

The rate at which images are produced is, alongside image quality, the second most important factor in determining material quality. In many cases, a frame rate of 25 fps (frames per second) is specified for surveillance systems, as it corresponds to the "live rate" used in television systems. However, with the emergence of megapixel cameras in the surveillance market, this video data parameter requires efficiency optimization. Such cameras do not always support 25 fps capture, and the resulting video streams are significantly larger, making transmission and storage design particularly challenging.

Therefore, frame rate optimization of source data is critical to overall system efficiency and should be adjusted according to the type of motion expected to be captured. In most applications, less than 12 to 15 fps is sufficient. For example:

- 25–30 fps for casinos or other environments where forensic identification may be required
- 12 to 15 fps for cash registers

- 5fps for indoor spaces, such as schools or offices
- 1 to 3 fps for outdoor areas, such as parking lots or streets

A third factor relevant to data source efficiency is the selection of compression standards and configuration of compression parameters. The choice of compression standard affects network design, data quality, and storage requirements.

It is important to note that the interdependencies between video quality, screen resolution, and bitrate must be considered, as these factors introduce additional challenges in designing an efficiency scoring system

3.2.2.3 Data acquisition and processing

In the context of video surveillance systems, data acquisition refers to the process of gathering relevant data. In most surveillance systems, data is retrieved from multiple types of sources, such as video and image content sensors, hardware sensors, contact closure sensors, as well as audio and biometric sensors. Although identifying relevant data is often subjective, there are well-established metrics available for evaluating data acquisition processes

There are multiple methodologies for assessing the efficiency of data acquisition systems. The most widely used are “ground truth”-based methodologies, where acquired data is compared to a reference dataset—also known as “ground truth data.” In video surveillance, this typically refers to manually identified data related to an event or object of interest. Ground truth-based methodologies can be applied to almost any data source, but they are most commonly used for evaluating video and image acquisition, as manual identification of video content is still easier for humans than for machines. However, because ground truth data must be created manually, this methodology is generally applied to individual video or image samples rather than entire libraries.

To address this limitation, some organizations provide pre-defined test datasets and sequences, along with benchmarking tools for acquisition system evaluation (IEEE International Workshop on Performance Evaluation of Tracking and surveillance, 2010), These tools allow users to input data into their acquisition systems and compare the system’s output with ground truth data using automated benchmarking.

Data acquisition efficiency is typically measured using precision (the proportion of relevant data within all retrieved data), recall (the probability that a randomly selected relevant item is retrieved), and F-measure or precision-recall curves, which combine both precision and recall. High efficiency is indicated by high values in these metrics. Other metrics also exist, often emphasizing either precision or recall. One such metric is SFDA (Sequence Frame Detection Accuracy), described by

Kasturi et al. (Kasturi, et al., 2009) and developed for the VACE (Video Analysis and Content Extraction) program by the U.S. Disruptive Technology Office (Disruptive Technology Office, 2007). SFDA not only captures precision-recall characteristics but also accounts for the spatial distance between the system-detected object and the ground truth object, aligning well with human visual perception.

Other notable metrics include MODA (Multiple Object Detection Accuracy) and MODP (Multiple Object Detection Precision), both introduced by Kasturi et al. as part of the CLEAR program evaluation system (CLEAR , 2007), which still provides a solid foundation for evaluating video data acquisition today.

However, there is no universal test dataset, and different video systems perform differently depending on the benchmarking data used. Comparing results across different datasets is especially important when verifying the expected data acquisition efficiency of a deployed system. Manohar et al. (Manohar, 2006) highlight this challenge by comparing testing environments and ground truth resources from two evaluation projects.

In conclusion, the benchmark data selected should reflect the context in which the system will be used. Fortunately, Pavlopoulou et al. (Pavlopoudou, et al., 2009) present a method for selecting test cases and datasets that reflect different operational contexts.

3.2.2.4 User interface and workflow support

The ultimate goal of the evaluation process is to deliver value to clients, end users, and customers. A well-designed interface can provide users with a valuable experience and enhance both the perceived system efficiency and the efficiency of system usage. Certain evaluation objectives, such as user interface responsiveness, have a direct impact on user satisfaction and perceived system quality and efficiency.

The user interface serves as the system's presentation layer and therefore significantly influences how users perceive system performance and functionality. Nielsen (Nielsen, 1993) highlights the importance of end-user knowledge by proposing a user testing-based approach to evaluating user interface quality. The case study presented an approach to improving interface quality through redesign driven by user testing.

Multiple metrics can be applied to measure system performance in the context of the end-user interface. In general, user interface efficiency should be assessed from two perspectives: functionality and performance.

Functional evaluation should determine how well the user interface supports user workflows. Most modern surveillance systems are workflow-oriented, with highly tailored interface requirements. The key objective is to evaluate how effectively the system supports users in performing individual tasks and sequential actions,

considering ergonomics and interface intuitiveness. Evaluating user interface efficiency is especially important when expanding integration with other systems. Therefore, it is a critical aspect of evaluation, particularly in cases where surveillance systems are integrated with large-scale traffic control systems, security systems, or other platforms.

MacLean et al. (MacLean, et al., 1990) identify two approaches to creating software systems: making tailoring mechanisms accessible to users or administrators and enabling interface customization by professional developers—either within open-source communities or through outsourced efforts. Tay and Cockburn (Tay & Cockburn, 1996) describe software development techniques for tailoring systems focused on workflow.

Performance evaluation typically aims to minimize user interface friction. The concept of User Interface Friction (UIF) refers to the resistance that the interface and its behaviour impose on user processes. UIF is measured using metrics such as response time for specific operations (e.g., connecting displays, selecting menu items), precision of mouse actions, and similar factors. It is important to note that UIF is not dependent on hardware processing power, assuming sufficient hardware resources are available. It is also independent of application functionality, assuming the required features are present. High efficiency and minimal UIF are critical for frequently repeated operations—high UIF in these cases can lead to significant productivity loss.

I have identified the following frequently repeated and delay- or precision-sensitive operations in surveillance systems:

- Camera control (sensitive to precision and delay)
- Visualizing cameras on selected displays
- Searching recorded material based on time criteria
- Exporting selected material to external media

The challenge of minimizing UIF is addressed in multiple publications and commercial benchmarks. Duis and Johnson (Duis, 1990) discuss the role of user interface responsiveness in evaluation and propose methods for improvement. Zuoxian et al. (Zuoxian, et al., 2009) introduce time performance metrics—active transition and active pattern—that can be used to build efficient, though somewhat complex, performance analysis algorithms

3.2.2.5 System intelligence

Albus (Albus, 1991) described system intelligence as the "function of generating and controlling actions increasing the probability of success in achieving high priority goals." He also listed the functional elements that an intelligent system (or Intelligent Control System) should contain:

- **System of Perception (Sensory Processing):** Responsible for filtering sensory input, detecting, recognizing, and interpreting.
- **Knowledge Representation:** Where the system environment is modeled. This includes the capability to collect, store, and organize knowledge to predict and simulate future scenarios.
- **Behaviour Generation:** Responsible for task decomposition, planning, issuing commands, and value-based decision-making (including computation of costs, benefits, and other attributes), as well as action generation.

Based on this definition, optimizing system intelligence directly contributes to improving overall system efficiency. We can observe that the definition of an intelligent system shares key elements with an efficient video surveillance system. It includes sensing capabilities (e.g., cameras, sensors) and the ability to perform data filtering (in its simplest form, such as a camera with a motion sensor). It can combine available data to support problem-solving and decision-making (e.g., license plate recognition systems integrated with access control). It may also apply automation mechanisms and possess learning capabilities or basic adaptive intelligence to develop alternative plans for future actions by evaluating their preferability (e.g., traffic monitoring systems with traffic learning features).

According to Krishnakumar (Krishnakumar, 2003), an intelligent system can be characterized by flexibility, adaptability, memory, learning, and the ability to manage imprecise information. An intelligent system should be capable of operating in uncertain environments, learning and adapting, transferring knowledge across domains (communication), and solving complex problems. Most modern video surveillance systems meet this definition

Similarly to video surveillance system evaluation, the essence of intelligence is qualitative in nature and nonmaterial, yet its quantitative measurement is becoming increasingly important. There have been multiple attempts to provide quantitative measures of system or solution intelligence. Meystel (Meystel, 2007) describes the Vector of Intelligence (VI), in which individual coefficients represent different aspects of intelligence. Special attention should be given to the publication by Horst (Horst, 2002), who proposes the Native Intelligence Metric (NIM). NIM measures

system intelligence based on the complexity of information managed within the system. It originates from Shannon's theory and is proportional to $-\log_2 p$, where "*p is the probability that the supposed specified complexity (of the system) arose purely through a logical chance hypothesis*" (definition by (Dembski, 1998)).

The main advantage of NIM is that it can be calculated prior to execution or testing within the actual system. However, a major drawback of NIM as a measure of intelligence is that it defines potential rather than the maturity of learning capability—which is a key expectation for intelligent surveillance systems. Additionally, the process of defining NIM is complex and difficult for an external auditor to perform. Unfortunately, most methodologies that represent intelligence as a derivative of information complexity are quite complicated to calculate.

Therefore, we focus on measuring intelligence as system efficiency in solving problems and achieving high-level goals. This can be simplified to measuring "success" in achieving incident detection objectives. Such success can be evaluated using metrics like frequency bias, the proportion of correctly performed actions (detected events), Probability of Detection (POD), also known as Hit Ratio (HR), and finally, the False Alarm Ratio (FAR).

All modern surveillance systems have some decision-making mechanisms built in. However, this aspect is particularly important in surveillance systems that incorporate video content analysis. Video content analysis improves decision-making speed, adds automation, and enhances reliability and quality when accessing or searching through material. Integrating video content analysis supports decision-making through predefined scenario alerts, forensic search capabilities, statistical analysis, traffic flow control, and more.

Video content analysis is especially critical in situations where decisions must be made quickly and reliably, such as in urban surveillance or access control in high-security facilities. Incorporating video analysis to support or enable decision-making should positively impact system efficiency. However, the deployment of video content analysis can also have negative consequences. If its efficiency is low and users rely heavily on automation, it may lead to poor overall system performance.

These challenges are discussed by Desurmont et al. (Desurmont, et al., 2004), who review general deployment issues of video content analysis; Ashani (Ashani, 2009), who focuses on urban environments; Finn (Finn, 2004) who examines public transport applications; and Lipton et al. (Lipton, et al., 2004) who explore forensic use cases. As a result, the deployment of video analytics-based intelligence is considered one of the highest-risk areas in the surveillance industry.

The topic of learning capabilities has attracted significant attention in international research. A comprehensive overview of learning methods in intelligent systems is provided by Linkens and Nyongesa (Linkens & Nyongesa, 1996). Their publication discusses various learning methodologies, including fuzzy logic, neural

networks, and genetic algorithm-based approaches in the context of intelligent systems.

Fuzzy logic-based learning is applicable to decision analysis. It extends traditional binary logic by assigning grades (weights) to elements, allowing it to handle imprecise and approximate information in decision-making. Neural network-based learning is suitable for tasks such as object or event (Ou, 2007). Research on genetic algorithm-based learning is currently applied in various domains, including multi-fault diagnosis, predictive diagnostics, schedule optimization, route planning, load distribution, and improving classification mechanisms (Rudas & Fodor, 2008). A solid overview of basic learning methods in computer systems is also provided by Kulikowski and Weiss (Kulikowski & Weiss, 1991).

In modern surveillance systems, learning approaches have been implemented early on in a few commercially available solutions, where scene learning is used to reduce configuration time and effort for incident detection systems. These learning approaches are implemented using iterative learning from operator experience, genetic algorithms, or neural networks.

System learning efficiency can be measured by the system's ability to generalize, its performance level in the specific task being learned, and the speed of learning—the latter being the easiest way to assess the maturity of a system's learning capability. In surveillance systems, learning can often be improved by using logged data for feedback and training of adaptive systems.

3.3 On multi-dimensional evaluation

Clearly, multi-dimensional evaluation is necessary when the goal is to build a multi-stakeholder, multi-use-case framework. However, it is also worth noting that recent developments in surveillance solutions further motivate a more holistic approach to evaluation. The more complex and intelligent a system becomes, the more difficult it is to capture its value. This complexity aspect of evaluation has also been explored in Publication I *Security Aspect of the Complexity of Modern Surveillance Systems - An Experience Report* (Karimaa, 2009). A good example of how complexity impacts evaluation challenges is presented in the publication by Gitto et al. (J-P. Gitto, 2016), which provides a strong analysis of how system quality is affected by system complexity. Based on this publication, I have added System Complexity metrics to reflect the impact of system transitions on overall system value.

Evaluation framework publications from the period after 2015 often focus on performance evaluation from architectural or complexity perspectives, as most video surveillance systems were transitioning to cloud-based architectures and undergoing redesign.

After 2018, video surveillance evaluations began to shift away from specific functional performance objectives and instead focused more on value reflecting operational excellence and broadly understood system intelligence. Only at this stage did a holistic approach to system evaluation begin to emerge, aiming to integrate multiple objectives.

A good summary of this phase of research is provided by V. Tsakanikas and T. Daguklas (V. Tsakanikas, 2018), who published a review of the current development status of video surveillance and related research. A notable example is the publication by Ming Liu and Zhi Xue (Ming Liu, 2021) in which fuzzy theory and a fuzzy multi-level evaluation method were applied to create an index system-based evaluation process. This method is similar to the scoring system proposed in this work but was limited to a specific phase of the system life cycle.

4 Evaluation framework proposal – scoring system design

This chapter presents the evaluation framework, including its requirements and the process behind creation of the framework.

First the rationale for the framework is presented in section 4.1 explaining the motivation and the process of creating the framework.

Next, Section 4.2 presents framework requirements reflecting on needs of the industry, stakeholders and the applicability.

Then, in Section 4.3 framework design including rationale behind it and in section 0 framework is presented including evaluation objectives and metrics.

4.1 Rationale for framework

The development of the framework published in this work was started based on industry need of providing objective, measurable and universal framework that would:

- provide simple and universally applicable way of measuring systems in different stages of product life cycle and different applications (use cases)
- capture multiple aspects of value relevant for different types of stakeholders
- encourage multi-stakeholder cooperation in improving current and developing future solutions, especially encouraging cooperation between the industry and the academia and development of new but relevant innovations

I believe this holistic yet practical approach to evaluation is needed by both the industry and the academia, yet both fora lacked cross-domain insights allowing to create it. In the industry, while video surveillance evaluation standard does exist, their focus is on assessing compliance with certain requirements of systems performance or security and focus on specific aspect of the system, so they don't provide holistic view allowing to be used across different group of stakeholders or use cases. In the academia, the evaluations typically focus on specific aspects or objective of the system to be improved, while missing the holistic view to the system value to system operators, users or buyers. Additionally, while system developments

drive system complex, intelligent and multi-purpose requiring this more holistic view of system value, the evaluation programs are clearly lacking behind the development offering very fragmented view on selected system objectives.
However, while both academia and the industry lack cross-domain tools, both have greatly contributed to different stages of this work by providing large selection of evaluation programs, evaluation objectives, and applicable metrics and insights.

The evaluation framework was created based on research approach as described earlier and generalized on the **Figure 3.**

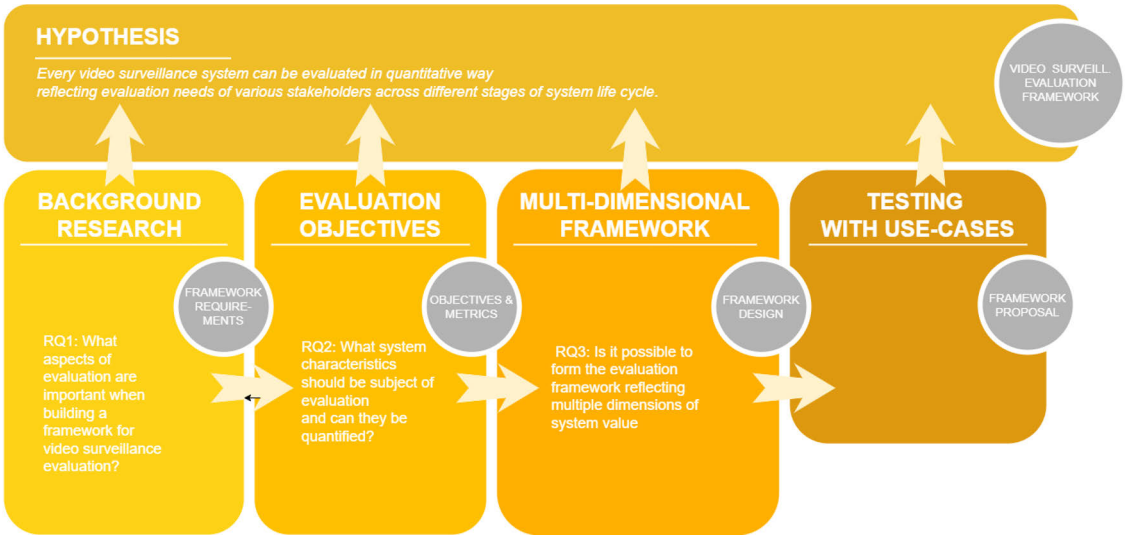


Figure 3. Research methodology simplified.

The framework was developed during different phases of research.

First, the background research (aimed to answer *RQ1: What aspects of evaluation are important when building an evaluation framework for video surveillance?*) provided general framework requirements. The requirements are presented in this chapter in section 4.2.

Next, the research on objectives and their quantification research (aimed to answer *RQ2: What system characteristics should be subject of evaluation, and can they be quantified?*) provided set of objectives and insights to their measurability (research on metrics).

This phase of the research was concluded by Publication V *Rating Functional Design: Evaluation and Scoring for Systems with Video Sources*, where I created a first draft of the framework aiming to capture functional value of the system using combination of evaluation objectives and measurable metrics (Karimaa, 2013a). This well received framework was a base improved during the next stage.

Finally, the last stage of research aimed to explore multi-dimensional aspects of "value" ((aimed to answer *RQ3: How to form the evaluation framework reflecting multiple dimensions of system "value" as relevant for various stakeholders, use cases and valid across various stages of system life cycle?*). The research on value aspects has been initiated by Publication VI *Value-Aware Approach to Management of Innovative Software Products and Services* (Karimaa, 2011), where I reflected on different aspects of value for different stakeholders and changes of value in different stages of product development and its life cycle. This allowed me to recap on value perspective of earlier research and different view on existing industrial evaluation programs and relevant developments in both industry and academia.

Following that, the framework was created as described in this chapter and finally it has been validated with two case studies.

4.2 Evaluation framework requirements

The biggest challenge behind the idea of universal evaluation criteria is to design guidelines for the process of selecting the metrics relevant for given types of the system. I have defined the following requirements for selecting such a methodology:

- Empiric methods are preferred - the process should be easy to adjust when more evaluation data is available.
- Methodology should accommodate various system designs and technologies by being focused on the functional objectives.
- Elements of the selected methodology can be easily decoupled and customized for different evaluation goals. A combination of such elements shall be used to provide a more complete picture of system evaluation. The motivation for this requirement is a need for partial system evaluation e.g. for the purpose of systems comparison in given functional area.
- Methodology shall be applicable in different phases of the system development – starting from system design concept and finishing on mature system products with wide experience from the field of system deployment.
    - In the case of evaluation of system concepts the evaluation result should also reflect the risk of inaccuracy when new technologies are not mature.
    - In the case of evaluation of existing systems, the evaluation results should point out areas of improvement to be implemented to achieve the best system efficiency.

- The concept needs to be easy to understand. Simplicity shall encourage feedback of multiple sites involved in process of system evaluation and not necessarily related to scientific environments.
- Methods should be familiar and accepted by organizations involved in system evaluation. New methods of calculation require additional explanations to be implemented and approved. There is not always a place for such justification (for example in calls for tenders or sales offering process).

The above requirements were fulfilled by e.g. a scoring system. Despite its criticism in scientific forum score -based methodologies are already well-presented in video surveillance. Score-based evaluation is well represented e.g. in retrieval of information, where in fact most Information Retrieval (IR) systems are based on idea of a numeric score reflecting how retrieved object is like the query. Moreover, even commercial offering tenders are traditionally based on scoring: the offered system competes with others with number of points collected by fulfilling tender requirements and conclusively scores, completed with price information are compared against each other. Also, score-based systems are easy to understand and easy to modify.

We have also utilized the advantage of balanced scoring, where weighting can be used to decouple individual elements and stress the importance of selected elements.

The proposed scoring system has been designed according to the following:

- Identification of key evaluation objectives (evaluation Targets) reflecting the system value
- Identification of metrics relevant for measuring each Target
- Metrics and Metric Score calculation, where Metric Score describes metric's impact on the system.
- Calculation of Target Score based on Metric Scores
- Calculation of system score based on the Target scores and selected relevant system specifications.

It is also important to underline that the scoring system needs to be adaptable for different types of system so relevant evaluation targets and suitable metrics choice should have some flexibility as not all the metrics might be possible to measure. At the same time, the system needs to be also objective.

The framework was created with the steps:

- Define a list of evaluation objectives important for the evaluation of surveillance systems.

- Identify a set of metrics applicable in context of the most common evaluation objectives of surveillance systems. Identify similarities in terms of functional objectives between surveillance systems and other systems to investigate the metrics applicable in context of surveillance systems.
- Build the guideline process for selecting the metrics. Give examples of the process for most common types of surveillance systems (identify the areas of evaluation focus).
- Design details of the scoring system.
- Test the scoring system with case study evaluating existing surveillance system (testing working hypothesis).
- Test the scoring system with case study concerning future concept evaluation.

4.3 Evaluation framework design

The scoring system has been designed as a sum of weighted individual scores created for every evaluation objective. System of weighted scores allows defining relevance of individual evaluation objective (characteristics selected for evaluation). Each evaluation objective score is calculated from set of metrics available for the system.

The scoring system has been designed in steps. First, evaluation objective (characteristics or system attributes reflecting system value) has been identified.

In demonstrated evaluation framework proposal, the choice of evaluation objectives was done based on study of evaluation objectives used in the industry and academic research. The goal was to focus first on evaluation objectives relevant for both groups, so the further development of the framework can be supported by stakeholders in both industry and academia.

These evaluation objectives can be also selected to be the targets (Targets) of the evaluation. Next, for each Target a set of metrics is listed. Only metrics relevant to given Target are included.

Next, metrics are calculated, and their metric scores are determined. Metric score illustrates impact the metric has on system efficiency. Finally, for each metric range of values is defined and Metric Score is calculated. Metric Score illustrates the impact on system value. Metric Score for ideal system value is equal to one, E = 1 (refer to Table 3).

Table 3. Metric Score value.

SCORE	SYSTEM VALUE
1	Ideal
0.8	Very good
0.6	Good
0.4	Medium
0.2	Moderate
0	Insufficient

Value of the system can vary from 0% to 100%. Metric Score assigns 0 to 0% and 1 to 100% and distribute remaining values in the range in suitable way where E equal to zero is insufficient and unaccepted efficiency and E equal to one is perfect (ideal) efficiency. Metric Score allows translating the result of metric testing into Metric Score. Metric Score is value between 0 and 1 and defines how the result reflects the system efficiency.

Target Score is calculated as average of Metric Scores available for given target. Ideal Target Score is equal to one.

System score is calculated based on Target scores. Customized system description is created based on weighs set by user defining the importance of system objectives relevant to his/her system. Final score combines Target Scores using score balancing with weights calculated based on customized system description.

To analyse the behaviour of our scoring system we will analyse the case of "perfect efficiency" of general system. In such case of perfect system all Metric Scores are equal to one. If the user defines all system objectives (such as system design or data performance) as equally important (general purpose system) the biggest contribution to system efficiency has System design (44%), then System Intelligence describing the performance of system in achieving its functional goals as well as system capability to learn (22%). Finally, the efficiency of selecting and configuring source data contributes to remaining scores of 11% per Target. These results are consistent with intuitive understanding of general video surveillance system – system design is defined with major system objectives and therefore contributes to high impact on system perceived value whereas the efficiency impact of selection and configuration of source data can be in some cases very small or in extreme cases even omitted (in example of manual monitoring systems).

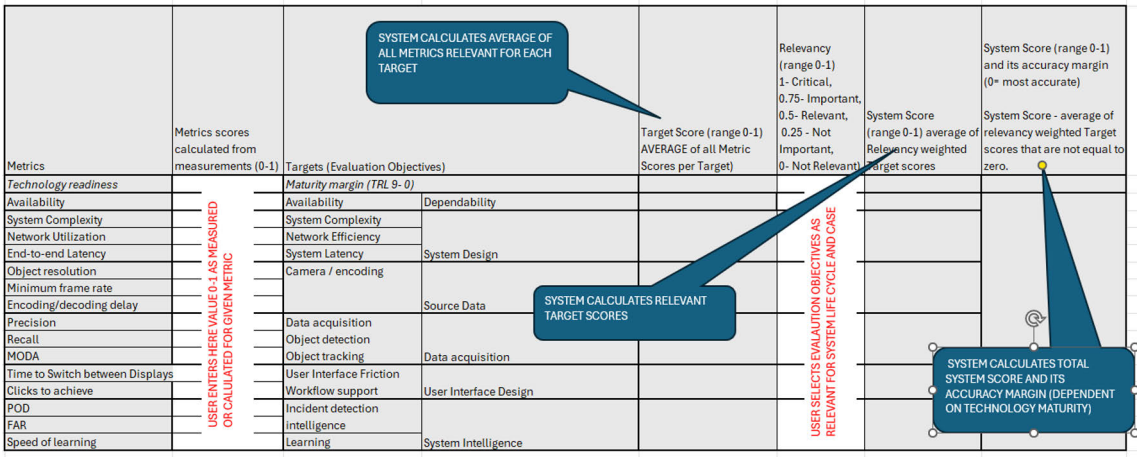
4.3.1 Scoring specification at a glance

The figure below presents scoring specification at a glance illustrating normalisation formulas for the metrics used in the case studies, default weights and their rationale, as well as definitions of the composite indicators used.

User is calculating or measuring each metric as demonstrated in next chapter. Measured values (metric scores) are entered in position marked with red text. Based on arithmetic average of relevant metrics Scoring system calculates Target Scores for each objective.

User can select the objectives of evaluation (evaluation target) as relevant for system life cycle. Relevancy is of the range 0-1, where 1 is Critical, 0.75- Important, 0.5- Relevant, 0.25 - Not Important, 0- Not Relevant. If preferred user can select the evaluation objective relevancy using only Critical and not Relevant and reiterate evaluation after applying specific design or implementation corrections and system improvements. By default, this scoring considers all evaluation objectives so the default value for Relevancy is one (1).

Finally, Total score based on arithmetic average of relevant Target Scores. Not relevant scores are not taken into consideration for this calculation.



4.3.2 Results uncertainty

This scoring system is early phase proposal that should be developed based on high volumes of cases from industry and research. However, at this early stage of scoring system maturity multiple aspects of results certainty should be considered. First, calculator embeds also scoring result accuracy margin which is dependent on TRL level of used technologies. Using non mature technologies impact's results reliability in terms of creating margin of possible results when technology is maturing. It is

possible that in future other aspects introducing performance, efficiency and value uncertainty will be identified, and this part of scoring system will be expanded.

Secondly, variations in metric weights and metrics measurement noise affect directly the final score. There is a risk that metrics scores would need to be modified in future- now for each metrics the impact on value is linear with described steps (E from 0 to 1 with step of 0.2 as presented in Table 3). It is possible that practical experience of using scoring system will present non-linear dependencies for at least some of the metrics.

Finally, the Relevance grading might need improvement in further use to limit the freedom of user selecting the relevancy. It should be possible to assign default relevancy for example for systems in specific life cycles phases. This is well illustrated in use cases presented in this work.

4.3.3 Ethics and compliance consideration

To ensure ethical and regulatory alignment, the framework can be further developed. It can be implemented by expanding quality objectives with metrics, such as privacy-by-design and auditability. These aspects reflect the system's ability to protect user data and provide transparent, traceable operations. They can be scored using metrics such as compliance with data protection standards (e.g., GDPR), presence of access controls, encryption protocols, logging mechanisms, and the ability to generate audit trails. Each such a new metric is evaluated and translated into a Metric Score, contributing to the overall Target Score for ethics and compliance. This integration ensures that systems are not only functionally efficient but also trustworthy and accountable.

4.4 Evaluation framework

4.4.1 Quality metrics

To illustrate selected aspects of system quality we have selected the following example metrics as presented in Table 4 below.

Table 4. Quality metrics.

OBJECTIVES	METRICS
DEPENDABILITY	Availability
MATURITY	Technology readiness
OTHER [1]	

[1] As selected from quality model.

4.4.1.1 Technology Readiness

Definition and calculation: the metric originates from research of NASA (Mankins, 1995) and illustrates the level of maturity of technologies used in the system. Technology Readiness Levels are defined as follows:

- *TRL 1 Basic principles observed and reported.*
- *TRL 2 Technology concept and/or application formulated.*
- *TRL 3 Analytical and experimental critical function and/or characteristic proof-of concept*
- *TRL 4 Component and/or breadboard validation in laboratory environment*
- *TRL 5 Component and/or breadboard validation in relevant environment*
- *TRL 6 System/subsystem model or prototype demonstration in a relevant environment*
- *TRL 7 System prototype demonstration in target environment*
- *TRL 8 Actual system completed and qualified through test and demonstration (ground or space)*
- *TRL 9 Actual system "flight proven" through successful mission operations*

Metric Score: The lower the level of TRL the higher the uncertainty of the system efficiency. Low levels of TRL for technologies would contribute to bigger margins and lower accuracy of the result, unless metrics measurable (or possible to estimate or simulate) in early TRLs can be used. In case of using any technology where TRL =1 the system design efficiency estimation accuracy originally estimated to E= 0.7 can vary from E=0.5 to E=0.9. Obviously, overestimated efficiency cannot exceed the level of System Score in any case. It is also worth to notice that introducing new technologies might have positive effect on overall system design efficiency (see positive margin of efficiency function).

Assuming the above, the System Score margin can be estimated as:

$$\text{System Score margin} = \frac{9 - \text{TRL}}{\text{A} * (\text{TRLmax})} = \frac{9 - \text{TRL}}{\text{A} * 9}$$

where A is the coefficient designed to keep System Score margin on the level not exceeding 0.4. This means the fact of introducing new, not proofed technologies should have (in worst case) such impact on system efficiency that:

- Ideal system might become good
- Very good system might become medium

- Good system might become moderate
- Medium system might become insufficient

Therefore, A can be calculated as:

$$A = \frac{9 - TRLmin}{0.4 * 9} = \frac{8}{0.4 * 9} = \frac{8}{3.6} = 2.22$$

Finally, the System Score margin can be estimated as:

$$\text{System Score margin} = \frac{9 - TRL}{2.22 * (TRLmax)} = \frac{9 - TRL}{20}$$

In case of multiple technologies being used the margins should be accumulated.

Measurements: Assessment shall be done based on descriptions available above.
Metric usability: Mandatory for systems where use of new technologies is evaluated. Mandatory for evaluation of system design for systems in concept phase.

4.4.1.2 System availability

Definition and calculation: System availability can be calculated by modelling the system as an interconnected group of individual components, where these components are defined as critical for systems availability. In general, the availability is calculated as:

$$CR = \frac{tu}{tu + td}$$

where

tu - system up time

td - system down time

In surveillance system where there is no server redundancy system availability is measured as availability of individual node responsible for system operation (usually server)

A= A1= ((365days-3.65)/365)*100%

Example: if server is unavailable 3.65 days per year its availability is 99%.

In surveillance systems where server redundancy is used total system availability is:

A = 1 - (1- A1*A2)

The system is unavailable when all components are unavailable which in practice is equal to take-over time

Example: if server1 is unavailable 3.65 days per year (A1 = 99%) and failover server 2 is unavailable 52 minutes per year (A2=99.99%) then system availability is:

$$A= 1 - (1- A1*A2) = 99.99\%$$

In surveillance systems where unavailability causes or can cause unavailability of other nodes, total availability should be calculated as so-called serial availability:

$$A = A1*A2$$

The system is available only if all the nodes are available. An example of such case is failure of the server node which causes stopping of recording activities at recording node. The system is unavailable when either of components is unavailable

Example: if server1 is unavailable 3.65 days per year (A1 = 99%) and recording node is not able to operate without server, then despite the fact the unavailability of recording is equal to only 52 minutes per year (A2=99.99%) the total system availability is:

$$A= 0.989901 = 99\%$$

Metric Score: Due to the critical character of availability in security-oriented surveillance system the Metric Score shall be calculated as equipotential function of availability:

- availability on the level of 99% should be considered as insufficient for most of the systems (E=0)
- availability of 100% should be considered as ideal and return E=1
- it can be assumed the highest availability of 99.99999% (Seven Nines) should be close to one (A=1) as determining the state of constant availability

Therefore, the Metric Score for Availability can be calculated as:

$$E = \frac{Ax - 2}{(\max Ax - \min Ax)} = \frac{Ax - 2}{5}$$

where Ax is approximated availability expressed in number of Nines (compare for A= 99 Ax=2)

Measurements: Available for every phase of system development:

- For systems in concept phase the availability should be calculated based on assumed downtime related to system maintenance actions (SW upgrades and configuration changes) for given set of server components.

- For systems with already available hardware the system downtime can be précised considering restart times of actual critical node machines and their estimated failure rates
- For system already deployed the numbers can be updated with actual measurements of downtimes

Metric usability: Mandatory for every system evaluation.

4.4.2 System design performance

System design performance metrics are summarized in Table 6 and described in below Sections.

Table 5. System design performance metrics.

OBJECTIVES	METRICS
SYSTEM DESIGN PERFORMANCE	System Software Complexity
	Video Network Utilization
	End-to-end Latency

4.4.2.1 System Software Complexity

Definition and calculation: System is complexity is dependent on number of components present on given abstraction level (SC is the number of components)
Metric Score: Metric Score can be calculated as:

- E= 0 for more than 14 (system complexity too difficult to manage),
- E = 1 for 7 and less

E=(14- SC)/7

Measurements: Available for every phase of system development:

- For system in concept phase the number of components can be evaluated based on number of components on business layer (system management functions), resources layer (number of components to be managed) and application layer (number of actual user interface applications components)
- For system with ready software structure actual number of software components can be calculated

Metric Usability: Mandatory for every system evaluation. Indicated an effort needed to maintain and further develop the system

4.4.2.2 Video Network utilization

Definition and calculation: The metric illustrates efficiency of network utilization and should be used especially for system communication paths with video traffic

$$\text{VNU} = \text{average}\left(\frac{\text{bo}}{\text{ba}}\right) = \text{average}\ \left(\frac{\text{n} * \text{c}}{\text{ba}}\right)$$

where
bo –bandwidth occupied by video content
ba - bandwidth available for video content
n – number of video streams
c – video capacity of video stream

The following can be assumed in context of calculating value of the metric:

- Typical MPEG-4 stream with parameters: 4CIF (704*576 pixel) resolution and 25fps frame rate and moderate amount of motion occupies approx. 4Mbps and the same parameters H.264 stream occupies about 2.5 Mbps
- Bandwidth available for video content is typically on the level of 50% of bandwidth available in given link

Metric Score: Value can be calculated as linear function of VNU where:

- E=1 for VNU=1
- E=0 for VNU =0

Usability: The metric should be evaluated for every type of system. The metric can be used to design required network capacity for network infrastructure links

4.4.2.3 End to End Latency

Definition and calculation: End to End Latency describes delays in video and transmission caused by encoding (including image capturing, digitising and compressing), transferring via the network, and decoding. End to End Latency is calculated for worst case delay path.

SL = encoding and delay + transmission delay + decoding delay

Metric Score

Scorecan be calculated as:

- E=0.8 if SD= 200ms
- E=0 if SD = 300ms

Usability: The metric should be evaluated for all the system especially ones with delay critical operations such as PTZ control or video synchronized with audio
Measurements: The metric can be used already in system design space as long as delays for encoding and decoding devices are known, and transmission delays can be estimated for known infrastructure design. In case of system where encoding and decoding devices as well as infrastructure paths are known the metric can be calculated by measuring encoding and decoding delay and calculating the transmission delay for given number of network devices on the path.

4.4.3 Source data performance

Source data metrics are summarized in Table 6 and described in below Sections.

Table 6. Source data metrics.

OBJECTIVES	METRICS
SOURCE DATA	Object resolution
	Minimum frame rate
	Encoding/decoding delay

4.4.3.1 Object resolution

Definition and calculation: The metric assess if the object size in camera picture is designed to support the context of camera usage. The object resolution (OR) is determined by object size on the picture of given resolution compared to actual size of the object. The value of the metric should be compared against experimentally appointed values of A (refer to Johnson tests described in previous chapters)

$$OR = \frac{Os * R}{Op}$$

where
Os - image space occupied by object t
R – image resolution
Op– object size (pixels)

Metric Score: Score can be calculated as:

- E=1 if OR= (A, 1.1A)- optimum resolution is achieved for object resolution from A to 1.1A (A with +10% margin)
- E=0 if abs OR= (0.5 *A or 1.5 A) (function of identification, recognition or detection does not perform under 0.5A, as well as image resolution too large which provides good accuracy of identification, recognition or detection, but introduced unnecessary bandwidth overhead)

A value (refer to previous chapter):

- Identification for forensic level A = 500 pixels/m
- Identification A= 250 pixels /m
- Recognition A= 100 pixels/m
- Detection A = 20 pixels/m

Usability: For object detection, recognition, identification and evidence producing systems. The metric can be classified as mandatory for systems producing evidence material for forensics purposes, such as license plate recognition or face recognition. The metrics defines object resolution for ideal conditions of scene illumination, and camera focus, lens quality and unlimited capabilities selecting optimal camera positioning and angle of view

Measurements:

- In case of evaluation of systems in concept phase the metric can be used to provide the minimal camera resolution which can be calculated based on OR formula where OR=A and Percentage of image space occupied by object is a value set according to CCTV recommendations (refer to previous chapter)
- When camera is already deployed the object resolution can be calculated based on OR formula.

4.4.3.2 Efficient Frame rate

Definition and calculation: The metric was created to improve the efficiency of selecting appropriate frame rate. The frame rate selection should consider minimum frame rate necessary to capture defined amount of motion. Minimum frame rate can be calculated as follows:

$$\text{MFR [frames per second]} = \frac{1}{\text{tc}}$$

where
tc - time slot required to capture the object

Time slot required to capture the object can be either known (for example 200ms is a timeslot to capture human activities) or can be calculated based on object moving speed and length of object path captured by camera:

$$\text{time slot} = \frac{\text{so}}{\text{vo}\ [\frac{\text{m}}{\text{s}}]}$$

where
so - length of object path captured by camera
vo -object speed

Example 1: max speed of e.g. human action is 200ms and it is also a time slow required to capture the movement. Minimal Framerate necessary to capture this movement is FR=1/0.2=5 frames per seconds

Example 2: Functional purpose of the camera setup it to detect human and cars on parking lot where max speed of car on parking is 60 km/h which corresponds to 16 m/s. Path captured by camera is approximately 5m therefore time slot required to capture the car on at least one frame within 0.3125 s which corresponds to minimal frame rate below 1fps.0.3 frames per second which means frame rate should have at least 4 fps.

The Efficient Frame Rate ratio can be calculated as follows:

$$\text{EFR} = \frac{\text{Fmin}}{\text{F}}$$

where
Fmin – minimum frame rate
F - measured frame rate

Metric Score: Score can be calculated as:

- E=1 if EFR =1
- E=0 if EFR = 0.04 (which responds to situation where 25fps is used instead of 1fps)

Usability: The metric should be evaluated for systems where cameras are installed with view of close proximity to the object and there is a risk of not capturing certain activity. It is especially important for systems generating evidence material. It is important to underline that selection of frame rate can be forced by legal

requirements of recording evidence- in such case the efficiency of frame rate selection should be ignored.

Measurements:

- The metric can be used to design required frame rate of camera or encoder against MFR requirements which is achievable as soon as scene conditions are known
- In case of deployed system designed frame rate should be verified against MFR in order to guarantee object appearance on the camera captured image

4.4.3.3 Encoding/Decoding Delay

Definition and calculation: Encoding/Decoding Delay is one of the most critical metrics in surveillance systems evaluation. In most of the cases the latency introduced by encoding/decoding process contributes greatly to the overall End to End Latency due to the fact of system being one IP network (in such cases transmission delays are to be omitted).

Metric Score

Score can be calculated as:

- E=1 if EDD =0
- E=0.8 if EDD= 150ms
- E=0.5 if EDD= 300ms

Usability: The metric should be evaluated for systems with delay critical operations such as PTZ control or video synchronized with audio

Measurements: Encoding/Decoding Delay is directly measurable as presented in test setup in **Figure 4.** The metric can be assessed as soon as encoding devices and decoding solution is known. Test can be design as illustrated on Figure 4. Analogue camera is used as video source. It points to a counter displaying the sequence of numbers with time step of one millisecond. The same signal sent to analogue display (which performs only visualization, but no decoding) and to encoder. From encoder, the signal goes to decoding component and then will be displayed on the second analogue monitor. Therefore, the time difference between numbers displayed on two monitors is the video latency of encoding and decoding. The results delay test can vary depending on coding buffers, encoding / decoding standards and parameters used techniques as well as on the amount of motion in the test material. Therefore, it is recommended that standard used are well documented, parameters are optimized to achieve minimum latency, and test sequences are used as video material.

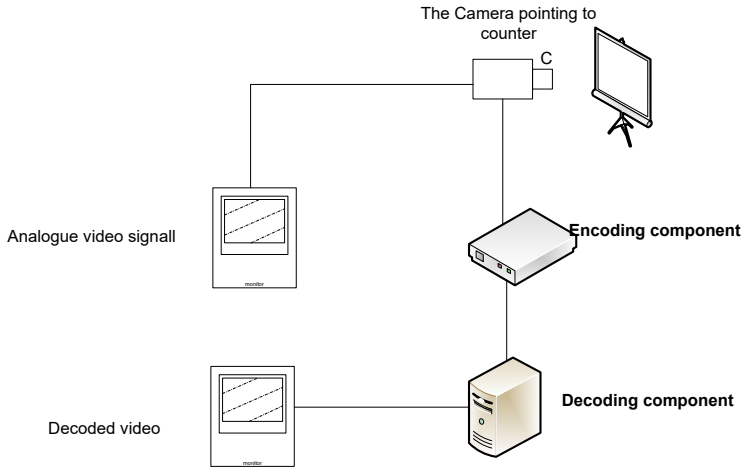


Figure 4. Encoding and Decoding Delay- test setup.

4.4.4 Data acquisition and processing

Data acquisition and processing metrics are summarized in Table 7 and described in below Sections.

Table 7. Data acquisition and processing metrics.

OBJECTIVES	METRICS
DATA ACQUISITION	Precision Recall
OBJECT DETECTION	MODA
OBJECT TRACKING	MOTA

4.4.4.1 Data acquisition - Precision

Definition and calculation: The precision describes amount of relevant data within all the data retrieved (see Equation 1) and indicates accuracy of retrieval relevant data

$$\text{Precision} = \frac{\text{Rr}}{\text{Ra}}$$

where: Rr – relevant retrieved data

Ra – all retrieved data

Metric Score: Score can be calculated as linear function of Precision

- E= 1 if precision =1
- E<1 for Precision >1 and
- E=0 if Precision =approx. 1/20 (typically when only one out of 20 alarms is relevant, in practice the acceptable ratio of false alarms should be specified by system requirements)

Usability: Practical approach of this metric is to illustrate ratio of false alarms in the system. This metric is recommended for incident-reaction focused systems where ratio of false alarm is critical for system efficiency. Good example of such system is incident detection system which relies on operator's capability of manual classification of incidents. As the operator is able to process only limited number of alarms in time the precision of event detection is critical.

Measurements:

- In case of evaluation of systems in concept phase the Precision values can be obtained as soon as detections algorithms are known and ready to test. Detection performance of such algorithms can be evaluated based on Ground Data comparison using test video sequences publicly available or even created manually (for coarse evaluation). More information about testing environment is available PETS project pages (PETS, 2004) .
- In case of evaluation of deployed system the measuring can be collected by comparing the number of events classified by operators as relevant with number of event unanswered/unclassified

$$\text{Precision} = \frac{\text{Ac}}{\text{Au}}$$

where

Ac – classified events/alarms

Au – unclassified events/alarms

4.4.4.2 Data acquisition - Recall

Definition and calculation: The Recall describes the amount of relevant data that has been retrieved within all the existing relevant data:

$$\text{Recall} = \frac{\text{Rr}}{\text{Rea}}$$

where:

Rr – retrieved relevant data

Rea – all existing relevant data.

Metric Score: Score can be calculated as linear function of Precision

- E= 1 if Recall =1
- E<1 for Recall >1 and
- E=0 if Recall =approx. 1/20 (when only one out of 20 alarms is relevant)

Usability: Practical approach of this metric is to illustrate system ability to identify all relevant data. This metric is recommended for incident-reaction focused systems where the ratio of missed relevant events is critical for system efficiency. Good example of such system is incident detection system where event related data is used for evidence and event investigation purposes. In case of such system no relevant events should remain undetected event with price of multiple false alarms.

Measurements:

- In case of evaluation of systems in concept phase the Precision values can be obtained as soon as detections algorithms are known and ready to test. Detection efficiency of such algorithms can be evaluated based on Ground Data comparison using test video sequences publicly available or even created manually (for coarse evaluation). More information about testing environment is available PETS project pages (PETS, 2004) .
- In case of evaluation of deployed system, the estimate measures can be collected by comparing the number of events classified by system with other incident data source (such as Police incident reports)

$$\text{Recall} = \frac{\text{classified events/alarms}}{\text{classified events} + \text{unclassified events from other source}}$$

4.4.4.3 Data acquisition –MODA

Multiple Object Detection Accuracy (MODA) has been proposed by Kasturi (Kasturi, et al., 2009) as a part of CLEAR program evaluation system (CLEAR , 2007).

Definition and calculation: MODA metrics illustrates object detection accuracy using numbers of missed detections (m) and false positive count (fp) as well as the number of ground truth objects (g). MODA can be calculated as below for specified amount of time and given sequence:

$$\text{MODA} = 1 - \frac{\text{m} + \text{fp}}{\text{g}}$$

Metric Score: Score can be calculated as linear function of MODA, where:

- E= 1 when MODA =1 number of ground truth object in the sequence
- E<1 for MODA<1

Usability: Practical purpose of this metric is to illustrate system accuracy to detect the objects in given frames or set of frames. This metric is recommended for evaluation of systems where object detection is critical and either number of missed detections or false positive count (false object detection alarms) is important for system efficiency. Good example of such system is incident detection system in e.g., airport system with focus on discovering unattended luggage. In such system luggage shall be discovered with missed detects close to zero and false detections should be minimum (ideally zero)

Measurements:

- In case of evaluation of systems in concept phase the MODA values can be obtained as soon as detection algorithms are known and ready to test. Detection efficiency of such algorithms can be evaluated based on Ground Data comparison using test video sequences. However, it is possible to make initial estimation by analysing manually recorded material and comparing it against results of automatic algorithms. More information about testing environment is available PETS project pages (PETS, 2004) and VACE project pages
- In case of evaluation of deployed system, the efficiency of object detection systems can be evaluated using percentage of objects lost where Metric Score can be calculated as follows

E=1 for 100% objects being discovered in defined conditions (time and visibility of the object by camera)

E=0 for defined number of objects being undiscovered (typically 50% undiscovered objects would lead to E=0) or number of false alarms exceeding given amount during testing process (typically number of ratio of 1/20 false alarms would lead to E=0)

4.4.4.4 Data acquisition –MOTA

Definition and calculation: Multiple Object Tracking Accuracy (MOTA) has been proposed by Kasturi (Kasturi, et al., 2009) as a part of CLEAR program evaluation system (CLEAR , 2007). MOTA metrics illustrates object tracking accuracy using numbers of missed detections (m) and false positive count (fp) and mismatches (mme) as well as the number of ground truth objects (g). MOTA can be calculated as below for specified amount of time and given sequence:

$$\mathrm{MOTA} = 1 - \frac{\mathrm{m} + \mathrm{fp} + \mathrm{mme}}{\mathrm{g}}$$

Metric Score: Score can be calculated as linear function of MODA, where:

- E= 1 when MOTA =1 number of ground truth object in the sequence
- E<1 for MOTA<1
- E=0 if MOTA= 0 which responds to situation when no object is tracked correctly in specified amount of time

Usability: Practical approach of this metric is to illustrate system accuracy of objects tracking in given time. This metric is recommended for evaluation of systems where object tracking is critical and either number of missed detections, false positive counts (false object detection alarms) or track mismatches is important for system efficiency. Good example of such system is incident detection system in e.g. airport system with focus on analysing the track of persons of interest (e.g. person involved in the incident). In such systems MOTA should be close to 1

Measurements:

- In case of evaluation of systems in concept phase the MOTA values can be obtained as soon as tracking algorithms are known and ready to test. Tracking efficiency of such algorithms can be evaluated based on Ground Data comparison using test video sequences. However, it is possible to make initial estimation by analysing manually recorded material and comparing it against results of automatic algorithms. More information about testing environment is available PETS project pages (PETS, 2004) and VACE project pages
- In case of evaluation of deployed system, the efficiency of tracking systems can be evaluated using percentage of objects lost where Metric Score can be calculated as follows:
    - E=1 for 100% objects being discovered in defined conditions (time and visibility of the object by camera)
    - E=0 for defined number of objects being undiscovered (typically 50% undiscovered objects would lead to E=0) or number of false alarms exceeding given amount during testing process (typically number of ratio of 1/20 false alarms would lead to E=0)

4.4.5 User interface efficiency

User interface efficiency metrics are summarized in Table 8 and described in below Sections.

Table 8. User interface efficiency metrics.

OBJECTIVES	METRICS
USER INTERFACE FRICTION	Time to Switch between Displays
WORKFLOW SUPPORT	Clicks to achieve

4.4.5.1 User interface friction

The process of identifying acceptable level of User Interface Friction metrics is very challenging. Obviously, optimum level of UIF is equal to zero, but acceptable level of UIF depends on user operation and context of system efficiency. However, in addition to UIF usability in context of scoring system these metrics can be used to produce direct comparison benchmarks

Definition and calculation: Time to Switch between Displays is the time needed to switch the display between two video sources. Delay of 100ms second is the limit giving impression the system is reacting instantaneously. Delay of 1 second is the limit for the user's flow of thought to stay uninterrupted, even though the user will notice the delay. If delay is over 1 second the user loses the feeling of operating directly on the data (Miller, 1968). If delay is longer than 3 seconds, the delay can affect the quality of operation if so, called fast sequences of camera images are displayed on any displays (typically these sequences consist of sequences of few seconds long videos from individual cameras)

Metric Score: Score can be calculated as:

- E=1 if TSD = 100ms and less
- E=0.6 if TSD= 1s
- E=0 if TSD = 3 s

Usability: The metric should be evaluated for all the systems where operations on live video are delay critical and where sequences of video sources are connected to display automatically

Measurements: the metric can be assessed when system is already deployed. Delay can be observed by connecting and recording testing the sequences of testing video or by retrieving information available in logs

4.4.5.2 Workflow– Clicks to Achieve

Definition and calculation: The metric describes number of steps needed to complete given operation which provides quantitative measure of effort needed to complete time and workflow sensitive operations.

Metric Score: Score can be calculated as:

- E=1 if number of clicks =1 (all critical functions are available within one-click)
- E=0 if number of clicks= 10 (high number of steps is not possible to be memorized)

Usability: The metric should be evaluated for all the systems where workflow is important for overall system efficiency
Measurements: the metric can be assessed when user interface is designed, and workflow critical operations are identified

4.4.6 System intelligence

System intelligence metrics are summarized in Table 9 and described in below Sections.

Table 9. System intelligence metrics.

OBJECTIVES	METRICS
EFFICIENT OPERATION	POD
	FAR
LEARNING	Speed of learning

4.4.6.1 Efficient Operation - False alarms

Definition and calculation: False Alarm Ratio (FAR) is one of the most popular metrics to describe intelligent data processing which relies on system efficiency of identifying relevant data.

$$FAR = \frac{data\ misinterpreted\ as\ relevant}{all\ identified\ data}$$

Metric Score

Score can be calculated as linear function of Precision

- E= 1 if FAR=0

- E<1 for FAR>0

Usability: Practical approach of this metric is to illustrate ratio of false alarms in the system. This metric is recommended for incident-reaction focused systems where ratio of false alarm is critical for system efficiency. Good example of such system is incident detection system which relies on operator's capability of manual classification of incidents. As the operator can process only limited number of alarms in time the precision of event detection is critical.

One should be treating with caution the interpretation of the results of system Score as they depend on the context of the captured data. A good example of such interpretation is precision and recall values. In general, the highest are the scores the better is the efficiency. However, achieving high scores for both precision and recall can be problematic and not always optimum from system Score point of view. There are several situations, where low precision is better (Menzies, et al., 2007): (a) when the cost of missing the target is expensive (mission critical applications), (b) when only a small fraction of the data is retrieved (selective sensors), (c) where there is little or no cost in checking false alarms. They should be considered when interpreting the measures for surveillance system.

Measurements: In deployed systems the measures can be collected by comparing the number of events classified by operators as false alarms (or unclassified) with overall number of events detected by the system

4.4.6.2 Efficient Operation – Probability of Detection

Definition and calculation: Probability of Detection (POD) describes the precision in identifying relevant data.

$$POD = \frac{identified\ relevant\ data}{all\ relevant\ data}$$

Metric Score: Score can be calculated as linear function of POD where

- E= 1 if POD =1
- E<1 for POD<1 and
- E= 0 when POD= 0.5

Usability: Practical approach of this metric is to precisely retrieve all the relevant data in the system. This metric is recommended for incident-reaction focused systems where the incident shall not be missed.

Measurements:
In deployed systems the measures can be collected by comparing the number of events identified by the system with overall number of incident (if available from police or other authority statistics)

4.4.6.3 Learning- speed of learning

Definition and calculation: System ability to learn can be described as the capability to improve system functional goals in time. Therefore, in case of surveillance systems where system goal is described by False Alarm Ratio (FAR) and Probability of Detection (POD) the learning efficiency can be described as function of FAR decreasing or POD increasing in time. The speed of learning describes system ability to achieve required (or ideal) POD or FAR in time:

$$SL = \frac{\Delta\, POD}{\Delta t}$$

or

$$SL = \frac{\Delta\, FAR}{\Delta t}$$

Metric Score: System ability to learn has major impact on system intelligence by providing weight to the system intelligence determined based on POD or FAR.

- If system is performing well, it does not have to display learning abilities to achieve its functional targets. If FAR=0 and POD= 1 and SLA for both =0 the overall system intelligence is still 1
- However, in case of intelligent system poor performance displayed with low POD and/or high FAR should be compensated with time

4.5 Original publications contribution

The framework described in this chapter has been created based on research published in original publications, where these publications contributed to creation of framework design as illustrated in Figure 5.

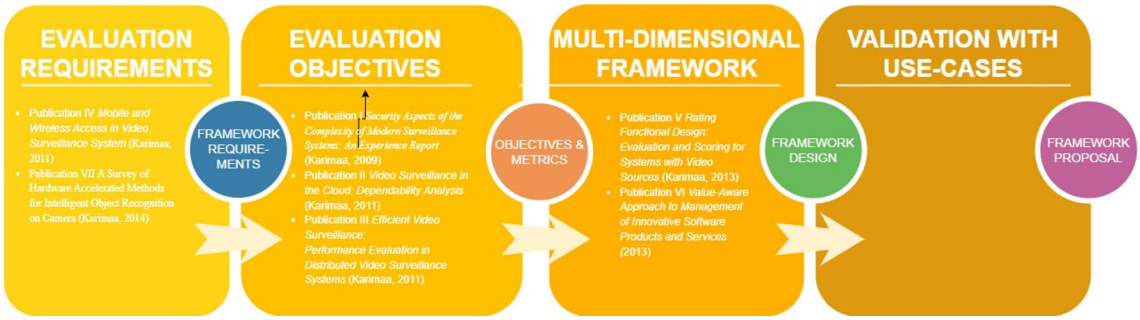


Figure 5. Original publications contribution to evaluation design.

Firstly, were background publications created during the phase of initial research : Publication IV *Mobile and Wireless Access in Video Surveillance* (Karimaa, 2011c) and also (later published) Publication VII *Survey of Hardware Accelerated Methods for Intelligent Object Recognition of Camera* (Karimaa, 2014). Both publications were focused on specific technical challenge, where the evaluation of the complete system solution was needed. These publications validated the importance of evaluation frameworks and provided good examples of situations, where evaluation frameworks would be helpful.

- Publication IV *Mobile and Wireless Access in Video Surveillance* (Karimaa, 2011c) analysed the challenge where the users expect to have seamless access independently on technology used and type of the access. The paper explores different access scenarios and discusses the evaluation of both solutions.
- Publication VII *Survey of Hardware Accelerated Methods for Intelligent Object Recognition of Camera* (Karimaa, 2014) reviews the challenge of online object recognition on camera device, review available object recognition methods, and address their applicability in the context of online processing applied on video camera. This publication contributed greatly to practical implications Source Data Performance as well as Data Acquisition and Processing part of the framework, as these has been proved to greatly affect the final performance and capabilities of the system.

It is worth to note, that mentioned two publications represent two types of evaluation (or evaluation needs) as referred in Chapter 3.2: Publication IV *Mobile and Wireless Access in Video Surveillance* (Karimaa, 2011c) focused on quality attributes, while (security, quality, dependability, usability), while Publication VII *Survey of Hardware Accelerated Methods for Intelligent Object Recognition of Camera* (Karimaa, 2014) focused on targeting specific functional objectives of the video surveillance system, namely concerning video retrieval and processing. This

background research publications contributed to understanding of context of use, and practical implications, contributing to better understanding of requirements for framework drafted in this chapter.

Second category of publications are the ones produced in second phase of the research, where I focused on specific evaluation objectives.

Publication I *Security Aspects of the Complexity of Modern Surveillance Systems: An Experience Report* (Karimaa, 2009) explored **security** aspects of evaluation for systems with growing security demands and growing complexity. The paper proposes the classification of security issues at various architectural levels including a review of the mechanisms to be applied. Based on this, classification the deployment examples are provided and the overview of possible solutions as well as a general conclusion on the security of future surveillance systems. This publication was the first attempt to evaluate system from one characteristic, namely security, point of view. However, no security evaluation was performed due to lack of video-surveillance standard combining both technical and social security aspects.

Publication II *Video Surveillance in the Cloud: Dependability Analysis* (Karimaa, 2011a) focused on system's **dependability.** The paper analyses the challenge of virtualization and new service-oriented architecture and brings a new perspective to the aspect of dependability of video surveillance solutions and other safety-critical applications. While the existing research is focused mainly on security challenges of cloud applications in general, this publication attempts to apply more holistic dependability (in opposite to more specific security) to very specific category of cloud-based IT systems- cloud-based video surveillance. The paper researches the dependability characteristics in context of transition of video surveillance architecture towards cloud solutions and proposes the areas of focus for system development and design. This publication contributed to creating Quality part of the framework, where dependability is identified as one of Quality objectives.

Publication III *Efficient Video Surveillance: Performance Evaluation in Distributed Video Surveillance Systems* (Karimaa, 2011b) researched the topic of **performance** evaluation. It is worth to refer again to most common evaluation types listed in Chapter 3.2: quality and functional performance evaluations. While Publication II contributed quality-objective research, Publication III contributed to functional performance area creating an input for performance-focused parts of proposed evaluation framework. Publication III addresses the problem of improving general efficiency of the system by evaluating and improving its performance. Whereas most of the current research was focusing on the performance of individual system components or system functionalities, the paper proposed to analyse the overall system performance by identifying the most critical areas- the very same concept I have used when creating evaluation framework in this chapter. It proposes different perspectives of performance starting from technical aspects (performance

of data retrieval, analysis and fusion) to more subjective ones, such as quality of user interface. The paper provides an example on how to define areas critical for improved system performance and for each of these areas we review techniques and methodologies for measuring the performance. This publication contributed not only to the performance part of evaluation framework but provided the first example of how we can define functional performance areas of interest for video surveillance systems. The publication proposed metrics for measuring the performance of these specific functional areas, which later became evaluation framework objectives in the framework proposed in this chapter.

Third category of publications are ones focused on **capturing multi-dimensional** aspect of system value.

Publication V *Rating Functional Design: Evaluation and Scoring for Systems with Video Sources* (Karimaa, 2013a), where I described the first proposal of functional evaluation aiming to combine earlier researched evaluation characteristics the research entered a next phase searching for a way to combine multiple evaluation characteristics. While in the industry the evaluation measurements are typically limited to few metrics, this article presented more holistic approach: sets of applicable metrics identifying the efficiency for specific selected (critical) system functional areas and then combined using scoring system. This publication presented a scoring system in the form as described in this chapter and demonstrated it on example video system in cloud environment. The paper presentation was well received and encouraged me to test the concept with multiple evaluation stakeholders and their feedback contributed to validation of the concept with selected use cases.

Publication VI *Value-Aware Approach to Management of Innovative Software Products and Services* (Karimaa, 2013b) summarized this phase research on value-based software concept and multi-stakeholder approach to evaluation of software product. This publication was relevant to validate concept against its usability for reflecting relevant aspect of value for multiple stakeholders. This publication contributed to improvement in framework as well as selection of the case studies for framework validation.

5 Testing hypothesis-case studies

This chapter describe case studies used to test our original hypothesis: Every video surveillance system can be evaluated in quantitative way reflecting evaluation needs of various stakeholders across different stages of system life cycle.

I test the proposed framework against the hypothesis above by demonstrating how framework (developed based on research questions RQ1, RQ2 and RQ3 and related publications) supports two practical cases demonstrating system evaluation in different stages of live cycle.

In earlier chapters I have proposed valuation metrics as well as areas of focus for valuation of different types of surveillance system. In this chapter we aim to test the theory of valuation for selected categories of surveillance systems and check if results are conclusive. To proof the theory is correct we aim to provide analysis of case system by evaluating metric its areas of focus.

5.1 Case study of deployed system

The structure of the system used as test platform is presented on Figure 6 Testing – platform system. Relevant parameters and device specific have been added to the **Figure 6**. The system is installed across one building and contains multiple cameras, system nodes and multiple user locations. Each encoder has at least one analogue camera connected, and it encodes the stream to one or multiple encoding standards (MPEG-4 or H.264 in this case). The system contains also IP camera which delivers the streams already encoded. All the system nodes are installed in “server-room” like location. The users are distributed between multiple locations.

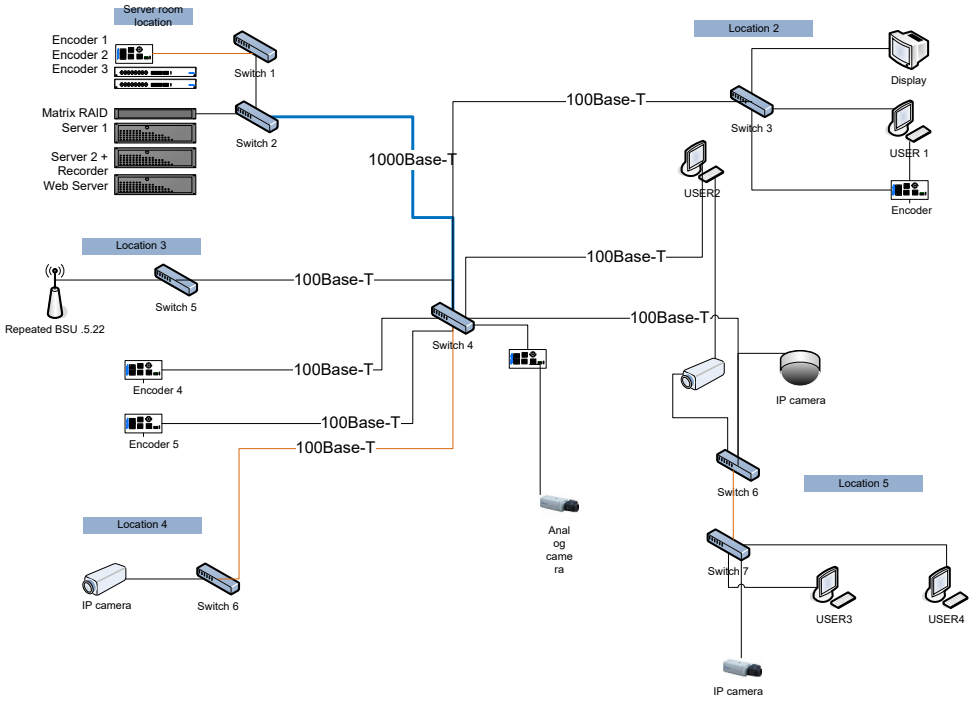


Figure 6. Testing - platform system.

5.1.1 Calculations

The following metrics were tested for VMX platform. The detailed description of tests is available under Measurement description of each metrics as described in Chapter 4.4.

5.1.1.1 System Quality

Technology Readiness

All technologies used in Testing System are available in TLR=9, therefore there is no need to add any margins to final System Score

$$System\ Score\ margin = \frac{9 - TRL}{(TRLmax)} = \frac{9 - 9}{20} = 0$$

System Availability

We define system (and surveillance service) as available when system can perform server functions as well as record the material. Redundant server is in the same machine as system recorder. Recoding operation is functionally independent from server (failure of server does not stop the recording). Software updates are performed twice a year. During software update unavailability of individual component is on the level of 30 minutes, therefore availability of single software components are:

$$A_{SERVER1} = A_{SERVER2} = A_{RECORDER1} = 99.99\%.$$

Both machines are based on Dell Power Edge Server, therefore according to information from Dell (DELL, 2018) availability of individual hardware build is:

$$A_{HW1} = A_{HW2} = 99.98\ \%$$

Therefore

$$A_{SERVER} = 1 - (1 - (A_{HW1} * A_{SERVER1}) * (A_{HW2} * A_{SERVER2}) = 99.99\%.$$

Recorder availability depends also on storage matrix which has been defined as

$$A_{MATRIX} = 99.9942\%\ (DELL, 2018).$$

Therefore

$$A_{RECORDER} = A_{HW2} * A_{RECORDER1} * A_{MATRIX} = 99.98 * 99.99 * 99.9942 = 99.98\%$$

System availability

$$A = A_{RECORDER} * A_{SERVER} = 99.98\%$$

Therefore, Metric Score for Availability Ax can be calculated

$$E = \frac{Ax - 2}{(max\ Ax\ - min\ Ax)} = \frac{4 - 2}{5} = 0.4$$

5.1.1.2 System Design performance

System Complexity

Video management system contains 7 types of software components: Server, Recorder, Proxy Server (for controlling encoders), Telemetry Server (for Pan Tilt

Zoom camera control), Client application, Web Server providing remote (web) client application access

Since the number of the components is less than 6 the metric is 7.

Metric Score can be calculated as:

$$E = \frac{14 - 7}{7} = 1$$

Video Network utilization

Efficiency of video network utilization is analysed for video path between video encoders or cameras and recorder or user machines or decoding displays.

VNU was calculated for following links

$$\text{VNU} = \max\left(\frac{\text{bandwidth occupied by video content}}{\text{bandwidth available for video content}}\right)$$

$$VNU = \max\left(\frac{\text{number of video streams} * \text{video capacity of stream}}{\text{bandwidth available for video content}}\right)$$

Table 10. Video network utilization.

LINK	STREAMS CONTRIBUTING TO VIDEO CONTENT (MAX)	METRIC SCORE
SWITCH 1- SWITCH 2 **1000 MBPS**	2.5Mbps (H.264 from Encoder 1 2x4Mbps (MPEG-4 from Encoder 2 and 3) 2x4Mbps (2 streams from recorder displayed on Decoding display 4x4x4Mbps (each user displays max 4 streams in workstation GUI displays)	$\frac{2.5 + 8 + 8 + 64}{0.5 * 1000} = 0.165$
SWITCH 4- SWITCH 3 **100 MBPS**	2x4Mbps (2 streams from recorder displayed on Decoding display 4x4Mbps (User 1 displays max 4 streams in workstation GUI displays)	$\frac{8 + 16}{0.5 * 100} = 0.48$

End to End Latency

The following path has been analysed to measure End to End Latency in Test System:

Encoder1/Encoder 2- Switch1-Switch2-Switch4-Switch3- User 1/Display 1

The biggest latency has been observed for MPEG-4 encoder stream decoded using software decoder installed on User workstation and was equal to190–210 ms.

Metric Score E=0.8.

5.1.1.3 Source Data performance

Object resolution

The object resolution has been identified as described in Section 4.4.3.1. Object resolution has been calculated as:

$$\mathrm{OR} = \frac{\text{Object size on image}}{\text{Object actual size}}$$

Results of tests are presented in Table 11

Table 11. Object resolution test results.

OBJECT TYPE	IMAGE SIZE	REAL SIZE	OBJECT RESOLUTION
CAR IN RED	144x47 (height 47)	1.60 height	30 pixel/m
CAR IN GREEN	287x104 (height 104)	1.60 height	65 pixel/m
PERSON	63x149 (height 149)	1.80 height	80 pixel/m

The tests show the system is configured to allow object detection, but does not allow object recognition

Efficient frame rate

The encoder responsible for coding of picture of camera facing parking lot from test is set to 25fps. Functional purpose of the camera setup it to detect human and cars on parking lot where max speed of car on parking is 60 km/h which corresponds to 16 m/s. Path captured by camera is approximately 5m therefore time slot required to capture the car on at least one frame within 0.3125 s which corresponds to minimal frame rate below 1fps.0.3 frames per second which means frame rate should have at least 4 fps.

Efficient frame rate ratio EFR =4/25=0.16

Metric Score E=0.2

Encoding/decoding delay

Measurement test was identical to one described in Section 4.4.3.3but the , network impact was minimized by replacing system network by single switch. The results of the test were comparable to the above test.

Encoding/decoding delay= 190-200ms

Metric Score E=0.7

5.1.1.4 Data acquisition performance

The tests performed to estimate data acquisition accuracy has been described in Appendix 1

Precision

Precision describes amount of relevant data within all the data retrieved. In our test all 8 object were detected contributing to high Precision value

$$\text{Precision} = \frac{\text{Retrieved relevant data}}{\text{All retrieved data}} = \frac{8}{8} = 1$$

Metric Score =1

Recall

Recall describes the amount of relevant data that has been retrieved within all the existing relevant data which illustrates

$$\text{Recall} = \frac{\text{Retrieved relevant data}}{\text{All existing relevant data}} = \frac{8}{8} = 1$$

Metric Score =1

MODA

The numbers observed during the test described in were as follows:

- missed detections (m) = 0
- false positive count (fp) =0
- g as the number of ground truth objects in the frame = 8

Therefore, MODA calculated as:

$$\text{MODA} = 1 - \frac{m+fp}{g} = 1$$

Metric Score =1

5.1.1.5 User Interface Efficiency

User interface friction- Time to Switch between Displays

The test designed Time to Switch between Displays was performed by connecting encoded streams to selected display in user interface application. First test was first stream coming from MPEG-4 encoder and then stream coming from H.264 encoder.

The time stamps relevant for given operations were retrieved from logs.

For MPEG-4 stream connection time was present in logs as:

```
2012.02.24_11.29.51.51> D XUR> ConnectDisplay ID=2, ThID=1078,
blocking=0
```

And video appeared on screen at:

```
2012.02.24_11.29.52.34> I [VMX 2] InitPlayer Stream detected
```

Therefore TSD $_{MPEG-4}$= 52.34-51.51 s =0.83 s = 830 ms

`H.264` stream connection and visualization time was present in logs as:

```
2012.02.24_11.47.58.49> D XUR> ConnectDisplay ID=2, ThID=1264,
blocking=0
```

```
2012.02.24_11.47.59.90> I [VMX 2] InitPlayer Stream detected
```

Therefore TSD $_{H.264}$ = 59.90 – 58.49 s =1.41 s

Observed TDS is above 1 second and therefore will be observed by user as delay in connection however it will not influence quality of operation

Metric Score E= 0.6

Workflow support- Click to Achieve

Table 12 list of functions requiring fast access together with number of operations needed to be performed on user interface to access, achieve or complete given function

Table 12. Workflow support metrics.

FUNCTION NAME	NUMBER OF OPERATIONS NEEDED TO ACCESS
CAMERA CONTROL	1 click (directly accessible from GUI)
VISUALIZING CAMERA ON SELECTED DISPLAY	1 drag&drop
SEARCH OF RECORDED MATERIAL BASED ON TIME CRITERIA	2: - Select the timeslot from time bar - Press search
EXPORT OF SELECTED MATERIAL TO EXTERNAL MEDIA	3

Number of operations for any individual function does not exceed 3and therefore Metric Score =0.8

5.1.1.6 System Intelligence

Functional performance: False Alarm Rate (FAR) and Probability of Detection (POD)

The tests performed to estimate data acquisition accuracy has been described in Appendix 1

Functional goal of the system was to detect and recording the events from parking lot and similar places. In case of incident detection systems False Alarm Ratio (FAR) is reliable measure of system intelligence. FAR was observed in the result of the test described in Section 4.4.6.1 and 4.4.6.2.

$$\mathrm{FAR} = \frac{\text{data misinterpreted as relevant}}{\text{all identified data}} = \frac{0}{8} = 0$$

Probability of Detection (POD) for the same test:

$$\mathrm{POD} = \frac{\text{identified relevant data}}{\text{all relevant data}} = \frac{8}{8} = 1$$

Therefore, the Metric Score for both metrics is E=1

5.1.2 Results and analysis

The analysis of testing platform using scoring system reveals:

- System Design – score 0.66 (out of max 1)

- User Interface - score 0.70
- Video Source Efficiency- score 0.37 – this indicates area for improvement
- Video acquisition and VIA accuracy – score 1, ideal
- System intelligence- score 1, ideal

These results are embedded to scoring system calculations with relevance of evaluation objectives as presented in Table 13.

Table 13. Evaluation results - deployed system case.

METRIC	RELEVANCE	DESCRIPTION
AVAILABILITY	Critical	Availability is relevant as the system should be available for customer presentation.
SYSTEM DESIGN	Critical	System design is importantly relevant for deployed system, Number of SW component to be updated and maintained is proportional to system maintenance effort. Since the system is presentation platform with frequent SW updates the effort should be minimized. Network design is important and e.g. link capacities and camera location should be considered
USER INTERFACE DESIGN	Critical	Deployed system should operate efficiently from user perspective therefore, interface capabilities and performance are critical
SOURCE DATA	Critical	While cameras and camera locations could be adjusted, the selection and camera settings impact to system value and functionality is critical for system performance
INCIDENT DETECTION INTELLIGENCE	Critical	System efficiency is achieving good level of incident detection should be possible to be presented

When such system description was applied the overall System Score was equal to 0.69. Calculations results are presented in Figure 7.

Based on above analysis we can conclude that to improve the system we should focus on improving the efficiency of camera setup, as well as Time to Switch between Displays (TSD) metric should be improved.

Metrics	Metrics scores calculated from measurements (0-1)	Targets (Evaluation Objectives)		Target Score (range 0-1) AVERAGE of all Metric Scores per Target)	Relevancy (range 0-1) 1- Critical, 0.75- Important, 0.5- Relevant, 0.25 - Not Important, 0- Not Relevant)	System Score (range 0-1) average of Relevancy weighted Target scores	System Score (range 0-1) and its accuracy margin (0= most accurate) System Score - average of relevancy weighted Target scores.
Technology readiness	*1*	*Maturity margin (TRL 9- 0)*		*0.0*	*1*	*0*	*0*
Availability	0.4	Availability	Dependability	0.4	1	0.4	
System Complexity	1	System Complexity	System Design	0.76	1	0.76	
Network Utilization	0.48	Network Efficiency					
End-to-end Latency	0.8	System Latency					
Object resolution	0.2	Camera / encoding	Source Data	0.37	1	0.37	
Minimum frame rate	0.2						
Encoding/decoding delay	0.7						
Precision	1	Data acquisition	Data acquisition	1	1	1	
Recall	1	Object detection					
MODA	1	Object tracking					
Time to Switch between Displays	0.6	User Interface Friction	User Interface Design	0.70	1	0.7	
Clicks to achieve	0.8	Workflow support					
POD	1	Incident detection	System Intelligence	1	1	1	0.70
FAR	1	intelligence					
Speed of learning	1	Learning					

Figure 7. Scoring system results (0.69) for deployed system case.

Case study – system concept

The structure of the system used as test platform on Figure 8. Relevant parameters and device specific have been added to the Figure. This multi-camera, multi-location system is to be deployed in wide area. User locations include remote users and Control Room users. The system is installed on local system server, but recording nodes are in cloud. Camera locations contain similar encoders and IP cameras as described in **Figure 8**.

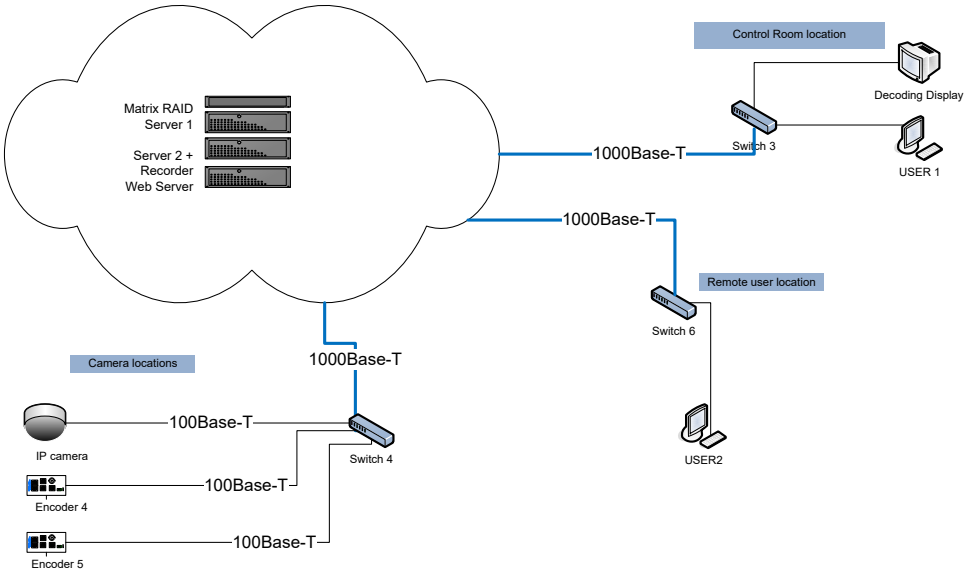


Figure 8. Concept system.

5.1.3 Calculations

5.1.3.1 System Quality

Technology Readiness

This system concept introduces two new technologies:

- Cloud computing technology allows running core applications in the cloud.
- Cloud storage technology allows storing recorded material in the cloud.

In case of surveillance solution cloud storage technology is available as TRL=7 as there are examples of systems able to demonstrate storing the material on virtual disks, but their technology is not in wide use (mainly due to security restrictions).

However, the technology of cloud computing is more immature. We assume that in case of surveillance systems the technology is available as TRL=3 (Analytical and experimental critical function and/or characteristic proof-of concept), the technology is not available yet as option for most of the systems, but some experiments can be conducted based on existing systems (Iveda Solutions, 2018).

Therefore, we can calculate:

- for cloud computing technologies:

$$\text{System Score margin} = \frac{9-3}{(\text{TRLmax})} = \frac{6}{20} = 0.3$$

- for cloud storage technologies:

$$\text{System Score margin} = \frac{9-7}{(\text{TRLmax})} = \frac{2}{20} = 0.1$$

If both technologies are utilized the System Score margin should be calculated as cumulative margin

$$\text{System Score margin} = 0.22 + 0.67 = 0.4$$

System Availability

We define system (and surveillance service) as available when system can perform server functions as well as record the material. Recoding functionality is independent from server availability (failure of server does not stop the recording). Software updates are performed once a year and down time of individual service caused by software update is less than 5 minutes what guarantees the availability of Five Nines:

$$A_{SERVER} = A_{RECORDER} = 99.999\%.$$

Availability of cloud services are expected to be on the level of Five Nines:

$$A_{CLOUD} = 99.999\ \%$$

Therefore,

$$A_{SERVER} = A_{CLOUD} * A_{SERVER} = 99.999\%$$

Similarly, the Recording service availability can be calculated as

$$A_{RECORDER} = A_{CLOUD} * A_{RECORDER} = 99.999\%.$$

System availability:

$$A = A_{RECORDER} * A_{SERVER} = 99.999\%$$

Metric Score can be calculated as:

$$E = \frac{Ax - 2}{(\max Ax - \min Ax)} = \frac{5 - 2}{5} = 0.6$$

Note: The above analysis displays stress of efficient software update on overall system availability. In order to benefit from high availability offered by cloud services the process of software upgrades should be designed to guarantee availability of Five Nines which means service down time should not exceed 3min per year.

5.1.3.2 System Design performance

System Complexity: Video management system contains 8 types of software components:

- Server application
- Recording application
- Proxy Service application (for controlling encoders),
- Telemetry Server application (for Pan Tilt Zoom camera control)
- Client interface for Control Room location
- Web Server providing remote (web) client application access
- Transcoding components for downscaling video traffic

Metric Score can be calculated as:

$$E = \frac{14 - SC}{7} = \frac{14 - 7}{7} = 1$$

Note: Immediate transition towards utilization of cloud services might contribute to additional software components. The architecture should be evaluated to minimize the number of functional component types to avoid complexity and minimize future maintenance and development effort.

Video Network Utilization

VNU metric helps to design the system in the efficient way. Where connection to cloud can be easily scaled using upload/download links capacity the Video Network Utilization can be used to optimize system efficiency.

The example of such design is as follows:

- Analysis of Camera location – Cloud link

Assuming the camera location contains single camera or encoder streaming MPEG-4 (25fps@D1) camera stream the bandwidth occupied by such stream is about 4Mbps. Assuming the optimal Video Network Utilization should be achieved the upload bandwidth on the link Camera location-Cloud should have 8Mbps capacity

Note: It is worth to notice the following solution introduce some challenges, e.g.: the buffer of encoding node should be maximized to allow fluent transmission- cloud services guaranteeing up-streaming of content are not available to the knowledge of the author of this work and / or up-streaming bandwidth is relatively more expensive than downstream bandwidth

- Analysis of Remote User location – Cloud link

Displaying of one camera stream of MPEG-4 (25fps@D1) consumes 4Mbps bandwidth. Therefore, available download bandwidth at remote user location should be on the level of 8Mbps.

- Analysis of Control Room location – Cloud link

Control Room location typically contains the functionality of displaying multiple streams. If four MPEG-4 (25fps@D1) are viewed by operators in control room the available download bandwidth at Control Room location, assuming VNU being on optimal level is 32Mbps (4 streams of 4Mpbs occupying 50% links capacity)

Metric Score = 1 at the phase of concept design but the measurement shall be repeated when solution matures, and all link capacities are known.

End-to-End Latency

There are two critical paths which should be analysed for proposed concept system:

- Camera location-Cloud- Remote User location
- Camera location-Cloud- Control Room User location

Estimates for End-to-End Latency for each of the paths can be created as sum of below delays available cloud services:

- Streaming latency for uploading the stream to cloud
- Streaming latency for downloading the stream from cloud.

The second one can be estimated to be on level below 200ms based on down streaming parameters available on cloud services, such (Spotify, 2018) or (YouTube, 2018).

However, the first latency is more problematic. Cloud services are based on file upload and latency-less services for up streaming video do not exist to the knowledge of the author. We can assume such delay will be on the level of few seconds.

Therefore, the End of End Latency for live video will be on the level of few seconds and will contribute to Metric Score of E=0.

5.1.3.3 Source data

Object resolution and Efficient frame rate ratio

These metrics can be designed in the way to achieve optimum efficiency. Metric Score E=1.

Encoding/decoding delay

The situation is identical to the one described in End-to-End System Latency. Metric Score E=0.

5.1.4 Data acquisition

We assume data acquisition efficiency of individual video sources can be optimized to achieve. Metric Score E=1.

5.1.4.1 User Interface

User interface friction- Time to Switch between Displays

Due to the fact the retrieving the live image from video sources towards cloud might introduce the delay of few seconds (refer to End to End Latency and Encoding/Decoding Delay metrics) it is correct to assume the Time to Switch between Displays can reach the level above 4 seconds. Metric Score E= 0.

Workflow support- Click to Achieve

We can assume the user application could be identical to one used for previously deployed systems. Therefore, workflow support performance can be assumed to be identical to one calculated in Section 5.1.1.5. Metric Score =0.8

5.1.4.2 System Intelligence

False Alarm Rate (FAR), Probability of Detection (PD) and Speed of Learning (SL)

We can assume the metrics are identical to one used for previously deployed systems as the characteristics of encoding devices remains unchanged. Therefore, workflow support can be assumed to be identical to one calculated in Section 5.1.1.6. Metric Score =1

5.1.5 Results and summary

The analysis of testing platform using scoring system reveals:

- System Design – score 0.67 (out of max 1)
- User Interface - score 0.40
- Video Source Efficiency- score 0.67
- Video acquisition and VIA accuracy – score 1, ideal
- System intelligence- score 1, ideal

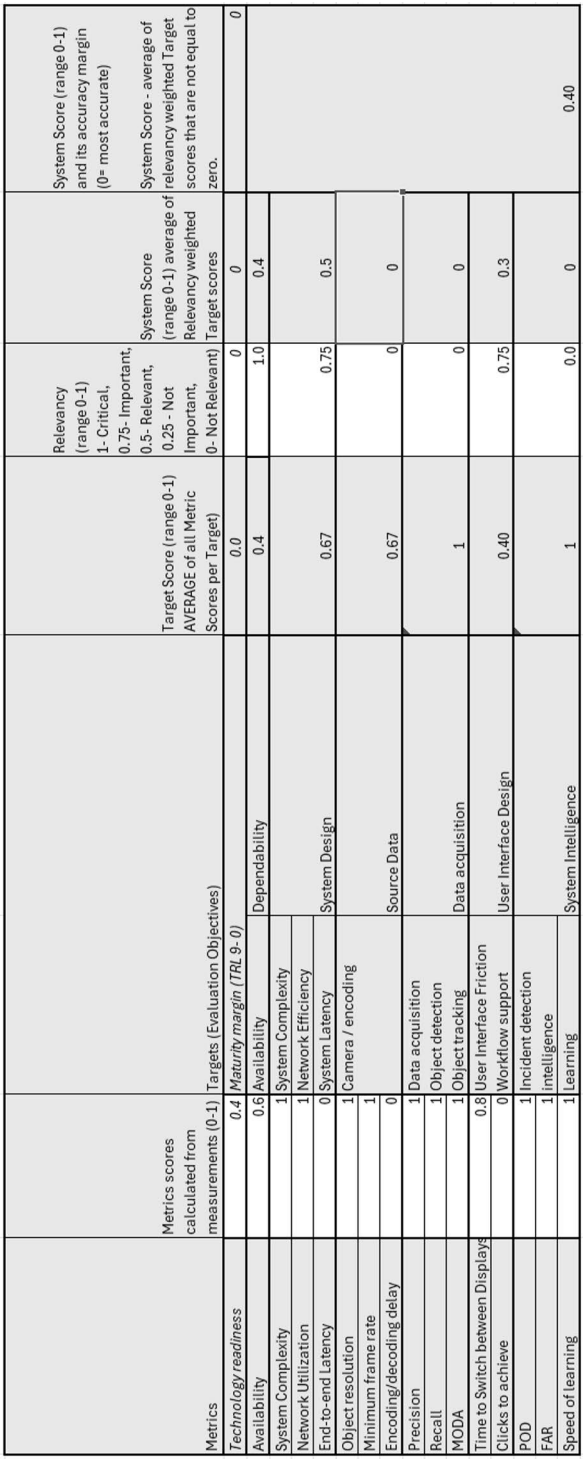
These results are embedded to scoring system calculations with relevance of evaluation objectives as presented in Table 14.

Table 14. Evaluation results - system concept case.

METRIC	RELEVANCE	DESCRIPTION
MATURITY	Not relevant	While some less mature technologies will be used, we choose to analyse general design at the early stage and take maturity into account in further evaluations
AVAILABILITY	Crucial	Availability is one of major arguments to go towards cloud-based scenario. It must be possible to be demonstrated
SYSTEM DESIGN	Important	While system design is important as such the metrics measuring system design might be difficult to measure or estimate at this stage of the concept
USER INTERFACE	Important	It is critical but can be improved (especially workflow support)
SOURCE DATA AND DATA ACQUISITION	Not relevant	Data source don't play very important role at this stage of concept design. We can consider them as more critical when reevaluating the concept with very specific camera sources
INTELLIGENCE & LEARNING	Not relevant	System intelligence does not have lot of information at this stage of concept design. We can consider them as more critical when reevaluating at the later stage

When such system description was applied the overall System Score was equal to 0.40. Results are presented in Figure 9

Based on the analysis we can conclude that when improving overall system efficiency, the biggest stress lies on minimizing the latency and delays in system operations on live streams.



Metrics	Metrics scores calculated from measurements (0-1)	Targets (Evaluation Objectives)		Target Score (range 0-1) AVERAGE of all Metric Scores per Target)	Relevancy (range 0-1) 1- Critical, 0.75- Important, 0.5- Relevant, 0.25 - Not Important, 0- Not Relevant)	System Score (range 0-1) average of Relevancy weighted Target scores	System Score (range 0-1) and its accuracy margin (0= most accurate) System Score - average of relevancy weighted Target scores that are not equal to zero.
Technology readiness	*0.4*	*Maturity margin (TRL 9- 0)*		*0.0*	*0*	*0*	*0*
Availability	0.6	Availability	Dependability	0.4	1.0	0.4	
System Complexity	1	System Complexity					
Network Utilization	1	Network Efficiency					
End-to-end Latency	0	System Latency	System Design	0.67	0.75	0.5	
Object resolution	1	Camera / encoding					
Minimum frame rate	1						
Encoding/decoding delay	0		Source Data	0.67	0	0	
Precision	1	Data acquisition					
Recall	1	Object detection					
MODA	1	Object tracking	Data acquisition	1	0	0	
Time to Switch between Displays	0.8	User Interface Friction					
Clicks to achieve	0	Workflow support	User Interface Design	0.40	0.75	0.3	
POD	1	Incident detection					
FAR	1	intelligence					
Speed of learning	1	Learning	System Intelligence	1	0.0	0	0.40

Figure 9. Scoring results for system concept case.

5.2 Reflection to hypothesis and Research Questions

This chapter presented two use cases of applying the framework. The aim is to present the use of the framework and support the hypothesis that *Every video surveillance system can be evaluated in quantitative way reflecting evaluation needs of various stakeholders across different stages of system life cycle*.

The proposed framework was built based on research on the following research questions:

Research on *RQ1: What aspects of evaluation are important when building a framework for video surveillance evaluation?*

This research provided framework design requirements

Publication IV *Mobile and Wireless Access in Video Surveillance* (Karimaa, 2011c) and Publication VII *Survey of Hardware Accelerated Methods for Intelligent Object Recognition of Camera* (Karimaa, 2014) played a great role in building the understanding for evaluation requirements. These two publications represented well two types of most typical evaluation challenges: Publication IV *Mobile and Wireless Access in Video Surveillance* (Karimaa, 2011c) focused on quality attributes, while (security, quality, dependability, usability), while Publication VII *Survey of Hardware Accelerated Methods for Intelligent Object Recognition of Camera* (Karimaa, 2014) focused on targeting specific functional objectives. Combined, these contributed to better understanding of basic requirements for proposed framework in general.

Research on *RQ2: What system characteristics should be subject of evaluation, and can they be quantified?*

This research provided selection of evaluation objectives and their metrics

This research has been captured in multiple publications, contributing not to identification of evaluation objectives research, but proposing relevant metrics to measure these objectives. Publication I *Security Aspects of the Complexity of Modern Surveillance Systems: An Experience Report* (Karimaa, 2009) explored **security** aspects, Publication II *Video Surveillance in the Cloud: Dependability Analysis* (Karimaa, 2011a) focused on system's **dependability** providing input to quality aspects part of the evaluation research**.** Finally, Publication III *Efficient Video Surveillance: Performance Evaluation in Distributed Video Surveillance Systems*

(Karimaa, 2011b) researched the topic of **performance** evaluation proposing performance objectives and the metrics to measure them.

Research on *RQ3: How to form the evaluation framework reflecting multiple dimensions of system "value" as relevant for various stakeholders, use cases and valid across various stages of system life cycle?*

This research resulted framework design, which was presented the framework in Chapter 4 and following that presented with use cases for supporting the hypothesis in Chapter 5.

Two publications were critical for this research and allowed me to capture **multi-dimensional** aspect of system value:

- Publication V *Rating Functional Design: Evaluation and Scoring for Systems with Video Sources* (Karimaa, 2013a), proposed the first draft of scoring-based functional evaluation framework and demonstrated it with the first case ever
- Publication VI *Value-Aware Approach to Management of Innovative Software Products and Services* (Karimaa, 2013b) summarized this phase research on value-based software concept and multi-stakeholder approach to evaluation of software product. This publication validated the evaluation concept against its usability for reflecting relevant aspect of value for multiple stakeholders. This publication contributed to improvement in framework as well as selection of the case studies for framework validation.

6 Conclusion

6.1 Academic conclusions

This chapter answers the Research Questions and provides insights to the content of the dissertation and research conclusions behind it.

RQ1: What aspects of evaluation are important when building a framework for video surveillance evaluation?

The Research Question RQ1 has been answered based on research on both industry and academia video surveillance evaluation programs, frameworks and processes, as well as industry research in form of professional discussions and experiences from using and benchmarking these. The research was represented by two research papers: Publication IV *Mobile and Wireless Access in Video Surveillance* (Karimaa, 2011c) and Publication VII *Survey of Hardware Accelerated Methods for Intelligent Object Recognition of Camera* (Karimaa, 2014). These publications provided examples of two challenges, where evaluation framework would make the difference: evaluation of system **quality** evaluation with different access method and evaluation of system **performance** with new object recognition solution. Presentation of practical use cases of these evaluation, combined with relevant literature review and industry research helped me to draft the framework requirements as presented in Section 4.2 Evaluation framework requirements.

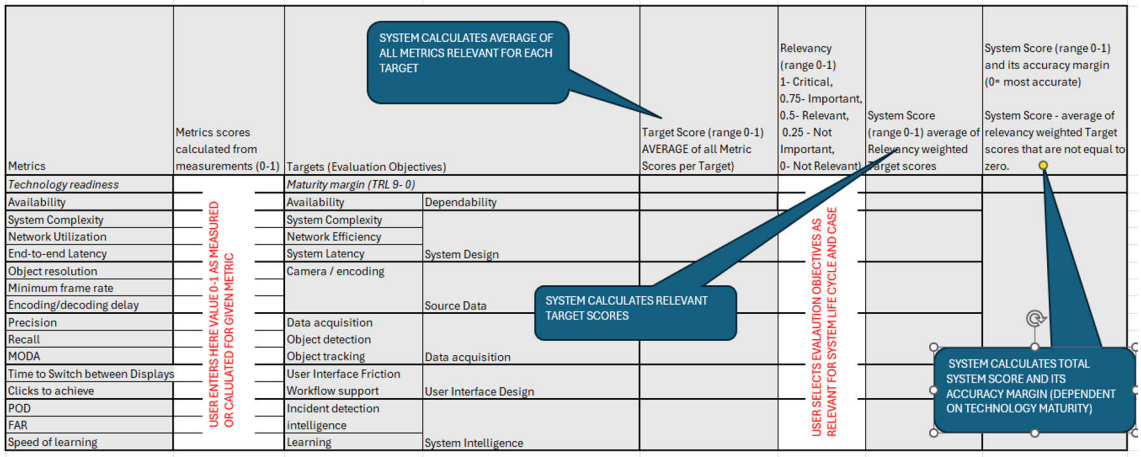
RQ2: What system characteristics should be subject of evaluation, and can they be quantified?

The Research Question RQ2 has been answered based on research presented in Section 3.2 *On video surveillance evaluation objectives*, where the objectives has been divided to quality and performance objectives and relevant publications on various evaluations objectives and metric for quantifications were listed. This research work was captured in multiple papers – each of them focusing on specific evaluation characteristics. Publication I *Security Aspects of the Complexity of Modern Surveillance Systems: An Experience Report* (Karimaa, 2009) explored **security** aspects, Publication II *Video Surveillance in the Cloud: Dependability*

Analysis (Karimaa, 2011a) focused on system's **dependability** and finally, Publication III *Efficient Video Surveillance: Performance Evaluation in Distributed Video Surveillance Systems* (Karimaa, 2011b) researched the topic of **performance**.

RQ3: How to form the evaluation framework reflecting multiple dimensions of system "value" as relevant for various stakeholders, use cases and valid across various stages of system life cycle?

The (previous) research on video surveillance evaluation characteristics created a good understanding what characteristics should be identified as critical from perspective of video system evaluation. This work allowed me to propose them as evaluation objectives and their metrics as listed in Section 0.



6.1.1 Results uncertainty

This scoring system is early phase proposal that should be developed based on high volumes of cases from industry and research. However, at this early stage of scoring system maturity multiple aspects of results certainty should be considered. First, calculator embeds also scoring result accuracy margin which is dependent on TRL level of used technologies. Using non mature technologies impact's results reliability in terms of creating margin of possible results when technology is maturing. It is possible that in future other aspects introducing performance, efficiency and value uncertainty will be identified, and this part of scoring system will be expanded.

Secondly, variations in metric weights and metrics measurement noise affect directly the final score. There is a risk that metrics scores would need to be modified in future- now for each metrics the impact on value is linear with described steps (E from 0 to 1 with step of 0.2 as presented in Table 3). It is possible that practical

experience of using scoring system will present non-linear dependencies for at least some of the metrics.

Finally, the Relevance grading might need improvement in further use to limit the freedom of user selecting the relevancy. It should be possible to assign default relevancy for example for systems in specific life cycles phases. This is well illustrated in use cases presented in this work.

6.1.2 Ethics and compliance consideration

To ensure ethical and regulatory alignment, the framework can be further developed. It can be implemented by expanding quality objectives with metrics, such as privacy-by-design and auditability. These aspects reflect the system's ability to protect user data and provide transparent, traceable operations. They can be scored using metrics such as compliance with data protection standards (e.g., GDPR), presence of access controls, encryption protocols, logging mechanisms, and the ability to generate audit trails. Each such a new metric is evaluated and translated into a Metric Score, contributing to the overall Target Score for ethics and compliance. This integration ensures that systems are not only functionally efficient but also trustworthy and accountable.

Evaluation framework. The first draft of the framework was including multi-objective evaluation framework was published in Publication V *Rating Functional Design: Evaluation and Scoring for Systems with Video Sources* (Karimaa, 2013a)

Additionally, Publication VI *Value-Aware Approach to Management of Innovative Software Products and Services* (Karimaa, 2013b) provided insights to multi-dimensional evaluations presented in Section 3.3 and multi-stakeholder perspective the system value. This work allowed me to design the framework as presented in Chapter 0 and validate its support of hypothesis as demonstrated in Chapter 5.

Hypothesis: Every video surveillance system can be evaluated in quantitative way reflecting evaluation needs of various stakeholders across different stages of system life cycle.

The hypothesis was confirmed by presenting the evaluation framework and demonstrated using two different case studies relevant for different type of stakeholders.

The evaluation framework presented in Chapter 4 is build based on scoring-system. Scoring system as an evaluation method follows the rationale identified in Section 4.1 and fulfils the requirements specified in Section 4.2 (based on RQ1 research).

Based on RQ2 research, including relevant own publications as well as literature research, I have identified the characteristics critical for evaluation of video surveillance and the same characteristics are then used as evaluation objectives in scoring-based evaluation framework. In the same research (RQ2 and relevant publications) I have identified metrics to be used for measuring these objectives.

Objectives and metrics have been combined using scoring system-based evaluation framework and in the Chapter 5 I have tested how this framework works for two different cases. These cases represented systems in different phases of life cycle and represented interest of different stakeholders: deployed system evaluation case would be in practice applied to existing system to provide minor improvements and they reflect more product engineering perspective, while the evaluation of system in concept phase would be more relevant in the process of system design so more relevant for sales engineering perspective.

Conclusively, in both cases both objectives and metrics were relevant and measurable and moreover, based on demonstrated result of evaluation we could draw the conclusions concerning system value and to get the insights on how to improve the system.

6.2 Practical implications

Current video surveillance market practices, including purchasing procedures and product development strategies, are characteristic of markets with immature technologies. Procurement processes—such as Calls for Tenders—often clearly favour single-system providers to avoid issues related to component incompatibility. As a result, many companies have adopted an expensive one-stop-shop strategy, aiming to provide all surveillance system components in-house, from cameras and sensors to video management, access control, and recording solutions. However, component interfaces in today's surveillance systems are already well standardized by organizations such ONVIF (ONVIF, 2018) or Physical Security Interoperability Alliance (PSIA, 2018) These architectural standards are widely adopted by major market players. This indicates that the market is approaching a transformation phase, where buyers will gradually gain confidence in the maturity of technologies and the interoperability of components. Consequently, buyer priorities will shift toward improving system performance and optimizing the cost of implementation and operation. These trends will gradually reshape the market, giving buyers the freedom to build systems using various fully interoperable components from multiple providers. We can expect to see increasing product diversity, with system components selected based on specific needs, budgets, and functional goals. This

shift highlights the urgent need for evaluation tools that serve both buyers and producers.

The results of this study demonstrate that the value of video surveillance systems can be expressed through numerical measures. Furthermore, the functional performance and quality of a system can be assessed within the context of its implementation environment and across different development phases. This numerical approach to system value holds significant potential for improving products, services, and processes in a maturing industry (see discussion above). As a result, this work has multiple practical implications for industry processes, product development, and the strategic offerings of market players. It is expected that this research, along with the framework proposed in Publication V *Rating Functional Design: Evaluation and Scoring for Systems with Video Sources* (Karimaa, 2013a) will contribute meaningfully to innovation in video surveillance by providing a platform for verifying the value of new solutions.

This study was conducted primarily from the perspective of system implementation. Therefore, the research presented here contributes most directly to practices related to system implementation, design and planning, and procurement processes. A key conclusion is that the ability to measure system value enables buyers to make more informed strategic decisions when selecting and purchasing multi-component systems. Understanding the value of system functionality helps buyers better address the challenges of acquiring individual components that deliver specific functions. This is expected to lead to more efficient system implementations. Similar trends have been observed in the IT security industry, where the outcome has been the development of various national standards and accreditation programs for system evaluation. These programs have produced multiple evaluation tools aimed at simplifying and commercializing the system evaluation process. A notable example is the Common Vulnerability Scoring System developed by the National Institute of Standards and Technology. This tool is now widely used in IT departments and is a popular method for evaluating system security, particularly in complex integration projects prior to procurement.

In addition to its contributions to buyer practices, this study offers valuable insights for producers of surveillance products and services. By influencing buyer decision-making, the research also impacts engineering processes, system development strategies, and general managerial practices. The study highlights functional differences between individual video surveillance solutions, leading to a more tailored approach to evaluating system components.

One of the most important managerial implications of this approach is the commercialization of individual components as independent, fully interoperable products with standard-compliant interfaces. The development of such components should focus on delivering optimal value within specific surveillance contexts. For

example, an event detection module designed for airport environments should aim to maximize the number of detected events—even at the cost of increased false alarms—while optimizing algorithms for specific conditions such as lighting, crowd density, and outdoor weather.

These insights also have implications for system engineering processes, which must be adapted to support continuous monitoring of component value throughout the development cycle. As a result, this research is likely to encourage the adoption of lean and agile product development methods. Moreover, monitoring the value of implemented solutions should become a key managerial responsibility during product development, ultimately strengthening the relationship between managers and customers. Managers should also have a clear understanding of their company's offering strategy, market positioning, and role within specific industry sectors. Consequently, greater emphasis is expected to be placed on creating and managing strategic partnerships to develop ecosystem-based offerings tailored to specific surveillance applications.

Finally, the managerial implications extend to product innovation. The tool presented in this work has strong potential to simplify and commercialize the innovation process. It can be used to assess system value at all stages of product development, even when dealing with new solutions, technological changes, or architectural redesigns.

6.3 Future work

However, while the objective of this work is to provide proof-of-concept in building universal evaluation tools, I expect there will be further developments in the proposed tools in at least the following areas:

- Range of applicable metrics - research concerning metrics and measurement is very dynamic and the selection should be updated with the most applicable metrics being available
- Metrics selection criteria - I assume the criteria of selecting appropriate metrics applicable for given type of the system will be discussed further especially when the tool can be verified when new system concepts appear.
- Finally, the score balancing - ratios describing the relevance of metrics for different categories of systems and formulas used in the tool should be verified when more statistical data is available for different types of systems. I assume the final score formula will be then adjusted.

The accuracy of proposed scoring system depends on a selection of metrics and their coverage in measuring given system area. I suggest that the process of selection of individual metrics could be improved as more metrics become available. When

expanding the scoring system with additional metrics, the correlation between selected metrics should be minimized as well as they should be selected in the way ensuring optimum coverage of system functionality in given area.

Additionally, when a scoring system is tested by more users, the new system objectives and evaluation targets might be identified. When such expansion is done the system should be re-evaluated. It might involve introducing additional coefficients to maintain correct relation between the number of system objectives and the importance of evaluation targets - now the most significant evaluation targets contain the biggest number of system objectives.

The system should be improved, when more evaluation data is available.

6.4 Summary

Cloud computing with storage virtualization and new service-oriented architecture brings new possibilities to the market of video surveillance solutions and other safety-critical applications. However, the criteria of efficient design such systems are not well established. The situation will become more critical when more innovation opportunities appear. Currently, possible innovations are investigated mainly in the context of improving existing solution, whereas the main advantages of today's innovation lay in new system concepts. To evaluate such concepts and build efficient prototypes, we need evaluation tools which are understood, utilized, and further developed by both business and research communities. This work provides an overview of research available today to construct such as tool and proposes the prototype of such tool in form of scoring system

List of References

A. T. Nghiem, F. B. M. T. a. V. V., 2007. ETISEO, performance evaluation for video surveillance systems. *IEEE Conference on Advanced Video and Signal Based Surveillance,* pp. 476-481.

Albus, J. S., 1991. *Outline for a Theory of Intelligence.* s.l., s.n., pp. 473-509.

Ashani, Z., 2009. *Architectural Considerations for Video Content Analysis in Urban Surveillance.* s.l., s.n., pp. 289-289.

CLEAR , 2007. *Classification of Events, Activities and Relationship: Evaluation and Workshop.* [Online] Available at: http://www.clear-evaluation.org/ [Accessed 22 02 2012].

DELL, 2018. *Dell homepage.* [Online] Available at: http://www.dell.com/content/topics/global.aspx/power/en/ps3q02_shetty?c=us&l=en [Accessed 25 02 2012].

Dembski, W. A., 1998. *The design inference: Eliminating chance through small probabilities..* s.l.:Cambridge University Press..

Desurmont, X. et al., 2004. *A seamless modular image analysis architecture for surveillance systems.* s.l., s.n., pp. 66-70.

Disruptive Technology Office, 2007. [Online] Available at: http://www.videorecognition.com/vt4ns/07/vace_brochure.pdf [Accessed 22 02 2012].

Duis, D. &. J. J., 1990. Improving user-interface responsiveness despite performance limitations. *Thirty-Fifth IEEE Computer Society International Conference on Intellectual Leverage*, pp. 380-381.

Finn, B., 2004. *Keeping an eye on transit.* s.l., s.n., pp. 12-16.

Geigenberger, P., 2017. Thioredoxins Play a Crucial Role in Dynamic Acclimation of Photosynthesis in Fluctuating Light. *Molecular Plant,* 3(10), p. 168–182.

Geigenberger, P., 2017. Thioredoxins Play a Crucial Role in Dynamic Acclimation of Photosynthesis in Fluctuating Light II. *Molecular Plant,* 3(10), p. 25–32.

Geigenberger, P., 2017. Thioredoxins Play a Crucial Role in Dynamic Acclimation of Photosynthesis in Fluctuating Light III. *Molecular Plant,* 3(10), p. 11–14.

Hauck, M., Kuperberg, M., Krogmann, K. & Reussner, R., 2009. *Modeling Layered Component Execution Environments for Performance Prediction.* s.l., s.n., pp. 191-208.

Horst, J. A., 2002. *A Native Intelligence Metric for Artificial Systems.* s.l., NIST.

Hsu, D., Sanford, B., Meyers, K. & Love, T., 2013. Response of the μ-opioid system to social rejection and acceptance. *Mol Psychiatry,* 18(11), p. 1211–1217.

IEEE International Workshop on Performance Evaluation of Tracking and surveillance, 2010. *PETS Benchmark data.* [Online] Available at: http://pets2010.net/ [Accessed 22 02 2012].

Inverardi, P., Mangano, C., Russo, F. & Balsamo, S., 2009. *Performance evaluation of a software architecture: a case study.* s.l., s.n., pp. 116-125.

ISO/IEC, 2017. *TS 25011:2017, Service quality models.* [Online] Available at: https://www.iso.org/standard/35735.html [Haettu 2023].

Iveda Solutions, 2018. *Iveda Solutions homepage.* [Online] Available at: http://www.iveda.com/ [Accessed 03 03 2012].

J-P. Gitto, M. B.-M. A. P. D. Z. I. G., 2016. A methodology for Complex System Quality Model Construction - First level. *IFAC Conference Paper Archive (Science Direct),* pp. 319-324.

Jun-Tao, L. &. X.-Y. J., 2009. An approach to performance evaluation of software architecture. *2009 First International Workshop on Education Technology and Computer Science*, 3, pp. 853-856.

Karimaa, A., 2009. *Security aspect of the complexity of modern surveillance systems: An experience report.* Potsdam, s.n., pp. 120-125.

Karimaa, A., 2011a. *Video Surveillance in the Cloud: Dependability Analysis.* s.l., s.n., pp. 92-95.

Karimaa, A., 2011b. Efficient Video Surveillance: Performance Evaluation in Distributed Video Surveillance Systems. In: *Video Surveillance.* s.l.:InTech, pp. 17-26.

Karimaa, A., 2011c. Mobile and Wireless Access in Video Surveillance System. *International Journal of Digital Information and Wireless Communications,* 1(1), pp. 267-272.

Karimaa, A., 2011. *Value-Based Software Product: Value assessment and management for complex software products,* Turku: Turku School of Economics, MBA thesis.

Karimaa, A., 2013a. Rating Functional Design: Evaluation and Scoring for Systems with Video Sources. *International Journal of Soft Computing and Software Engineering,* Osa/vuosikerta 3.

Karimaa, A., 2013b. Value-Aware Approach to Management of Innovative Software Products and Services. *Journal of Economics, Business and Management,* Volume 1, pp. 66-71.

Karimaa, A., 2014. *A Survey of Hardware Accelerated Methods for Intelligent Object Recognition on Camera.* s.l., Springer, Cham, p. 523–530.

Kasturi, R. et al., 2009. Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 319-36.

Krishnakumar, K., 2003. *Intelligent systems for aerospace engineering – an overview,* s.l.: s.n.

Kulikowski, C. A. & Weiss, S. M., 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems..* s.l.:Morgan Kaufmann Publishers.

Linkens, D. & Nyongesa, H., 1996. *Learning systems in intelligent control: an appraisal of fuzzy, neural and genetic algorithm control applications.* s.l., s.n., pp. 367 - 386.

Lipton, A. et al., 2004. ObjectVideo forensics: activity-based video indexing and retrieval for physical security applications. *IEE on Intelligent Distributed surveillance Systems,* pp. 56-60.

Mabrouk, A. B. a. E. Z., 2018. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications ,* pp. 480-491.

MacLean, A., Carter, K., Lovstrand, L. & Moran, T., 1990. *User-tailorable systems: pressing the issues with buttons.* New York, s.n., pp. 175 - 182.

Mankins, J. C., 1995. *Technology readiness levels.* s.l.:s.n.

Manohar, V. S. P. R. H. G. D. K. R. &. G. J., 2006. Performance evaluation of object detection and tracking in video. *Asian Conference on Computer Vision* , pp. 151-161.

Meneely, A. S. B. &. W. L., 2013. Validating software metrics: A spectrum of philosophies. *CM Transactions on Software Engineering and Methodology (TOSEM)*, pp. 1-28.

Menzies, T., DEkhtyar, A., Distefano, J. & Greenwald, J., 2007. Problems with Precision: A Response to "Comments on 'Data Mining Static Code Attributes to Learn Defect Predictors'. *IEEE Transactions on Software Engineering,* pp. 637-640.

Meystel, A., 2007. *Measuring Performance of Systems with Autonomy: Metrics for Intelligence of Constructed Systems,* s.l.: NIST.

Miller, R. B., 1968. *Response time in man-computer conversational transactions.* s.l., s.n., pp. 267-277.

Ming Liu, Z. X., 2021. *A Fuzzy Post-project Evaluation Approach for Security Video Surveillance System,* s.l.: Cornell University.

Nielsen, J., 1993. Iterative user-interface design. *IEEE Computer,* pp. 32-41.

ONVIF, 2018. *ONVIF homepage.* [Online] Available at: http://www.onvif.org [Accessed 5 11 2012].

Ou, G. &. M. Y. L., 2007. Multi-class pattern classification using neural networks.. In: *Pattern recognition.* s.l.:s.n., pp. 4-18.

Pavlopoudou, C., Martin, D. & Yu, S. J. H., 2009. *Learning from disagreements: Discriminative Performance Evaluation.* s.l., s.n., pp. 1-8.

PETS, 2004. *Benchmark Data for PETS-ECCV.* [Online] Available at: http://www-prima.inrialpes.fr/PETS04/caviar_data.html [Accessed 25 02 2012].

PSIA, 2018. *PSIA homepage.* [Online] Available at: http://www.psialliance.org/ [Accessed 5 11 2012].

Rudas, I. J. & Fodor, J., 2008. *Intelligent Systems.* s.l., s.n., pp. 132-138.

Sharma, V. S. J. P. &. T. K. S., 2005. *Evaluating performance attributes of layered software architecture.* Berlin: Heidelberg: Springer Berlin Heidelberg..

Sjaardema, T. A., Smith, C. S. & Birch, G. C., 2015. *History and Evolution of the Johnson Criteria.* Albuquerque, NM (United States): Sandia National Lab. (SNL-NM).

Spotify, 2018. *Spotify - a word of music.* [Online] Available at: http://www.spotify.com/ [Accessed 03 03 2012].

Tay, A. & Cockburn, A., 1996. *User interfaces for workflow systems: designing for end-user tailorability.* s.l., s.n., pp. 302-303.

Ulrich, K., 1995. *The role of product architecture in the manufacturing firm,* s.l.: Elsevier.

V. Tsakanikas, T. D., 2018. Video surveillance systems- currect status and future trends. *Computers & Electrical Engineering,* pp. 736-753.

Woodside, M. F. G. &. P. D. C., 2007. The future of software performance engineering. *Future of Software Engineering (FOSE'07),* pp. 171-187.

Yin, R. K., 1994 (2003). *Case Study Research.* London: SAGE Publications.

YouTube, 2018. *YouTube - Broadcast Yourself.* [Online] Available at: http://www.youtube.com/ [Accessed 03 03 2012].

Zuoxian, N., J.Xin-hua, L.Jian-cheng & Shen, J., 2009. *Performance analysis of workflow instances.* s.l., s.n., pp. 865-869.

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2025

ISBN 978-952-02-0480-8 (PRINT)
ISBN 978-952-02-0481-5 (PDF)
ISSN 2736-9390 (Painettu/Print)
ISSN 2736-9684 (Sähköinen/ Online)