NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

GRADUATE PROGRAM
INFORMATION AND COMMUNICATION TECHNOLOGY

MSc THESIS

A survey on Quality of Experience of Virtual and Augmented Reality environments

Leonidas M. Paragioudakis

Supervisor: **Lazaros Merakos,** Professor

ATHENS

JUNE 2021

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επισκόπηση Ποιότητας της Εμπειρίας σε περιβάλλοντα Εικονικής και Επαυξημένης Πραγματικότητας

Λεωνίδας Μ. Παραγιουδάκης

Επιβλέπων **Λάζαρος Μεράκος,** Καθηγητής

ΑΘΗΝΑ
ΙΟΥΝΙΟΣ 2021

MSc THESIS

A survey on Quality of
Experience of Virtual and Augmented Reality environments

Leonidas P. Paragioudakis
S.N.: M1586

Supervisor: **Lazaros Merakos,** Professor

June 2021

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επισκόπηση Ποιότητας της Εμπειρίας σε περιβάλλοντα Εικονικής και Επαυξημένης Πραγματικότητας

Λεωνίδας Μ. Παραγιουδάκης
Α.Μ.: Μ1586

Επιβλέπων **Λάζαρος Μεράκος**, Καθηγητής

Ιούνιος 2021

ABSTRACT

Virtual and augmented reality applications constitute a new and promising technology, with applications in medicine, education, gaming, e-commerce and many more. This technology poses a new challenge to application designers, as well as network service providers, because it is resource intensive in order to be able to provide the immersive experience to its users. This task becomes even more challenging in mobile networks, due to factors that are difficult to be modeled and predicted, such as mobility, handoff strategy and resource allocation. This thesis aspires to provide a review of Quality of Experience (QoE) estimation and provision techniques and methods that have been developed around these applications.

In the first section, we review QoE provisioning strategies for virtual reality applications. This section examines some corner cases of augmented reality applications, such as a heavy machinery simulation software, an educational application, and many digital immersive applications. The scope and diversity of applications and implementation methods lead to some conflicting conclusions in relation to QoE evaluation methods.

In the next section we refer to augmented reality applications, again with a reference to a wide variety of applications such as an augmented reality task assistant, augmented reality video games and digital immersive applications. The conclusions in this section are more robust and peoples' feelings can form more meaningful aggregations.

In the last section we investigate the QoE in virtual and augmented reality applications when these applications are implemented in mobile devices. This part is concerned with more technical aspects such as mobility management, handoff strategies and resource allocation algorithms and their impact on users' experience.

SUBJECT AREA: Quality of Experience

KEYWORDS: Quality of Experience, Mean Opinion Score, Virtual Reality, Augmented Reality

ΠΕΡΙΛΗΨΗ

Οι εφαρμογές εικονικής και επαυξημένης πραγματικότητας αποτελούν μια νέα και πολλά υποσχόμενη τεχνολογία, με εφαρμογές στην ιατρική, την εκπαίδευση, τα βιντεοπαιχνίδια, το ηλεκτρονικό εμπόριο και πολλές άλλες. Αυτή η τεχνολογία θέτει μια νέα πρόκληση για τους σχεδιαστές εφαρμογών, καθώς και για τους παρόχους υπηρεσιών δικτύου, επειδή είναι εντατική ως προς τους απαιτούμενους πόρους ώστε να είναι σε θέση να παρέχει συναρπαστική εμπειρία στους χρήστες της. Αυτή η πρόκληση γίνεται ακόμη πιο δύσκολη στα δίκτυα κινητής τηλεφωνίας, λόγω παραγόντων που είναι δύσκολο να μοντελοποιηθούν και να προβλεφθούν, όπως η κινητικότητα, η στρατηγική μεταβίβασης και η κατανομή πόρων. Αυτή η διπλωματική φιλοδοξεί να παράσχει μια ανασκόπηση των τεχνικών και μεθόδων εκτίμησης της ποιότητας της εμπειρίας (QoE) και των μεθόδων που έχουν αναπτυχθεί γύρω από αυτές τις εφαρμογές.

Στην πρώτη ενότητα, εξετάζουμε τις στρατηγικές παροχής QoE για εφαρμογές εικονικής πραγματικότητας. Αυτή η ενότητα εξετάζει αρκετές περιπτώσεις εφαρμογών εικονικής πραγματικότητας, όπως ένα λογισμικό προσομοίωσης βαρέων μηχανημάτων, μια εκπαιδευτική εφαρμογή και άλλες ψηφιακές εφαρμογές εμβύθισης. Το εύρος και η ποικιλία των εφαρμογών και των μεθόδων εκτίμησης της ποιότητας της εμπειρίας οδηγούν σε αντικρουόμενα συμπεράσματα σχετικά με τις μεθόδους αξιολόγησης QoE.

Στην επόμενη ενότητα αναφερόμαστε σε εφαρμογές επαυξημένης πραγματικότητας, και πάλι με μια αναφορά σε μια μεγάλη ποικιλία εφαρμογών, όπως ένας βοηθός επαυξημένης πραγματικότητας, βιντεοπαιχνίδια επαυξημένης πραγματικότητας και άλλες ψηφιακές εφαρμογές. Τα συμπεράσματα σε αυτήν την ενότητα είναι πιο ισχυρά και τα συναισθήματα των ανθρώπων μπορούν να σχηματίσουν πιο ουσιαστικά πορίσματα.

Στην τελευταία ενότητα διερευνούμε την QoE σε εφαρμογές εικονικής και επαυξημένης πραγματικότητας για κινητές συσκευές και κινητά δίκτυα. Σε αυτό το μέρος ασχολούμαστε με πιο τεχνικές πτυχές όπως η διαχείριση της κινητικότητας, οι στρατηγικές μεταβίβασης και οι αλγόριθμοι κατανομής πόρων και ο αντίκτυπος που έχουν αυτοί στην εμπειρία των χρηστών.

SUBJECT AREA: Ποιότητα της εμπειρίας

KEYWORDS: Ποιότητα Εμπειρίας, Εικονική Πραγματικότητα, Επαυξημένη Πραγματικότητα

To my family

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The present work, entitled “A survey on Quality of Experience of Virtual and Augmented Reality environments”, is a survey on Quality of Experience estimation methods and techniques that apply to virtual and augmented reality applications. This thesis has been written to fulfill the graduation requirements of the Information and Telecommunications technologies program at the National and Kapodistrian University of Athens. It was written in Athens, from January to June 2021.

The motivation for this thesis was the post graduate course of “Mobile and wireless networks” where I first encountered the term Quality of Experience from my current co-supervisor, Eirini Liotou. The research was difficult but conducting extensive investigation has allowed me to answer the main questions of the thesis. Fortunately, Eirini was always available and willing to answer my queries with clarity.

I would like to thank my supervisors Prof. Lazaros Merakos and Dr. Eirini Liotou for their excellent guidance and support during this process. I also wish to thank all of the respondents; without whose cooperation I would not have been able to conduct this analysis.

I also benefitted from debating issues with my friends and family. If I ever lost interest, you kept me motivated. My parents deserve a particular note of thanks: your wise counsel and kind words have, as always, served me well.

I hope you enjoy your reading.

1 INTRODUCTION

The estimation and measurement of the impact of the technology in the everyday life plays a crucial role in the advancement of technology and the spread of the technology to the world, and many times is equally important as the technology itself. This is particularly important when it comes to new forms of technology, such as Virtual reality (VR) and Augmented reality (AR) applications.

In the past, this challenging task had been approached with a technocratic conception, in terms of Quality of Service (QoS). According to this conception, the measurement of and provision for Key Performance Indicators (KPI's), such as bit rate or jitter was enough to ensure that the end users enjoy a service, or an application.

On the other hand, modern day application and service providers have concluded that this implicit way of estimating the annoyance or delight of the final user does not reflect the actual user experience and can lead to unsatisfied users and service or application providers that do not have a clue about their customers' feelings.

Quality of Experience (QoE) is a more human-centric approach to address this situation, namely, to be able to express in more precise terms the degree at which a customer is satisfied with a product or a service. This common ground between providers and customers is valuable in the modern-day markets, where competition for products and services is fierce. This was the initial motivation for the current thesis.

The estimation of the users' delight or annoyance from a product or a service can be measured with many methods. The most popular and accurate group of these methods is the direct measurement of the users' feelings. This approach is the base upon which every other method is built. Although this method is accurate and provides a clear estimation of users' feelings, it is not always easy to be implemented, nor is it scalable. Most of the studies reviewed in this thesis use this method, either directly, or as a validation for the results of other methods.

On the other hand, the estimation of QoE does not always show positive user feelings. This is especially true for VR and AR applications, where users often experience sickness feelings. The estimation of this negative experience is very important for application designers and network providers, in order to be able to improve aspects of the applications or services.

This estimation becomes even more important if we consider that these types of services and applications have been created to be used with mobile devices. Applications like remote medical services or on the job training for error prone or dangerous tasks can prove the importance of QoE provisioning in mobile networks.

QoE provision is a very important task for modern networks and applications. Many network operators already implement QoE provisioning subsystems in the form of Software Defined Networking (SDN). Also, it is the dominant proposal of the research attempts on the subject. In a glance, such systems provide a fair tradeoff between QoE provisioning and management on the one hand and network and application overhead on the other hand.

2 ABOUT QOE

2.1 The concept of Quality of Experience

2.1.1 Definitions

Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of her or his expectations with respect to the utility and/or enjoyment of the application or service in light of the user's personality and current state, as defined by EU Cost Action 1003 Qualinet [1]. The QoE concept has emerged in the telecommunications field with the basic motivation that QoS is not powerful enough to fully express everything nowadays involved in a communication service.

An analogy to the interface design field is the emerge of the UX field the latest years. As UX deals with studying, designing and evaluating the experience that the user has through the interaction with a graphical interface, some aspects of QoE can be considered equivalent to the UX. Nevertheless, QoE is a concept that is not only limited to systems or parts of systems (such as the interface) but takes into advantage the content itself.

The importance of understanding service quality from the end user's perspective has been recognized. A survey done by Accenture [2] showed that around 82% of customer defections is due to frustration over the product or service and the inability of the provider/operator to deal with this effectively. Furthermore, for each person that calls with a problem, 29 never call, according to technical reports published by ETSI Technical Committee Human Factors. According to the same report, 90% of the customers abandons a service without even complaining about it.

Beside its definition itself, QoE can be seen from other perspectives, notably as the science of QoE or the usage of QoE in an application scenario. Naturally, while some concepts and definitions may have a wide application, their modeling and implementation in different areas may have to differ to consider specific contexts. The science of QoE regards the study of QoE, e.g. what forms QoE, which is intrinsically multidisciplinary and skill demanding, and designing methods for QoE assessment. Moreover, the usage of QoE in an application scenario regards using QoE in designing applications, products, service or producing content, objectively evaluating QoE and also delivering services/content at a certain QoE.

2.1.2 Experience: Perceived and Reference

It is said that computer programs are one of the most complex things that humans make. Humans are complex by nature. As a result, a field that aims to describe the level of annoyance or delight that a human being is experiencing from a service has to take multiple parameters into account.

Furthermore, humans tend to interpret any stimulus or event by objective criteria. Roto et al. in [add] suggest that the actual quality formation process consists of two paths. The *reference path* associates the perceived stimulus with past experiences and objective values of a person. The *perception path* handles the process of filtering the perceived physical signal through the persons special traits. Every sub-step of the two paths feeds the sub-steps of the other path in a n:n fashion.

Each path leads to its corresponding end. *Reference path* leads to the desired quality features and *perception path* leads to the perceived quality features. Through comparison and judgment process these paths lead to the experienced quality outcome. This outcome constantly changes through time, space and character and is called experienced quality event. This event however happens in a totally personal and happens inside every person. Information about this process can only be obtained in a distant and indirect manner. The following figure describes the above procedure:

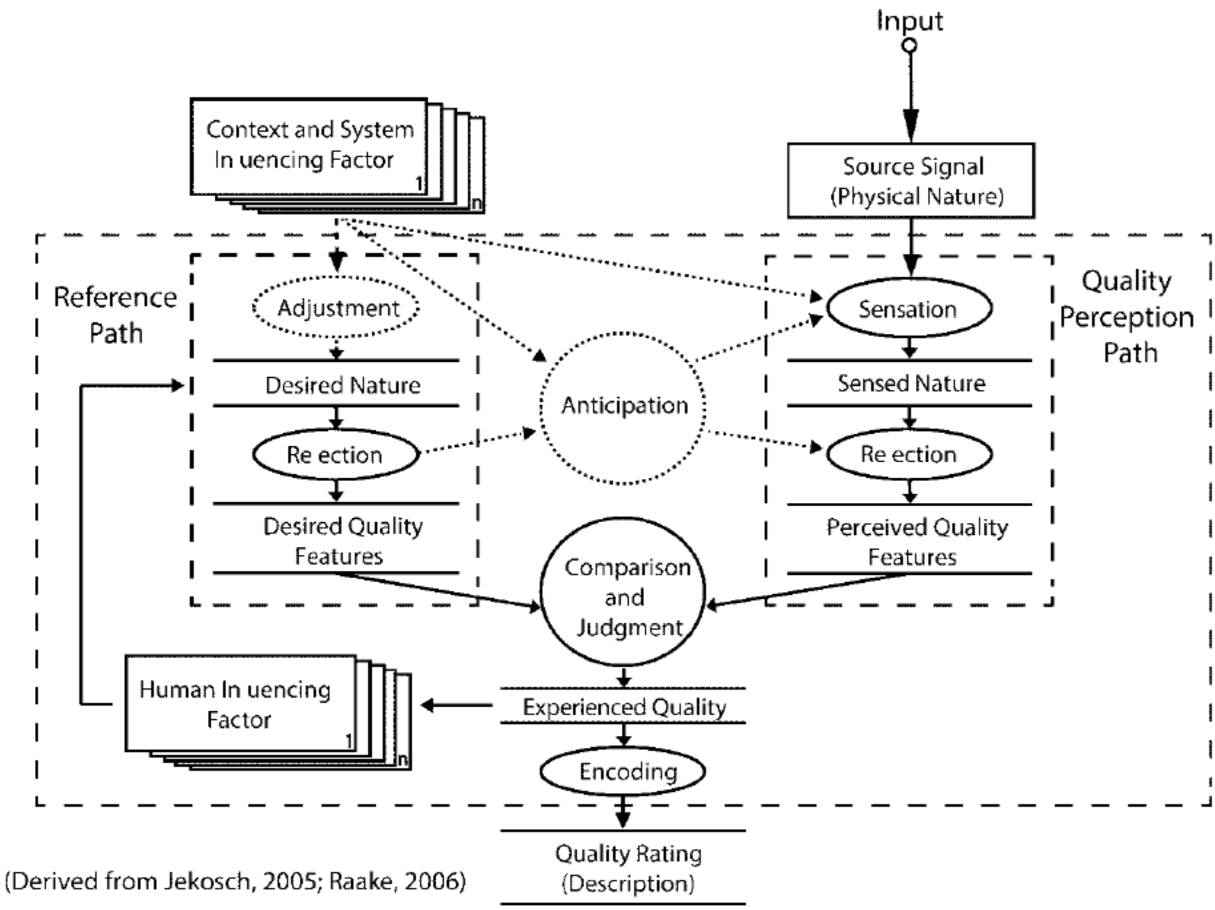


Figure 1 Reference and Perception Path

2.1.3 Factors that affect QoE

In the previous section we saw that the experience that a user enjoys by using a service or a product is directly related to her personality, past experience and values. That fact makes the QoE multidisciplinary and dependent to a lot of factors:

Personal factors – These factors describe the socio-economic background, the mental and educational state, the emotional and welfare situation and the demographic positioning of a user. These factors can be:

- High level processing, including cognitive ability, interpretation, judgment etc.
- Low level processing, including physical, emotional and mental constitution of the user.

System factors – System factors are the more technical aspects of a service or a product. Such factors may be physical, network, video or multimedia specific. Physical factors involve:

- SNR
- Bit Rate
- Bit error probability
- Throughput
- Spectral efficiency

Network factors involve:

- Congestion period
- Jitter
- Packet loss ration
- Round trip delay

Multimedia factors involve:

- Frame rate
- Video bit rate
- Video resolution
- Video content
- Terminal type
- Codec type and implementation
- Number of stalling events
- Duration of stalling events
- Multimedia Adaptation protocols or not (HAS)
- Time on highest layer in case of adaptation protocols

Context - Webster's Online Dictionary (n.d.) defines context as a "the set of facts or circumstances that surround a situation or event." According to Jumisko-Pyykkö et al. [3] context in telecommunication services consists of the following physical, temporal, social temporal, task and technical contexts one at a time.

- Physical context includes location, gradient and altitude, physical objects, orientation and weather and lighting conditions.
- Temporal context has to do with the time available for service use and also the time that passed and during which the service has been used.
- Task context involves multitasking through the service usage, interruptions during the service usage and possible relationships between the user and other service users.
- Social context describes the relationship between users and other people.
- Technical context focuses on devices, network and systems available.

Context is a very important concept. It allows for a service to be able to simulate the reference path of an individual. For example, knowing the physical context for two individuals, allows for a service provider to be able to find a more suitable route (e.g. close proximity communication can be accomplished with device-to-device communication). Most of the working individuals come home around the same time given a temporal context. On this context, system and network administrators should expect irregular bursty traffic during these hours.

2.2 Impact of QoE

Obviously, enhancing the experience that the end user perceives is crucial for all the service stakeholders. Andreas Sackl et al. in [add] conducted an experiment with which they tried to shed light in the intricate relationship between content selection, quality decisions and payment strategies. One of the most interesting findings was that only 4.6 % of the participants did not want to spend any money on quality enhancement. The amount of money that the customers were willing to pay after the experiment is shown in the following illustration:

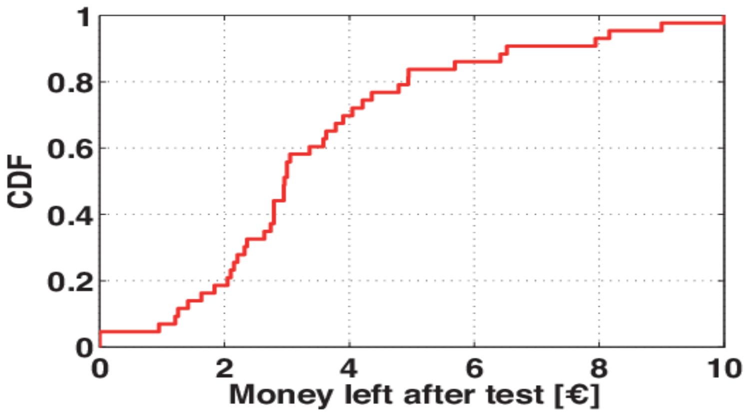


Figure 2 Willingness to pay for a service that a customer chose

Another interesting experiment shows that users are more likely to have a more positive experience for a service that they chose to pay for [5]. A graph from this experiment can be seen in the following illustration. The two plots show that the impact of QoE affects recursively the customers' opinion for the service and willingness to pay for it.

QoE is the absolute way to evaluate a service because it encapsulates the person's reference path along with technological factors. However, it does not affect only the user of the service. It greatly affects service providers that care about the economic sustainability of the service. It also affects network operators, administrators and designers by providing them an indicator about the network requirements. It also helps them to identify and prioritize the network factors that affect the user experience. QoS has great impact in marketing and customer service sectors because it can improve customer experience, reduce complaints and lead to more meaningful service level agreements.

IQX hypothesis plays a dominant role in the field of correlation QoS – QoE [49]. IQX hypothesis suggests that QoE is an exponential function of *n* disturbance factors. Further improvement of the disturbance factors above point x_1 yields quality improvements that cannot be noticed by the users. Therefore, this point is the optimal point of operation, as it minimizes the costs and maximizes the profit. Turning point x_2 is the point after which the deterioration of disturbance factors makes the service unusable. The graphical representation of the IQX hypothesis is shown in the next figure.

Khorsandroo et al. in [6] use the IQX hypothesis for QoE to export quantitative results about the relationship QoS – QoE. The authors focus on a single influence factor and conduct a Mean Opinion Score (MOS) experiment every time, in order to be able to establish a quantitative relationship between QoS and QoE.

For the first disturbance factor, namely *Voice Quality Affected by Loss, Jitter, and Reordering* the authors asked the MOS from the participants for various levels of this disturbance factor. Specifically, the packet loss ration varied from 0 to 90 percent in steps of 0.9 percent. The result could be projected in an exponential curve with equation:

$$QoE = 3.01 \cdot e^{-4.473} \cdot Ploss + 1.065$$

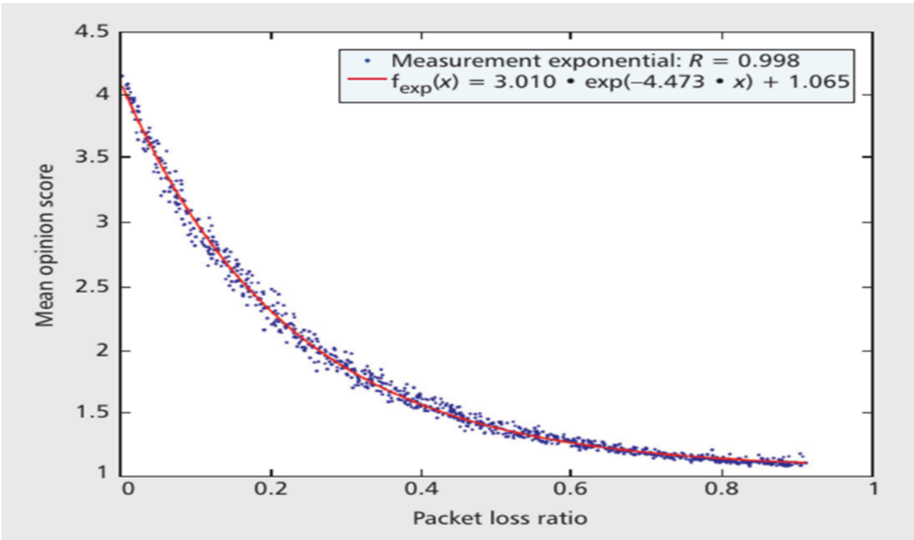


Figure 3 **Measurement results and obtained mapping function between packet loss ratio and QoE**

For the second disturbance factor, namely *Weighted Session Time to Perceived Web Browsing Quality* the users were asked to type in the same search query in three sessions where the network context was drastically different. The network context (fast, medium, and slow network) was considered through different maximal session times, in particular 6 s, 15 s, and 60 s. In total, 49 experiments were conducted for each of the three network contexts with varied response and download times. Here, the testing users were distinguished into two separate groups, trained experts and untrained naive users. The authors found that a logarithmic function could fit the results as proposed in [7], but the exponential curve was a better fit, confirming the IQX hypothesis. One remarkable finding in this case was that the logarithmic function yields a clean 5 score for waiting time below 0.62 sec, while the exponential yields a clean 5 score for waiting times below 0.5 sec. This is the x_1 threshold of the IQX curve. The results of the second case can be seen in the below figure:

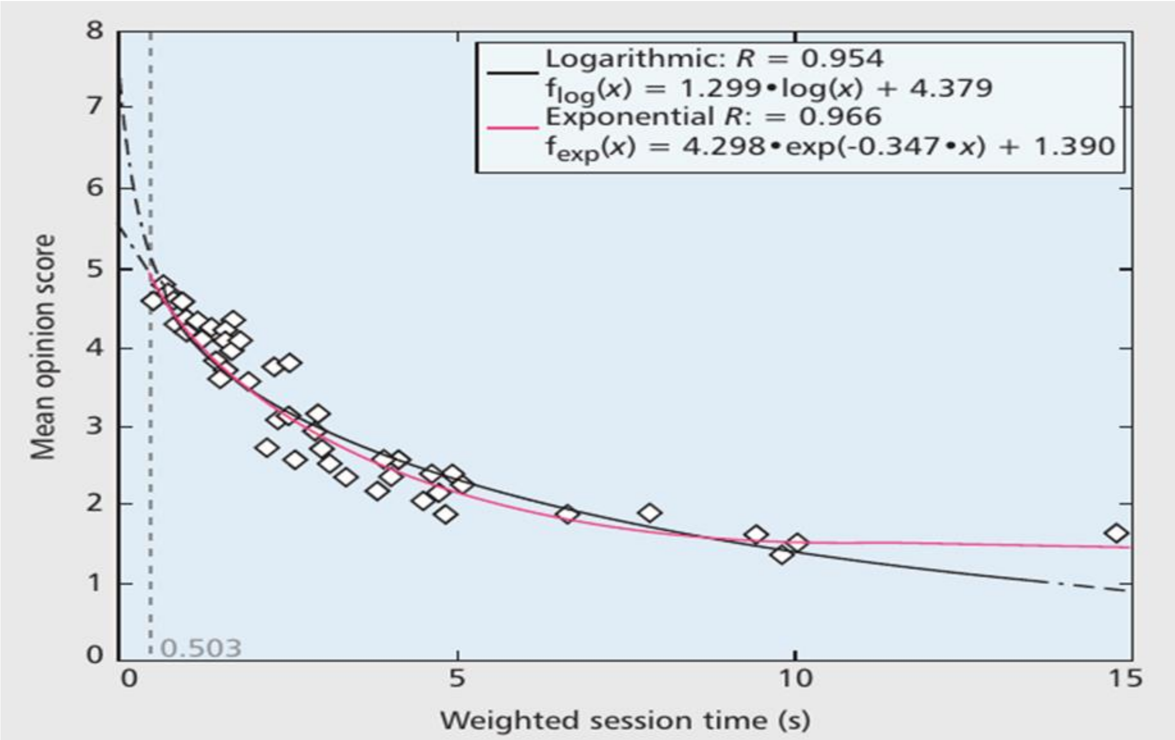


Figure 4 Measurement results for web browsing in a fast network taken from G.1030 [3] and comparison of logarithmic model flog(x) and exponential model f exp(x)

2.3 How can we measure Quality of Experience

In the previous section we saw that QoE is a valuable tool for all the stakeholders of a service. Inherently it is a very broad and complicated concept, as it tries to describe and measure the human attitude to external stimuli. Initial studies about QoE were focused on specific services at a time (for example VoIP, IPTV). Furthermore, these studies require the signal to be transmitted in order to evaluate QoE in the user side.

Opposed to these models, Liotou et al. in [8] propose that parametric QoE models are more appropriate for QoE estimation in domains such as mobile networks, or resource demanding applications as video streaming and video on demand.

The authors propose a primary classification of the QoE parametric models based on whether QoE is evaluated directly by humans or by using algorithms or mathematics formulas to extract human opinions indirectly. The first approach is called *Subjective tests*, while the second is called *Objective tests.* A graphic representation of the classification of the parametric models is the following:

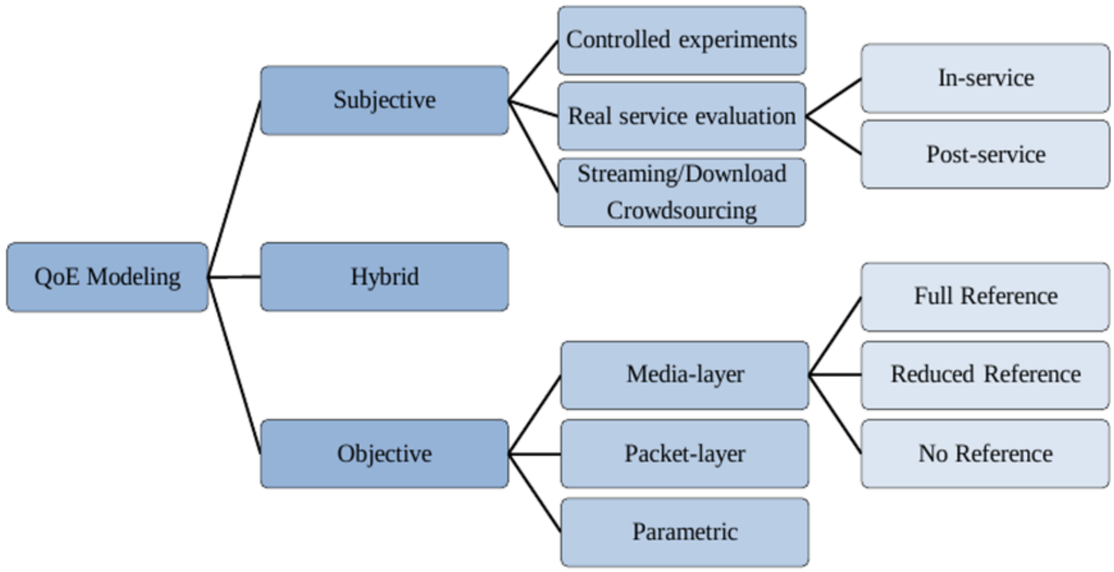


Figure 5 Classification of QoE parametric models

2.3.1 Subjective Tests

Subjective tests are real life experiments in which human participants express their experience about a service or application. The participants in these experiments may be passive receivers of the application or service output (e.g. listening to audio or watching video) or may be actively a part of the experiment (for example using simulators or being a user of a 3D application). These tests are meticulously designed to incur a certain level of annoyance every time the experiment is repeated. For example, in the case of video streaming an experiment may introduce a growing number of stalling events every time the experiment is repeated. These tests need to be thoroughly designed in advance and the user group needs to be properly selected based on guidelines and recommendations by standardization bodies. Perhaps the most important recommendation towards that direction is the ITU-T P.800 [4]. Various methods can be used for quality evaluation of the outputs. For example, users may be asked to evaluate the input on an absolute numeric scale. The results are filtered through the reference path for each user and are evaluated according to users' past experiences, capabilities, perceptions, education level etc. and primarily quantify the quality, effectiveness and delight that the user enjoyed through this service.

These tests have high value in evaluating the degree of delight that a user enjoys through the service or application, in a sense that they can take into account any conscious or unconscious aspects of human quality evaluation. Perceptual tests are exceptionally good at capturing the internal state of the human factor, but in order to be reliable they have to be thoroughly designed and the users must be unbiased and objective.

One disadvantage of the aforementioned method is that even though its results are valuable for laboratory testing, it is not ideal for real time QoE evaluation. One way to overcome this is by prompting the user of an application or a service to rate the service either in-service or after the service has ended (after-service).

The guidelines that should be followed in subjective tests are very broad and include room conditions (e.g. isolation), audio equipment and generally any equipment needed for listening, viewing, talking. They also include guidelines for selection of the background of the users (experts or non-experts), their age distribution and in general the randomness in selection of the participants.

Another branch of the subjective methods of evaluation suggests that the users should be able to download the test in their own equipment and conduct it in their own familiar environment. This more relaxed approach is considered to be more realistic and more general as it is naturally broader and is addressed to a larger community. These techniques are called *crowdsourcing* techniques in a sense that they outsource the job of evaluating the QoE to anonymous users. One such example is the Google Microworkers platform as well as the Amazon Mechanical Trunk, where an Internet user may conduct QoE experiments designed by other parties (such as researchers), who require a general public for an evaluation task.

2.3.2 Objective Tests

While subjective tests are a great means to evaluate and quantify the quality that a user perceives from a service or an application, it is costly, time-consuming and not reproducible on demand. Also, they cannot be used for real time user experience monitoring with credibility. These reasons have led to the development of objective tests. Objective tests try to predict the quality that the end users perceive without their intervention by using some objective metrics. They can be classified in various categories:

- Reference signal utilization: Whether the reference signal or part of it participates in the QoE evaluation. According to this criterion we distinguish the Full Reference (FR) or reference-based or double-ended models, the Reduced Reference (RR) models and the No Reference (NR) or single- ended models, where “reference” refers to the original signal.
- Evaluation method: Depending on the kind of input that is used for QoE measurement we can differentiate between Media-layer (signal-based), Packet-layer / Bitstream, and Parametric models. Media layer models make use of signals and may be full-reference, reduced reference, or no reference. Packet layer models extract information from packet headers, while bitstream models extract information from packet headers and payload.
- Model mode: Depending on how we inject a test signal to the system under test or not we have active and passive models.
- Usage purpose: Depending on the purpose under which we conduct the test.

Parametric QoE models are derived by conducting subjective tests in a controlled environment where one or more of the aforementioned parameters are injected to the user input. Statistical methods are used to come up with a direct equation that calculates the QoE as a function of these parameters.

Hybrid models lie between the subjective and the objective evaluation models. They rely on previously acquired subjective scores which are used to train machine learning models. This model then maps network parameters to MOS values. Later, this model can be used for real time predictions and can be adjusted using real time data.

2.4 QoE provisioning

In the previous sections we saw that QoE is the main goal behind any application or service. This task is challenging for operators of such services and applications for two main reasons. The first one is that operators have been focusing on purely technical aspects for a long time. The reason behind this is that operators have the ability to measure many if not all the technical aspects of the provided service (packet loss, jitter, loading times). So, it has been easy for them to stick to these technical factors and *assume* the perceived quality based on these. The second one has to do purely with the lack of the term QoE and the studies that show how QoE can quantify, manage and improve the experienced user quality. The main challenge for the operators now is to be able to make the transition between the QoS metrics to QoE real time provisioning.

Liotou et al. in [9] suggest that the key objectives in this transition are:

- Monitor QoE for the end user
- Enhance their experience
- Improve the network efficiency

The authors introduce a QoE management entity with embedded sub-entities capable of accomplishing specific tasks towards the completion of the above goals. These sub-entities are:

- the QoE controller which stands in between the network or the service and its main job is the data collection from the service operation.
- The QoE monitor has the role of interpreting the data gathered from the QoE controller through appropriate models and representations.
- Finally, the "control room" of the framework is the QoE manager sub-entity. It has the role of reacting to the monitor output and take the appropriate actions towards the accomplishment of the goals. A graphical representation is shown in the next figure:

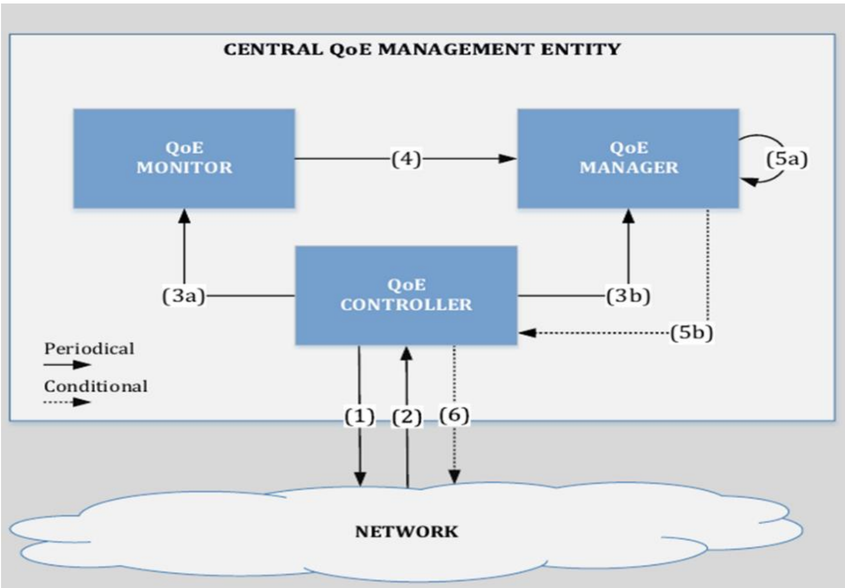


Figure 6 **QoE management entity**

The above framework introduces some challenges for the service or the application operators. More specifically, about the Controller the operators should decide what data to collect, how to collect these data, how often the data should be collected and in which manner will they be collected. For example, a real time application running on distributed infrastructure should be really challenging to monitor. Also, the operators should be able to justify what will be the overhead of this operation in the application service.

About the QoE monitor, the operators should be able to answer which model should be used, how often it should be calculated and what should be the output to the Manager entity. Finally, in the management layer it should be defined what will be the reaction to what input, where and when will be the intervention.

Other approaches attempt to place QoE management entities closer to the medium. In [10] a QoE utilized the Media-Aware Proxy (MAP) entity to gather QoS data across the network. These data are stored in the Media Independent Information Service (MIIS) database. In the center of this approach lies the QoE controller, which uses the data from MIIS database to feed ML-based models for QoE estimation. The output of these models is the perceived QoE, so this output can be used as a feedback to the MAP entity in order to maximize the perceived video QoE by selecting the appropriate stream optimization method. An overview of the described process is the following:

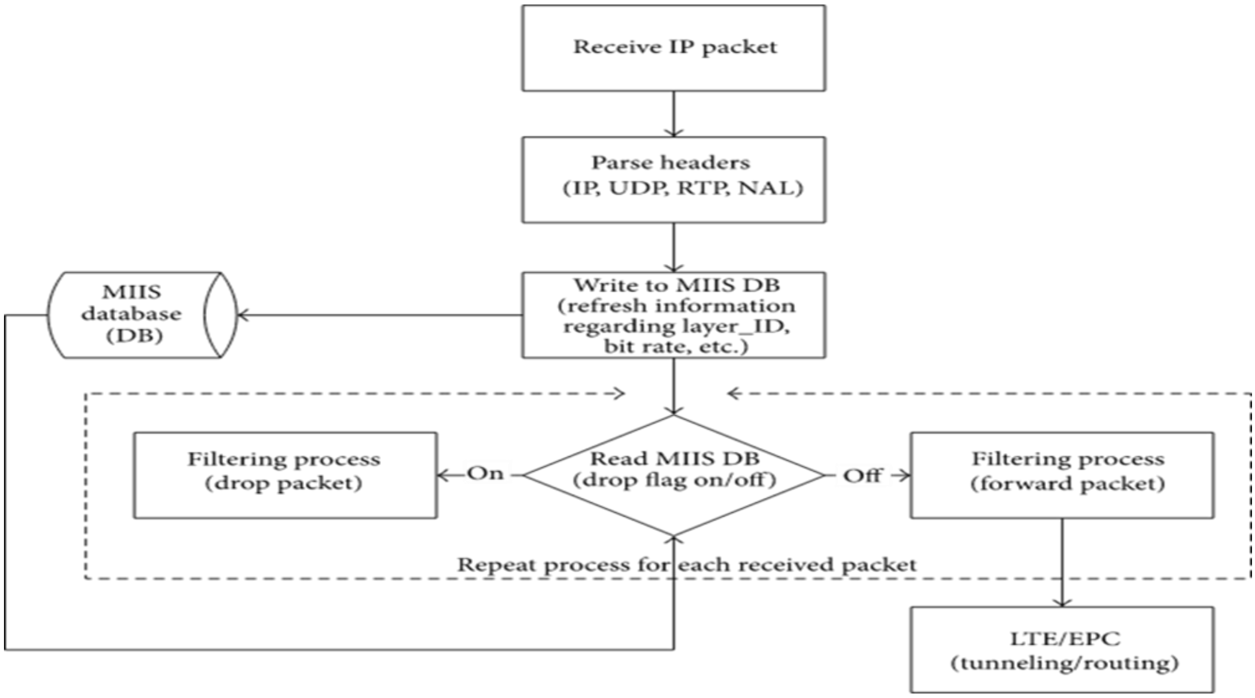


Figure 7 MAP trained ML models for QoE enhancement

2.5 Challenges related to QoE transition

In the previous chapters we saw that QoE is the ultimate metric for an application or a service and it is welcomed by all the stakeholders of a service or an application. We also saw some early attempts and recommendations towards QoE adaptation. This section is concerned with the challenges related to this transition.

2.5.1 Technical challenges

The first set of challenges is the purely technical aspects of QoE transition implementation. In the contrary of the centralized QoS information gathering (information regarding the network in general), QoE implementation must gather per-user, per-application and per-terminal information in a real-time manner.

Also, this plethora of data points poses a challenge on how these data should be modeled. Efficient collection and scalability are also a concern. Finally, the way the data should be grouped in order to present an overview of the situation is also a big and interesting challenge. This task becomes even broader due to the diversity of the technologies involved in these applications or services. One should be aware of different types of devices, underlying network technologies, applications, hardware etc. in order to be able to monitor and model properly the perceived quality.

Finally, energy consumption and environmental issues should be taken under consideration, provided that QoE provisioning and monitoring exists in all the phases of the service or application delivery process and incurs an overhead in these processes.

2.5.2 Economic Challenges

While the technical aspects of implementing QoE provisioning are notable, there is also the economic aspect of it which should be addressed and calculated in the equation. In [11] the authors claim that the axiom of Network Neutrality is at stake during QoE provisioning and monitoring, especially from the over-the-top (OTT) service providers.

More specifically, one should consider the pricing and the strategies derived from user that have different requirements. In [12] Reichl et al. suggest a fixed-point model for QoE-based Charging. According to the authors, the QoS-based pricing follows a fixed model according to which consumers buy directly a set of technical characteristics. In such a model the price follows the classic supply-demand cycle. More specifically, the set of the specifications (QoS) that a provider can guarantee is directly related to the demand for the service. More customers will create excess demand which grows faster than the resources development, leading to quality adaptations. On the other hand, price is directly dependable on the set of specifications a customer buys. Finally, the demand is directly dependent on the price. To sum up, the QoS-based pricing follows the virtuous cycle that is shown in the next figure.

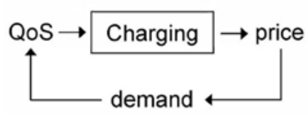


Figure 8 QoS-based pricing

On the other hand, QoE-based pricing models are different because price is part of the context of the perceived quality. So, in this case price defines the set of specifications that a customer buys (prices rise on improved sets of specifications) but also impacts the perceived quality itself as a part of context. The higher the price, the higher also the users' expectations concerning the offered service quality. In this case, the pricing follows the following cycle.

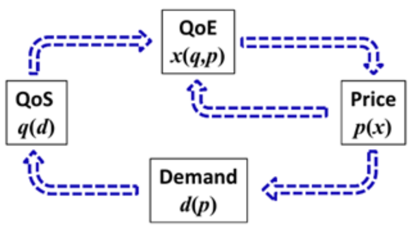


Figure 9 **QoE based pricing**

2.5.3 Legal – Ethical challenges

Legal and ethical challenges are another set of challenges that arise before the adaptation of QoE model. More specifically, Network Neutrality axiom may be jeopardized in the face of different quality levels as noted previously.

Another discussion has to do with the conflict between network operators and OTT services. Network operators claim that the amount of data traffic increases and big part of this increase is due to the OTT services. In fact, operators lose money. Their profit is unaffected while the operations costs is higher. On the other side, OTT service providers develop and maintain state of the art applications which are served on top of a default best-effort bearer. Also, these apps, designed to provide different set of specifications per user and per subscription fall on the same pool of resources.

Finally, a big legal challenge in QoE provisioning is the way in which sensible personal information about a customer can be distributed, shared and reviewed in a public end to the end path.

3 QUALITY OF EXPERIENCE IN VIRTUAL REALITY APPLICATIONS

3.1 Introduction

Virtual reality (VR) applications are applications that use immersive sensory experience that digitally simulates a virtual environment. This environment could be similar to or completely different from the real world. VR applications have been developed in a variety of domains such as education, architectural and urban design, digital marketing and activism, engineering and robotics, entertainment, fine arts, healthcare and clinical therapies, heritage and archaeology, occupational safety, social science and psychology.

Currently standard VR systems use either VR headsets or multi-projected environments to generate realistic images, sounds and other sensations that simulate a user's physical presence in a virtual environment. A person using VR equipment can look around the artificial world, move around in it, and interact with virtual features or items. The effect is commonly created by VR headsets consisting of a head-mounted display with a small screen in front of the eyes but can also be created through specially designed rooms with multiple large screens. VR typically incorporates auditory and video feedback but may also allow other types of sensory and force feedback through haptic technology.

VR headsets create virtual environments with a process known as rendering. This process takes as input:

- The geometry of the surrounding virtual environment
- Materials which are applied to the geometry and specify the appearance of the surface.
- Light sources to illuminate the scene.
- A virtual camera on which to form the image.

Illumination can be of local or global type. Local illumination is the direct reflection of the light source on the camera. Global illumination consists of all the possible inter-reflections between objects in the scene before entering the camera. The reflected radiance from any visible point is the solution to the following equation, known as the Rendering equation:

$$L_0(x,\omega_0) = L_e(x,\omega_0) + \int_{\Omega} L_i(x,\omega_0) f_r((x,\omega_i,\omega_0)|N \cdot \omega_i| d\omega_i)$$

This equation means that the reflected radiance L_0 from any point in the scene x in a direction ω_0 is the sum of the emitted radiance L_e and the reflected radiance L_i from every visible point in the scene. It is only natural the reflected radiance from every visible point of the scene to be expressed as an integral over the hemisphere Ω. The term $f_r(x,\omega_i,\omega_o)$ is called the Bidirectional Reflectance Distribution Function and expresses the reflected radiance as a function of the material. In order to compute the images (frames) used for the video, this equation must be solved constantly to provide the correct reflected radiance for every point. This task is usually accomplished using numerical methods and most commonly the Monte Carlo method.

3.2 Related Works

There are several related works concerning the quality evaluation of VR applications. All the works mentioned here use some form of subjective evaluation for QoE estimation.

3.2.1 Multimedia based QoE assessment methods

In this chapter we discuss methods that try to correlate the QoE that a user enjoys with the multimedia influence factors. In general, these methods introduce system annoyance factors in VR content and conduct objective tests in this gradually degrading quality content.

Brunnström et al. in [13] investigate the QoE that a user perceives for a simulator that utilizes VR. This simulator creates a virtual environment that approximates the operation of a forestry crane. In order to estimate the QoE that a user of the simulator enjoys, the authors designed a subjective evaluation method. In the first section of the method, the users are called to use the crane simulator in order to complete a specific task. This task consisted of loading two piles of 16 logs onto the truck. After the completion of one pile, the participants were asked to take a short break after which they had to continue to the other pile. The task took about 15 minutes for one pile of logs.

In the second section of the experiment, various delays are introduced in the baseline experiment. Specifically, two kinds of delay are added every two minutes approximately: joystick delay and screen delay. After the completion of the two sections of the experiment, the participants are asked to answer the following questions:

- How would you rate the picture quality?
- How would you rate the responsiveness of the system?
- How would you rate your ability to accomplish your task of loading the logs on the truck?
- How would you rate your comfort (as in opposite to discomfort)?
- How would you rate the immersion of the experience?
- How would you rate your overall experience?

The results of the questionnaires are presented in the next figures. Light grey bars are showing the baseline-delay MOS and dark grey the delay MOS:

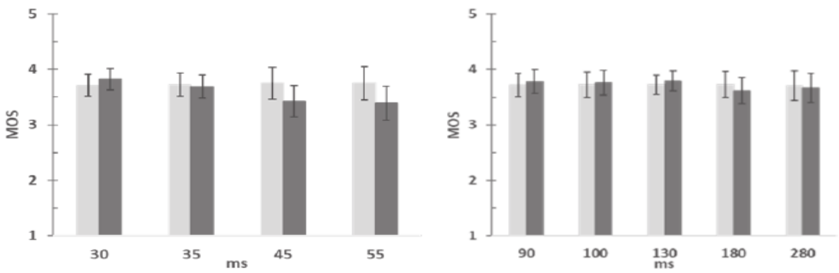


Figure 10 The Mean Opinion Score (MOS) for Overall Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms) [13]

The results of subjective tests show a clear degradation in user experience for increasing duration of delays. More specifically, the display delay seems to have a more direct impact on users' experience such as 10 ms in screen delay can cause 1 grade in MOS scale.

The impact of joystick delay seems to have lighter effects on users' QoE, such as even a 200 ms delay cannot degrade the users MOS by one grade. This measurement is contradictory to the fact that this simulator concerns a working example of crane loading VR application.

A different pattern is followed when people are asked about the task accomplishment quality for different display and joystick impairments. It seems that no clear trend can be distinguished for added levels of impairment both in joystick and the screen delay.

On the other hand, when people are asked about the comfort quality, the screen delay plays a crucial role. The MOS can be altered for even one degree in a Likert scale for screen delays as small as 20 ms.

Immersion quality also seems to depend heavily on-screen delay. Comfort and immersion quality are directly connected to vision as well as spatial presence so it is only natural to be influenced more heavily from visual impairments.

Also, the contextual parameters also play a role, with the users in this subjective experiment to perform a repetitive and mundane task of crane operation. Finally, in the overall quality index it is clear that degradation in screen delays affect far more the MOS than the joystick delays.

In a second level, the authors examine the sickness symptoms that a user feels after the usage of the simulator. The symptoms under consideration are very broad and include general discomfort, fatigue, stomach disorders etc. The symptoms have been grouped into four categories, namely Nausea (N), Oculomotor (O), Disorientation (D). They also participate in the total score (TS). The total symptoms taken under consideration are the following: *General Discomfort, Fatigue, Headache, Eye Strain, Difficulty Focusing, Increased Salivation, Sweating, Weight, Nausea, Difficulty Concentrating, Fullness of Head, Blurred Vision, Dizzy (Eyes Open), Dizzy (Eyes Closed), Vertigo, Stomach Awareness, Burping.* The results of the simulator sickness questionnaire are presented in the following diagrams:

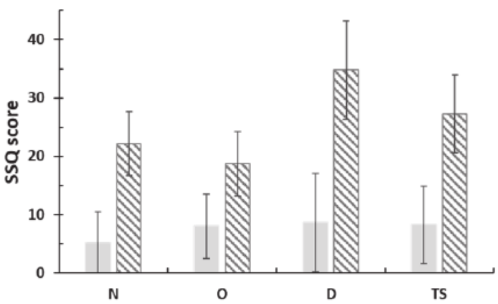


Figure 11 simulator sickness questionnaire for the delay experiment [13]

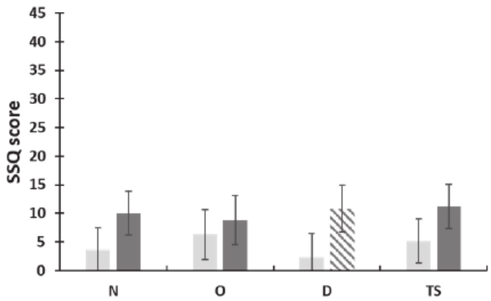


Figure 12 simulator sickness questionnaire for baseline experiment [13]

The light grey bars illustrate the symptoms' intensity before the experiment, and the dark grey and dashed bars represent the symptoms intensity after the experiment.

The authors' conclusion is that the comfort and immersion quality are strongly affected by screen delays, while the impact of joystick delays have very small impact. According to the simulator sickness questionnaire, the conclusion is that delays and especially the display delays strongly influence the sickness feeling of the participants, causing some of them not to continue the experiment.

Tran et al in [14] follow a subjective approach to estimate the quality that the users perceive through 360-degree video. Contrary to the previous works, the authors adjust the content to be able to estimate the QoE. Specifically, three video excerpts are examined with duration 30 seconds each; The first one is the most static, having static camera, medium object motions, few moving objects and static background. The second one has medium camera motion, medium object motion, many moving objects and dynamic background. The third one has fast camera motion, fast object motions, many moving objects and dynamic background.

The videos were encoded with H264 encoder. By combining different quantization parameters (the ratio between I, P and B frames in H264) and screen resolution, the authors created 60 different versions of the videos. For the reproduction of the videos, two device sets were used. The first one is a Samsung Galaxy S6 and a Samsung gear VR Head Mounted Display (HMD). The second one is a Samsung Galaxy S5 and a Google Cardboard.

Each user is shown 20 versions randomly and then displayed two times: one in VR mode and one in non-VR mode. In particular, each viewer watches the chosen version in the non-VR mode in the first display time, and then gives answers to questions Q1 and Q3. Then, the viewer watches the same version in the VR mode in the second display time, and then gives answers to questions Q1–Q4. The questionnaire was the following (Q1-Q7):

- How is your assessment about the perceptual quality of the video on the scale from 1 to 5?
- How is your assessment about the sense of presence in VR environment on the scale from 1 to 5?
- Is this viewing acceptable to you? (1 means that you accept and are willing to watch until the end of the session, and 0 means that you do not accept, feel annoying, and want to quit the session).
- Which do you prefer, non-VR rendering mode or VR rendering mode? (0 is non-VR and 1 is VR).

- How is the level of dizziness or nausea during VR viewing experiment on the scale from 1 to 5?
- How much does wearing a VR HMD affect your experience in VR environment on the scale from 1 to 5? (1 means very cumbersome and annoying, 5 is absolutely no problem).
- How much does the FoV of the device affect your sense of presence in VR environment on the scale from 1 to 5? (1 means very limited, 5 is absolutely no problem).

Regarding the encoding technique, the experiment concludes that the MOS for the 4K videos is 4.22, 4.33, 4.28 for the first, the second and the third video, respectively. The perceptual quality scores for the same videos are 4.44, 4.44, 4.56, respectively. These results, according to the authors, show that the perceived quality for 4K video is quite acceptable.

About the cybersickness, the results showed that 89% of the users reported nausea and sickness after watching the VR version of the videos in the device set 1. For the device set 2 the corresponding number was 94%.

Importing the designed levels of annoyance, the authors found that high resolution in combination with lower number of I frames lead to lower MOS. Another interesting result is that the degradation of quantization parameters leads to faster drop of MOS for the larger resolution videos. The same trend seems to appear in the perceptual quality and acceptability rate of the three videos examined. It is remarkable that the acceptability rate MOS fall faster for the degradation of the quality than the other quality aspects.

By further analyzing the impacts of resolution, we can conclude that when the quantization parameters are lower than 32, the difference in QoE between the 3480 X 1920 resolution and the 2560 X 1440 resolution is insignificant. On the other hand, the QoE scores fall rapidly as the resolution drops to the 1920 X 1080 and 1280 X 720. Also, the decrease of the perceptual quality is from 78 % to 89 %. It is also worth noticing that all the versions encoded in 1280 X 720 cannot surpass the 60 % threshold for the acceptability rate MOS. Therefore, regarding the screen resolution, the authors conclude that VR video should be provided only in the highest levels of resolution.

In the further analysis of the encoding parameters, the authors notice that in the higher resolution level, the differences in MOS for quantization parameters are insignificant. On the other hand, quantization parameters level below 40 retrieve MOS below 30 %, in other words, that level of dilution in I frames makes the result annoying and difficult to attend.

About the effects of content motion on QoE the authors found no impact of it in the acceptability of the video. Furthermore, perceptual quality and presence was little affected by content motion. On the other hand, cybersickness was greatly affected by the content motion.

The paper concludes by analyzing the effects of VR and non-VR rendering mode in the MOS. The authors cite the results of a video sample with quantization parameter equal to 22. It is clear from the results that both the perceptual and the acceptability scores are in non VR-mode are higher than those in VR mode. In addition, it is clear that the lower the resolution, the larger the differences between the two modes.

In a very interesting diagram is examined the percentage of the participants that would rather a VR rendered video from a non-VR as a function of the resolution and the content. The diagram is displayed in the figure below. It is interesting that for resolutions 1920 X

1080 and above more of the 50% of the users prefer the VR rendering mode, even though the acceptability and perceptual quality scores are lower. This is a clear indication of the fact that the concept of QoE can express better the interests and different traits of people than a set of technical specifications and SLAs.

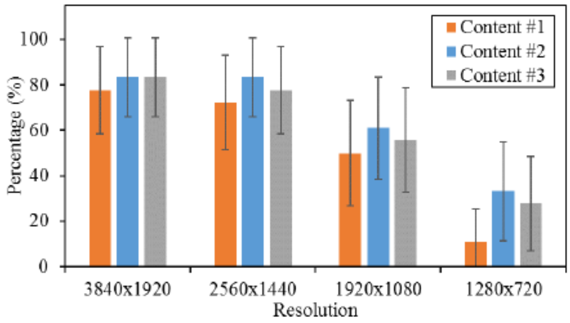


Figure 13 Percentage of viewers who prefer a VR version from a non-VR version [14]

Anwar et al. in [15] use a subjective evaluation technique for the estimation of the quality of service in VR video. The authors chose the content of the video samples in order to contain a wide variety of spatial intensive (SI) and temporal intensive (TI) videos. For all the video samples 4K (3840 X 1920) resolution is used and the bit rate varies between 1, 5 and 15 Mbps. A 5 second stalling event is introduced in a random point of each video sample using a spinner to indicate the stall. In total, 208 video samples are produced with this process, including 16 source videos, 48 compressed without stalling, 48 initial stalling, 48 middle stalling, and 48 videos with multiple stalling.

The data retrieved from the subjective test are displayed in the next table. The table is organized as following: For every bit rate level (1, 5, 15 Mbps) and stalling event it contains the MOS. The first three rows contain the MOS for no stalling, initial stalling, middle stalling and multiple stalling. The last three contain the difference in MOS for two of the above rows, in other words it is a measure of the deterioration of the MOS between the two states.

Table 1 Stalling events at various Bit rates as used in the experiment.

Stalling event	1 Mbps	5 Mbps	15 Mbps
No stalling	3.94	6.3	8.67
Initial stalling	3.49	5.66	6.94
Middle stalling	3.3	5.04	6.22
Multiple stalling	2.79	4.38	5.44

No stalling – initial stalling	0.45	0.64	1.73
No stalling – middle stalling	0.64	1.26	2.45
No stalling – multiple stalling	1.15	1.92	3.23

By analyzing the above data, the authors claim that stalling in VR video always affects user experience. Also, multiple stalling events are more annoying to the users than single stallings in the beginning or the middle of the video. Another interesting finding is that the level of annoyance, or the MOS drop is strongly related to the bit rate of the video. One practical explanation for this fact is that the users in VR video have higher expectations while watching high quality content, so they become frustrated when the video stalls.

The authors of this paper proceed one step ahead of the subjective test result and propose a QoE prediction method. For this purpose, a Bayesian inference method is used in order to estimate the influence of the two factors of bit rate and stalling events. In order to increase the accuracy of the classifier, the authors introduce 50 hidden layers on the level of QoE. QoE takes values between 1 and 10, so the hidden layers translate to 0.2 better QoE. In this sense, every level of the posterior QoE has to belong in one of the 50 levels with probability between 0 and 1. The Bayesian logic is well fitting in this kind of estimations, because it takes into account the role of the prior probability for these QoE levels.

Vlahovicet al. in [16] assess the impact of different locomotion methods used for VR applications in users' QoE, immersion and sickness feelings. The navigation techniques that were examined in this work are:

- Controller movement: This navigation method is similar to the first-person shooter games. The initialization, speed and duration of the navigation is controlled by a game controller, while the direction is controlled by a HMD. During this movement, all the video aspects of the game are visible to user, as opposed to the next navigation categories.
- Controller movement with tunneling: This navigation method is equivalent to the previous one, but while the user is moving, the field of view narrows down. This movement rendering is responsible for motion sickness.
- Teleportation: In this movement category, the user must select a desired point of movement and the video location will change without displaying the motion part, with a short blackout.
- Finally, the method of human joystick is examined. In this method the user defines every aspect of the movement by leaning to any direction, and the VR application corresponds accordingly.

For the assessment of these methods, the authors develop a VR application that depicts a village that offers a plethora of navigation and field of view options for the users. The equipment used was the HTC vine platform. After familiarization with the apparatus the participants dive in the experiment. After the experiment completion, the users are asked to provide their opinion about the experience they had. This evaluation concerns:

- The immersion. The feeling of absence from the present world and the sense of physical presence in the virtual world. This evaluation is performed on a 5 levels Likert like scale.
- Overall QoE. Also rated in a 5 levels Likert like scale.
- Physical discomfort. This question can have three possible answers: no discomfort, mild discomfort and strong discomfort.

For the first part, the users rated the traditional Controller movement way of navigation as the most immersive of all, with a MOS of 4. The participants rated the tunneling vision controller movement as the least immersive of all with a MOS of 3.86. The other two movement methods are in between those two. In general, we can say that the immersion does not get affected greatly by the movement method.

The results of the second question are quite different. The overall QoE seems to be more affected by the navigation methods. More specifically, the teleportation method seems to present the best overall experience for the users with a MOS of 4.31. The worst experience seems to be provided by the human joystick method, with the average rating of 3.2.

As far as the sickness feelings are concerned, it seems that the traditional controller-based motion scenario causes the most sickness feelings. 10.34 % of the participants reporting strong discomfort feelings and 62 % reporting mild discomfort feelings. The teleportation navigation scenario seems to be the best in terms of sickness feelings. 6.9 % of the participants report mild discomfort.

By examining the results of the subjective tests, one can observe the contradictory outcomes of the immersion and the overall QoE questions. The best navigation method for immersion appears to be the worst for the overall experience. This can be explained by the fact that this navigation method presents the highest sickness feelings among all the navigation methods.

Also, a non-expected result is that the navigation method that is biased to cause sickness feelings, namely the controller movement with tunneling is rated as the second-best method for sickness feelings avoidance.

These contradictory results show the importance of the subjective tests versus the Key Performance Indicator (KPI) – based methods for the avoidance of bias and the optimization of users' experience. Also, it is important that the subjective test can adapt better in specific use cases such as the use case of this work.

Madhusudana et al. in [17] propose a quality estimation model, namely the Stitched Image Quality Evaluator (SIQE) for stitched panoramic images that are used for VR and AR applications. The model is based on objective metrics and is validated by subjective tests.

The authors create an image database with images such as the ones mentioned in the previous paragraph. The authors observe three main impairments in these images:

- Ghosting and blur
- Color distortion

- Geometric distortion

The images are evaluated in terms of MOS by 35 participants in the experiment in a scale of 0 – 100. The device used for the evaluation is a Samsung Gear VR HMD.

After the completion of the subjective test, the authors try to fit the results of the people's opinion to subjective parameters and to come up with a quality prediction methodology. This methodology has two parts: The first one evaluates objective quality elements in both the stitched image and the original ones. These elements are the following:

- Steerable pyramids: The images are decomposed using the algorithm of steerable pyramids. This allows for capturing structural impairments, as well as spatial deformations. The authors argue that these kinds of impairments are not isotropic, thus such algorithms should make it easier to isolate and detect them accordingly.
- Divisive normalization. During this step, image is represented as simple nonlinear map, where each component in a cluster of coefficients is divided by the square root of a linear combination of the squared amplitudes of its neighbors [18]. According to the authors, this step reduces the statistical dependencies between neighboring coefficients.
- Bivariate modeling: This step aims to spot and report the correlation between the adjacent coefficient that is caused from ghosting.
- Patch weighting: During this step, the non – smooth overlapping features of the images are isolated using a simple texture estimation algorithm, namely the Gray Level Co-occurrence Matrix (GLMC) and these regions are weighted more than the smooth ones.

These steps lead to metrics of feature extraction differences. These metrics are compared to the known database that was designed and implemented in the previous steps where equivalent metrics are stored and associated with MOS. This comparison leads to the actual prediction of the overall quality of the image.

By cross validating the results, the authors argue that the proposed quality prediction method has better confidence interval and is more focused on the panoramic image quality prediction, compared to other equivalent metrics such as BRISQUE and DIIVINE.

The works that were mentioned in this chapter are concentrated in the table below:

Table 2 A summary of the works mentioned in this chapter.

Method	**Description**	**Low-level processing**	**High-Level processing**	**Multimedia**	**Network**	**Physical**	**Eval. Method**	**Results**
Egan et al [19]	Objective and subjective QoE estimation using heart rate and subjective tests	Yes	No	No	No	No	Subjective and Objective	Both subjective and objective tests showed that QoE was higher in the VR video sample compared to non-VR.

Method	Description	Low-level processing	High-Level processing	Multimedia	Network	Physical	Eval. Method	Results
Brunnström et al [13]	Log loading VR simulator in which the authors intervene with a controlled delay	Yes	Yes	Yes	No	Yes	Subjective	Subjective tests have shown that every degradation in visual or physical factors deteriorates the QoE but the multimedia IF are more important
Tran et al [14]	Subjective evaluation of side effects of VR	Yes	No	Yes	No	Yes	Subjective	Video resolution proves to be more important than quantization optimization for VR application as higher video resolutions manages to reverse the MOS.
Anwar et al [15]	Subjective evaluation of the impact of stalling events and bit rate in VR video.	Yes	Yes	Yes	No	No	Subjective	Initial and middle stalling events are less disturbing to the user than multiple stalling events. Also, the annoyance of the users is more intense for stalling events in higher bit rates
Zheleva et al [20]	Subjective evaluation of the impact of resolution in QoE based on electroencephalographic sensoring	Yes	Yes	Yes	No	No	Subjective	QoE of the subjective test is not conformed to the electroencephalographic monitoring

Method	Description	Low-level processing	High-Level processing	Multimedia	Network	Physical	Eval. Method	Results
Vlahovic, Suznjevic	Subjective evaluation of the impact of navigation methods in VR games	No	Yes	Yes	No	No	Subjective	Contradictory results about navigation impact on users QoE
Madhusudana and	QoE prediction from image analysis metrics for panoramic images use in VR	Yes	Yes	Yes	No	No	Subjective and Objective	The proposed image analytics algorithm seems to be able to predict more accurately the final image quality

3.2.2 Human factors based QoE assessment methods

This section deals with QoE assessment methods that are based on human influence factors. These methods introduce gradually degrading quality content to users and assess their opinion about it based on human body metrics, such as electroencephalographic activity, electrodermal activity etc.

Egan et al. in [19] try to correlate the heart rate and electrodermal activity to the perceived quality that a user enjoys from a service. The participants in the experiment are asked to watch a scene from a virtual city in two environments. One is a VR environment, and the other is a usual 2D environment. The objective metrics are captured with simple paramedical instruments. In the second phase of the experiment the users were asked to fill a questionnaire about the experience they had during the watching of the two excerpts.

There is a clear distinction between the VR mode and the non-VR mode. More specifically, participants present higher mean heart rate while consuming VR content, compared to the mean heart rate during the consumption of non-VR content. An equivalent result is achieved in the mean electrodermal activity between VR and non-VR mode.

The multiple choices asked in the experiment were the following:

1. I was immersed in the environment.
2. I enjoyed experiencing the virtual environment.
3. The virtual environment was realistic.
4. I did not feel a strong sense of presence whilst experiencing the system.
5. The system was easy to use.
6. I would have liked more time in the virtual environment.
7. I did not feel any discomfort while using the application.
8. I needed to learn a lot of things before I could get going with this system.

9. My experience did not meet the expectations.

In all but two of the questions the participants had higher opinion rating about VR content, than the equivalent non-VR content. The questions that stood out were the question 5 and the question 6. Both of these questions express the users' confusion about the VR as well as the users lack of familiarization about the VR technology. This trend appears in a plethora of works about VR and AR and is ranked as a highly active field of research. In most of the remaining questions, users definitely ranked the VR experience higher that the non-VR one. This fact is even more clear when we consider that these questions have to do with the immersion in the user level.

The authors conclude that electrodermal activity is more representative of user QoE than heart rate. They also conclude that the electrodermal activity more closely reflects the user QoE as this is represented in the objective results. Also, regarding the subjective evaluation, the authors conclude that in 7 of the 9 questions the users experienced significantly better quality from the VR environment.

Zheleva et al. in [20] conduct a subjective test in an attempt to correlate the alpha activity of the human brain to the QoE perceived by VR video. Alpha activity is the dominant oscillatory activity of the human brain. It has been associated with more complex cognitive process such as attention, memory and divergent thinking.

For the subjective test, the participants are called to attend a video sample from the movie INVASION!. The authors have produced four variants of this video; The first one (Q1) has high resolution (2469 X 2743), the second one (Q2) has lower quality (1808 X 2009), the third one (Q3) is (1169 X 1298) and the lowest quality one (Q4) has resolution 512 X 549. Each participant attends the three video samples in a resolution descending order. During the video attendance, the electroencephalographic activity of the attendants is recorded.

After the video watching, the participants are asked to fill a form about their experience. Specifically, it is examined the absolute rating of each individual video quality, the immersion that the participants enjoyed and the simulation sickness that the users felt.

The results of the subjective tests showed that higher video quality levels were rated higher from the participants with the best quality video, namely the Q1 to be rated with 4 in a Likert – like scale for the questions concerning the absolute category rating. The middle quality level video samples were rated with slightly less, yet in the same grade as the best one. The worst one is rated with 2 points in the same scale.

On the other hand, in questions that concern the overall sensory immersion, all the samples were rated in approximately the same levels. The results were as expected in simulator sickness – related questions, with the lower resolution sample causing less simulation sickness feelings than the higher resolution ones.

The mean alpha values per quality sample do not follow the previous trend and their values remain constant around 2.0 Hz.

By analyzing the results, the authors conclude that whereas the immersion level should be projected in the mean alpha values for each category, this is not the case in this study. The authors elaborate by stating that VR is a complex experience that relies on more than the isolation of the participants senses from the outside world.

Katsigiannis et al. in [21] create and use a micro controller-based board to enhance the QoE reporting capabilities of a smart exercise bike system. More specifically, the system has three parts: the actual physical stationary exercise bike, a commercial head mount display (HMD) and a computer attached to them. The system is classified as an exergame, in other words a video game that involves physical activity. The proposed extension to the system has the ability to capture and report physiological responses to the exergame stimuli. These responses involve electrocardiography (EEG) and galvanic skin responses (GSR). For the purposes of the experiment, the authors created a virtual environment in which the user were able to roam freely using the stationary bike and the HMD.

The participants experienced 6 different quality levels of the developed set:

- high resolution (1024X1024) and high frame rate (60 fps) – (HH)
- medium resolution (512X512) and high frame rate (60 fps) – (MH)
- low resolution (256X256) and high frame rate (60 fps) – (LH)
- high resolution (1024X1024) and medium frame rate (30 fps) – (HM)
- high resolution (1024X1024) and low frame rate (15 fps) – (HL)
- random objects in low resolution (256X256) and high frame rate (60 fps) – (RH)

Participants use the service twice in each of the quality levels shown above. Between the switching of quality levels participants take a break and use the highest quality to avoid remembering the previous settings. The results of the subjective tests concerning the MOS and the sickness scores have two parts; The first one has to do with the actual MOS in varying combinations of resolution and frame rate. The second one has to do with sickness feelings.

More specifically, the MOS rating were found to be equivalent with most of the quality settings, by spanning in the 2,6-3,47 interval. The video settings with higher resolution were rated, as expected, higher than the ones with lower settings. Also, the random resolution videos follow this trend, for reasons that will be analyzed in the next paragraphs.

The results that concern simulation sickness feelings follow the same pattern. All the quality settings seem to cause same levels of discomfort to the participants. The total scores range around 60.

By analyzing the data from the subjective tests, we can see that the most negative experience for the users is correlated with the lowest resolution quality settings. One interesting fact in the results is that the quality settings in which random objects are projected with low resolution, achieve the same MOS as the high resolution and high frame rate settings. This finding can be explained by the fact that the users were paying more attention to the road and the fast moving virtual environment so they did not notice some random low resolution objects.

In the simulation sickness questionnaire, the results follow the pattern of the MOS. The highest total score of sickness symptoms is reported in the lowest resolution settings. Another interesting finding of the sickness scores is that the low frame rates do not affect the sickness feelings more than the resolution.

In an attempt to correlate the subjective tests results with the physiological results, the authors found that the peaks in galvanic skin responses and electrocardiography responses had no correlation to the quality settings.

Keighrey et al. in [22] assess the perceived quality in a VR and AR speech and language therapy (SLT) application which is used for diagnosis and mitigation of speech and language difficulties. This test is comprised of slides that assist the process of

Comprehensive Aphasia Test (CAT) [23]. The test presents the users a slide in which exists correct object described and some incorrect ones.

The experiment participants were asked to take the Comprehensive Aphasia Test and submit a questionnaire of 14 questions after the completion of the test. The questions' purpose was to evaluate the QoE of participants according to user interaction, immersion, discomfort and enjoyment.

During the experiment, some physiological metrics of the participants are recorded in order to be correlated with the objective test results. These metrics include: electrodermal activity, heart rate, response times and incorrect responses and miss-clicks.

The mean heart rate per slide seems statistically the same for VR and AR mode. The interesting part is the common trajectory of the two modes as participants proceed from slide P to 10.

Regarding the electrodermal activity, it is interesting that AR mode stimulates the participants in a greater extent than the VR mode. On the other hand, VR mode's electrodermal activity remains stable throughout the experiment only to reach EED levels of AR mode in the last slides. The authors associate this finding with the performance stress, which is more intense in AR.

It is worth noting that for 6 out of 11 slides the AR mode clearly outperforms the VR mode both in the meantime of accomplishment and in stDev of it. Another interesting fact about these numbers is that in VR mode the larger response time for slides 7 – 10 can be projected in the electrodermal activity for the same slides. This fact is interpreted as an increase in cognitive load as the user tries to find the correct answer something that is associated with anxiety feelings.

In the next table we can see the results of the incorrect responses and miss-clicks

Table 3 Miss- clicks and incorrect responses in AR and VR [22]

	AR		VR	
	Other	SD	Other	SD
Miss-Clicks Incorrect Response	.20 .35	.523 .933	.65 1.15	1.137 1.461

The authors state that a statistically significant result was found between the incorrect responses between the AR mode and the VR mode. This is reflected in the fact that on average the users gave 1.15 incorrect responses, while in AR the users gave 0.35 incorrect responses. A smaller but noticeable difference can be observed between the miss clicks between the AR (0.20) mode and the VR mode (0.65). The authors suggest that the results regarding the incorrect responses and the response times should be given as a feedback to the designers of the VR mode in order to increase the window for error and reduce the response times.

3.2.3 Network factors based QoE assessment methods

This section deals with QoE evaluation methods that try to correlate network influence factors with the QoE the user perceives. These methods conduct objective test with gradually deteriorating network quality settings and try to correlate the level of which each network factor deteriorates the result.

Doumanoglou et al. in [24] evaluate the impact of the network parameters in a 3D VR Tele - immersive (TI) game. TI games are a special video games type in which the motions of the participants are captured inside special rooms (TI stations) equipped with multi-cameras setup. This allows for these game applications to define two roles according to which the game can be used. The first one is the ‘player’ role, in which the users are immersed into a virtual environment through their realistic appearance using their local TI station. The second role is the ‘spectator’ role, which is watching the live session through a client application.

For the purposes of the current evaluation, the authors selected a use case in which two ‘players’ are interacting through LAN connection, while a third participant is spectating the game from a remote location. In this point, one must take into account the decoupling of interaction data transmission from the 3D appearance data transmission. Video games in general are designed to minimize the interaction time with the game objects, so the payload in the interaction data stream has very small payload. On the other hand the 3D reconstruction stream data has larger payload and falls behind the interaction data stream. In the current experiment settings, the QoE (which the experiment tries to assess) of the spectator depends:

- on the players’ 3D reconstruction geometry resolution
- on the players’ 3D reconstruction texture resolution
- on the players’ 3D reconstruction’s lag with respect to the game state

The authors recorded two live gaming sessions: one in a high quality resolution and one in a low quality geometry resolution. From these two sessions they derive the four samples used in the objective evaluation experiment. The samples settings are presented in the next table:

Table 4 Video samples settings [24]

<table>
<tr><th>Sequence</th><th>Session</th><th>Geometry Resolution</th><th>Texture Resolution</th><th>Visual Quality</th><th>Stream Rate (MBit/s)</th><th>Frame Rate (fps)</th><th>Duration</th></tr>
<tr><td>1</td><td>2</td><td>High</td><td>1920 × 1080</td><td>a</td><td>47.5</td><td rowspan="2">13</td><td rowspan="2">52s</td></tr>
<tr><td>2</td><td>2</td><td>High</td><td>640 × 360</td><td>b</td><td>24</td></tr>
<tr><td>3</td><td>1</td><td>Low</td><td>960 × 540</td><td>c</td><td>44</td><td rowspan="2">17</td><td rowspan="2">2min 27s</td></tr>
<tr><td>4</td><td>1</td><td>Low</td><td>480 × 270</td><td>d</td><td>8.5</td></tr>
</table>

These four quality settings videos are projected to the spectator by varying some network influence factors. These network factors are:

- The spectator is located at 50 ms Round Trip Time away from the players and the game server.
- The spectator is located at 100 ms Round Trip Time away from the players and the game sever.

- The transportation level protocol is TCP.
- The transportation level protocol is UDP.

All the combinations of the network parameters with the various video transmission qualities can be found in the next figure:

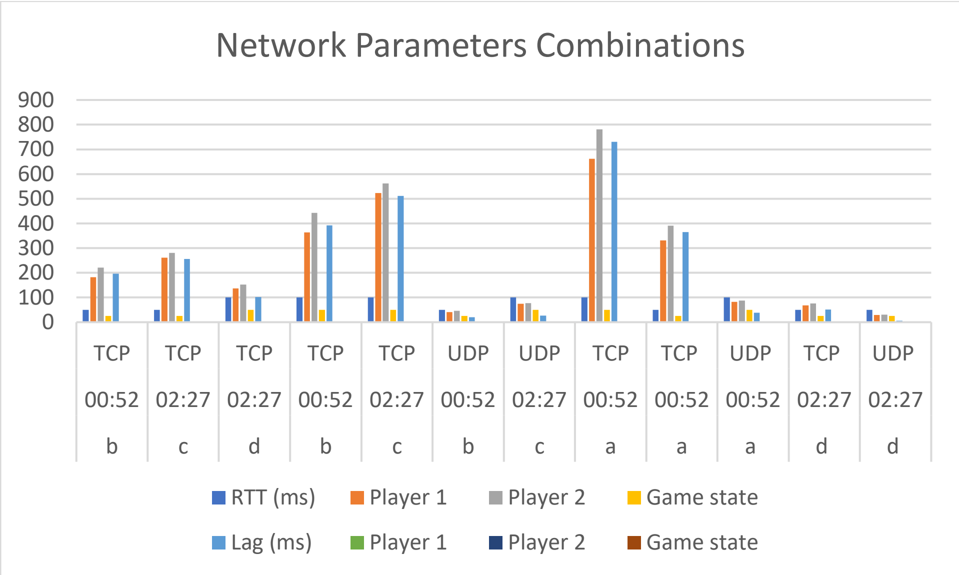


Figure 14 Network parameters combinations used in the experiment

The lag column in the table refers to the maximum time difference between the frame arrival of the players appearance data and game state data. The users are presented the different versions with a five-minute break between two versions. After the watching of the videos, the users are asked the following questions:

1. How would you judge the appearance of the players?
2. Did you find the navigation within the virtual environment easy?
3. Did you feel comfortable during the spectating sessions?
4. Was the movement and position of the players consistent with how you would imagine such a game being played in the real world?

The users rated the video sequences in the MOS scale (1 to 5). The next figure illustrates the ratings of the video samples from experienced and all users.

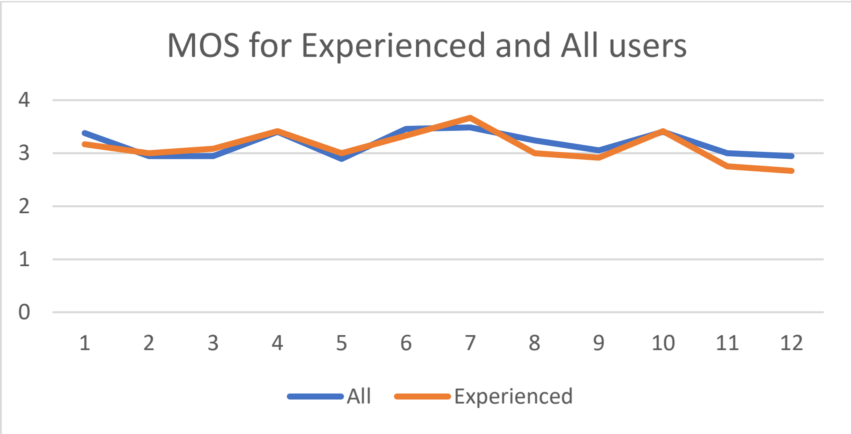


Figure 15 **MOS for experienced and all users for the video samples**

The next figure depicts the MOS for the results of the quantitative questions of the users' questionnaire for experienced and non-experienced users, respectively.

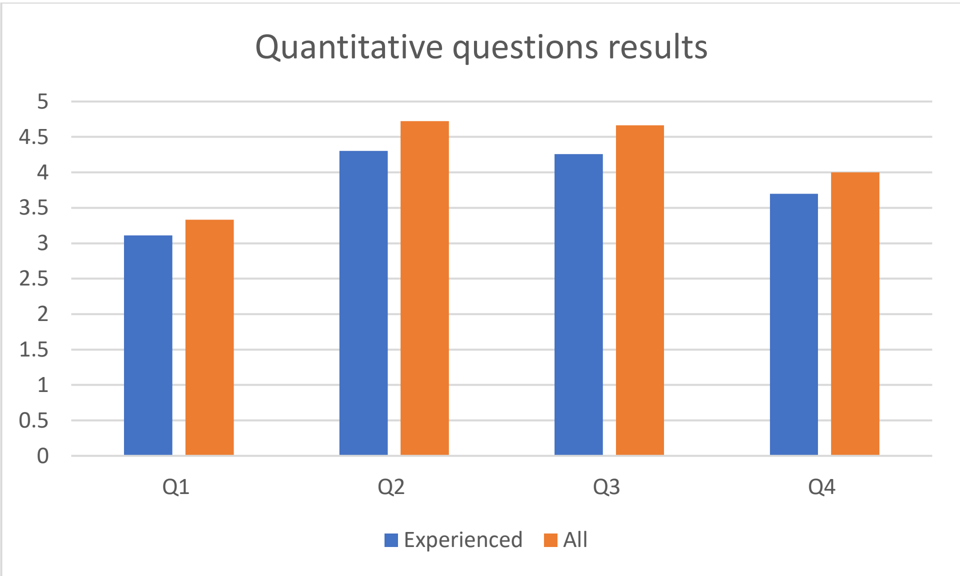


Figure 16 Quantitative questions results

The MOS for each individual quality and user group are depicted in the next figures. An interesting fact is that quality "b" receives higher rating than "a" both in experienced users and in all users group. This can be explained by the fact that frame rate in both qualities is equal. Also, the texture resolution is not evaluated as equally annoying. This fact is confirmed by MOS referring to geometry resolution, which is higher for high resolution both in experienced and in all group.

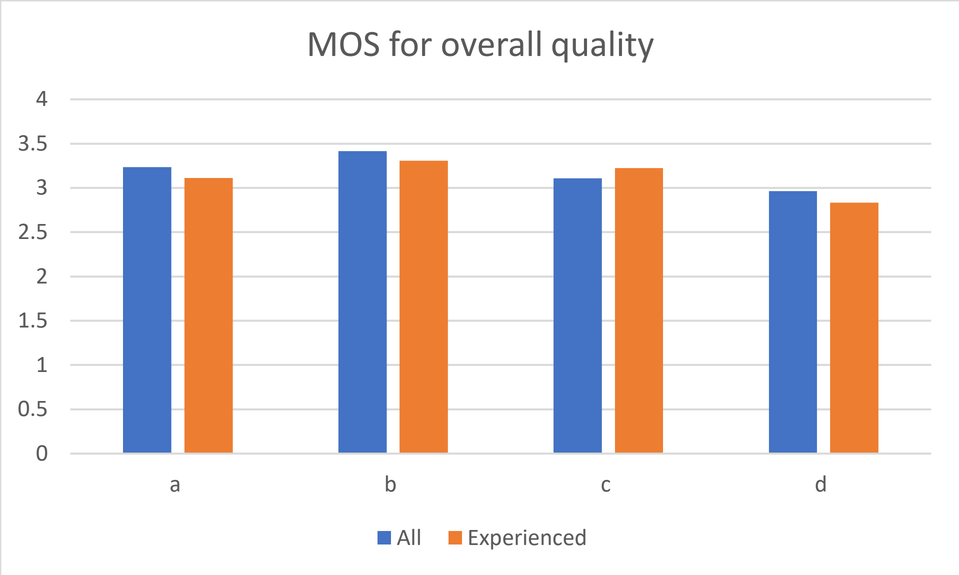


Figure 17 MOS for overall quality for experienced and all users

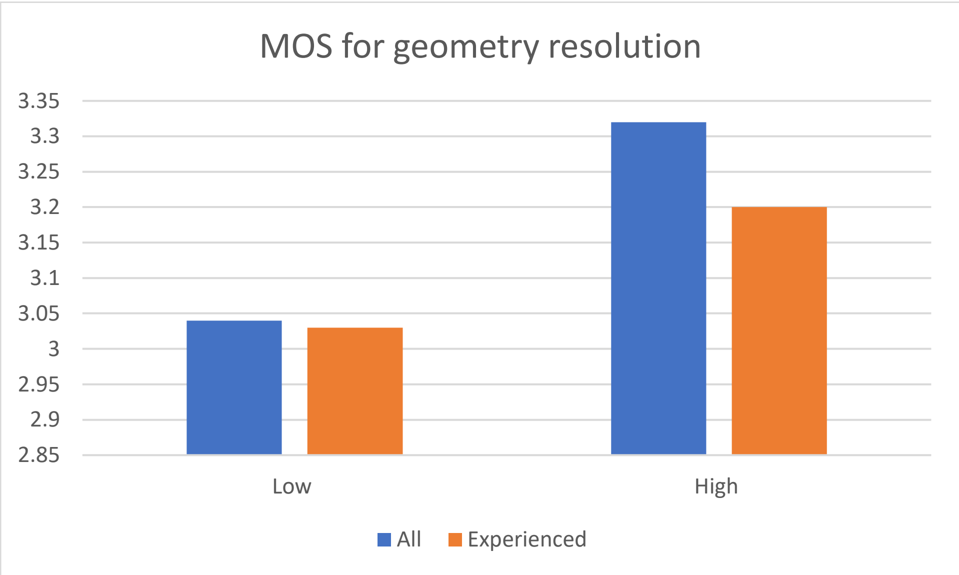


Figure 18 MOS for geometry resolution for experienced and all users

Regarding the impact of the network protocols in the QoE on the VR immersion, the MOS is the following. UDP protocol is well known for its better latency performance over TCP, due to its best effort design. However, this comes with the cost of unreliable transmission. The users' MOS confirm that the latency in UDP retrieves better scores. From a QoE perspective, this means that the users prefer the reduced latency with a cost of some lost frames.

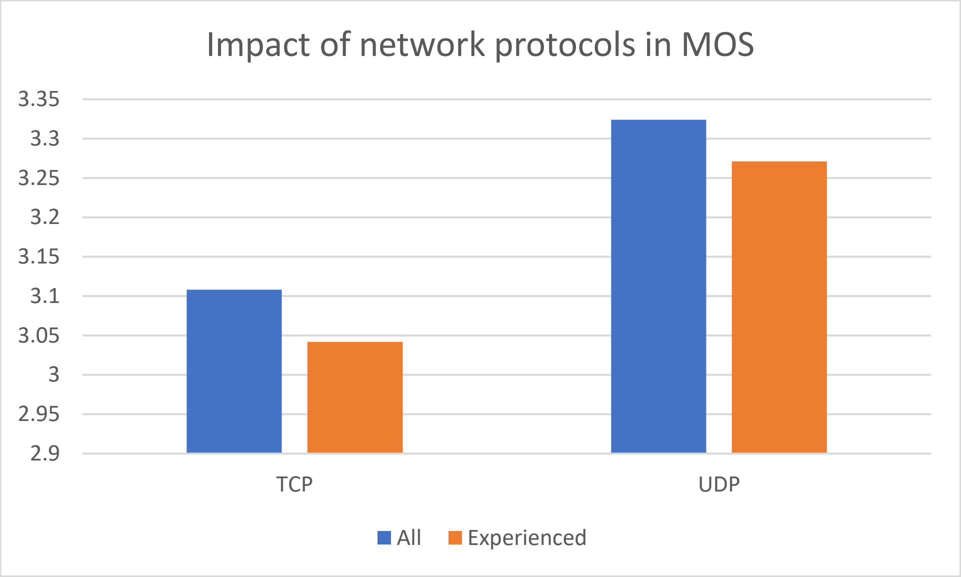


Figure 19 Impact of network protocols in MOS for experienced and all users

Another interesting abstraction of the users' responses was the grouping around high or low latency. The threshold between the high and low latency is empirical and is selected by the authors at 250 ms, based on typical values of game latency, average reaction time the capture rate of the human motion from TI stations. The results are shown in the next figure:

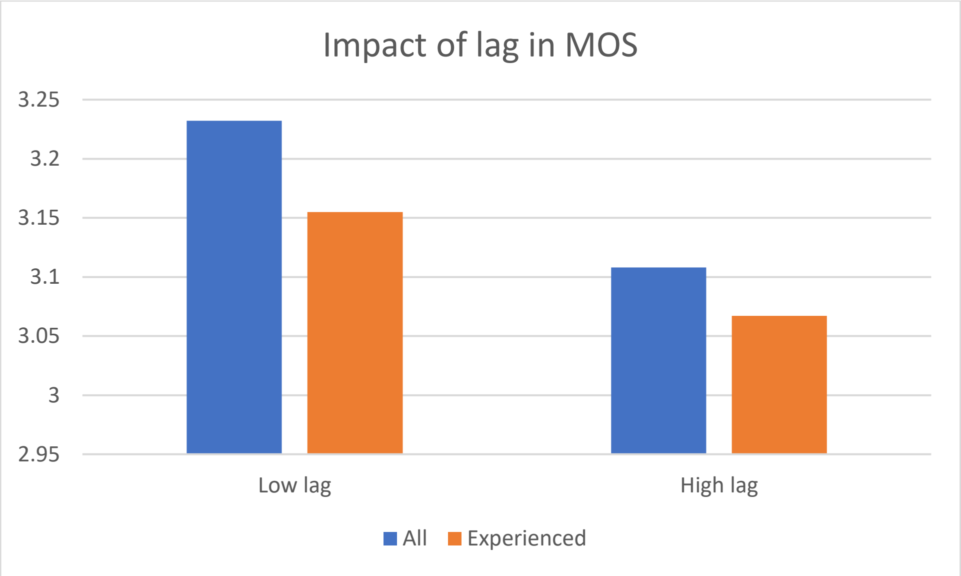


Figure 20 **Impact of lag in MOS for experienced and all users**

3.3 Subjective tests

In this chapter, we have created a comparative chart of the works that have been examined in the previous chapter.

Table 5 Variety of subjective tests used in VR QoE estimation

Work	subjective metrics	Objective Metrics	ITU recommendation	Simulation / real time	Reference Measurements
Egan et al. [19]	MOS	Heart Rate Electrodermal activity	P.910	Simulation	NR
Brunnström et al [13]	MOS SSQ	Delay	BT.500-13[5] P.910[6] P.913[8]	Simulation	FR
Tran et al [14]	MOS SSQ	Quantization Bit rate Resolution	ITU-T P.800.2 ITU-T P.913	Simulation	FR
Anwar et al [15]	MOS	Stalling Bit rate	Not mentioned	Real Time	NR
Zheleva et al [20]	MOS SSQ	Resolution Alpha frequency Heart rate	Not mentioned	Simulation	NR
Katsigiannis et al [21]	MOS SSQ	Frame Rate Resolution	ITU-T P.910	Simulation	FR
Keighrey et al [22]		Heart Rate Electrodermal activity	Not mentioned	Real Time	NR
Doumanoglou et al [24]	MOS	Transport Layer Protocols Bit rate	ITU-T P.910	Simulation	FR

From the comparison it is clear that all the related works use the MOS metric to conclude about the QoE of users. Sickness feeling is a very important issue in VR applications and engineers try to understand its root and limit it, this is why the SSQ (Simulator Sickness Questionnaire) is present in four of the works mentioned here.

3.4 Correlation between QoS Metrics and QoE

In this chapter we will discuss the extent at which the QoS metrics of the subjective experiments correlate to the QoE results from the users' questionnaire. The general result is that multimedia and network QoS metrics reflect better the measured QoE. In the other hand, human body response metrics have statistically insignificant correlation with the recorded QoE and lead to contradictory conclusions. In order to render this relationship, we created the next table:

Table 6 Correlation of QoS influence factors to the QoE results for VR

Work	QoE	Objective metrics
Egan et al. [19]	Users' immersion and satisfaction is higher for VR than for non-VR	Heart rate is higher for VR applications, while electrodermal activity is higher for non - VR applications
Brunnström et al [13]	MOS is reduced and sickness feeling is increased in the presence of multimedia annoyance factors	Screen and joystick delays degrade the quality of the VR application
Tran et al [14]	MOS is reduced and sickness feeling is increased in the presence of multimedia annoyance factors	Quantization, bit rate and resolution annoyance factors degrade the VR video samples
Anwar et al [15]	MOS is reduced and sickness feeling is increased in the presence of multimedia annoyance factors	Stalling and bit rate degrade the quality of the VR video samples
Zheleva et al [20]	MOS is reduced and sickness feeling is increased in the presence of multimedia annoyance factors	Heart rate and electrodermal activity are not related to the MOS results
Katsigiannis et al [21]	MOS is reduced and sickness feeling is increased in the presence of multimedia annoyance factors	Physiological factors are not correlated with the users' perceived quality
Keighrey et al [22]	MOS is reduced and sickness feeling is increased in the presence of multimedia annoyance factors	Heart rate and electrodermal activity reflects the users perceived quality

Doumanoglou et al [24]	MOS is reduced and sickness feeling is increased in the presence of multimedia and network annoyance factors	Bit rate, resolution and different transport layer protocols create a combination of 3D game quality

From the table above we can conclude that human physiological metrics are highly contradictory and in most of the works do not relate to the subjective tests results, with the exception of Keighrey et al [22]. On the other hand, most of the network and multimedia QoS factors are directly related to user experience. Works like Katsigiannis et al [21] try to correlate between both without success.

3.5 Discussion

What is the common ground among these works? What do the subjective and the objective tests have to offer to the estimation of how happy is a VR user?

One main common point between all the works mentioned in the previous section, is the contradictory results of the objective methods. Electroencephalographic data in [20] does not conform to the MOS. Users prefer a high-resolution VR video from a conventional video of the same resolution in [14]. People would rather consume a VR video in [19] even though most of them has motion sickness feelings. All the above artifacts prove why the common ground of the opinion of the users of a service, or an application is the ultimate performance metric. As noted before, human beings are unpredictable and most of the high level activities of the human brain are highly affected by the individuals' past experience.

Also, most of the works of the previous section confirm that resolution plays an important role in the satisfaction that a user gains from a VR service. More of the half of the users in [14] prefers to consume VR content if this is provided in high definition, despite the fact that general acceptability levels for the same quality video is the same for VR and non-VR video samples. Heart rate of the users of VR and AR applications is constantly increasing in the presence of new slides in [22].

Another interesting fact is that even though resolution is important for the users, it can be a double edge sword. MOS falls rapidly to unacceptable in [14] with the degradation of the video resolution. The same thing happens in [15] where stalling events lead to larger reduction of MOS for higher resolution videos. Also, the importance of resolution depends on the application. VR gaming spectators evaluate frame rate as more important in [24].

Furthermore, since VR applications involve all the senses, it is clear that vision has a dominant role in user satisfaction. MOS was practically unaffected for joystick stalls in [13] while was drastically reduced for video stalls.

Finally, it appears that UDP is more suitable in high frame rate interactive applications, while TCP seems more suitable for VR environments where image resolution plays an important role.

To sum up, in order to be able to provision the QoE of users in VR Applications, one should be able to monitor and gather data about:

- The application: screen resolution, frame rate, lag
- The network: packet loss, bit rate, jitter
- The user: physiological metrics and feedback

3.6 Validation and proposition

The bibliographical review regarding QoE for VR applications reveals that a lot of work has to be done yet in order to be able to estimate and satisfy the users' needs. QoS data collected in the previous steps should be able to transform with clarity to human experience. For example, the QoE provisioning system should be able to tell that the user has been immersed in the virtual world, the precision the user enjoys through the interaction with the application is satisfactory, the interaction with the game is realistic and so on.

In order to achieve these goals, at least two entities should be added to the traditional VR application model: a client-side entity and a server-side entity. The client side should be able to monitor the application and the data and the server-side entity communicating with the client-side entity should be able to monitor the network.

The data collected from the server side and the client-side entities should be used as input to an artificial neural network which will be trained to match the data input to the human quality sentences mentioned in the previous paragraphs. The results of the neural network should be cross validated with other sources.

When the overall quality of the application drops, the administrative entity should take action in the direction of quality assurance. An analogy from Video on Demand QoE assurance mechanism is the adaptive streaming mechanism. Similar mechanisms can be used in VR applications by placing the annoyance factors in hierarchical order and trying to avoid the worst ones, even without overall control over the Network.

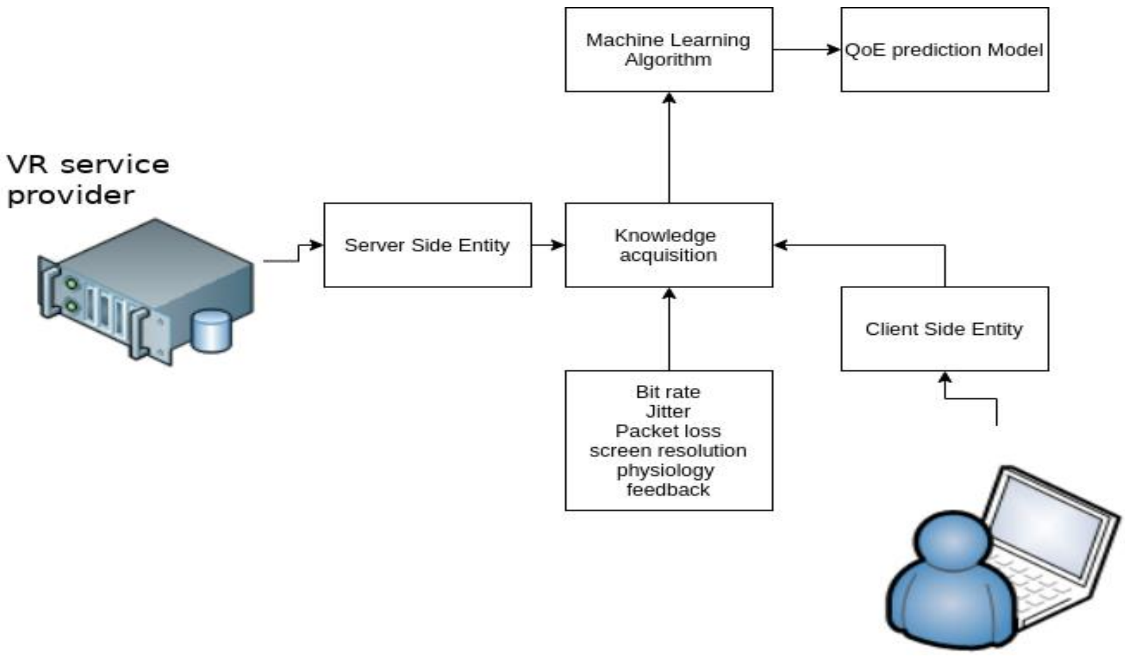


Figure 21 QoE provision for VR applications

4 QUALITY OF EXPERIENCE IN AUGMENTED REALITY APPLICATIONS

4.1 Introduction

Augmented reality (AR) is an interactive experience of a real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information, sometimes across multiple sensory modalities, including visual, auditory, haptic, somatosensory and olfactory. AR can be defined as a system that fulfills three basic features:

- a combination of real and virtual worlds
- real-time interaction
- accurate 3D registration of virtual and real objects.

The overlaid sensory information can be constructive (i.e. additive to the natural environment), or destructive (i.e. masking of the natural environment). This experience is seamlessly interwoven with the physical world such that it is perceived as an immersive aspect of the real environment.

In this way, AR alters one's ongoing perception of a real-world environment, whereas VR completely replaces the user's real-world environment with a simulated one. AR is related to two largely synonymous terms: mixed reality and computer-mediated reality.

One of the most powerful characteristics of the AR is that it is used to enhance the physical world with enriched, artificially generated items. This fact poses a challenge to the AR designers because AR systems have to act as an intermediate between the physical environment and the digital one. This is opposed to other multimedia systems that have to take care only of the digital part of the experience. AR also has a lot of potential in the gathering and sharing of tacit knowledge. Augmentation techniques are typically performed in real time and in semantic contexts with environmental elements. Immersive perceptual information is sometimes combined with supplemental information like scores over a live video feed of a sporting event. This combines the benefits of both AR technology and heads up display technology (HUD).

Possible applications of AR include:

- Archaeology
- Architecture
- Urban design & planning
- STEM education
- Industrial manufacturing
- Commerce
- Literature

- Visual art
- Fitness

One excessive use case of AR that is worth noting is that of a guidance and support system for workplace environments. This case is particularly innovative because an AR system could act as a real time virtual helper for many professionals and replace the traditional written instructions.

4.2 Related works

There is a number of works that tries to quantify QoE in AR. Most of them try to correlate network, multimedia and physical annoyance factors to MOS and sickness feelings subjective tests. The content and the conclusions of these works will be the concern of the next subsections.

4.2.1 Multimedia factors-based methods

Seeling in [25] conducts a subjective evaluation experiment for an AR image system. The equipment used here is inexpensive and consists of the binocular Epson Moverio BT–100 mobile viewer and NeuroSky Mindwave Mobile headset which is a user level equipment capable of measuring various bands of Electroencephalographic (EEG) activity.

Image database Tampere Image Database from 2013 (TID2013) is used in order to have a pool of different quality images. Tampere Image Database been employed in several such subjective experiments and every image in various quality levels has a MOS associated with it. For the particular experiment, only JPEG encoded samples are used.

The participants in the experiment rate the images in various quality levels in AR and non-AR mode and rate their experience in a five-point Likert scale. In the contrary to the subjective evaluation, Tampere image database has a 1-10 rating system. The author regularizes the MOS taken from the database with the formula below:

$$\overline{M_i^l} = \frac{4 \cdot TID2013 - MOS(I_i^l)}{9} + 1$$

where M represents the Likert MOS in the TID2013 and the I represents the 1-10 scale rating for the same image and quality level. Also the author aggregates the MOS for a quality level by calculating the mean of the MOS for all the images on the same quality level.

The MOS for AR mode is calculated on a granularity level per image per impairment level. In general, MOS is different within the same group for every specific image, and the differences in MOS is ranging up to 2 grades in the Likert – like scale.

Also, the first 3 impairment levels, namely the 0, 1 and 2 are equivalent in terms of users' MOS, while the next impairment levels present a linear drop in MOS. One obvious observation is the steeper decline in MOS as the image quality is degraded in comparison to the MOS for the non – AR mode for the same pictures.

The results are slightly different for non – AR mode. The MOS results have lower deviation for the same impairment level. Also, the overall MOS is lower for the same impairment level compared to AR mode MOS. Another interesting fact is that the MOS is dropping linearly for all the impairment levels. This means that in non-AR mode there is not a point after which the image degradation becomes unbearable, but the MOS drop for every impairment level. This means that the participants can understand every drop in image quality and respond negatively to it.

The authors attempt a direct comparison between AR and non – AR mode, by calculating the difference on the MOS between the two modes. The result is the following:

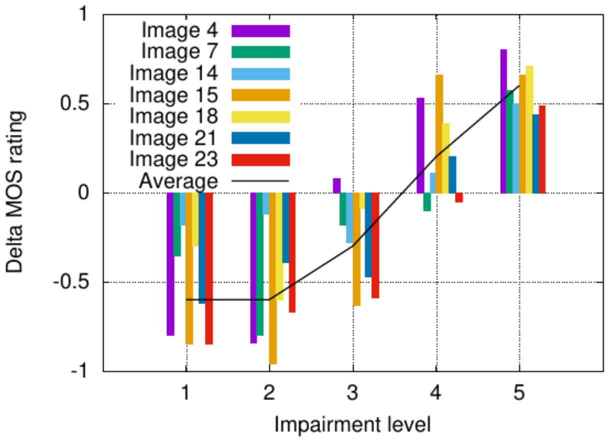


Figure 22 Difference between AR and non - AR modes [25]

The immediate observation is that worse quality levels receive higher MOS in non-AR mode than in AR mode. Also, the difference between the MOS is in favor of higher quality levels than the lower quality levels.

Next, the authors present the relationship between the MOS and the EEG activity and focus specifically in the area of the Low gamma frequencies which has been recorded to be highly correlated to the user visual perception [26]. The results of the measurement of the aggregated Low gamma frequencies for different image quality levels are depicted in the following diagram. The low gamma frequencies are recorded for a time span of 10 seconds:

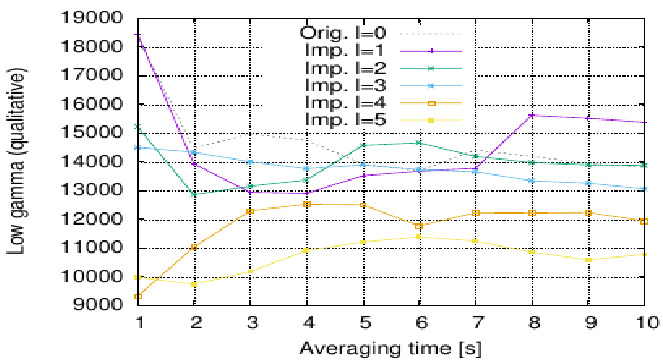


Figure 23 Gamma frequency over time for different quality levels [25]

An obvious
observation is that the aggregate levels of the Low gamma frequency has instantaneous

rise in very short time span for high quality levels. In the contrary, low gamma levels take some time to rise in the lower quality levels and these levels remain in low levels overall.

Some overall conclusions from all these measurements are:

- QoE levels for upper medium image quality levels are practically the same as the QoE levels for the best quality levels in AR mode. There is little improvement by raising the image quality levels above upper medium.
- The QoE is lower for low quality AR images than for low quality opaque mode images.
- High quality images in AR mode seem to stimulate directly the low gamma activity of the human brain something that is related to visual perception.
- The high variance of MOS in the AR scenarios result in a strong support of the objective estimation methods for estimating the user experience. This estimation should be directed towards objective QoE evaluation for each individual.

Longyu et al in [27], evaluate the experience that users perceive using HoloLens AR smart glasses based on a fuzzy inference system. Then, they validate the estimation results with actual participants opinions and try to generalize the results.

Fuzzy inference systems use a selection of parameters and combine them in order to create a cohesive metric towards the direction of measuring something that is not inherently quantitative. In this sense, fuzzy inference systems are a good match for measuring QoE. The authors analyze the input parameters as depicted in the next figure:

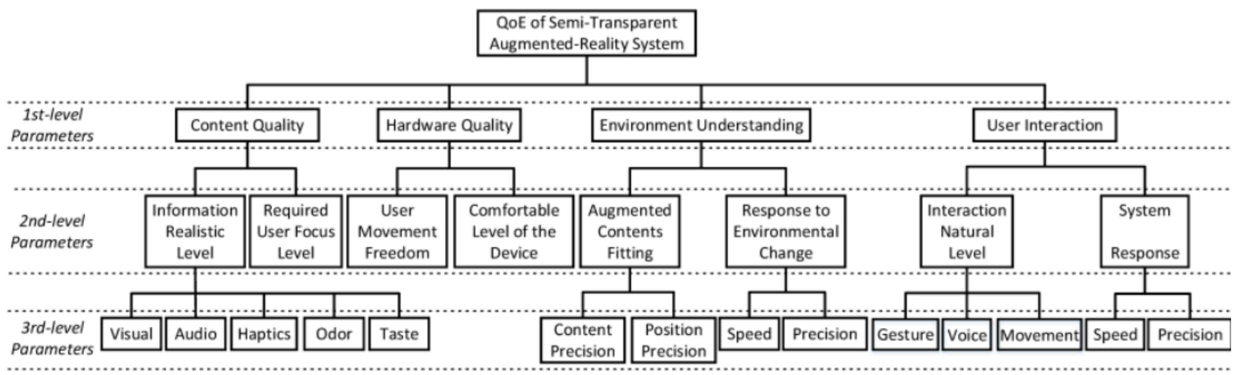


Figure 24 fuzzy inference parameters [27]

The authors use member functions that map the user input in the [0,1] space for the first level parameters namely, Content quality, Hardware Quality, Environment Understanding, User Interaction. The input to the member functions is the users' responses from the questionnaire and have been normalized in the [1,5] space.

After the member functions have been defined, the users' answers can be used to derive fuzzy rules according to which we can conclude about the overall user experience. For example, if a user rates content quality, hardware quality, environment understanding, user interaction and the overall rating as 75, 50, 100, 75, and 85, respectively, these values can subsequently be interpreted into member functions as Good, Fair, Excellent, Good, and Good, respectively, and a fuzzy rule can be derived as follows:

IF content quality is Good,

AND hardware quality is Fair,

AND environment understanding is Excellent,

AND user interaction is Good,

THEN the overall rating is Good.

The inference system used in this work has 5 membership functions that implement the fuzzy rules for all the basic inputs of the experiment. These inputs are:

- content quality
- hardware quality
- environment understanding
- user interaction.

By combining these inputs, the inference system creates 43 fuzzy rules, the answers to which produce the overall result.

The AR applications used for the experiments were a first-person shooting game "RoboRaid" and an adventure game "Young Conker". After a familiarization period, the users are asked to dive into these games. After the game, the participants are handed a questionnaire designed based on the fuzzy inference system that is proposed. The results of the questionnaire and the corresponding fuzzy inference system results are depicted in the next table. QoE_u is the overall user rating and QoE_f is the estimated user rating using the fuzzy inference rules systems:

Table 7 Comparison between the user reported and framework calculated MOS [27]

Testing Application	QoE	Mean	Lower Bound	Upper Bound	Std. Error	Me-dian	Std. Dev	RMSE (%)	T- Test P-Value
RoboRaid	QoE_u	85.133	82.776	87.491	1.099	85.0	4.257	3.895	0.493
	QoE_f	85.853	83.450	88.257	1.121	85.0	4.340		
Young Conker	QoE_u	86.867	82.282	91.452	2.138	90.0	8.280	5.791	0.728
	QoE_f	87.413	84.427	90.399	1.392	91.9	5.392		

This aggregate table shows that the framework approaches very well the mean of the opinion scores for every game. On the other hand, there is a high stdDev and RMS that indicates that a deeper analysis should be made to understand better how this framework approximates the users' experience.

The figure below shows the actual MOS for each individual user and for each individual game:

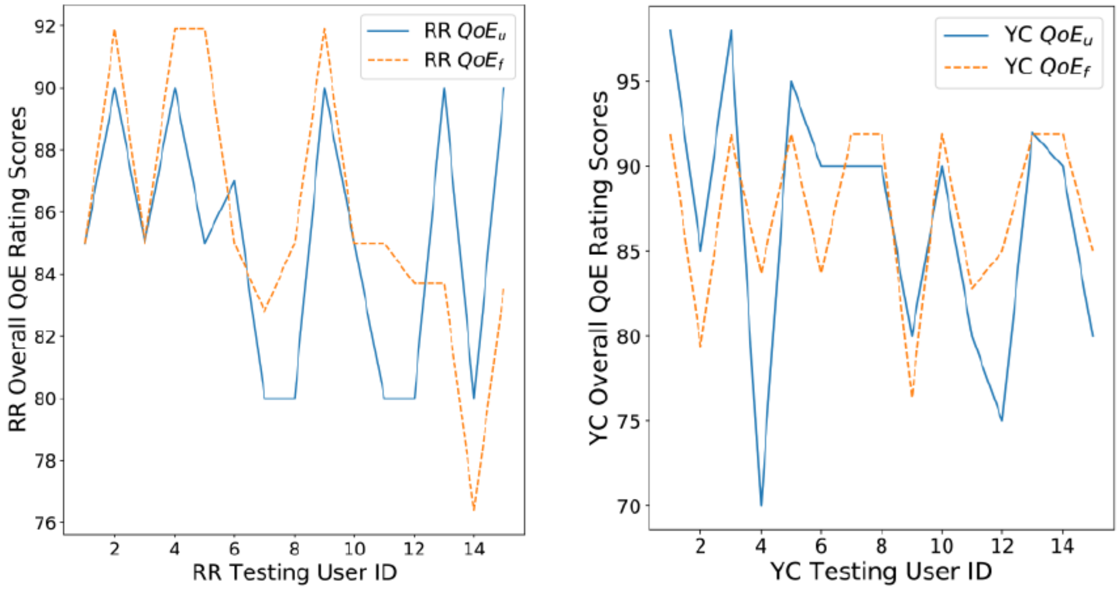


Figure 25 MOS for each user for each game [27]

This more descriptive view reveals that the framework actually underestimates the MOS for individual users. In this context, the relatively high standard Deviation and Root Square Error values actually make sense by giving a more pessimistic perspective of the user experience.

Bauman and Seeling in [28] use the BRISQUE metrics [29] and EEG metrics to conclude about the quality that a user perceives by watching AR enhanced images in various degrading qualities. The results are validated using subjective tests and MOS. For the purposes of the experiment professionally generated spherical images are used. The images are displayed in a custom viewer application that displays them on a head-worn mobile AR viewer.

The authors use the Epson Moverio BT-200 without shades head-worn AR viewer. Also, they employ a commercial-grade EEG headband, used for capturing EEG data for TP9, Fp1, Fp2, and TP10 positions. These positions denote standardized electrode position in the human skull. For these positions common brain wave data are gathered at 10Hz sampling. The following EEG values are gathered:

- low ɩ^p at 2.5-6.1 Hz
- delta δ^p at 1-4 Hz
- theta θ^p at 4-8 Hz
- alpha α^p at 7.5-13 Hz
- beta β^p at 13-30 Hz
- gamma γ^p at 30-44 Hz

In addition, users are called to rate the quality of each image presented to them in a 5-point Likert scale. The MOS for a quality level is the mean of the opinion scores of all users for this quality level.

The participants watch the aforementioned images in 5 impairment levels and two modes: Regular content (AR) and Spherical mode (SAR). The initial BRISQUE metrics for the images can be found in the next table:

Table 8 BRISQUE scores for the initial images [28]

Impairment	Garden	Bamboo	Mosque	Ocean	Golf	Beach
Orig., 0	27.602	11.325	8.643	10.107	11.632	3.402
1	34.082	31.327	15.609	15.020	31.747	23.422
2	33.384	41.784	27.429	22.109	32.648	25.361
3	35.943	51.389	46.592	33.902	34.323	32.967
4	49.366	57.931	61.083	43.216	46.32	52.790
5	80.646	90.495	86.582	65.843	86.237	85.074

The users are asked to rate the images in the two modes of view (AR and SAR) and all the impairment levels. These ratings are called SAR-MOS and AR-MOS denoting the MOS for each view mode. The authors also include the difference in MOS between the two modes, which is denoted as SAR-DMOS. In the next figure presented are the MOS results and the BRISQUE measurements in the various impairment levels:

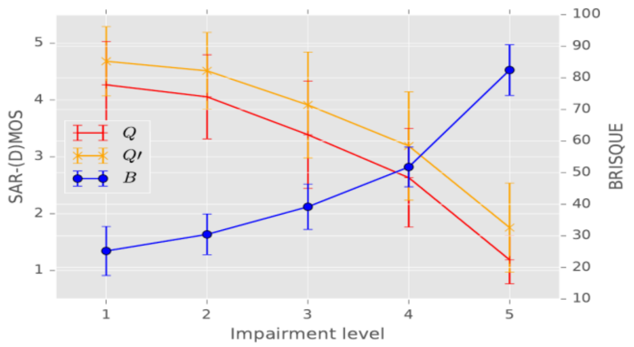


Figure 26 MOS and BRISQUE metrics for the various impairment levels. Q denotes the SAR-MOS and Q' denotes the SAR-DMOS [28]

Observing the results in MOS, we can understand that the MOS follows a logarithmic-like fall in the degradation of the quality levels. Also, the difference falls in a slightly faster manner for the worst quality levels. In the contrary, BRISQUE metrics follow a logarithmic rise while the quality is degrading which can be explained by the fact that BRISQUE is a spatial metric.

The authors came up with a fitting curve given by $Q = 5.3661 - 0.0493B$ with $R^2 = 0.861$ where B denotes the BRISQUE metric value. for the SAR-MOS case colored in red. In addition, the difference between the AR spherical mode and the simple AR mode follows a nearly parallel curve given by $Q = 5.7843 - 0.0474B$ with $R^2 = 0.886$. These results reveal that a somewhat linear relationship exists between perceived quality and BRISQUE values, but with a lot of exception points to it.

To strengthen further this point, the authors try to fit the IQX equation (Chapter 2.4) in the current metrics. More specifically, the QoE is represented by MOS, while the QoS is represented by BRISQUE. According to IQX hypothesis, the QoS and QoE are connected with an exponential relationship in the general form of $QoE = a \cdot e^{-b \cdot QoS}$. Nonlinear regression for the determination of coefficients a, b using the data of the experiment results in:

$$Q = 130.021 \cdot e^{-3.872 \cdot 10^{-4} \cdot B} - 124.632$$

$$Q' = 8489.479 \cdot e^{-5.589 \cdot \cdot 10^{-6} \cdot B} - 8483.694$$

with R^2 values 0.861 and 0.886 for the SAR-MOS and the SAR-DMOS, respectively. This proves that the BRISQUE metric fits equally well both in linear and in non-linear equation as a variable upon which the QoE is dependent.

Motivated by the results, the authors try to find a polynomial relationship between the BRISQUE metric and the QoE that a user perceives. Several degrees are applied in order to approximate the best fitting in the recorded data and the results are cross-validated to provide *a-posteriori* information about the quality of fitting. The results for polynomial degrees up to 3 are presented in the table below. MSE denotes the mean square error, MAE denotes the mean absolute error and MedAE denotes the median absolute error. By examining the results, we can tell that in the first-degree approximation the logistic regression produces better results, but with higher errors. In the second-degree approximation the mean errors are reduced for both methods.

Alexiou et al in [30] evaluate the effects of geometrical degradation to the users' perceived quality in the context of AR 3D point cloud. A point cloud is a common and practical way of storing and rendering 3D models in AR. Point clouds are a viable solution for the user to perceive 3D digital objects in a more immersive way.

A subjective test is conducted in order to assess the impacts of geometric distortion and noise in the point cloud. During the experiment, the participants are using a HMD in order to be able to assess the quality of the point cloud. Five different contents are presented to the participants, in the shape of easy to recognize shapes.

The five shapes are subjected in two kinds of degradation: the first one is the introduction of noise in the point cloud, while the second one is a sparser and more regularized representation of the point cloud. For this reason, no color was added to point cloud representation, nor complex scenes were used, since it would be difficult for the participants to evaluate and rate the impact of geometric degradation in such settings.

Regarding the first annoyance factor, the introduction of Gaussian noise affects every point in the point cloud by altering its position. The metric of this disposition is expressed by a target standard deviation.

The second annoyance factor is achieved by using a standard compression algorithm used in multimedia, the Octree-pruning. The application of this algorithm to the point cloud leads to visible distortion of the objects and incorrect density of the point cloud.

The participants of the experiment are using a software and hardware combination developed by the authors. For the display of the results is used an Occipital Bridge AR headset. The objects are projected in an AR manner in a test table covered in a medium gray tissue. The subjects were instructed initially to stand in front of the test table at the distance of 1 meter and they were free to change their position after the beginning of each evaluation session. After inspecting the objects, the users provide their scores.

Since two types of degradation were assessed, the evaluation procedure was split in two different sessions. For every session, a training phase was initialized, where the participants were informed for the general characteristics of the type of content they are going to be presented and the type of degradation that is about to be introduced. For each session, 25 distinct stimuli were used.

The effects of the first annoyance factor, namely the Gaussian noise is quite normal in its distribution and affects all the point cloud shapes in equivalent manner as far as MOS is concerned. For example, for a standard deviation of 10^{-3} of the Gaussian noise, the MOS spans in an interval of less than 1 degree in a Likert scale. This means that for the same levels of degradation, each shape receives equivalent MOS.

The effects of the second annoyance factor, namely the compression algorithm, the results are quite different, because the Octree-pruning algorithm causes spatial impairment in a less normal way than the Gaussian noise. The subjective tests' results for this degradation factor varied greatly depending on the image and the content. More specifically, for the lowest compression level, the MOS spans in an interval of 1.5 degrees in a Likert scale.

It is obvious that while the standard deviation of the Gaussian noise is reduced, the MOS is decreasing in logarithmic manner. Another thing worth noting is that participants perceive equivalent deduction in quality regardless of the content with the introduction of noise. This can be easily explained since users are able to recognize more easily the degradation and the distortion of the geometric shapes.

On the contrary, when the content is subjected to compression, like distortion, the content affects the final rating. We can observe that cube has significantly higher ratings than the dragon for the same distortion level. An obvious explanation for this is that more complex (or rounder) shapes are easier to be rated lower. Also, since people tend to rate by comparison, irregular shapes such as the vase tend to be rated lower when a normalizing algorithm is applied to them. Another useful corollary of the compression subjective tests is that regular shapes such as the sphere and the cube have more concentration points in their original form and a small compression level does not affect their rating because it just normalizes the denser areas.

Apart from the subjective tests, the authors use objective quality metrics used in point cloud models and attempt to correlate the objective metrics with the MOS for each setting. The objective quality metrics used in point cloud models are the following:

- Point-to-point (p2point): The point-to-point error is calculated by connecting each point of the point cloud under evaluation to the closest point of the reference point cloud.
- Point-to-plane (p2plane): The p2plane error measures the projected error along the normal of the closest point of the reference point cloud.
- Peak-to-Signal Noise Ratio: The ratio of the squared maximum distance of the nearest neighbors.

For each of the above measures, geometric errors can be calculated either by using root mean square (RMS) difference, or the Hausdorff distance. The authors combine all the above objective metrics to create a total of 8 objective metrics for the point cloud representation. Then, a fitting to the MOS results is attempted in a linear, logistic and cubic manner. The results of the fittings in terms of Pearson linear correlation coefficient (PCC), the Spearman rank order correlation coefficient (SROCC), the root-mean-square error (RMSE) and the outlier ratio (OR) between MOS and MOS_p are presented in the following figure:

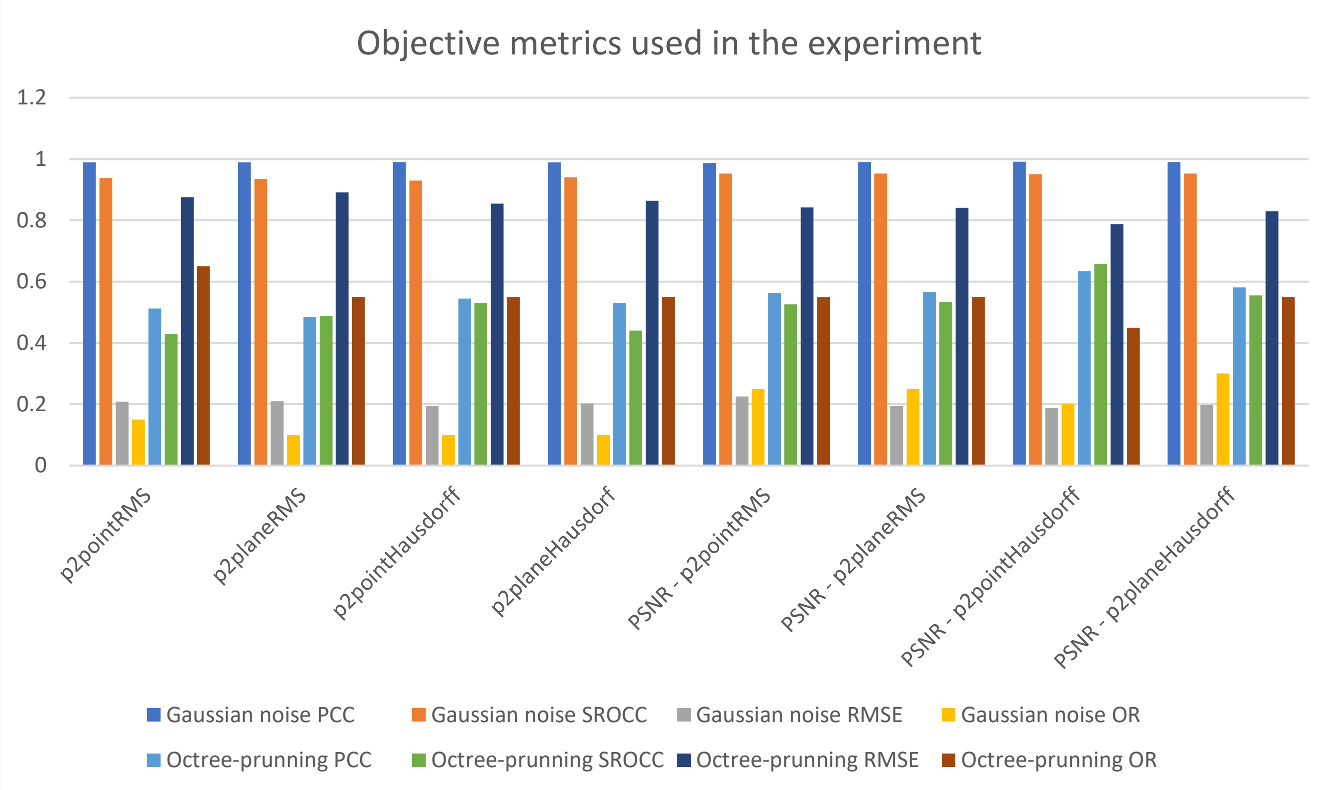


Figure 27 Objective metrics combinations used in the experiment.

The next figure depicts the relationship between the objective metrics and the MOS for various levels of degradation:

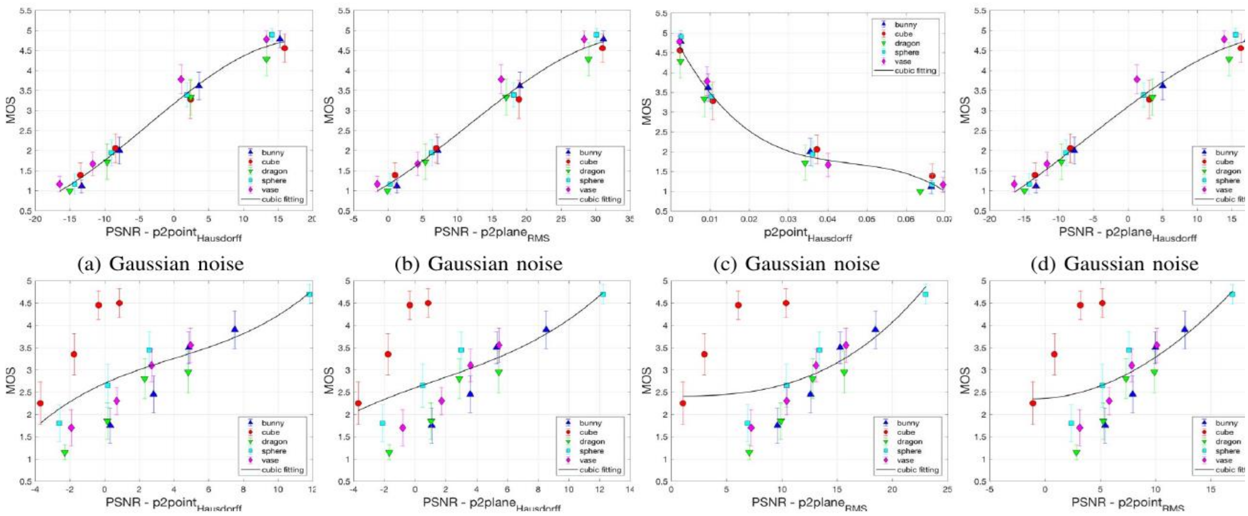


Figure 28 Fitting between objective metrics and the subjective results [30]

It is obvious that for the Gaussian noise case the correlation is strong between the objective metrics and the subjective scores. One interesting point is that distances used in objective metrics are full referenced, hence reducing the accuracy of the final fitting. On the other hand, the participants were able to observe noisy point clouds alongside with the original ones, hence they were able to tell the difference between the original ones and the degraded ones. These facts explain the strong correlation between the objective and subjective results in the Gaussian noise case.

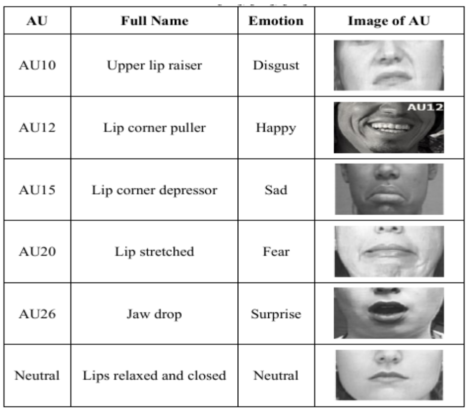
On the contrary, the correlation between the MOS and the objective results is weaker in the case of compression. According to the authors, this has to do with the specific attributes of each shape. For example, curved surfaces are represented by less and less points during the compression. As a consequence, less detail and more rough representation are observed, which is rated lower by the participants. On the other hand, planar shapes as the cube do not present this attribute and we can see that they are better fitted to the MOS regarding the objective metrics. Also, higher compression rates do not affect structured shapes as the cube so much, since the simplification of their point cloud and the loss of points is not perceived by the participants as equally annoying.

4.2.2 Human factor-based evaluation methods

Eoghan et al in [31] use the lower facial micro expressions to evaluate the perceived quality that a user enjoys. The authors use a subjective evaluation test in which participants are selected according to ITU P.913 specifications. In the experimental setting, the participants use an AR application that guides them towards the resolution of a Rubik's Cube. This application is developed using the Kociemba algorithm [32] to solve the standard 3X3 Rubik’s cube in the fewest possible number of moves. By using a prototype AR HMD, the application scans the initial state of the Rubik’s cube and guides the user towards the most efficient solution of the problem according to the Kociemba’s algorithm.

The experiment begins with the cube in the same initial position, namely the superflip position that requires the maximum 20 steps to solve using the optimal algorithm. The test participants are divided in two groups, the group that solves the Rubik’s cube without the application assistance and the group that uses the application as an assist to solve the problem. Both groups are handed instruction on how to solve the problem and the non-AR assisted group is provided with one-instruction-per-page format.

A desk mounted camera in 1080p resolution is used for the lower facial recognition of the participants. The camera uses the OpenFace recognition application [33], from which the facial micro expressions were categorized into cognitive characteristics. The categorization done by the authors is the following:



AU	Full Name	Emotion	Image of AU
AU10	Upper lip raiser	Disgust	
AU12	Lip corner puller	Happy	AU12
AU15	Lip corner depressor	Sad	
AU20	Lip stretched	Fear	
AU26	Jaw drop	Surprise	
Neutral	Lips relaxed and closed	Neutral	

Figure 29 lower facial expressions and emotions [31]

After the training phase, the participants are asked to solve the Rubik's cube, each team with the appropriate manner (AR directed or not). After the Rubik's cube resolution, the participants are filling a self-assessment manikin (SAM), which is a post-experience questionnaire about the users' affective state. The SAM questionnaire had three scales, one for each dimension of affect (arousal, valence and dominance). The participants also fill a 5-point Likert scale questionnaire consisting of 14 questions that aim to cover the feelings of utility, interaction, aesthetics, usability and efficiency, as well as their acceptability of the assistance of the AR application.

Alongside with subjective questionnaires, the authors try to assess the QoE implicitly by interpreting the participants' lower facial micro expressions, both for paper-based instructions and for AR based instructions as discussed in the previous paragraph.

The authors present this assessment as a graph representing the deviations from the baseline for the normal expressions. From this presentation it can be observed that AR mode instructions have a more intense impact on facial expressions, compared to non-AR mode instructions.

The expressions that seem to be more stimulated during the experiment are:

- The facial expression associated with the sad emotion.
- The facial expression that represents the fear
- The neutral expression is also more dominant for AR mode.

An immediate observation is that the "bold" feelings (happy – sad) have more positive deviation for the AR group and have negative or no deviation in the paper-based instructions group. This fact, in combination with the negative deviation of the surprise and neutral feelings shows that the AR instructions were clearer and were a steady helper in the solution process.

The authors also evaluate the deviations from the baseline of micro expressions for the same emotions. The lower facial micro expressions figure differs only in the feeling of the surprise for the AR assisted group.

Also, the feeling of disgust is significantly lower in the lower facial micro expressions picture for the paper-based instructions group, while the feeling of sadness is significantly lower for AR group when evaluating micro expressions. The statistical differences are depicted in the table below:

Table 9 Statistical differences between the test groups for normal and micro expression of five basic emotions [31]

Emotion	Normal Expression	Micro expression
Fear	0.085	0.185
Disgust	0.081	0.613
Sad	0.884	0.621
Neutral	0.032	0.001

Happy	0.046	0.053
Surpris e	0.002	0.070

Finally, the authors present the results of the SAM that the users filled before the tests. There are no significant statistical differences for the feelings of valance, arousal and dominance between the two groups. One interesting point is the correlation between the AR group's lowest valance and the highest deviations of surprise in the facial expressions. Also, the paper-base instructions group has a higher mean valance feeling and this is reflected to the group's higher deviations of disgust feelings in the facial expressions results.

When it comes to the results of the subjective questionnaire, the results are tremendously different. The participants found the written instructions more useful and the MOS was higher for several questions. More specifically, in the direct question whether the instructions were useful, AR group responded with a Mean Rank (MR) value of 22 in the Mann – Whitney U – test scale, while the paper instructions group responded with a MR of 27. Another important question that highlights the users' annoyance is question 8 in which the AR users respond with a MR of 28.33, while the paper assisted group responds with a MR of 20.67. A summary of the questionnaire results can be found in the next figure:

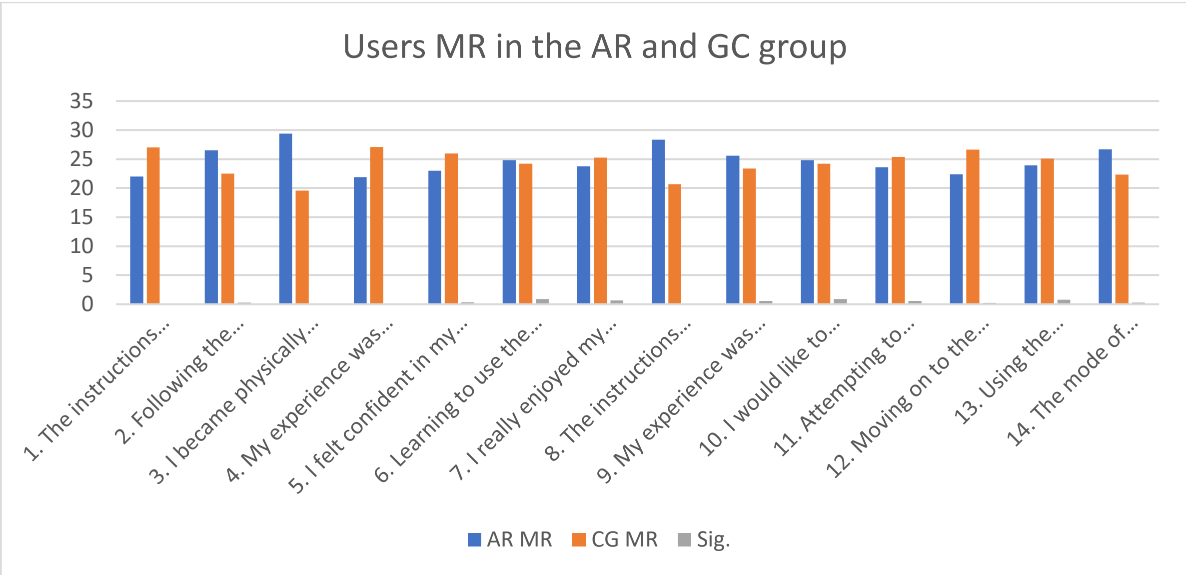


Figure 30 Users MR in the AR and GC group

Bauman and Seeling in [28] use the electroencephalographic data gather from consumer level device to estimate the user experience. The authors utilize a commercial-grade EEG headband, used for capturing EEG data for TP9, Fp1, Fp2, and TP10 positions. These positions denote standardized electrode position in the human skull. For these positions common brain wave data are gathered at 10Hz sampling. The following EEG values are gathered:

- low $\imath^p$ at 2.5-6.1 Hz
- delta δ^p at 1-4 Hz
- theta θ^p at 4-8 Hz

- alpha α^p at 7.5-13 Hz
- beta β^p at 13-30 Hz
- gamma γ^p at 30-44 Hz

The authors consider two main scenarios for QoE estimation from EEG. In the first, the existing EEG profile of the user is not considered to the final regression. In the second one, the individual EEG profile for each user is utilized as bias for regression between QoE results and EEG activity during the experiment.

In the first case, three individual approximations are performed: The first one estimates the final QoE directly from EEG activity. The fitting is applied with linear regression, as well as logistic regression with degrees up to three. The results are presented in the next figure as a linear prediction performance for each individual for various degradation degrees:

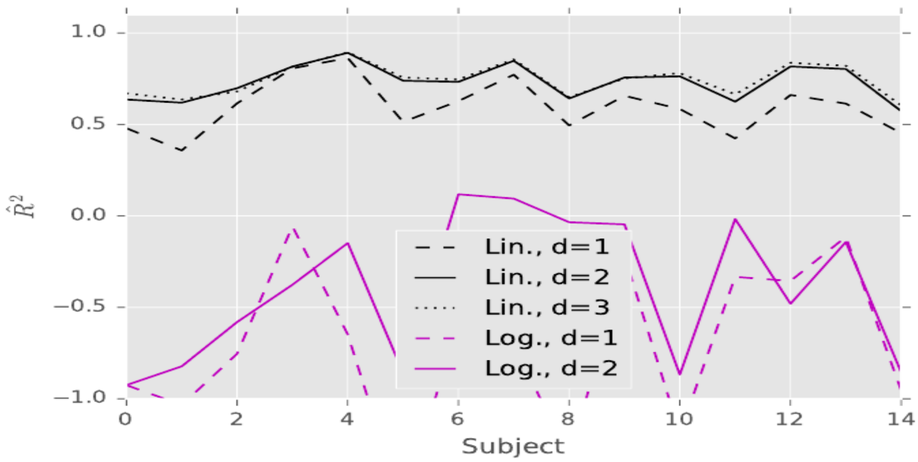


Figure 31 QoE estimation from EEG activity. Various fittings [28]

It is worth noting that linear regression of degree 2 fits better the experiments' data points. It also describes better the most individuals in the experiment. In addition, an obvious result is that logistic regression fails to capture the underlying relationship between the EEG activity and the QoE subjective results.

In order to cross validate the relationship between QoS and QoE, but also the relationship between BRISQUE metrics and QoE results (see Chapter 4.1.1), the users perform regression between the EEG activity and these two variables, namely the QoS parameters expressed as compression rate and the BRISQUE metrics. The general pattern is the same; linear regression expresses better the relationship between the EEG activity and the QoS parameters or the relationship between the BRISQUE metrics and the EEG activity. One notable difference is that logistic regression fits better these relationships than the EEG - QoE relationship.

In the second case, where everyone's EEG activity is known and used as a bias for the regression process, the results appear to be different. The initial values of the EEG activity can affect the regression the same way a real-world case scenario would. The results of the regression regarding the relationship between EEG activity and QoE are illustrated below:

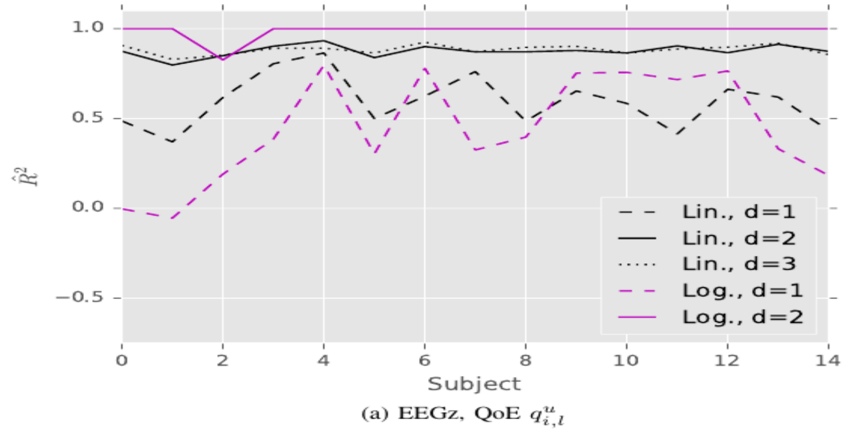


Figure 32 Regression between EEG and QoE with prior knowledge of users EEG signature [28]

The worth mentioning element here is that even though the *a posteriori* outcome has higher values, it is more variable compared to the results of the plain EEG regression. The variability narrows down for higher polynomial degrees. Another interesting result here is that the logistic regression results show that we can predict QoE more accurately with this method when we employ *a priori* knowledge about the EEG activity of an individual. In the case of logistic regression of degree 2 the result is almost perfect (if we exclude participant 2). This result is promising because the modern equipment allows for building the profile of users' EEG.

Equivalent results are observed in the relationship between QoS parameters and EEG and BRISQUE metrics and EEG. A common point with the previous results is the almost perfect prediction of the image objective quality parameters when using logistic regression of higher degree. The prediction results of the other regression methods are greatly improved as well.

The same results are observed in the relationship between BRISQUE metrics and EEG activity with prior knowledge of user profile. The trend of reduction of the variation according to which the BRISQUE metrics is correlated to EEG with the rise of degree of the polynomial.

In addition to the results about traditional image display for the general field of view, the users measure the correlation between the EEG footprint of a participant and the QoE when using spherical images in the head worn AR-viewer. As done in the previous settings, the authors distinct the prediction methods to the ones performed without prior knowledge of participants' EEG activity and the ones that take into account this activity.

In the first case, similar results to non-spherical images are observed. More specifically, the initial linear regression results are as expected highly variable and lower in prediction results than the corresponding results of the non-spherical image regression. Higher degree polynomials seem to improve the prediction rate and narrow the variability. Interestingly, an inversion in prediction rates is observed using higher degree in the logistic regression. As performed in the previous settings, the authors attempt to cross-validate the relationship between QoE and QoS and BRISQUE metrics (Chapter 4.1). The relationship between QoE and QoS, as expected, follows the general pattern of prediction rates. In general, the higher degrees pose a fair compromise between prediction accuracy and computational cost.

In the contrary, the fitting between BRISQUE metrics and EEG presents a reduction of the prediction rate compared to the non-immersive flat mode of the images. Also, a reversal effect in prediction rates is observed when a degree higher than 2 is used.

In addition to the solely EEG based inference of users' experience, the authors attempt a QoE estimation utilizing the knowledge for the user EEG signature. We can observe that the general behavior of the prediction rate is similar to the ones followed in the non-immersive mode. In general, the increase in the degrees results in direct increase for the prediction rates and the reduction of the variability. Also, in this case second degree logistic regression outperforms linear regression models.

Regarding the QoS prediction, the prior knowledge yields better results in all the regression methods. The logistic approach outperforms the other methods here. It is worth mentioning that these relationships are validating means towards the goal of estimation of the users QoE out of EEG activity. In that sense, a computational intense method like the logistic regression is not preferred in real time scenarios.

Finally, the BRISQUE metrics' predictions follow the trend of the previous metrics and is greatly improved by the knowledge of prior EEG activity of each participant.

Vovk et al in [34] investigate the sickness feelings that users develop during the usage of AR educational content and more importantly, what percentage of these feelings are due to the AR immersion. The authors evaluate the sickness feelings in three different use cases of AR:

- Aviation training use case
- Medical use case
- Astronaut training uses case

For the purposes of the experiment the authors use the Microsoft HoloLens headband. In order to estimate the sickness feelings in such diverse test subjects, the authors use two applications:

- a recording application which is supporting the experience capturing and the recreation of the AR feedback. The recording application provides activity guidance, for example how to perform an aircraft assembly procedure by directing attention to relevant parts, overlaying annotation to explain step-by-step what needs to be done.
- a player application, which engages the users in AR applications. This application immerses the participants in a training scenario, where they need to follow a sequence of recorded steps to perform a certain procedure, e.g. ultrasound diagnostics.

After the completion of the subjective tests, the users are asked to complete a SSQ with 16 questions – symptoms that are rated by the participants on a Likert-type scale. The SSQ is comprised of three subsections:

- Nausea (N), containing symptoms that are related to gastrointestinal distress
- Occulmotor (O), containing symptoms that relates to eyestrain, difficulty in focusing, blurred vision, and headache
- Disorientation (D), containing vestibular disturbances

The results of the experiment are displayed below:

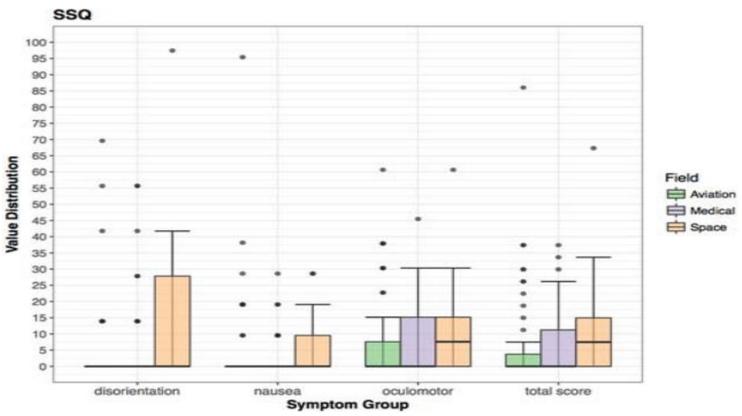


Figure 33 Sickness feelings for each use case [34]

The most frequent feeling was eyestrain followed by headache and general discomfort. The authors claim the effects of the context in each of these feelings. For example, in the aviation use case the pilots are constrained in a plane, so the body move is confined by the relatively small space. In the space trial, the participants are called to step on and off a small platform in order to perform a task. On the contrary, during the medical experiment the participants were sitting on a chair. Also, some of the participants had to wait 2-4 hours before experiencing the experiment.

4.3 Subjective tests

In order to be able to summarize the subjective tests used in all the works mentioned in Section 4, we created a comparative table. It is obvious that the authors trust the MOS the most, so this is the reason it is used in most of the works. Also, in two of the works SAM and SSQ are used to estimate emotional aspects of the QoE.

Table 10 Summary of the related works and their characteristics

Work	subjective metrics	Objective Metrics	ITU recommendation	Simulation / real time	Reference Measurements
Seeling [25]	MOS	Electroencephalographic frequency	P.910	Simulation	NR
Eoghan et al [31]	MOS SAM	Facial expressions	P.913 P.910	Simulation	NR
Longyu et al [27]	MOS	Fuzzy rules	Not mentioned	Real time	FR
Bauman and Seeling [28]	MOS	BRISQUE metrics Electroencephalographic activity	P.910	Simulation	NR

Alexiou et al [30]	MOS	Gaussian noise Octree-pruning compression	P.1401	Simulation	FR
Vovk et al [34]	SSQ	No		Simulation	FR

4.4 Correlation between QoS Metrics and QoE

In this chapter we will discuss the extent at which the QoS metrics of the subjective experiments correlate to the QoE results from the users' questionnaire. The general result is that multimedia and network QoS metrics reflect better the measured QoE. On the other hand, human body response metrics have statistically insignificant correlation with the recorded QoE and lead to contradictory conclusions. The correlation between the QoS metrics and QoE is depicted in the next table.

Table 11 OoE compared to objective metrics in AR

Work	QoE	objective metrics
Seeling [25]	Image quality plays a crucial role in QoE in AR. High image quality enhances the user experience, while low image quality degrades it.	EEG activity is directly related to the user satisfaction in high quality levels
Eoghan et al [31]	Users find the AR assistance annoying and distracting. Paper written instructions receive better MOS	AR assistance is steadier and people feel happier using it than paper-based assistance
Longyu et al [27]	Users rate AR games	Fuzzy inference QoE evaluation systems offer a close estimation of users' perceived experience.
Bauman and Seeling [28]	QoE is higher in higher quality AR images	BRISQUE metrics has a linear relationship to the users perceived quality. Also, EEG metrics are able to predict users perceived quality with low prediction error probability.

Alexiou et al [30]	QoE is affected from Gaussian noise and the Octree-pruning compression in a logarithmic manner	Point cloud quality metrics fail to capture the impact of content in the QoE
Vovk et al [34]	Users from different use cases feel oculomotor feelings in AR sessions	Contextual parameters play an important role in the sickness feelings of the users

The results of the works that refer to the perceived experience of users in AR applications are more consistent than the ones referring to VR applications.

One common conclusion among all works is that multimedia parameters play an important role in users' satisfaction when using AR applications. MOS remains high for high images in [25]. Also, network, multimedia and context parameters seem to be able to compensate for each other in [27], creating fuzzy rules that keep the user satisfaction high.

Another conclusion from these works is that spatial image quality is very important in the quality perceived by the users in AR applications [28], [30]. This is particularly important because of the way multimedia are compressed in order to be distributed efficiently nowadays. It is a challenge to the AR content creators to be able to compress efficiently the content without compromising the user's quality.

While new compression mechanisms would be useful for efficient distribution of AR applications, another conclusion of these works might help to maintain users' satisfaction in high levels. More specifically for same small levels of compression, MOS seems to remain stable in [28], [30]. Also, it is worth noting that these results were recorded while using a very interventional compression algorithm, the octree pruning.

While high quality multimedia is essential for maintaining users' satisfaction, the lack of it creates the opposite results. Participants prefer traditional opaque images in [25], [28]. This conclusion can be observed and in VR works (Chapter 3). This conclusion is very useful for AR applications, because they can easily alter their function from AR to plain reality when necessary (multimedia or network parameters degradation). This could be implemented as an adaptive mechanism, similar to HTTP adaptive streaming.

Another point that many of this chapters works seem to converge is that electroencephalographic activity can be related to the quality that the users perceive. High quality images in AR mode seem to directly stimulate the low gamma activity of the human brain something that is related to visual perception [25], [28].

In the contrary, lower facial expressions do not seem to be related to users perceived quality in [31]. This is particularly useful because AR applications are often intended to assist with tedious, repetitive tasks like the one in the experiment. Participants seem to find paper printed instructions more useful than the AR ones.

To sum up, in order to be able to provision the QoE of users in AR Applications, one should be able to monitor and gather data about:

- The application: resolution, frame rate, compression techniques

- The network: packet loss, bit rate, jitter
- The user: physiological metrics and feedback
- Streaming techniques that keep the application in specifications that satisfy the users and being able to adapt to different conditions.

4.5 Validation and proposition

The results of the works reveal that AR applications have to be treated differently from general purpose applications and certainly it is a field that has open challenges for future works.

In these applications a plethora of data should be gathered in order to monitor the user quality and to be able to manage it. A fuzzy inference rules entity should be in the epicenter of quality provisioning system. This entity could take as input data about all the key factors referred in the previous sections. Through these, this entity should be able to decide about the quality levels currently perceived by the users.

Another important entity that should be added to application model is an adaptation entity, with the ability to sense all the KPIs involved in user experience. This is particularly important in AR applications because in lower quality KPIs the AR could become plain reality application (no augmentation at all) and according to the previous section this could maintain the user satisfaction for a period of time until the application provider is able to provide again higher quality services.

These two entities should get their inputs from user side and provider side management entities. These entities should cooperate in order to gather the required information regarding all the KPIs involved.

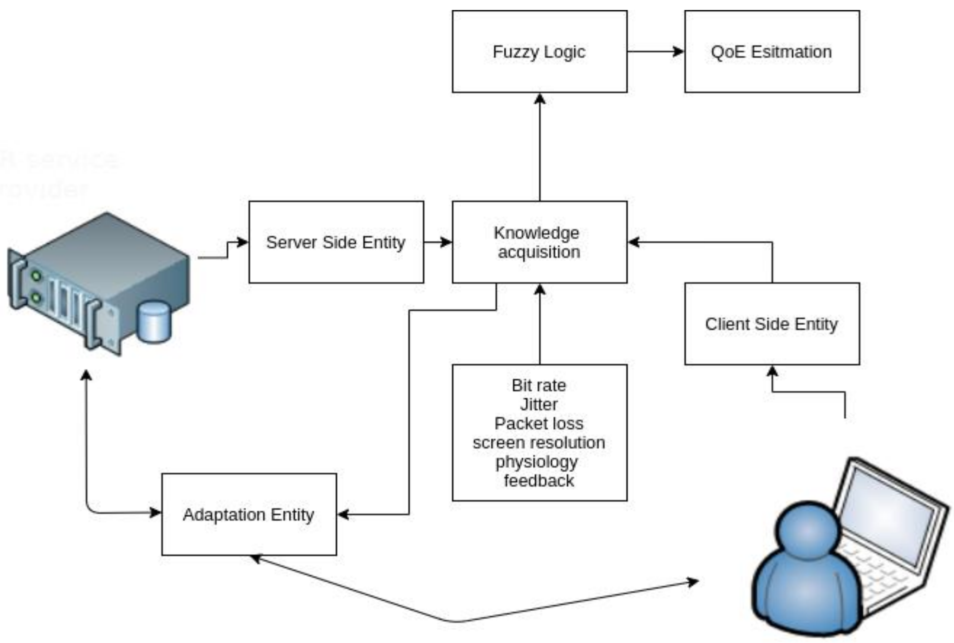


Figure 34 A high level description of a QoE provisioning system for AR

5 QUALITY OF EXPERIENCE IN MOBILE NETWORKS FOR VR/AR APPLICATIONS

5.1 Introduction

As we saw in the previous chapters, QoE provisioning and awareness is a complex issue that often needs, complex calculations, reasonable trade-offs and involves a plethora of parameters that often conflict with one another.

In mobile networks, this sensitive but highly important aspect of modern-day communications is subject to many other unpredictable factors such as multiplexing type, handoff algorithms, users' mobility and many more.

The challenge of QoE provisioning for mobile users becomes even bigger for applications such as AR applications and VR applications. These types of applications are very often resource intensive, have low latency requirements and are responsible for the greatest amount of IP traffic in modern networks.

Furthermore, release 16 of 3GPP describes a new – built-in entity in 5G that aims to unify and standardize the way in which various parts of the network cooperate in order to provide the best possible experience for the users. This standardization shows that the new era mobile networks should provide a standard way of calculating the users experience and to be aware of it.

The works mentioned in this chapter are concentrating on different aspects of this procedure and optimize various parts of the mobile networks. The common ground among all the referred works is that a QoE management system is proposed by all of them. This shows the importance of cooperation and coordination of different parts of the network.

The rest of the chapter is divided in 4 sections: Chapter 5.2 refers to related works and tries to explore different methods of QoE provisioning in mobile networks. Chapter 5.3 discusses the common ground in the works of the chapter 5.2 and tries to elaborate about the most important parts of these works. Chapter 5.4 concludes, by providing a high level description of a proposed QoE provisioning framework that is based on the works of this section.

5.2 Related Works

5.2.1 QoE in LTE and anticipated QoE in 5G

The realization of QoE through the LTE period is of great concern for many research works. Although many works describe the enhanced experience in terms of technical attributes, it is only natural for contemporary researchers to focus on the human side of the service, especially in the era of 5G and Software Defined Networking (SDN). M. Suryanegara in [35] uses a questionnaire to estimate the QoE over the advancement of similar mobile technologies in the past (2G,3G,4G). After the completion of the questionnaire, the author aggregates the responses into 6 groups that summarize the public opinion about the transition from 2G to 4G. These groups are namely:

- Same – Same: People that state that the QoE has not been improved in the past years and do not expect it to improve with the advancement of 5G.

- Better – Same: People who believe that 4G is better that 3G, but do not expect 5G to be better.
- Worse – Better: This group perceives 4G as worse than the previous technologies but trusts that 5G will be better.
- Same – Better: These users feel that 4G has not significantly improved their experience over the previous technologies but expect 5G to be better.
- Better – Better: 4G feels better for these users and expect the same level of enhancement from 5G.
- Users that expect 5G to be worse.

The two main concerns of participants are infrastructure support and the actual service and the way this service impacts their life. These two service quality concerns can be analyzed to more specific domains such as service coverage, security and privacy, accessibility, impact on society and economy, etc. The author uses these norms to estimate the anticipated amelioration from the 5G advent and translates these anticipated advancements into technological advancements. These aspects are presented in the next figure:

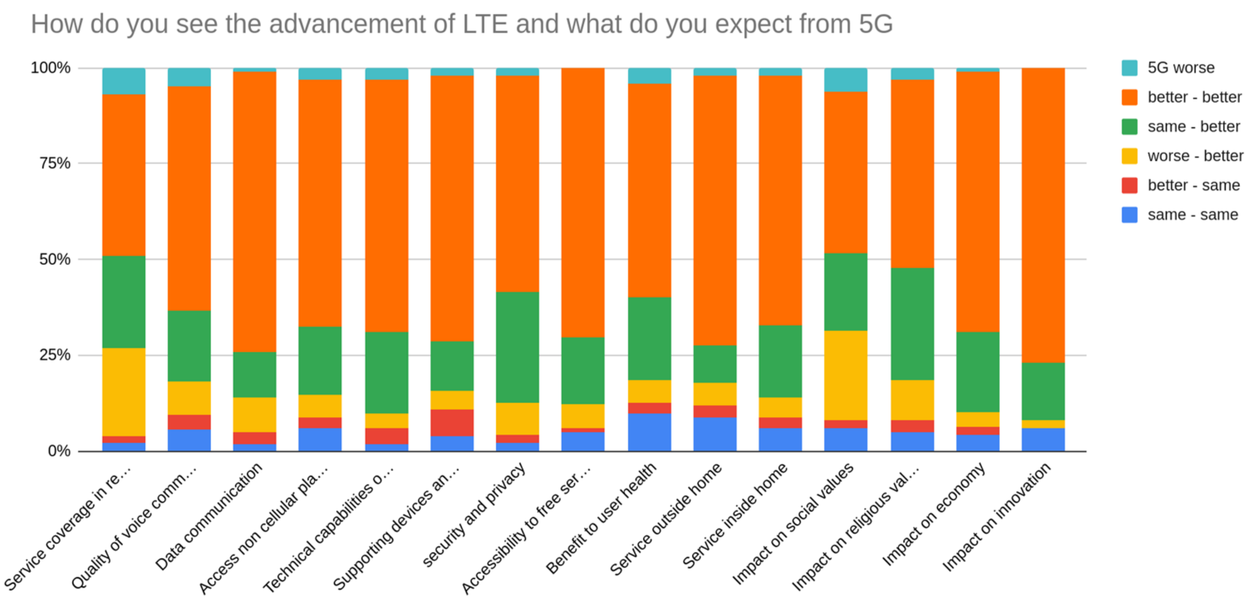


Figure 35 MOS about LTE advancements and expectations from 5G

A similar set of quality requirements and expectations are presented in [36] where Liotou et al. list the built-in aspects needed for QoE assurance, and the corresponding technological advancements that can be used to build these aspects into the 5G network. According to the authors these aspects concern:

- Consistency. According to the authors this refers to the seamless operation of the service with limited fluctuations.
- Transparency. The user should not realize the efforts and enhancements of the underlying network to provide the best possible experience, nor should be able to intervene or provide direct input.
- User personalization and service differentiation.

- Resource and energy efficient QoE awareness.

It is obvious that all the resources available for this topic agree that Consistency is of great importance. Users should feel a continuous and uninterrupted service that should span to various devices, ecosystems, access protocols and services. This is why service coverage, data communication, access on non-cellular platforms and other service specific questions are answered with a positive vibe in [35] and users expect further improvement in 5G.

The efforts and mechanisms that the underlying network is utilizing to support better user experience are observable in all the service – specific answers in [35]. This is particularly useful in areas like service coverage and quality of voice communications, but also in service outside home. Users feel that the network has been able to support better experience without really realizing the mechanisms under which this enhancement is possible, and without even making decisions about it (one typical example is cell handoff or adaptive streaming). These continuous improvements have created mostly positive MOS and high expectations about 5G.

By a more macroscopic view, we can see that the dominant trend in all service specific questions is that the improvements towards 4G have improved the QoE that users perceive. The even more interesting point though, is that the aggregation of clusters that expect to be better is dominant in every question. This realization creates high expectations about the advent of 5G and pushes the research for built-in QoE in 5G.

Although a minority, disappointed users should be treated with specialized solutions as their perception of improvement may be different from the average. This is why authors of [36] agree that QoE should be evaluated in personal level and specialized innovation with variations should be provided appropriately in order to further marginalize the disappointed users.

5.2.2 QoE estimation based on objective parameters

Petrangeli et al in [37],[38] propose an architecture for HTTP based adaptive streaming that aims to enhance the users experience on VR and AR applications for mobile users. More precisely the architecture proposes the following:

- On the contrary to the traditional HTTP/1.1 tiling where the client needs to request for every particular tile specifically, the authors propose an HTTP 2 based server push procedure. Furthermore, the client requests are viewport – specific and further reduces the volume of requested data.
- A future viewport prediction algorithm in order to minimize the unnecessary data transactions.

The evaluation of the proposed architecture involves the human factor because the movement and reactions of individuals impacts the viewport and consecutively the requested content. For this purpose, 10 individuals are asked to participate in the evaluation in order to generate the corresponding tiling traces. Furthermore, the authors differentiate between “real” 4G bit rates and fixed low bit rate at 5 Mbps in order to demonstrate the effectiveness of the proposed architecture. For every case of the above and given the viewport traces recorded by the users the authors measure the time spent on the highest quality for the following settings:

- HTTP 1.1
- HTTP 2 without the proposed architecture
- HTTP 2 using the proposed architecture
- No tiling at all

In order to estimate the effects of the proposed method in more human terms, the authors are feeding the aforementioned (and more) QoS metrics to a machine learning model that projects these QoS metrics to human perceived quality.

From the results it is clear that viewport prediction and server push has clearly benefited the time that a user spends in higher qualities. The result is further improved in higher duration transmission, making this solution ideal for AR/VR applications. Also, the time spent in the better quality is higher in low bit rate settings.

In [39] the authors propose a solution for the estimation of the QoE of 5G network end users for streaming high resolution video used for AR / VR purposes over millimeter wave. The 5G millimeter wave network should be used to obtain the higher bandwidth and bit rate gains. More specifically, the authors examine the effect of QoS factors in the users feeling of delight or annoyance.

Since the use of millimeter wave is not publicly accessible, the authors use the simulation software NS3 to estimate the QoS parameters over the millimeter wave transmission and feed these values into a machine learning model to estimate the people’s perception.

The QoS parameters that are calculated as input to the model are the following:

- The Peak-to-signal-noise ratio (PSNR)
- Jitter
- Packet loss
- Delay
- User profile

The authors are using the fuzzy ARTMAP (FAM) algorithm in order to classify the QoS input and understand about the user delight.

M. Jalil Piran et al in [40] discuss a new channel allocation scheme for 5G that adopts channel allocation according to users QoE expectations, minimizes the latency, provides seamless multimedia service, improving QoE for resource intensive applications like AR and VR applications.

More specifically, the proposed framework:

- Tries to predict the arrival of licensed users (LU) and to make the best usage of the channel. For this purpose, the framework retains an index-based scheme in order to collect channel quality information. This scheme is associated with QoE expectations in order to be assigned for appropriate usage upon LU arrival.
- Makes channel reservation according to the previous scheme and assigns channels with a priority-based logic.
- Conducts channel estimation in the time domain using hidden Markovian models (HMM).
- Splits the video transmission in two parts: the base layer (BL) and the enhancement layer (EL). This way, Cognitive users (CU) that allocate large portion

of channel when it is idle, do not experience a violent handoff event because they receive at least the base layer of the transmission while searching for additional resources upon the LU arrival. This split in the two aforementioned layers is achieved using the SVC codec.

The authors evaluate the performance of the proposed framework by calculating the PSNR, average number of collisions and MOS. The MOS is derived from QoS parameters. The proposed framework is contrasted to similar frameworks and dynamic allocation strategies against arrival rate of LUs. The framework is also evaluated for its various traffic classes and the impact of LU arrival rate for each of the traffic classes is taken under consideration.

According to PSNR, the authors state that even with high LU arrival rate, the video reconstruction is improved comparing to similar technologies. This fact has to do with the better adaptation of cognitive users to better portion of the resources.

A very interesting measurement is the impact of varying LU arrival rate to every quality class that the framework forms. Apart from the fact that the proposed scheme has improved MOS compared to other schemes, it is obvious from the measurements that the classes that are used for interactive applications, hence are sensitive to delay and jitter are almost not affected by the high arrival rate of LU. This is particularly important for AR and VR applications. Also, all classes show mild decrease after an arrival threshold, which proves that the handoffs are performed more mildly and seamlessly for the users.

The problem of resource allocation in the physical layer proves to be of critical importance for VR and AR applications in 5G networks. In [41] M. Chen et al examine the problem of resource allocation in an unmanned aerial vehicle (UAV) – enabled LTE – U network for users that communicate in VR context. The enhanced version of resource allocation that the authors propose is evaluated by using MOS and KPI metrics.

On the contrary to the robust way in which multiplexing is achieved in wired communications, in wireless networks and especially in 5G, resource allocation is a great point of arguments among researchers. The authors' contribution in this argument consists of the following points:

- Resource blocks are allocated from licensed and unlicensed areas, respectively. The quality of the content is adjusted according to the extent of this allocation. The ultimate goal behind this is the maximization of the QoE for the users.
- The resource allocation problem is modeled as an echo state network with leaky integrator neurons.

The setup is evaluated using KPI measurements and MOS. The results clearly show that:

- The number of UAV plays a crucial role in the KPI and especially in average delay, a very important metric for VR/AR applications.
- The proposed allocation scheme performs better in mean delay measurements compared to other similar resource allocation schemes for the same UAV density. It is remarkable that the proposed scheme is more resilient to the lack of UAVs, in other words, the lack of UAV's affects the delay less than other resource allocation schemes.
- Another interesting factor is that the proposed scheme is intelligent enough to be able to realize when the delay requirements are met, hence it should allocate resources towards other directions in the goal of enhancing users' experience.

- The MOS is directly dependent to the density of UAV's and is increased in a linear manner in all allocation schemes as the number of UAVs increases. This is caused by the fact that the denser the UAVs, the more resources are able to allocate for their perspective users.

Chang Ge et al in [42] examine the QoE that a 5G Satellite backbone should provide in order to support the existing VR and AR applications. Three KPI are measured, that have proven to impact QoE in more conventional network settings:

- Initial startup delay
- buffering
- live stream latency

5.2.3 QoE estimation using machine learning

Several research attempts have focused on the enhancement of estimation of users' experience by using machine learning algorithms that learn through experimental data. These algorithms often have the ability to modify technical aspects of wireless communications such as resource allocation, hand-off strategy, content quality adaptation and many more.

In [43] Zhou et al propose an enhanced hand-off algorithm (Comp-HO) in order for 5G to be able to utilize the advantages of Multi-access Edge Computing (MEC) for AR and VR applications. The authors estimate the improvement in user experience by using the NS-3 simulator and projecting the results to user experience using an existing model.

More specifically, this algorithm performs the following operations to ensure quality:

- Each User Equipment (UE) reports back to its serving base station the signal quality measurements of all nearby base stations (RSRQ), as well as information about MEC applications that are currently running.
- Each base station collects this information in parallel from the UE's connected to them as well as load information about MEC servers nearby.
- Based on these inputs, the base station starts the handoff procedure if its RSRQ fails to meet the threshold required for this kind of application.

The algorithm is tested in a simulation environment (NS3) against other algorithms for a lot of different settings. These settings concern:

- Handoff rate. This is achieved through UE's speed variation
- FPS for the requested services
- Mobility

The main performance metrics measured by the authors are median absolute deviation (MAD) of delay, jitter and packet loss percentage. Furthermore, these metrics are used as input in an existing AR task impairment model [44] to estimate the final impact of these parameters to the end users' degree of delight or annoyance. The measured results are projected to their corresponding probability density, thus fitting a probability density function from the results.

From the QoS perspective, the proposed algorithm improves delay, especially in the "edge" cases where the improvement has more impact to the users. The same is valid for jitter, where the probability density function is smooth without large deviation areas. This fact shows that users (and application designers) can expect a zone of delay and jitter which will be met in real time usage with extremely high probability. The results are even more impressive when the current algorithm competes with the other ones with regards to the UE speed. In general Comp – HO outperforms other algorithms by 70% – 80 % in user delay and different UE speeds, while the improvement is even bigger in different frame rates.

The trade-off to these significant improvements in delay and jitter is the algorithm decreased performance on Signal-to-interference-plus-noise ratio and transmission delay and jitter. The authors attribute this reduced performance on the algorithm attempt to optimize the user delay and jitter.

The impact of the aforementioned measured facts to the users' experience is estimated using a known impairment model [44]. The authors compare the impairment scores' distribution for the algorithm and compare to impairment scores for other algorithms. An impairment score is a normalized score that quantifies the reduction in user experience in a given task, i.e., here, an AR task. The authors claim that the impairment scores' distribution measure shows that the Comp – HO algorithm reduces the fraction of packets with full impairment (thus the lowest user experience) which should translate to actual QoE gains during these types of AR tasks.

A fixed set of KPIs is always guaranteed to be able to describe the level of delight or annoyance of users. On the contrary to other research attempts that estimate QoE through machine learning models that take as input a fixed and immutable set of parameters and do not change these parameters over time, Schwarzmann et al in [45] propose and test a machine learning model that uses and constantly reevaluates a plethora of KPI indications from the network, user equipment, application, edge servers and other parts of user experience.

The authors utilize to the maximum the new analytics entity described in release 16 of 3GPP, the NWDAF. According to this release, the analytics entity:

- Should be connected to Service Based Interface (SBI), in order to be able to collect data from Application Functions (AFs) and from other network 5G control plane Network Functions.
- Should be able to collect data from the 5G management plane.
- Should be capable of generating analytics based on machine learning models. These analytics should be standardized in order to be consumable by Network Functions, Application Functions and other network entities.
- Is implemented internally in vendor specific manner.

According to this, the authors propose a procedure that aims to reveal the most critical parameters for QoE using the analytics and the processes derived from NWDAF. More specifically, the proposed architecture consists of three phases:

- In the first phase, third party Application Functions reveal information about the users' QoE. During this phase, the NWDAF database is enriched with true users' QoE data.
- In the second phase, the NWDAF entity, which is already monitoring the network, generates a vast number of network features and their corresponding significance to the users' QoE. Multiple sets of these features, in combination with the QoE feedback realized in the previous step are used to train machine learning models for QoE estimation. This process is repeated for multiple sets of parameters and machine learning models until a desired level of accuracy is reached.

- Once identified, the features set that better describe QoE is communicated in other parts of the network. This process is repeated and updated in order to enhance the users' experience.

The measurements and MOS is obtained from OMNeT ++ in combination with INET and SimuLTE frameworks. The authors use a simple network topology of a single eNB and a plethora of mobile users around it scattered in an area of 500 X 500 meters around it. In this setting, the authors vary the vast majority of network features and apply the aforementioned algorithm to extract the most valuable features set. The results of their efforts are represented with a composed cumulative distribution function fitting the users MOS during the framework application.

The personalization of user experience and the proper adoption of network and application specifications to that direction is of great concern in other research efforts as well. In [46] Y. Wang et al propose a QoE management system that uses machine learning model to capture the underlying relationship between the user, and the various users' states (e.g. walking, indoors) and preferences (both content preferences and network preferences).

The proposed architecture consists of two parts, an online and an offline one. The offline part has mainly data mining responsibilities and prepares the data for real time management. This module has a detail level per user and per service, in other words collects data for every user and for every specific service. The main data that the offline part collects are:

- QoS monitoring data such as Device, Infrastructure and Network specific data
- Context monitoring data such as Location and Mobility
- Experience monitoring data such as Survey and feedback from users

After data collection, the offline part has the responsibility of pre-processing and cleaning these as well as to store them in the appropriate databases. In the next step, these data are used to train machine learning models in order to be able to understand user preferences when used in real time.

The online part uses the previously acquired data and models derived from them to enhance user experience in real time. The online model consists of three parts:

- Real Time Data Collector: This entity gathers real time data about users' identification, the kind of service currently in use and the network status.
- Preference Prediction Component: This entity utilizes the offline data and models gathered in the offline step and uses them alongside with real time data from the Real Time Data Collector in order to further personalize the provided services.
- QoE Management Component: This entity maintains a model that eventually maps all the factors available to users' delight or annoyance.

The authors evaluate the proposed framework with a subjective test. More specifically, the authors gather all these sorts of data mentioned in the previous paragraphs including offline data (such as age, gender, occupation) and online data (network specs, content quality, contextual parameters). In order to validate the proposed framework, the authors use a two-step QoE modeling, which depends not only in network parameters, but also on user preferences. In the first step, the user preferences are modeled, while the second step is using this model of user preferences combined with network parameters. The problem in its final dimensions is modeled as a three layered Bayesian graphic and uses the Monte Carlo expectation-maximization algorithm to train this model. After the model training the derived users QoE is calculated using the sigmoid function:

$$QoE = \frac{\theta}{1 + e^{(-aS + \beta r_{ij} + \gamma)}}$$

where α, β, γ, θ are parameters constraining the quantization of QoE.

Simulation results of the architecture shows improved performance of the systems with regards to the users QoE, when compared to other resource allocation systems for similar settings. The authors claim that QoE can be improved by 20 percent, while 96 percent of the participants report a better experience.

5.2.4 Proposed QoE management systems

The researchers attempt to enhance user experience in mobile networks shows that the subject of users perceived quality is very important to the cellular mobile networks and in fact is described as a built-in attribute in current and future releases of 3GPP. This fact leads to the necessity of managing this very complex and multifactorial aspect of modern mobile networks under organized and standardized network entities.

In [9] Liotou et al suggest a QoE management system for heterogeneous networks. The management system is simulated using software simulators. The authors discuss the tradeoffs and overhead included in this management system and the necessity for such systems to be built-in the LTE networks.

According to the authors, a QoE management system should contain the next entities:

- The QoE Controller: It is the interface between the actual network and the higher level QoE management entities. The controller acquires data by selecting the appropriate resources depending on the model used by the system. The controller also controls the frequency of the parameters sampling, therefore the frequency of QoE assessment. It also has the responsibility of modifying the network parameters in the direction imposed by the management system.
- The QoE monitor: It is responsible for monitoring the QoE in real time with granularity per flow. QoE monitor uses known statistical models to calculate the QoE per traffic flow and reports back to the QoE manager the results.
- The QoE manager: Takes input from the QoE controller for the current network state, from the QoE monitor estimates about the current QoE scores and operator specific information such as Service Level Agreements (SLA) or network policies and takes actions towards the improvement of users QoE.

The authors evaluate the proposed framework with the LTE - sim software [47] that was extended with the QoE management system. The simulation environment is a heterogeneous network consisting of different kinds of cells (macro-cells, femto-cells) and uniformly distributed user equipment (UE).

The overhead implied to the network for the implementation of the management system is of great concern for the authors. This is calculated by varying QoE estimation periods from QoE controller during different simulation sessions. The base for the QoE calculation accuracy is the QoE calculated by the system with a period of 0.1 seconds. The visual representation is depicted in the following diagram.

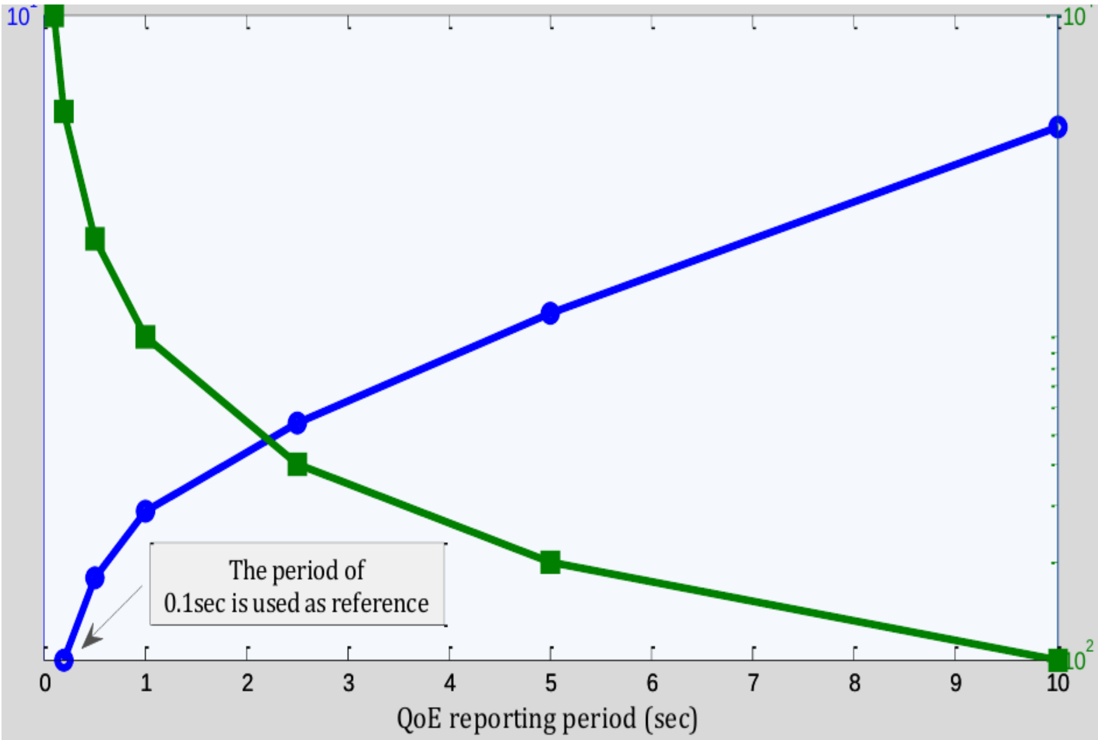


Figure 36 Trade-off between the network overhead and achieved accuracy in the QoE prediction [9]

Through the simulation is proved that the QoE management system is fully aware of the network and QoE expectations, by being able to predict when QoE is close to a critical threshold, therefore it should allocate resources from less congested parts of the network in order to maintain the QoE levels.

During 2021 the video traffic accounted for 82% of the total IP traffic. With this concern Nightindale et al in [48] propose a built-in QoE awareness system for 5G, that has the accuracy and the low complexity needed to provide real time indication for UHD flows required by AR/VR applications. The authors base their approach on two very important technological advancements. The first one is the scalable H.265 encoding and the second one is the development of a scalable and robust real time QoE prediction model.

The proposed model contains the following contributions:

- A 5G QoE framework. This entity is responsible for monitoring the whole lifecycle of UHD flows, as well as the aggregation and concentration of this information around usable data points.
- A low-complexity QoE estimator and future projector. This entity is especially important because it allows for real time and limited resources estimation and future projection of QoE.
- A 5G aware QoE system that is capable of extracting UHD video metadata and flows QoS.
- A UHD capable, scalable H.265 system which is QoE aware.

In order to evaluate the framework and the accuracy of QoE estimation, the authors perform a subjective test in which the participants watch a series of UHD videos encoded in scalable H.265 encoding, both in the original form and in streaming version. Later levels of impairment are added to these samples in order to evaluate the credibility of the QoE estimation system by comparing the predicted QoE and the actual QoE obtained by the participants.

A very important aspect of the framework is the first one, namely the ability to gather and aggregate QoS parameters about UHD flows. The authors suggest that a magnitude that summarizes these aspects is the video flow Congestion Index (CI). This metric summarizes the maximum level of congestion across all the network interfaces a flow traverses.

After obtaining the Congestion Index per flow from the first step, the authors utilize the MOS in order to perform regression and to be able to correlate CI to the actual QoE. The regression results show that these two magnitudes have the following relationship:

$$QoE = -0.892 + \frac{5.082}{\sqrt{CI}}$$

The results of the prediction method are cross validated with the rest of the participants MOS. The predicted QoE seems to vary from the actual MOS by 0.9 in a 5 steps Likert-like scale.

5.3 Discussion

It is common ground among the researchers that study the users' experience in mobile networks that 5G should be a technological advancement that places the users in the center. This realization is even stronger with VR and AR applications and ultra-high-definition multimedia content that require more and smarter distributed resources in their mission for immersing users from the reality.

The most common and most challenging problem that all researchers are facing is the problem of resource allocation, especially in the highly congested and constantly altered medium of mobile communications. In [41] leaky integrator neurons are used in order to manage this problem in UAV enabled 5G communication for AR purposes. Also, in [9] this problem is handled with the aim of network analytics tools and QoE prediction models which are able to tell whether QoE falls below a threshold and can trigger the decision to alter the current resource allocation. A similar method is used by [48], in combination with a scalable compression algorithm. In [43], a smart hand off algorithm is used for that matter.

Also, another common point of concern in the research for QoE in 5G is the usage of meta information about the network, how this knowledge can lead to meaningful conclusions about the users QoE and how can one be aware that users enjoy high quality of services most of the time. In [9] this is tackled by a QoE management system that uses machine learning in order to assess the current quality that the users enjoy, through network metrics. In [48], a similar approach is taken but with a simpler assessment method, namely the Congestion Index which is proved to be a reliable indication about the users' delight of annoyance for high-definition video transmission.

The problem of mobility and handoff strategy is another common point in discussion about the users' QoE in mobile networks. In [40], the problem of cognitive users is examined and how the quality of their experience can be affected by licensed users that share the same cell. Machine learning is once again used in this case in order to predict this arrival and be ready to handoff users effectively with minimum impact on multimedia experience.

In [43], the authors are using a smart handoff algorithm that takes into account the nearby edge servers and their capabilities that they provide towards the achievement of QoE goals.

Finally, all the research efforts propose a QoE management system that combines analytics from the network, users and many other contextual parameters with a machine learning model that is capable of exploiting these data towards the prediction of users' QoE. The results of these models are usually managed by another entity that has the authority to intervene and alter network settings in a smaller or larger extent, thus being capable of constantly improving users' experience.

5.4 Validation and proposition

The current research state of art reveals that the task of maintenance of users' immersion in VR and AR applications in mobile networks is a challenging task and lots of ground has to be covered in order to be able to provide seamless quality to the end users. This difficult task should be delegated to a built-in network entity that has the capabilities and capacity of accomplishing such tasks with the minimum possible overhead. This entity should consist of three distinct sub-entities. These sub-entities are the QoE controller, the QoE monitor and the QoE manager.

The QoE Controller is the data collector of the framework and provides the analytical and persistent attributes that the framework needs. This entity has the task of collecting data from multiple sources, very often unstructured and "dirty" and has to structure and clean it, as well as store it in appropriate structures for the other parts to use. Such data may include contextual data such as the location of the users, the mobility of the users, the device type and more. It also may include user specific data such as bias about the content. Furthermore, it should gather data about network KPIs and application KPIs such as delay, jitter, frame rate, edge servers state and many more.

The QoE monitor is the mind of the framework. It utilizes the data gathered in the previous step and uses or trains QoE models, in other words models that can predict the users' delight or annoyance under the specific conditions. Furthermore, the monitor can project QoE in the future and assist in decisions regarding future admission in the network.

The decision-making part of the framework is the QoE manager that is constantly receiving feedback from QoE controller and QoE monitor and is constantly aware of QoE levels. The QoE manager can trigger decisions that can maintain critical levels of QoE or can maximize it. These decisions may concern the content itself: frame rate adaptation, compression adaptation, prefetching items from the edge servers and many more. Also, it can control the way in which new User Equipment can enter the network. Such decisions may concern the handoff strategy, the allocation strategy and the spectrum allocation strategy.

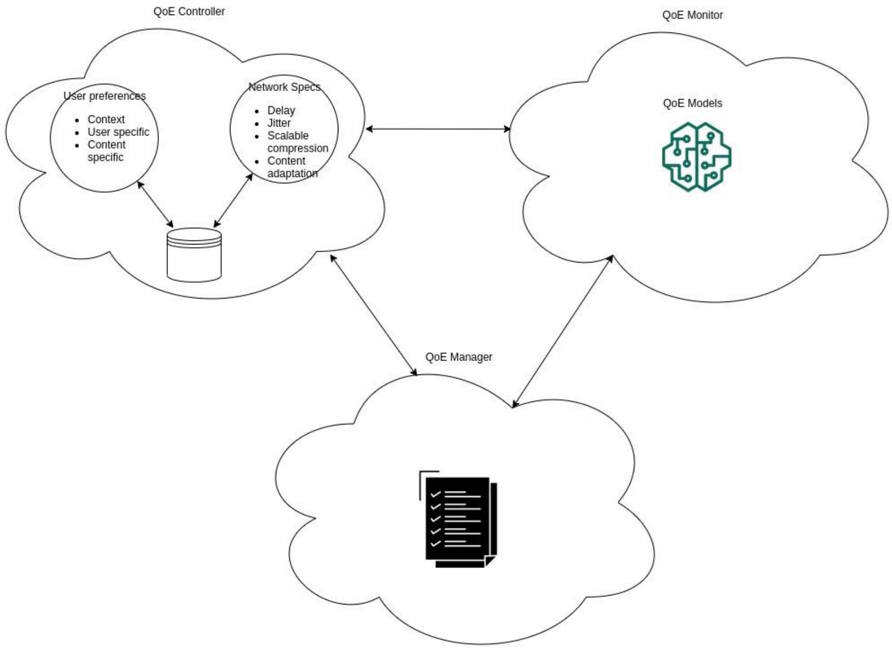


Figure 37 A high level description of a QoE provisioning system for VR/AR applications in mobile networks

6 CONCLUSION

This thesis has conducted a survey of recent works in the area of QoE measurement in AR and VR environments. It reveals the depth of this subject and the important milestones that must be made in order to achieve the goal of seamless QoE provisioning.

The first and most important conclusion of this thesis is that QoE is the common language between a modern-day application or service provider and its customers. The reason behind this statement is the fact that QoE is a multifactorial metric that overcomes the strict technical, or business way of understanding things. On the contrary, QoE provides a way of understanding the human aspect of technology consumers.

The most credible way of understanding this valuable information is well known from the ancient years. This method is to directly ask a person with the appropriate questionnaire. Subjective tests are the main concern of this study and prove to be irreplaceable for understanding peoples' feelings. Unfortunately, this method is expensive to implement, it is prone to contextual parameters and does not scale well.

People are imperfect and diverse by nature and so are their answers on the same questions. In chapter 3 we saw contradictory results between different metrics in the same study. Electroencephalographic data does not follow MOS. Users prefer a high-resolution VR video from a conventional video of the same resolution. People would rather consume a VR video even though most of them have motion sickness feelings.

On the contrary, users of AR seem to be more concise about what they like and what annoys them. AR applications are less immersive than VR ones because AR combines multimedia with the real world. One very common conclusion among the researchers is that this co-existence of the real and the digital world should remain uninterrupted. In this direction, network parameters seem to play a more important role in retaining users' satisfaction. Also, network, multimedia and contextual parameters seem to be able to compensate for each other.

A common disadvantage of both VR and AR applications is that they are resource intensive, and the lack of these resources can lead to worse user experience compared to non-VR or AR equivalent. People seem to prefer conventional equivalents whenever the AR lacks the necessary resources. This is a double edge sword for application designers because they need to be able to constantly provide such resources so that the users' experience can remain above a critical threshold. This corollary can also help application designers and network providers to come up with adaptive solutions, much like adaptive streaming which has the capabilities of sensing the environment and adapt their specifications to maximize the users' satisfaction, or at least reduce the users' annoyance.

In mobile networks, the researchers and network operators try to cover this need for resources with smart resource allocation mechanisms, advanced time-portions sharing strategies and improved handoff algorithms. A common ground among all the studies on the subject is that a QoE management system is necessary for addressing these issues and constantly provide the best possible experience for users. Such management systems play a dominant role in the network and have advanced sensing and modeling capabilities.

QoE management systems are a common research proposal among studies that deal with QoE in VR and AR applications too. Most of these systems have three discrete parts:

One that collects the necessary data from various sources, including users' data, environmental data and contextual data. These entities are usually central network entities, and their usage impose a level of overhead in the network data. The design and efficiency of these systems is a research challenge, especially given the heterogeneity of modern-day networks.

Also, the raw data of a QoE management entity usually are dispatched to an analytics entity that utilizes these data to extract intelligence and create meaningful aggregations. Many research works propose the usage of machine learning in these entities, to extract conclusions out of the data. The development and optimization of appropriate machine learning models for these entities is a research challenge for future works, especially where same parts of the network serve different purposes and network operators need to be very versatile.

These network operations should be part of new advanced operations that are performed by software parts of the network. This trend, namely the SDN, is already implemented in modern-day networks and such operations are a logical extension of these functions.

The challenge of increasing network utilization and optimization with the same resources is always a trending topic in network development. This challenge becomes even more relevant in the advent of VR and AR applications. Topics such as new resource allocation algorithms, advanced handoff strategies and mobility management need further exploration and optimization.

ABBREVIATIONS

QoE	Quality of experience
TI	Tele - immersive
LAN	Local Area Network
RTT	Round Trip Time
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VR	Virtual Reality
AR	Augmented Reality
EED	Electro Dermal Activity
MOS	Mean Opinion Score
SSQ	Simulator Sickness Questionnaire
EEG	Electroencephalography
HMD	Head mounted display
SAM	self-assessment manikin
MSE	mean square error
MAE	mean absolute error
MedAE	median absolute error
p2point	Point-to-point
p2plane	Point-to-plane
p2mesh	Point-to-mesh
KPI	key performance indicators
MEC	Multi-access Edge Computing
PSNR	Peak-to-Signal-Noise Ratio

LU	Licensed Users
CU	Cognitive-enabled Users
HMM	Hidden Markovian Model
UE	User Equipment
MAD	median absolute deviation
SINR	Signal-to-interference-plus-noise ratio
AF	Application Functions
NF	Network Functions
SBI	Service Based Interface
SLA	Service Level Agreements
CI	Congestion Index

REFERENCES

[1] K. Brunnström *et al.*, *Qualinet White Paper on Definitions of Quality of Experience*. 2013.

[2] "Deterministic Modeling for Evaluating Consumers' Attitude towards Telecommunication Service in Bangladesh," *Bus. Ethics Leadersh.*, vol. 3, pp. 72–82, Jun. 2019, doi: 10.21272/bel.3(2).72-82.2019.

[3] S. Jumisko-Pyykkö and T. Vainio, "Framing the Context of Use for Mobile HCI," *IJMHCI*, vol. 2, pp. 1–28, Oct. 2010, doi: 10.4018/jmhci.2010100101.

[4] "ITU-T: Recommendation E.800 (2008)," 2008.

[5] A. Sackl, P. Zwickl, S. Egger, and P. Reichl, "The role of cognitive dissonance for QoE evaluation of multimedia services," in *2012 IEEE Globecom Workshops*, Dec. 2012, pp. 1352–1356, doi: 10.1109/GLOCOMW.2012.6477779.

[6] khorsandroo, sajad, Md. Noor, Rafidah, and khorsandroo, sayid, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Netw.*, vol. 24, no. 2, pp. 36–41, Apr. 2010, doi: 10.1109/MNET.2010.5430142.

[7] "G.1030 : Estimating end-to-end performance in IP networks for data applications," 2005.

[8] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "A Survey on Parametric QoE Estimation for Popular Services," *J. Netw. Comput. Appl.*, vol. 77, Jan. 2017, doi: 10.1016/j.jnca.2016.10.016.

[9] E. Liotou, D. Tsolkas, N. Passas, and L. Merakos, "Quality of Experience Management in Mobile Cellular Networks: Key Issues and Design Challenges," *IEEE Commun. Mag.*, vol. 53, Jul. 2015, doi: 10.1109/MCOM.2015.7158278.

[10] I. Politis, A. Lykourgiotis, and T. Dagiuklas, "A Framework for QoE-Aware 3D Video Streaming Optimisation over Wireless Networks," *Mob. Inf. Syst.*, vol. 2016, p. 4913216, Mar. 2016, doi: 10.1155/2016/4913216.

[11] L. Skorin-Kapov, M. Varela, T. Hossfeld, and K.-T. Chen, "A Survey of Emerging Concepts and Challenges for QoE Management of Multimedia Services," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, pp. 1–29, May 2018, doi: 10.1145/3176648.

[12] P. Reichl, P. Maillé, P. Zwickl, and A. Sackl, *A Fixed-Point Model for QoE-based Charging*. 2013, p. 38.

[13] K. Brunnström, M. Sjöström, M. Imran, M. Pettersson, and M. Johanson, *Quality of Experience for a Virtual Reality simulator*. 2018.

[14] H. T. T. Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang, "A Subjective Study on User Perception Aspects in Virtual Reality," *Appl. Sci.*, vol. 9, no. 16, 2019, doi: 10.3390/app9163384.

[15] M. Anwar, J. Wang, W. Khan, S. Ahmad, and Z. Fei, "Measuring Quality of Experience for 360-degree Videos in Virtual Reality," *Sciece China Inf. Sci.*, vol. 63, Mar. 2020, doi: 10.1007/s11432-019-2734-y.

[16] S. Vlahović, M. Suznjevic, and L. S. Kapov, "Subjective Assessment of Different Locomotion Techniques in Virtual Reality Environments," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2018, pp. 1–3, doi: 10.1109/QoMEX.2018.8463433.

[17] P. C. Madhusudana and R. Soundararajan, "Subjective and Objective Quality Assessment of Stitched Images for Virtual Reality," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5620–5635, Nov. 2019, doi: 10.1109/TIP.2019.2921858.

[18] J. Malo, R. Navarro, I. Epifanio, F. Ferri, and J. M. Artigas, "Non-linear Invertible Representation for Joint Statistical and Perceptual Feature Decorrelation," in *Advances in Pattern Recognition*, Berlin, Heidelberg, 2000, pp. 658–667.

[19] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2016, pp. 1–6, doi: 10.1109/QoMEX.2016.7498964.

[20] A. Zheleva, W. Durnez, K. Bombeke, G. Wallendael, and L. Marez, *Seeing is Believing: The Effect of Video Quality on Quality of Experience in Virtual Reality*. 2020, p. 4.

[21] S. Katsigiannis, R. Willis, and N. Ramzan, "A QoE and Simulator Sickness Evaluation of a Smart-Exercise-Bike Virtual Reality System via User Feedback and Physiological Signals," *IEEE Trans. Consum. Electron.*, vol. PP, pp. 1–1, Nov. 2018, doi: 10.1109/TCE.2018.2879065.

[22] C. Keighrey, R. Flynn, S. Murray, S. Brennan, and N. Murray, *Comparing User QoE via Physiological and Interaction Measurements of Immersive Augmented Reality and Virtual Reality Speech and Language Therapy Applications*. 2017, p. 492.

[23] K. Swinburn, G. Porter, and D. Howard, *Comprehensive Aphasia Test*. .

[24] A. Doumanoglou *et al.*, “Quality of Experience for 3-D Immersive Media Streaming,” *IEEE Trans. Broadcast.*, vol. 64, pp. 379–391, Jun. 2018, doi: 10.1109/TBC.2018.2823909.

[25] P. Seeling, “Image Quality in Augmented Binocular Vision: QoE Approximations with QoS and EEG,” *Period. Polytech. Electr. Eng. Comput. Sci.*, vol. 61, Sep. 2017, doi: 10.3311/PPee.9454.

[26] P. Seeling, *Augmented vision and Quality of Experience assessment: Towards a unified evaluation framework*. 2015, p. 1740.

[27] L. Zhang, H. Dong, and A. El Saddik, “Towards a QoE Model to Evaluate Holographic Augmented Reality Devices: A HoloLens Case Study,” *IEEE Multimed.*, vol. PP, pp. 1–1, Oct. 2018, doi: 10.1109/MMUL.2018.2873843.

[28] B. Bauman and P. Seeling, “Towards Predictions of the Image Quality of Experience for Augmented Reality Scenarios,” May 2017.

[29] A. Mittal, A. Moorthy, and A. Bovik, “No-Reference Image Quality Assessment in the Spatial Domain,” *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.*, vol. 21, Aug. 2012, doi: 10.1109/TIP.2012.2214050.

[30] E. Alexiou, E. Upenik, and T. Ebrahimi, “Towards subjective quality assessment of point cloud imaging in augmented reality,” in *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2017, pp. 1–6, doi: 10.1109/MMSP.2017.8122237.

[31] E. Hynes Msc, R. Flynn, B. Lee, and N. Murray, *An Evaluation of Lower Facial Micro Expressions as an Implicit QoE Metric for an Augmented Reality Procedure Assistance Application*. 2020.

[32] T. Rokicki, H. Kociemba, M. Davidson, and J. Dethridge, “The Diameter of the Rubik’s Cube Group Is Twenty,” *SIAM Rev.*, vol. 56, no. 4, pp. 645–670, Jan. 2014, doi: 10.1137/140973499.

[33] S. J. Henderson and S. K. Feiner, “Augmented reality in the psychomotor phase of a procedural task,” in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2011, pp. 191–200, doi: 10.1109/ISMAR.2011.6092386.

[34] A. Vovk, F. Wild, W. Guest, and T. Kuula, “Simulator Sickness in Augmented Reality Training Using the Microsoft HoloLens,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada, Apr. 2018, pp. 1–9, doi: 10.1145/3173574.3173783.

[35] M. Suryanegara, “Managing 5G Technology: Using Quality of Experience (QoE) to Identify the Innovation Enhancement Pattern According to the Indonesian Market,” *IEEE Access*, vol. 8, pp. 165593–165611, 2020, doi: 10.1109/ACCESS.2020.3022365.

[36] E. Liotou *et al.*, “Shaping QoE in the 5G ecosystem,” in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, May 2015, pp. 1–6, doi: 10.1109/QoMEX.2015.7148089.

[37] R. I. T. da Costa Filho *et al.*, “Predicting the Performance of Virtual Reality Video Streaming in Mobile Networks,” in *Proceedings of the 9th ACM Multimedia Systems Conference*, New York, NY, USA, 2018, pp. 270–283, doi: 10.1145/3204949.3204966.

[38] S. Petrangeli, V. Swaminathan, M. Hosseini, and F. De Turck, “An HTTP/2-Based Adaptive Streaming Framework for 360° Virtual Reality Videos,” in *Proceedings of the 25th ACM International Conference on Multimedia*, New York, NY, USA, 2017, pp. 306–314, doi: 10.1145/3123266.3123453.

[39] A. Lounis, F. Alilat, and N. Agoulmine, “Neural Network Model of QoE for Estimation Video Streaming over 5G network,” in *6th International Workshop on ADVANCEs in ICT Infrastructures and Services (ADVANCE 2018)*, Santiago, Chile, Jan. 2018, pp. 21–27, [Online]. Available: https://hal.archives-ouvertes.fr/hal-01777376.

[40] M. Jalil Piran, N. H. Tran, D. Y. Suh, J. B. Song, C. S. Hong, and Z. Han, “QoE-Driven Channel Allocation and Handoff Management for Seamless Multimedia in Cognitive 5G Cellular Networks,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6569–6585, Jul. 2017, doi: 10.1109/TVT.2016.2629507.

[41] M. Chen, W. Saad, and C. Yin, “Echo State Learning for Wireless Virtual Reality Resource Allocation in UAV-Enabled LTE-U Networks,” in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6, doi: 10.1109/ICC.2018.8422503.

[42] C. Ge *et al.*, “QoE-Assured Live Streaming via Satellite Backhaul in 5G Networks,” *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 381–391, Jun. 2019, doi: 10.1109/TBC.2019.2901397.

[43] P. Zhou *et al.*, *5G MEC Computation Handoff for Mobile Augmented Reality*. 2021.

[44] B. Krogfoss, J. Duran, P. Perez, and J. Bouwen, “Quantifying the Value of 5G and Edge Cloud on QoE for AR/VR,” in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2020, pp. 1–4, doi: 10.1109/QoMEX48832.2020.9123090.

[45] S. Schwarzmann, C. Cassales Marquezan, M. Bosk, H. Liu, R. Trivisonno, and T. Zinner, “Estimating Video Streaming QoE in the 5G Architecture Using Machine Learning,” in *Proceedings of the 4th Internet-QoE Workshop on QoE-Based Analysis and Management of Data Communication Networks*, New York, NY, USA, 2019, pp. 7–12, doi: 10.1145/3349611.3355547.

[46] Y. Wang *et al.*, “A Data-Driven Architecture for Personalized QoE Management in 5G Wireless Networks,” *IEEE Wirel. Commun.*, vol. 24, no. 1, pp. 102–110, Feb. 2017, doi: 10.1109/MWC.2016.1500184WC.
[47] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, “Simulating LTE Cellular Systems: An Open-Source Framework,” *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 498–513, Feb. 2011, doi: 10.1109/TVT.2010.2091660.
[48] J. Nightingale, P. Salva-Garcia, J. M. A. Calero, and Q. Wang, “5G-QoE: QoE Modelling for Ultra-HD Video Streaming in 5G Networks,” *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 621–634, Jun. 2018, doi: 10.1109/TBC.2018.2816786
[49] M. Fiedler, T. Hossfeld and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," in IEEE Network, vol. 24, no. 2, pp. 36-41, March-April 2010, doi: 10.1109/MNET.2010.5430142.