



INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Quality of Experience Assessment in Internet TV

Joana Sofia Cardoso Palhais

Dissertação para obtenção do Grau de Mestre em
Engenharia Electrotécnica e de Computadores

Júri

Presidente: Prof. Doutor Nuno Cavaco Gomes Horta
Orientador: Prof. Doutor Mário Serafim dos Santos Nunes
Co-orientador: Prof. Rui António dos Santos Cruz
Vogal: Prof^a Doutora Maria Paula dos Santos Queluz Rodrigues

Maio de 2011

Acknowledgments

I want to thank all those who, directly or indirectly, contributed to the preparation of this dissertation. Without them, this journey would have been much more arduous.

To my supervisors, Prof. Mário Serafim Nunes and Prof. Rui Santos Cruz, for guiding me throughout these months, always showing the willingness to receive me and the knowledge imparted which served to enrich this work.

To all the collaborators who helped me, especially Eng. Nuno Magalhães, INESC/INOV, who created the application for the subjective tests. To Prof. Paula Queluz who kindly lent the light meter and various advises. To Maria Barradas, delegate of MEEC, for advertising the snack made for the assessment sessions, which resulted in filling up quickly the testing room of IST Alameda.

Finally I want to thank my family for all the support given. Without them, this step would not have been possible. To my parents who provide the effort that allowed the conclusion of my studies. To my sister, Catarina, for getting me out of home whenever it was necessary to relieve stress. To my grandmother for her unconditional support. To my godfather, Eng. Nelson Cardoso, for being my inspiration in this course. To my godmother for the hours spent reviewing my English. Finally but not the least, my boyfriend, Miguel Rosario, for all the support, understanding, motivation, patience and assistance provided at all times, especially in bad ones. Without him, many things would have been impossible to accomplish.

To all of you, thank you very much!

Abstract

The concept of Quality of Experience is attracting nowadays a growing attention and, increasingly, becoming a subject of concern for Service Providers. When talking about Internet TV or WebTV, the concern is even higher due to the unreliable nature of the network, which provides no guarantees of delivery.

Motivated by the European Project My e-Director 2012, which will provide the coverage of the London Olympic Games via the Web, this dissertation intends to investigate the influence that the interest level on a particular sport has in the subjective quality assessment of the corresponding broadcasted media.

Therefore, a subjective test was performed, where each observer visioned a set of six sports, each encoded in four pairs of bitrate/resolution. In these six sports, three correspond to the sports that the observer likes more and three to those he/she likes less. So, each observer evaluated a set of twenty-four videos, with duration of 30 s each. After each video, the observer had to rate the perceived video quality on an eleven-grade scale, with values between 0 and 10, keeping in mind that the scenario was of WebTV.

By analyzing the data collected, the influence of the interest level in the subjective assessment was inferred, with very positive results, and an empirical formula deduced to estimate the Mean Opinion Score (MOS) as a function of bitrate and interest level.

Keywords: Mean Opinion Score, Objective Video Quality, Quality of Experience, Subjective Video Quality

Resumo

O termo Qualidade de Experiência está a ganhar grande terreno nos dias de hoje e, cada vez mais, é alvo da preocupação por parte dos Fornecedores de Serviço. Quando se fala em televisão por Internet – Internet TV ou WebTV – a preocupação é acrescida, devido à falta de fiabilidade da rede, que não assegura garantias de entrega.

Motivada pelo Projecto Europeu My e-Director 2012, que vai efetuar a cobertura dos Jogos Olímpicos de Londres via Web, esta dissertação pretende verificar a influência que o grau de interesse num dado desporto, tem na classificação subjetiva da difusão desse evento.

Procedeu-se, portanto, à realização de um teste subjetivo, onde cada observador visionou um conjunto de seis desportos, codificados em quatro pares de débito binário e resolução. De entre estes seis desportos, estavam os três que o observador menos gostava e os três de que mais gostava. Cada observador avaliou então um conjunto de vinte e quatro vídeos, tendo cada vídeo a duração de 30 s. Após cada vídeo, o observador teve que classificar a qualidade percebida, numa escala de 0 a 10, tendo sempre em mente que estava num cenário de WebTV.

Com os dados recolhidos, inferiu-se sobre a influência do grau de interesse na classificação subjetiva, que se revelou significativa, e procedeu-se à construção de uma fórmula empírica para estimar o Mean Opinion Score (MOS) em função do débito binário e do grau de interesse.

Palavras-chave: Mean Opinion Score, Objective Video Quality, Quality of Experience, Subjective Video Quality

Index

Acknowledgments	3
Abstract.....	5
Resumo	6
List of Figures	9
List of Tables	11
List of Acronyms	12
Chapter 1 – Introduction	13
1.1. Motivation and Objectives.....	14
1.2. Contributions.....	15
1.3. Dissertation Layout	15
Chapter 2 – My e-Director 2012	17
2.1. Introduction	18
2.2. What is My e-Director 2012	18
2.3. Architecture	19
2.4. How it Works.....	21
2.5. Summary	21
Chapter 3 – From Quality of Service to Quality of Experience	23
3.1. Introduction	24
3.2. QoS and QoE	24
3.3. QoP – A New Concept or a Revised One?	25
3.4. The Co-existence of Concepts	25
3.5. Summary	27
Chapter 4 – Video Quality Metrics.....	29
4.1. Introduction	30
4.2. Objective Metrics	30
4.2.1. Mean Square Error (MSE)	30
4.2.2. Peak Signal-to-Noise Ratio (PSNR)	31
4.2.3. Structural Similarity (SSIM).....	31
4.2.4. Video Quality Metric (VQM)	32
4.2.5. Perceptual Evaluation of Video Quality (PEVQ)	33
4.3. Subjective Metrics.....	33
4.3.1. Double-Stimulus Impairment Scale (DSIS).....	35
4.3.2. Double-Stimulus Continuous Quality Scale (DSCQS).....	35
4.3.3. Single-Stimulus (SS)	36
4.3.4. Single-Stimulus Continuous Quality Evaluation (SSCQE)	36
4.3.5. Simultaneous Double-Stimulus for Continuous Evaluation (SDSCE)	37
4.4. Summary	37

Chapter 5 – Methodology for Subjective Test	39
5.1. Introduction	40
5.2. Selection of Test Materials	40
5.3. Selection Criteria for the Observers	41
5.4. Video Quality Assessment.....	44
5.5. Architecture Used for the Assessment Tests	45
5.5.1. Streaming Server	45
5.5.2. Database.....	46
5.5.3. Client	46
5.6. At the Session Day	47
5.7. Summary	48
Chapter 6 – Data Analysis.....	51
6.1. Introduction	52
6.2. Data Analysis from the Initial Survey	52
6.3. Data Analysis from the Video Quality Assessment	54
6.4. MOS Estimation.....	59
6.5. Summary	70
Chapter 7 – Conclusions and Future Work	71
7.1. Conclusions	72
7.2. Future Work	72
References	73
Annex 1 – FFmpeg Comands	75
Annex 2 – Application’s Appearance.....	77
Annex 3 – Ishihara Plates.....	78
Annex 4 – Temporal Activity MATLAB® Code	82

List of Figures

Figure 1 - My e-Director 2012 appearance	18
Figure 2 – My e-Director 2012 Simplified System Architecture.....	20
Figure 3 – QoS/QoP/QoE three layer model.....	26
Figure 4 – TCP/IP model and parameters relationship with the QoS/QoP/QoE three layers model	27
Figure 5 – Additional evaluative scales	35
Figure 6 – Continuous five-grade scale to DSCQS method.....	36
Figure 7 – Visual and color acuity tests.....	42
Figure 8 – Diagram used for selection of the observers	43
Figure 9 – Video quality assessment star-scale.....	44
Figure 10 – Test session architecture for Media and Stream Adaptation	45
Figure 11 – UI appearance when the assessment windows pops-up.....	47
Figure 12 – Age distribution over all the age groups.....	52
Figure 13 – Distribution by gender	53
Figure 14 – Distribution of the sports interests by the levels low, medium and high	53
Figure 15 – Percentage of videos shown by Interest level.....	54
Figure 16 – Sports in the low interest levels.....	55
Figure 17 – Sports in the high interest levels	55
Figure 18 – Comparison between the average values assigned to each interest level	56
Figure 19 – Standard Deviation for each interest level curves.....	57
Figure 20 – Level 1 vs. Level 2 for Boxing	58
Figure 21 – Comparison between the three sports with low interest	58
Figure 22 – Comparison between the three sports with high interest.....	59
Figure 23 – Logarithmic trend lines for each interest level.....	60
Figure 24 - Comparison between MOS(R) and experimental curves	61
Figure 25 – MOS as function of interest level for each available bitrate	62
Figure 26 – Average curve vs. MOS estimation each for level 1	63
Figure 27 – Average curve vs. MOS estimation each for level 2	63
Figure 28 – Average curve vs. MOS estimation each for level 4	63
Figure 29 – Average curve vs. MOS estimation each for level 5	64
Figure 30 – Global temporal activity vs. 99% percentile	65
Figure 31 – Difference between consecutive frames to Floor Exercises	65
Figure 32 – Temporal activity for the 32 available sports.....	66
Figure 33 – Athlete's presentation.....	67
Figure 34 – Exercise zoom.....	67
Figure 35 – Zoom to athlete's hands along the pommel horse	67
Figure 36 – Athlete overview	67
Figure 37 – Athlete's coach	67
Figure 38 – Athlete waiting for the scores	68

Figure 39 – New athlete presentation	68
Figure 40 – Difference between consecutive frames to Pommel Horse	68
Figure 41 – Application login screen	77
Figure 42 – Application logout screen	77

List of Tables

Table 1 – QoS and QoS characteristics	27
Table 2 – Available videos for the subjective video quality assessment sessions.....	41
Table 3 – Bitrate/Resolution pairs for the transcoded videos.....	41
Table 4 – Characteristics of the Streaming Server	46
Table 5 – Characteristics of the used computers/monitors at each test session	47
Table 6 – Woman/Man low and high sports interests	54
Table 7 – FFmpeg command syntax and description	75

List of Acronyms

ACR	Absolute Category Rating
CIF	Common International Format (resolution: 352 × 288)
CRT	Cathode Ray Tube
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DSCQS	Double-Stimulus Continuous Quality Scale
DSIS	Double-Stimulus Impairment Scale
FR	Full Reference
HD	High Definition
HVS	Human Visual System
IP	Internet Protocol
LCD	Liquid Crystal Display
MOS	Mean Opinion Score
MSE	Mean Square Error
NR	No Reference
PEVQ	Perceptual Evaluation of Video Quality
PSNR	<i>Peak Signal-to-Noise Ratio</i>
QCIF	Quarter CIF (resolution: 176 × 144)
QoE	Quality of Experience
QoP	Quality of Perception
QoS	Quality of Service
RM	Rapid Movement
RR	Reduced Reference
SDSCE	Simultaneous Double-Stimulus for Continuous Evaluation
SM	Slight Movement
SS	Single-Stimulus
SSCQE	Single-Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity
TCP	Transmission Control Protocol
UI	User Interface
VQM	Video Quality Metric
WAN	Wide Area Network

Chapter 1 – Introduction

Chapter index

- 1.1. Motivation and Objectives, 14
- 1.2. Contributions, 15
- 1.3. Dissertation Layout, 15

1.1. Motivation and Objectives

The European Project My e-Director 2012 aims to develop an architecture for interactive and personalized WebTV. With this new architecture users will be able to choose the events they want to watch, the cameras that best capture the selected events or the athletes that they want to follow. After selecting the most interesting events and athletes, users will receive alerts in their terminals, through messages and pictures, about the events that meet the selected criteria.

Due to the complexity of the architecture and the rich user interface of the terminal player, the evaluation by users of the media being displayed becomes difficult. It is thus necessary to develop specific assessment methodologies in order to define the so called quality of experience (QoE). To estimate the overall QoE two steps must be implemented: in the first, a suite of tests with human evaluators needs to be performed to enable the collection of the corresponding subjective data. Such data should be further validated, using statistical methods, to obtain curves of Mean Opinion Scores (MOS) as a function of the evaluated parameters. After the assessment test session, the second step is performed which consists on a questionnaire that is made to catch the observers global satisfaction on the experience with the My e-Director 2012 platform. The next questions illustrates the type of questions that can be asked: "Would you like to use My e-Director 2012 again?", "Do you use the Play/Pause function in last view?", "Are you satisfied with the selecting mode in camera function?", etc.

However, since it was impossible to use the real environment of My e-Director 2012, an environment test was develop to play/assess the selected videos. So, with this new test platform only the perceived video quality (QoP) can be estimated. Concluding, the main goal of this dissertation is the investigation of the influence that the interest level on a particular sport has in the subjective assessment of a visioned media containing sports. The hypothesis is that content have a positive influence in their subjective assessment, i.e., as long as the content desirability increases, its subjective assessment will also increase for the same quality and bitrate. To validate or refuse the hypothesis, a subjective test was designed to collect the necessary data. The validation of the collected data turns possible the estimation of MOS as a function of both the QoP and the interest level. As QoP is heavily dependent on the bitrate of the media, MOS is estimated as a function of bitrate (R) and interest level (IL). The resulting empirical formula for that estimation is an additive formula, where the first term is a function of R and the second term a function of IL .

Note that the QoP is influenced by many factors from the quality of service parameters (packet loss, delay, jitter) until the encoder parameters, such as the quantization step, bitstream and resolution. Beyond these technical factors, as the QoE, the QoP is also influenced by emotional factors, which cannot be physically measured and may vary from user to user. For instance, if a user is more irritable his/her reaction to the unavailability of the service would be different from the reaction of a relaxed user. For a very demanding user a service could never be perfect because he/she wants even more details, options, functionalities, etc. So, the users' personality and humor has an important influence but cannot be measured or estimated.

1.2. Contributions

This dissertation aims to contribute in two ways:

- To develop a specific assessment methodology to define the QoE;
- To develop an empirical formula to estimate the MOS based on subjective parameters, such as the interest level for a given content.

1.3. Dissertation Layout

This dissertation consists of seven chapters and it is structured in three main stages:

1. Introduction: Chapter 2, Chapter 3 and Chapter 4
2. Main body: Chapter 5 and Chapter 6
3. Conclusion: Chapter 7

The same structure is applied to each of the chapters. All start with a brief introduction on the topic followed by its development, concluding with a summary and the key ideas of the chapter.

Chapter 2 – My e-Director 2012: provides an introduction to this project that somehow served as motivation for this dissertation research.

Chapter 3 – From QoS to QoE: makes a review of the quality concepts as QoS, QoP and QoE and relates them in a three layer model, making the bridge with the TCP/IP model.

Chapter 4 – Video Quality Metrics: identifies the differences between the subjective and objective methods for video quality measurement and the metrics that are more commonly used in each category; MSE, PSNR, VQM, SSIM and PEVQ for the subjective methods and DSIS, DSCQS, SSCQE and SDSCE for the objective methods.

Chapter 5 – Methodology for Subjective Test: to achieve the objective, it was necessary to perform a subjective test. As such, this chapter explains the entire process of the test session's preparation, as well as the implementation at the session day.

Chapter 6 – Data Analysis: after the subjective test, the collected data was validated and analyzed to verify if the practical results were in line with the expectations. Also in this chapter, an empirical formula was designed to estimate the MOS as a function of the interest level and bitrate.

Chapter 7 – Conclusions: presents a summary of the main contributions of this dissertation, the conclusions derived from the research and the future work that can be developed to further improve the assessment methodology.

Chapter 2 – My e-Director 2012

Chapter index

- 2.1. Introduction, 18
- 2.2. What is My e-Director 2012, 18
- 2.3. Architecture, 19
- 2.4. How it Works, 21
- 2.5. Summary, 21

2.1. Introduction

In live sports events such as the Olympic Games, many athletes from different countries and modalities compete at the same time. The problem is that the spectator can only see the camera angle that the operator shows.

My e-Director 2012 puts the spectators in the director's seat and provides the power that enables them to choose what camera angle, athletes and objects they want to see at that moment, based on their personal preferences.

2.2. What is My e-Director 2012

My e-Director 2012 [1] is a R&D project involving key personnel of European institutions representing Industry, Research Institutions and Universities, which are partially funded in this project under the 7th Research Framework Programme (FP7) [2] of the European Union.

The main goal of My e-Director 2012 is to research and develop a unique interactive broadcasting service that enables end users to select focal actors and points of interest within real-time broadcasted scenes. The service will resemble an automated ambient intelligent director that will operate with minimal or even without human intervention based on the user's profile and preferences. With My e-Director 2012 end-users can choose what camera angle they want to see, track their favorite athlete during the whole game, change modalities and so on. All of these actions can be possible due to the "Multi-camera Selection" and "Person and Object Tracking", which are the main innovations of My e-Director 2012 system. These types of dynamic mechanisms take into account face recognition and identification by detecting and filtering movement across frames, even in outdoors venues, high activity scenes and environments involving many athletes.

Figure 1 shows the appearance of My e-Director 2012 platform. As can be seen, users can select the event that they want to see (1); pause and rewind the live stream (2); make zoom, select the specific camera to choose the angle to watch the event, view/change their profile and watch in full screen mode (3).



Figure 1 - My e-Director 2012 appearance

My e-Director 2012 will be available to all users' profile in general:

1. TV users: people who use the television in terms of their amusement, education and acquiring information and who also exploit the services that are provided, for example teletext (stock market, weather forecast, television guide, etc.).
2. Portable computer users: uses a portable computer that is designed to be moved from one place to another, such as personal digital assistants (PDA) and Pocket PCs which can offer most of the functionalities of desktop computers.
3. Mobile phone users: prefer to use mobile phones. These terminals support many services and accessories, such as SMS, email, GPS and MMS.

To ensure that all users have access to the best quality for each profile, a HTTP adaptive streaming solution, based on Microsoft Smooth Streaming technology is used to provide large scale HTTP based video streaming to end users. The Smooth Streaming enables an uninterrupted streaming experience, guaranteeing, for each user, the best possible video quality reception. With this technology, multimedia sources (video and audio) are encoded in various bitrates and a HTTP progressive download method provides the capability to react to both bandwidth and terminal conditions, in order to seamlessly switch the video quality that the user is receiving, maximizing the users' Quality of Perception (QoP) and Quality of Experience (QoE).

Other important feature are the network portability (vertical handover) and terminal portability capabilities, allowing a flexible user experience in terms of either a continuous uninterrupted playback during network handover with the same terminal or a suspend/resume session playback, either with the same terminal or with a different one (with the same or different capabilities like screen resolution or access network attachments).

2.3. Architecture

The My e-Director 2012 platform will be a heterogeneous system distributed over three different stages:

- Content creation and annotation;
- Service provision;
- Media access and playback.

Content creation and annotation is responsible for collecting the video and sport information data from their original sources, processing and delivering it in a format ready to be used for service provision.

In *service provision* stage, all the personalized media streams are prepared and delivered to the end users by matching recommendations to their preferences for switching the viewing channel.

Finally, the *media access and playback* will provide the interface between the rest of the system and the end user by offering personalized media streaming.

This brief presentation of My e-Director 2012 platform shows how complex the architecture will be. A simplified architecture of My e-Director 2012 is illustrated in Figure 2.

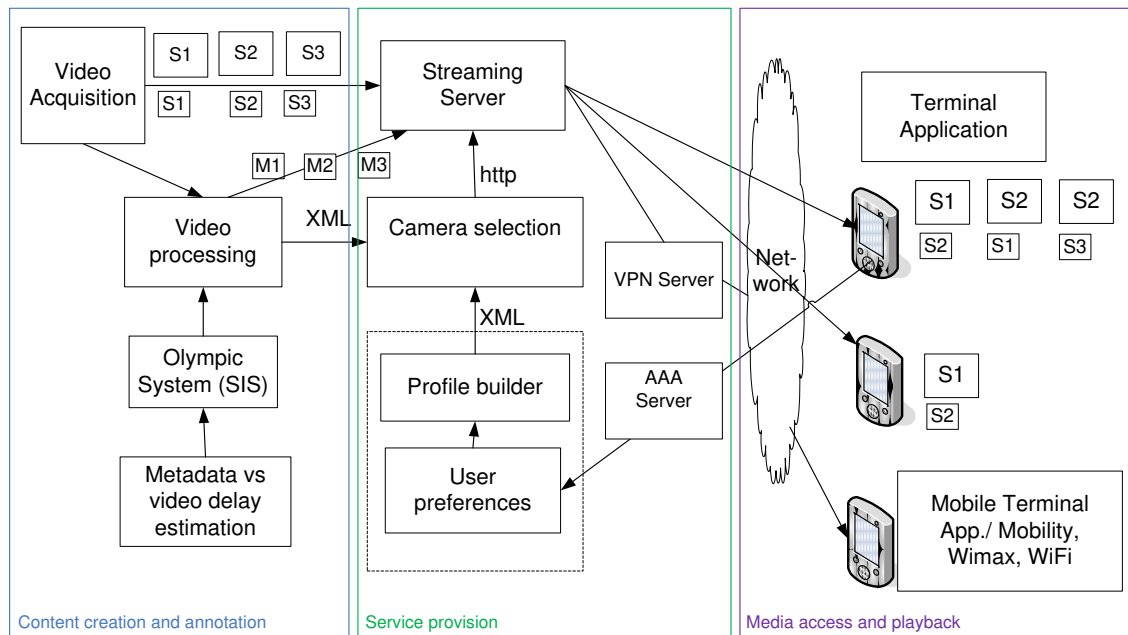


Figure 2 – My e-Director 2012 Simplified System Architecture

From the architecture diagram of Figure 2 it is possible to identify the mentioned three stages and all its processes. All starts with the video acquisition by the cameras. Then, the raw video streams are sent to the video-processing block that produces a stream of metadata for each camera with all relevant information that can be identified. With this information and the users' preferences the camera selection module can select the camera angle or the event that best matches with the users' profile.

In view of the fact that this is a network centric model, the users' device requires only a standard browser, plus a standard media player client or a thin My e-Director 2012 client to have access to the service. Some requisites/recommendations must also be assured to guarantee the best performance on a computer machine:

- Microsoft Silverlight Plug in;
- Internet Explorer 6, 7, 8 for Windows (2003, XP SP2 or greater, Vista and 7);
- Firefox 2, 3.x for Windows (2003, XP SP2 or greater, Vista and 7);
- Safari 3, 4 for Macintosh (Intel only);
- 68 kbps (or superior) broadband Internet access (xDSL like);
- 2.4 GHz Pentium 4 or Intel-based Macintosh, 1 GB RAM, hardware acceleration-capable graphics card or superior hardware.

In contrast, at the Service Provider side, the complexity will be huge, as at every time, there needs to be always a camera that better matches to the user preferences. For that reason, the request for new camera selection will be continually received. As this is extremely difficult to process for each

user, users may be grouped in profiles with similar preferences and execute the camera selection for each profile.

2.4. How it Works

To access My e-Director 2012 service, users must register in the service website. The registration will be pretty standard, where a minimal data set like age, gender and email address is requested. In the registration process, users must also specify their preferences for sports and countries (location). After the registration, each user profile is created and users can then have free access to the service. Once inside the application, users may change their preferences at any time and may also choose the athletes they want to follow.

When users enter their session, they will start to see a default event, like as if they had just turned on their traditional TV set. Based on each user's profile, the camera selection block will identify the event that best fulfills the user interests.

It is possible to easily find the event that best fits each user due to the variety of existing metadata. After selecting the event to send, the Streaming Server sends the recommendation to the user terminal and, depending on the user accepting the suggestion or not, it proceeds to the camera change. If the user does not accept the recommendation and wants to select a new event, he/she just needs to select the new event and wait for it, as if it were a channel change.

This is the simplified *modus operandi* between the stages *Media Access and Playback* and *Service Provision*. The exchange of information is therefore bidirectional, since both stages communicate with each other. In contrast, the communication between the *Content Creation and Annotation* and *Service Provision* is unidirectional, since *Service Provision* only receives the video streams and Metadata from the first stage as described in 2.3. Architecture.

2.5. Summary

My e-Director 2012 is an innovative application that aims to revolutionize the way spectators watch sports events, like the Olympic Games, where many games and athletes compete at the same time.

This new experience sets the users in the director's seat, since they can choose what they want to see. However, the application has the ability to choose the events that best fit the profile of each user without human intervention. This is possible due to an intelligent system based on the users' profile that can identify all the events that meet the users' needs.

Although the architecture at the Service Provider side is fairly complex, on the users' side it is very simple with a small set of basic requirements. The user interface (UI) is also very simple and intuitive to use.

Chapter 3 – From Quality of Service to Quality of Experience

Chapter index

- 3.1. Introduction, 24
- 3.2. QoS and QoE, 24
- 3.3. QoP – A New Concept or a Revised One?, 25
- 3.4. The Co-existence of Concepts, 25
- 3.5. Summary, 27

3.1. Introduction

In the past, Service Providers were concerned about measuring the Quality of Service (QoS) for the audio and video data sent to their consumers. However, nowadays, more and more people are choosing their own platform to watch video content. The traditional television sets are being replaced by mobile devices and personal computers, allowing users to access video contents anywhere. Regardless of the type of device, content viewed, or network used for access, each person still has some basic expectations about the viewing experience. This means that a new concept called Quality of Experience (QoE) is growing up as rapidly as the Quality of Perception (QoP), although the concept of QoS continues to be important.

This chapter tries to define these three concepts and study their relationship based on a three layers model that can also relate with the TCP/IP model.

3.2. QoS and QoE

The first thing to clarify is that Quality of Service (QoS) and Quality of Experience (QoE) are two distinct concepts that cannot be ignored and are both important. Despite their differences, these two concepts are related and one cannot exist without the other.

QoS has been in use for a long time and has reached a high level of common understanding. ITU-T Recommendation E.800 [3] defines QoS as *“the totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service”*.

The QoS concept is based on technical performance and typically measures the network performance at the packet level. Therefore, the most common parameters used are:

- Packet loss;
- Delay;
- Jitter;
- Throughput.

The concept of QoE is relatively new and is attracting growing attention. Therefore, different definitions of the QoE are stated throughout the literature, as exposed in [4]. In spite of all these definitions, ITU-T Recommendation P.10/G.100 [5] defines QoE as *“the overall acceptability of an application or service, as perceived subjectively by the end-user”*.

The QoE concept is based on the global enjoyment and satisfaction of the end-user. Typically, the parameters more commonly used are:

- Fidelity of information: audio and video perceived quality;
- Usability: ease of use;
- Responsiveness: System start-up and channel zapping time;
- Availability: how many times the system was off.

The QoE concept differs from QoS because the latter is more concerned with the performance of the network while QoE is concerned about the overall experience that the user has when accessing and using the services. Because of that, it is common to refer to QoE as user-centered and QoS as a technology-centered.

3.3. QoP – A New Concept or a Revised One?

The concept of Quality of Perception (QoP) emerged just after the QoE concept but focused in the detection of a change in quality or in the acceptance of a quality level. However, this is not a new concept; the QoP was already known as the user-perceived QoS (QoSE). ITU-T [3] defines QoP (QoSE) as *“a statement expressing the level of quality that customers/users believe they have experienced”*. With the introduction of QoE, and in order to avoid ambiguities or exchange of concept meaning, it has been defined that the three quality concepts would be QoS, QoP and QoE instead of QoS, QoSE and QoE.

The QoP is strongly dependent of the QoS, since the loss of a packet has a destructive visual effect. However, this type of errors can be masked by using the concealment techniques. Each codec has its own technique that can be more or less complex. Although there are several codecs, the most common is the MPEG2, which is gradually being replaced by MPEG4/AVC. More important than the codec choice, is the choice of the encoder parameters – bitstream, quantization step, resolution, etc. the next topic will enhance on this theme.

Typically, the QoP is measured with a subjective rating scale such as the Mean Opinion Score (MOS) [6]. The MOS is a numeric scale value between 1 and 5, where 5 represents the highest quality and 1 the lowest one.

When the concept of MOS was introduced, it was calculated in a manual way. Groups of people were recruited and invited to watch a piece of content. Then, each person would give a numeric score to the quality of content just watched and the average of the scores would become the MOS for that piece of content.

Today, the method of generating MOS values is made using algorithms that compute estimates of the MOS based on the characteristics of both the media stream and the network. This is an automated method that produces results, which are usually as reliable as the results that have been generated by humans as stated in [7]. The MOS metric will be described more deeply in Chapter 4 of this dissertation.

3.4. The Co-existence of Concepts

As previously described, QoE and QoS are distinct and both important and are related. QoS is a technical approach whereas QoE and QoP are user-centered approaches.

Figure 3 shows the relationship between these concepts in a three layer model. QoS is the lowest layer because it operates at packet level and QoE the highest because it is related with the user opinion.

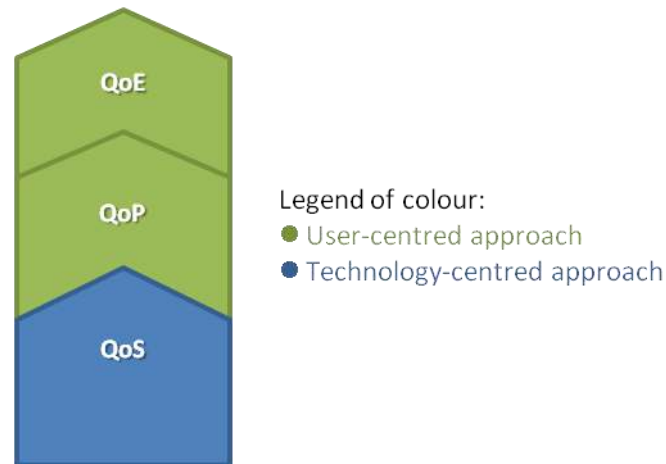


Figure 3 – QoS/QoP/QoE three layer model

This three layer model can also be related with the TCP/IP model. As is of common knowledge, the TCP/IP model consists of five layers – physical, data link, network, transport and application – running over the application layer the various services offered by Service Providers. Each layer has different functions and therefore has different performance parameters.

In the network and transport layers, the parameters more used are delay, packet loss and jitter whereas in the application layer the encoder parameters are used, such as resolution, quantization step and bitstream.

In the past, Service Providers have used the parameters of the network and transport layers to infer about the quality of service offered. Therefore, in recent years, the importance of monitoring the encoder parameters, in the application layer, has increased. And so, the concept of quality of user perception was born. Similarly, at the service level, the concept of user's enjoyment and satisfaction became also important and the concept of quality of experience was introduced.

Figure 3 shows the relationship between the TCP/IP model and the QoS/QoP/QoE three layers model. By observing Figure 4 it is quite intuitive that QoS should be the last layer, QoP the second and QoE the first due to their relationship with the TCP/IP layers.

At the QoS level the parameters more used are those from the network and the transport layers. These parameters could help Service Providers to measure the performance of its network.

To measure QoP and QoE levels, Service Providers must perform surveys over their client base to catch their perception on the quality of the service and the global satisfaction. By analyzing the results, Service Providers are able to know the maximum quality they can deliver and what is the sufficient

level of quality that can be accepted by their viewers. It is also important to notice that there is a strong correlation between the desirability of movie content and subjective ratings of video quality [8].

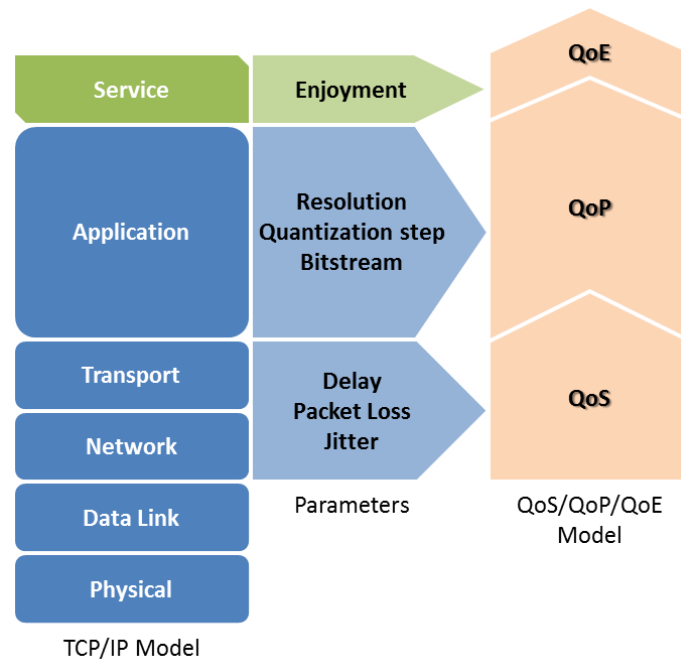


Figure 4 – TCP/IP model and parameters relationship with the QoS/QoP/QoE three layers model

3.5. Summary

Although the QoS continues to play an important role, in recent years Service Providers pay more attention to the way end-users catch the performance of their services. This leads to the QoP and QoE concepts which are user-centered approaches in contrast to QoS which is a technology-centered approach. Table 1 illustrates the main characteristics of QoS and QoE concepts.

The relationship between these three concepts can be expressed in a three layer model, where QoS is at the lowest level and QoE the highest one. This model can also be related with the TCP/IP five layers model allowing identifying each quality concept and respective parameters in the corresponding TCP/IP layer.

Table 1 – QoS and QoS characteristics

	QoS	QoE
Approach	Technology-centered	User-centered
Parameters	Packet loss Delay Jitter Throughput	Fidelity of information Usability Responsiveness Availability
Measurement	Exact measurement: When the parameters are fixed, the final result is always the same.	Subjective measurement: Depends on users' opinion and expectations.

Chapter 4 – Video Quality Metrics

Chapter index

- 4.1. Introduction, 30
- 4.2. Objective Metrics, 30
 - 4.2.1. Mean Square Error (MSE), 30
 - 4.2.2. Peak Signal-to-Noise Ratio (PSNR), 31
 - 4.2.3. Structural Similarity (SSIM), 31
 - 4.2.4. Video Quality Metric (VQM), 32
 - 4.2.5. Perceptual Evaluation of Video Quality (PEVQ), 33
- 4.3. Subjective Metrics, 33
 - 4.3.1. Double-Stimulus Impairment Scale (DSIS), 35
 - 4.3.2. Double-Stimulus Continuous Quality Scale (DSCQS), 35
 - 4.3.3. Single-Stimulus (SS), 36
 - 4.3.4. Single-Stimulus Continuous Quality Evaluation (SSCQE), 36
 - 4.3.5. Simultaneous Double-Stimulus for Continuous Evaluation (SDSCE), 37
- 4.4. Summary, 37

4.1. Introduction

There are two fundamental methods for video quality measurement: objective and subjective. The objective method involves the use of information contained in the image without the need of a human observation. The subjective method relies on the human judgment to infer the quality of the video. As long as the subjective method becomes more complex, i.e., takes into account not only the stream characteristics but also the Human Visual System, the results obtained are closed to the ones obtained when using subjective methods and they are usually reliable and correlated [7].

This chapter intends to clarify the differences between objective and subjective methods and to identify the metrics that are commonly used for the video quality assessment in both measurements.

4.2. Objective Metrics

Objective video quality measurements do not need human intervention for classification of the video, as it is an automated method, based on algorithms, able to estimate the video quality by just analyzing the characteristics of the media stream.

The objective video quality metrics are usually classified in three classes [4]:

1. Full Reference (FR): the original video is totally available as well as the decoded one;
2. Reduced Reference (RR): some characteristics of the original video are used to compare with the decoded video;
3. No Reference (NR): the original video is not available, only the decoded video.

Of these three metric classes, the most common and accurate approach is the FR. The NR has been, so far, the less used.

Generally, Mean Square Error (MSE) and Peak Signal-to-Noise ratio (PSNR) are the techniques more used for objective metrics due to their simplicity. Both techniques indicate the differences between the received video signal and the reference video signal. Other metrics can also be used for this purpose, such as the Perceptual Evaluation of Video Quality (PEVQ), the Structural Similarity (SSIM), the Video Quality Metric (VQM), and a few others. The next subtopics will try to describe these objective metrics more deeply.

4.2.1. Mean Square Error (MSE)

MSE is a very popular metric, both in video and in image quality and it is defined as:

$$MSE = \frac{1}{XYT} \sum_{t=1}^T \sum_{y=1}^Y \sum_{x=1}^X [p(x, y, t) - p'(x, y, t)]^2 \quad (1)$$

In MSE, the original video is represented by $p(x, y, t)$ and the distorted one by $p'(x, y, t)$.

MSE is applied to videos with a frame size of $X \times Y$ pixels and T frames. The estimate is made by comparing pixel by pixel in both original and distorted frames. This type of comparison, pixel-by-pixel, gives a quick evaluation of the video quality. However, it has some drawbacks. Due to its *modus operandi*, MSE does not consider the influence of image content and viewing conditions. This means that, for some situations, MSE values can be higher, noticing video degradation, although the human system is unable to identify the differences.

4.2.2. Peak Signal-to-Noise Ratio (PSNR)

PSNR is a very popular metric both in video and image quality and it is defined as:

$$PSNR = 10 \log_{10} \frac{(2^M - 1)^2}{MSE} \quad (2)$$

In which, $2^M - 1$ is the maximum value that a pixel can take for an M -bit image. MSE is defined as previously.

This dependency for MSE shows that the PSNR value approaches infinity as the MSE approaches zero. So, a higher PSNR value provides a higher image quality. On the other hand, if MSE is too high, means that there exists a high numerical difference between images, so PSNR takes small values.

As for the MSE drawbacks, PSNR also operates on a pixel-by-pixel basis, and so, the Human Visual System (HVS) is not considered in the solution. Because of that, PSNR is also a quick evaluation of the video quality.

4.2.3. Structural Similarity (SSIM)

SSIM is a more advanced measurement metric than PSNR. This metric considers that the HVS is highly specialized in extracting structural information but it is not specialized in extracting errors. For that propose, SSIM divides each frame in 8×8 pixels blocks, $b_i(x, y, t)$ and $b'_i(x, y, t)$, and measures the luminance, contrast and structural distortion, as well as differences between the two blocks, by using the means (μ and μ'), variances (σ and σ') and covariance (cov) of original and distorted sequences.

The SSIM index for a block $b_i(x, y, t)$, according to [9] and [10], is calculated as:

$$SSIM_i(t) = l_i(t) \times c_i(t) \times s_i(t) \quad (3)$$

where $l_i(t)$, $c_i(t)$ and $s_i(t)$ are defined as follows:

$$l_i(t) = \frac{2 \times \mu_i(t) \times \mu'_i(t) + C_1}{\mu_i^2(t) + \mu_i'^2(t) + C_1} \quad (4)$$

$$c_i(t) = \frac{2 \times \sigma_i(t) \times \sigma'_i(t) + C_2}{\sigma_i^2(t) + \sigma_i'^2(t) + C_2} \quad (5)$$

$$s_i(t) = \frac{cov_i(t) + C_3}{\sigma_j(t) \times \sigma_j'(t) + C_3} \quad (6)$$

All of these parameters give the difference between the original blocks and the distorted ones. $l_i(t)$ refers to luminance, $c_i(t)$ to contrast and $s_i(t)$ to the structure difference. C_1 , C_2 and C_3 are positive constants introduced in equations in order to avoid a null denominator.

Although SSIM indexes are calculated only for properly selected R blocks, they still provide good experimental results. By not calculating the entire frame, the computational costs of SSIM indexes are significantly reduced [9].

Based on each $SSIM_i(t)$ a quality index, $Q(t)$, is calculated for every frame by using a weighting value w_i :

$$Q(t) = \frac{\sum_{i=1}^R w_i(t) \times SSIM_i(t)}{\sum_{i=1}^R w_i(t)} \quad (7)$$

So, the SSIM for the entire sequence is calculated as a sum of frame quality indexes, weighted by $W(t)$:

$$SSIM = \frac{\sum_t W(t) \times Q(t)}{\sum_t W(t)} \quad (8)$$

For no perceived impairments SSIM returns one. In contrast, for a rising level of impairment it returns zero. Notice that SSIM is smaller for a higher level of motion, because the spatial distortion is less visible in a fast moving video. Therefore, more tests must be performed in order to verify the validity of this measurement on motion videos.

Although the main idea of SSIM is simple and gives better results than the other metrics, its mathematical complexity is comparatively higher, which becomes a drawback.

4.2.4. Video Quality Metric (VQM)

VQM measures the perceptual effects of video impairments, i.e., blurring, jerkiness, global noise, block and color distortion, by using a Discrete Cosine Transform (DCT) based on the simplified human spatial-temporal contrast sensitivity model. The distortion is estimated in four basic steps [9]:

1. Each frame from the original and the distorted video is divided in blocks of 8 x 8 pixels, $b_i(x, y, t)$ for the original frame $p(x, y, t)$ and $b_i'(x, y, t)$ for the distorted frame $p'(x, y, t)$, applying on each the DCT.
2. Then, the DCT coefficients are converted into local contrast values $LC_i(u, v, t)$ and $LC_i'(u, v, t)$ by using each block DC component.
3. The local contrast values are converted too, now into just noticeable difference values, $JND_i(u, v, t)$ and $JND_i'(u, v, t)$, by using static and dynamic spatial contrast sensitivity function (SCSF).

4. Finally, the JND coefficients are subtracted to produce difference values, $Diff_i(T)$, and then, the VQM score is calculated as a sum of mean of difference values, $Dist_{Mean}$, and the weighted maximum difference value of all frames, $Dist_{Max}$:

$$Dist_{Mean} = 1000 \times \text{mean}(\text{mean}(Diff_i(T))) \quad (9)$$

$$Dist_{Max} = 1000 \times \max(\max(Diff_i(T))) \quad (10)$$

$$VQM = Dist_{Mean} + 0.005 \times Dist_{Max} \quad (11)$$

Note that the weight parameters 1000 and 0.005 are chosen based on several primitive psychophysics experiments and in the standardization ratio, respectively. The VQM score decreases as the quality of compressed video rises and it is zero for the lossless compressed video.

4.2.5. Perceptual Evaluation of Video Quality (PEVQ)

PEVQ is a FR algorithm since it requires two input signals, the original and the distorted one. Both inputs must be in the same resolution (VGA, CIF or QCIF) and must also follow the YCbCr color representation.

It is a very robust model that estimates the MOS based on the HVS [11]. By using the perceptual masking properties of the HVS, it is possible to know how much a signal can be distorted until the HVS notices the distortion.

The algorithm starts by performing the alignment steps and collecting the information about frozen or skipped frames. Then, the synchronized and equalized images are compared for visual differences, in the luminance and chrominance domain, taking into account the masking effects based on the HVS. The result of this process is a set of indicators which can describe certain quality aspects. Finally, the integration of the indicators by non-linear functions can derivate the MOS score.

As ITU-T Recommendation [J.247] lists, the key features of PEVQ are:

- (Fast and reliable) temporal alignment of the input sequences;
- Full frame spatial alignment;
- Color alignment algorithm based on cumulative histograms;
- Detection and perceptually correct weighting of frame freezes and frame skips;
- Perceptual estimation of degradations;

4.3. Subjective Metrics

Subjective metrics are concerned about collecting data directly from end users, once it is very user centered. This type of assessment method is recognized as the most reliable to quantify user perception since human beings are the ultimate receivers in most applications.

There are two approaches for measuring the subjective video quality:

- Testing the user's perceived quality (QoP);
- Surveying end users about their global experience (QoE).

For testing the user's perceived quality, a group of observers must be recruited to obtain their opinion when asked to rate a sequence of videos or to detect a change in quality. The biggest advantage of this approach is that data is collected in a laboratory with a high level of control. This methodology allows the simulation of real environments through controlled set of parameters such as transmission delay or packet loss. On the other hand, one of the disadvantages is that the measures are only concerned about the human ability to detect changes in quality. This means that the user's behaviors and interaction will not be measured.

To bridge this gap, after testing the user's perceived quality, a surveying could be done to catch the user's engagement and pleasure. This type of work methodology is extremely expensive in terms of time and cost since a group of observers must be validated as well as the test environment and the video contents. Another problem is that it is not an automated mechanism and it is important to notice that if the same test is repeated with different observers, the results must have to be similar.

The methodology followed to perform subjective tests to catch the user's perceived quality is standardized in the Recommendation ITU-R BT.500 [12] and in the Recommendation ITU-T P.910 [13]. The first Recommendation has been the reference for the subjective quality evaluation of television pictures in large formats, when displayed in the classical CRT screens. Recommendation ITU-T P.910 adapts the ITU-R BT.500 to multimedia applications such as videoconferencing, focused in reduced pictures formats, e.g., CIF and QCIF, and in new types of display screens such as LCD.

In spite of the methodology chosen, the grading scale that is widely used to measure the subjective quality is typically the Mean Opinion Score (MOS). The MOS is a numeric scale between 1 and 5 values, where 1 represents the lowest quality and 5 the highest one.

Traditionally, the scale is a five-grade scale, as shown in Figure 5a). However, additional evaluative scales can be used, such as a nine-grade numerical quality scale (Figure 5b)), an eleven-grade numerical quality scale (Figure 5c)), or a quasi-continuous scale for quality ratings, (Figure 5d)), according to [12] [13].

The next subtopics explain the most popular metrics for subjective video quality and the methodologies followed according to the referred recommendations.

Typically, still picture sequences of 3 to 4 s with five repetitions may be appropriate, with voting made in the last two. If the test contains a video, voting during the second repetition may use sequences of 10 s with two repetitions. It is a good practice to limit the duration of sequences to less than 10 s to meet the display time requirement.

Once observers assess the two versions of each pair of pictures, the grading scale is printed in pairs holding the double assessment of each test pair. A continuous five-grade scale is used in order to avoid quantizing errors, as Figure 6 illustrates. To avoid confusions between the scale divisions and the rated results, the scales are printed in blue and observers will mark their assessment in black.

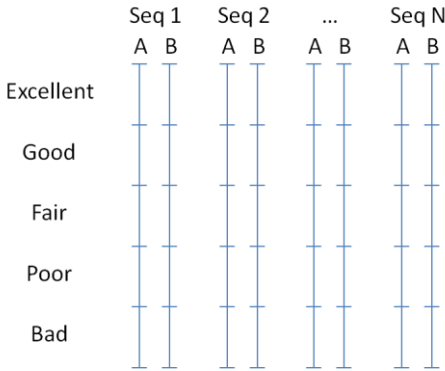


Figure 6 – Continuous five-grade scale to DSCQS method

4.3.3. Single-Stimulus (SS)

The SS method consists in showing a single image or sequence of images to observers that must provide an assessment for the entire presentation. The test session consists in a series of assessment trials, which should be presented in a random way, if possible, in a different order for each observer. Since each sequence is presented one at a time and is rated independently, this method is also called Absolute Category Rating (ACR), according to [13].

After each presentation, observers are asked to evaluate the quality of the sequence shown. The presentation time may be reduced or increased according to the content viewed. In general, the presentation and voting lasts about 10 s each. A five-grade quality scale should be used in this methodology, as illustrated in Figure 5a). However, if a higher discriminative is wanted, a nine-level scale may be used too, as illustrated in Figure 5b).

4.3.4. Single-Stimulus Continuous Quality Evaluation (SSCQE)

This single-stimulus methodology intends to evaluate the impairments of the digital television compression, which are scene-dependent and time-varying. For that purpose, a programme segment (PS), e.g., news, and a quality parameter (QP), e.g., bit rate, are chosen and combined in PS/QP pairs. Note that, each PS should be of at least 5 min long.

In the test session, one or more different combinations of PS/QP pairs should be seen without separation and arranged in a pseudo-random order. Based on this condition, the test session should last between 30 and 60 min.

For this method, the grading scale must be continuous like the handset slider mechanism shown in Figure 6.

4.3.5. Simultaneous Double-Stimulus for Continuous Evaluation (SDSCE)

The SDSCE has been developed starting from the SSCQE by making slight deviations concerning on:

- *The way of presenting the images:* a reference must be used to evaluate the fidelity of the content seen. (Note that in the SSCQE there is no reference in order to reproduce a viewing condition that is as close as possible to the client's home)
- *The rating scale:* a slider-scale should be used, where the lowest value must be 0 and the highest 100.

In the SDSCE test, observers will watch two sequences at the same time, the reference and the content under the test condition. These sequences could be shown side-by-side on the same monitor or in two aligned monitors.

By knowing which of the sequences the reference is, observers are requested to check the differences between both and to assess the fidelity by moving the slider of a handset-voting device. If the fidelity is perfect, the slider should be at the top of the scale range and when the fidelity is null, the slider should be at the bottom.

4.4. Summary

There are two distinct methods to perform video quality assessment, the objective and the subjective method.

In the objective method the measurements are made without human intervention. Despite the exits of various metrics, in this chapter only five of the metrics were described: MSE, PSNR, VQM, SSIM and PEVQ.

The differences between all these metrics are in the parameters that are used. The simpler metrics use only the differences between frames, as are the cases of MSE and PSNR. These metrics are useful to make a quick estimate of quality. In contrast, the more complex metrics use not only the differences between frames, but also mechanisms to take into account the HVS and the perceptual effects of video impairments, in order to estimate how much a signal can be distorted until the human eye notices it. These metrics are useful when the level of requirement increases for a more refined estimate.

In the subjective method the human presence is required, since the measurements are based on ratings given by a group of people. As in the objective method, there are several metrics that can be used, however, only five of the subjective metrics were described: DSIS, DSCQS, SS, SSCQE and SDSCE.

The differences between these metrics consist in showing or not a reference and in the type of rating scale used. For instance, in DSIS, observers see the reference video and the decoded one always in the same order. However, in DSCQS the reference and the decoded videos are shown in a pseudo-random way. As such, observers do not know which of the videos is the reference. SS only shows the decoded video and uses a five-grade scale (a discrete scale). In contrast, although SSCQE only shows the decoded video, it uses a five-grade scale slider, therefore, a continuous scale.

Chapter 5 – Methodology for Subjective Test

Chapter index

- 5.1. Introduction, 40
- 5.2. Selection of Test Materials, 40
- 5.3. Selection Criteria for the Observers, 41
- 5.4. Video Quality Assessment, 44
- 5.5. Architecture Used for the Assessment Tests, 45
 - 5.5.1. Streaming Server, 45
 - 5.5.2. Client, 46
 - 5.5.3. Database, 46
- 5.6. At the Session Day, 47
- 5.7. Summary, 48

5.1. Introduction

This chapter describes the methodology that has been followed to prepare and setup the subjective tests sessions, covering the materials and logistics, selection of observers, the assessment rating and the test architecture used.

The subjective tests were performed to infer the influence of content on the overall QoP. For this purpose, in each test session, a suite of six sports (from a set of 32) encoded in four different bitrates were shown to each of the selected observers. At the end of each video clip the observers ranked, their perceived video quality on a scale ranging from 0 to 10.

In the sessions, the panel of observers experienced an environment similar to what they are used to at their homes, where the original video is not available for comparison with the received decoded video, i.e., on the test session the observers had no reference video for comparison.

5.2. Selection of Test Materials

As each observer would be watching a suite of six sport modalities codified in four different bitrates, a total of twenty-four videos were prepared for each test session and for each observer. The reason behind the choice of the sports theme for the videos was to create a more realistic test environment since My e-Director 2012 will be focused on large athletic events, such as the coverage of the coming 2012 Olympic Games in the City of London. Moreover, it is unclear how end users will react to degradation in quality of this type of contents since users are, in general, very demanding in this field.

All of the test videos to be used were made public on the Internet (on a web server). The original videos selected for the tests are in high definition (HD) 720p format (resolution of 1280×720 pixels), coded with H.264 codec (in baseline profile) with 25 fps and a bitrate of 2 Mbps,. A total of thirty-two sport modalities, listed in Table 2, were selected covering those typically viewed in the Olympic Games.

In order to test the degradation of quality based on bitrate oscillation, the original videos were transcoded in four bitrate/resolution pairs (listed in Table 3) using the FFmpeg tool functions (see Annex 1 for more details on the usage of the tool). The resulting videos had no audio and were cut to a fixed duration of 30 seconds each.

The reason for having no audio in the video sequences was to ensure that all the attention of the observers was focused on the quality of the videos. Although it is recognized that audio usually increases the engagement in video contents, in most sports, images speak for themselves and no audio is necessary. Another reason that led to the elimination of audio was that the bitrate required to transmit audio is negligible, when compared with the video bit rate, for current broadband access networks (the audio in the original videos, was encoded in 64 kbps).

Table 2 – Available videos for the subjective video quality assessment sessions

Aquatics	Platform diving Swimming	Badminton
Athletics	Hurdles Javelin Long Jump Medley Relay Pole Vault Steeplechase Triple Jump Walk	Basketball
Gymnastics	Balance Beam Floor Exercises Horizontal Bar Pommel Horse Rings Trampoline Uneven Bars Vault	Cycling BMX
Martial arts	Boxing Judo Taekwondo Wrestling	Football
		Handball
		Hockey
		Table Tennis
		Tennis
		Volleyball
		Weightlifting

Table 3 – Bitrate/Resolution pairs for the transcoded videos

Bitrate (kbps)	Resolution (pixels)
1450	848 × 480
600	424 × 240
350	320 × 176
190	320 × 176

Note that these four bitrate-pairs selected for the transcoding correspond to the average bitrates commonly used in this type of streaming applications. The resolutions are not the standards commonly used, e.g. CIF and QCIF, because they are calculated by the IIS Smooth Streaming, which gives the best resolutions for each bitrate.

5.3. Selection Criteria for the Observers

The test group should be formed by at least 15 observers according to ITU-R [12]. However, observers have to meet a certain set of prerequisites in order to be selected to participate.

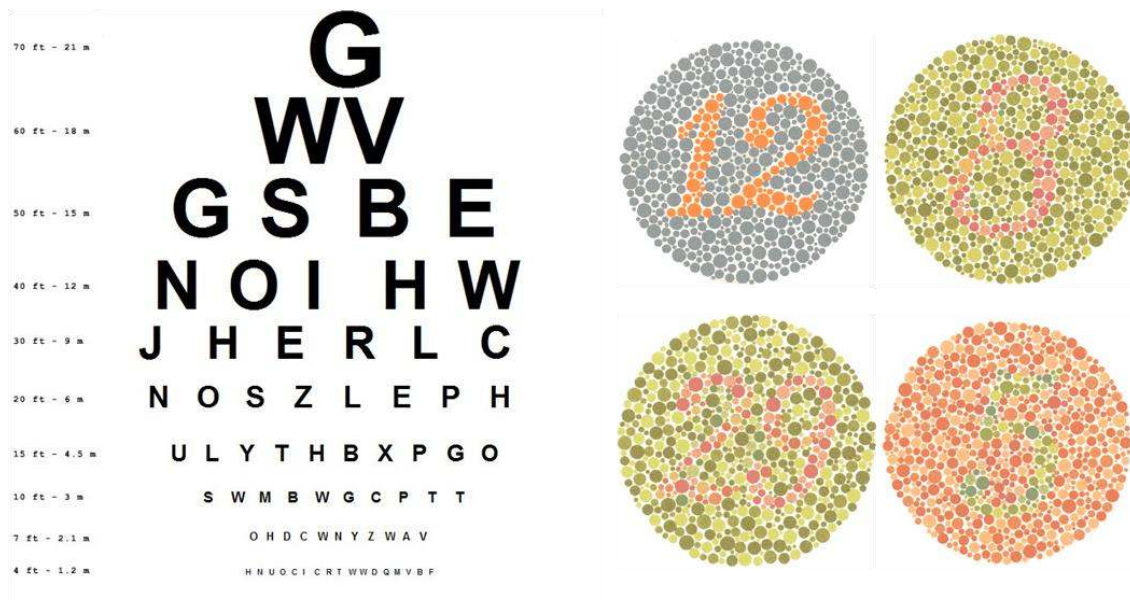
The selection was made in two steps:

4. Initial survey;
5. Visual acuity.

The initial survey was an online survey sent to a wide group of students from the two *campi* of a University, where the candidates were asked about their name, gender, age, profession and e-mail address. In this survey the candidates were asked to rank from 1 to 5 their interest level on the thirty-two typical sport events available. This information was useful to:

- Name: identify the observers at the session day;
- Gender: validation of the universe of answers. It is expected that the distribution among genders has more men than women, in a proportion of 80/20 due to the nature of the typical gender distribution in the courses at the University;
- Age: distribute the observers by age range;
- Profession: identify video quality expertise. As referred in ITU-R [12], observers should be non-expert, in the sense that they are not directly concerned or related with video quality as part of their normal work. Preliminary findings suggest that non-expert observers may yield more critical results with exposure to higher quality transmission and display technologies.
- E-mail: schedule the test session with the candidates whose responses were validated;
- Rank their interest level: create a database as described in the topic 5.5. Architecture.

For the visual acuity (including normal color perception) the candidates are subject to a simple vision test, at the test session day, by using the Snellen chart and the Ishihara plates, as depicted in Figure 7. This vision test is of surmount importance since observers must be in their perfect visual conditions to ensure adequate video assessment test results. If observers fail either the Snellen or the Ishihara tests they should not be accepted for the QoP group test.



a) Snellen Chart

b) Ishihara Plates

Figure 7 – Visual and color acuity tests

The workflow presented in Figure 8 describes, in a simple way, all the process tasks for the selection of observers for the subjective tests.

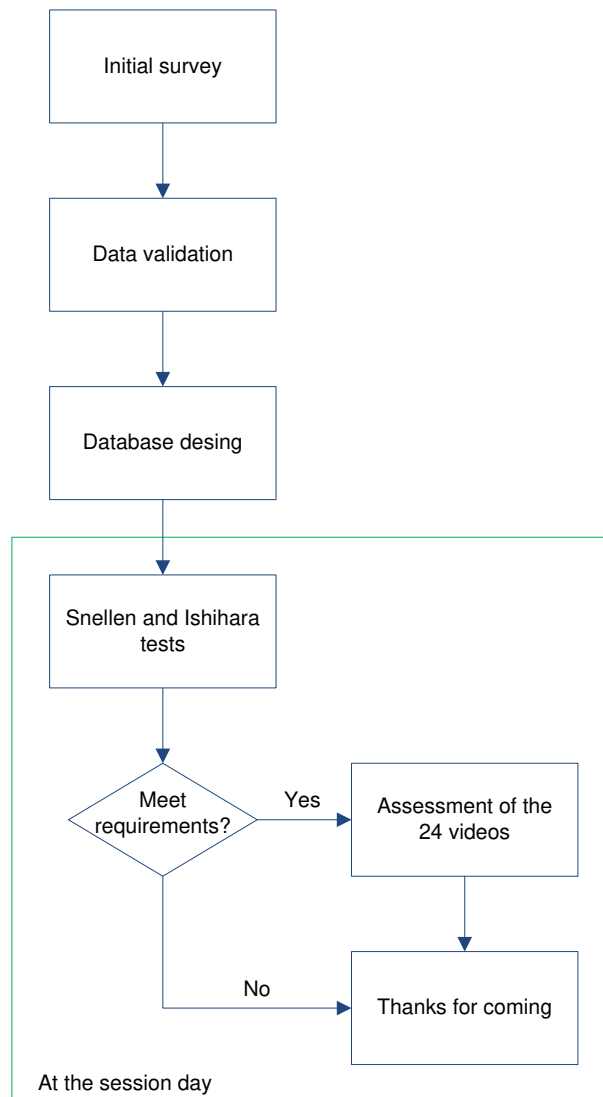


Figure 8 – Diagram used for selection of the observers

The first action was the initial survey for pre-selection. This was an online survey sent by e-mail to the group of students. The survey lasted a week and responses collected using forms from Google Docs service.

After the period for accepting responses finished, the second task was the validation of the collected data. All the data was analyzed to validate the responses in order to avoid fake answers and candidates whose profession was related with video quality analysis/assessment.

Then, with the validated answers, a database was created containing the list of the twenty-four sport clip sequences that each observer should watch in the subjective video quality assessment session.

At the session day, the observers started by testing their visual acuity and color perception using the Snellen and Ishihara tests. Observer that revealed no visual problem, were invited to the room where the subjective video quality assessment would take place.

The selected observers would then watch the sequence of 24 video clips and assess each video at every 30 seconds. This last stage lasts about 15 minutes.

5.4. Video Quality Assessment

During the video quality assessment tests, the observers rated the perceived video quality at the end of each video clip, i.e., at every 30 seconds by selecting a rank value on a small window that pops-up over the client user interface (UI) showing a star scale between 0 and 10, as illustrated in Figure 9. When the ranking window pops-up, the video sequence is paused to allow the observer to judge the viewed sequence and to rate it, avoiding therefore any type of pressure.



Figure 9 – Video quality assessment star-scale

The assessment star scale is a one-click scale. Observers just need to choose the rating value by clicking over the respective star. After clicking, the ranking window automatically closes and the video sequence restarts for another 30 seconds clip. Each assessment process is cycled during the whole test session until the whole video sequence is watched and rated.

With this method, all observers are capable of rating the videos that they have just watched, with the advantage of their assessment being immediately sent back to the streaming server database for further statistical analysis at the end of the test session.

This type of subjective analysis is a NR method, since the observer only has access to the decoded video.

Among the subjective metrics mentioned in Chapter 4, the method used for these tests is a trade-off between the SS and SSCQE metrics. As previously described, these metrics only require available the decoded video, despite of having different rating scales. Another difference is that in the SSCQE videos must correspond to a PS/QS pairs.

The option was therefore to use a discrete scale, such as the one used by SS, and the videos arranged with a PS of sports and a QS of bitrate. For the subjective test, these PS/QP pairs are watched in a pseudo-random order.

5.5. Architecture Used for the Assessment Tests

The architecture for the assessment session is very simple and is composed by a web streaming server and N client computers, as illustrated in Figure 10.

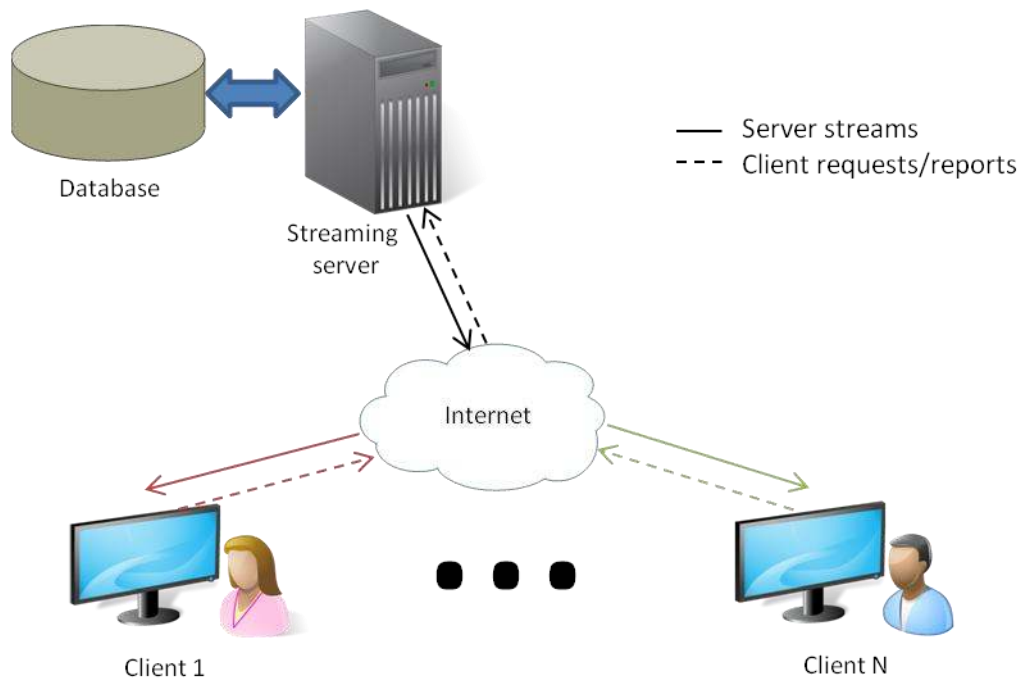


Figure 10 – Test session architecture for Media and Stream Adaptation

The web streaming server stores all the available video files used for the tests and provides a database to register all the ratings given by each observer via the respective client user interface.

The communication between the streaming server and each client was bidirectional and in unicast mode, allowing the video streams to be distributed to all clients and to collect client data using the same network link.

The streaming server is a web server and clients use a web browser to request streams from the server. There will be as many client computers as the number of observers for the tests to be performed individually. There was no special requirement for the client computers, as any current desktop type is suitable for the purpose. The computers used in the test environment were however all-equal in their characteristics and were connected via a Fast-Ethernet network switch.

5.5.1. Streaming Server

The Streaming Server is setup as a Web Server with the characteristics specified in Table 4.

For the Streaming Server, it is required the installation of Microsoft's IIS 7.0 with IIS Smooth Streaming extensions.

Table 4 – Characteristics of the Streaming Server

Operating System	Windows Server 2008, 32 bits
Processor	Intel Core i7 @ 2.67 GHz
Memory	6 GB

5.5.2. Database

A database (DB) was built to store the classification given by each observer to each of the twenty-four video clips. Before the test session, the DB was populated with the following data:

- The ID of each observer;
- A list with the six sports that each observer must see;
- The order of visualization of all the twenty-four videos.

The list with the six sports contained the three sports that each observer likes more and the three that he/she likes less. This is possible since, from the initial survey, each observer gave his/her preference on the 32 sports available. Thus, each user watched the video sequences that best fitted their profile.

This decision of showing the three sports that the observer likes and dislikes more, follows the work described in [8]. It was expected that the observers would tolerate more errors (greater image degradation) in their favorite sports and give lower scores to the sports less desired.

This type of behavior is present nowadays, for example, in all users of YouTube. They do not mind watching a video with a quality that leaves much to be desired because they need to do so, since otherwise they would not have access to the contents that are available.

5.5.3. Client

For the Client computers, as already mentioned, any common standard desktop computer, supporting any current web browser is considered adequate. However, the following configuration is necessary:

- A web browser – to access the Web Streaming Server;
- The Microsoft® Silverlight plug-in installed – to play the available video streaming;

For the display/assessment of the videos, an application was developed by using the Microsoft® Silverlight framework. The application has three different stages:

1. Log in;
2. Player;
3. Log out.

The log in to the test session just required introducing the ID of the given observer. This allowed the application to identify the database record with the assessment video sequence for that specific observer. The observer is only informed about his/her ID after successfully perform and pass all visual acuity tests.

After log in, the observer enters in the viewing environment, which is a player that shows the videos in full screen mode and has no trick-function buttons for interaction. The interaction with the application occurs only at the end of each video, when a window pops-up with the assessment scale, as shown in Figure 11. The log in and log out screens are depicted in Annex 2.



Figure 11 – UI appearance when the assessment windows pops-up

The log out is done automatically by the application, as soon as the observer classifies the last video clip in the sequence.

5.6. At the Session Day

There were performed two test sessions in different rooms, as such, computers and monitors used in each session were different, as Table 5 shows.

Table 5 – Characteristics of the used computers/monitors at each test session

	Monitor Gateway, 4/3 ratio
Room 1	Processor Intel® Pentium® 4 @ 3.00 GHz
	1.50 GB RAM
	Monitor Acer, 16/9 ratio
Room 2	Processor Intel® Core™ 2 Quad @ 2.83 GHz
	3.00 GB RAM

The ambient light was monitored by using a light meter – Center ® 337 Light meter – and was maintained at an average of 200 lux, as stated in the ITU recommendation [12] for Home Environment.

To ensure that all monitors were equally graded, the automatic mode, AUTO, was chosen. Thus, all monitors were with a brightness of 80% and a contrast of 50%.

Prior to each test session, the observers have tested their visual acuity and, if the tests results were successful, they received their own ID to log in the test application. The observers were then carefully introduced to the methodology of video quality assessment and instructed on the grading scale, the actions to take and the duration of the test session. Also, in this first stage, a training sequence with four videos was shown to clarify any doubts that might arise. The data issued from these training sequences was not considered in the results of the effective test sessions. This introduction stage took around five minutes.

After the introduction stage, the assessment session began. Each assessment session lasted 15 minutes during which the observers were only concerned about the video quality assessment they rated at the end of each video sequence. The twenty-four video clips were shown sequentially, but with short intervals between each one to allow the observer to assess the content just watched. The four pairs rate/resolutions were shown in a random way to avoid ranking by impulse. If the videos were always shown with the same sequence, the observer would figure out which would be the next video. Assuming that the first video was the best one and the quality was degraded successively, the observer would rank a lower value in the next video because he/she would have “learned” the sequence order.

5.7. Summary

In order to carry out the subjective tests, a lot of decisions were required in order to be able to perform it.

The first decision was related to type of video material to be shown. And the decision has been:

- Sports videos would be shown because My e-Director 2012 will cover the Olympic Games;
- Each video would have 30 seconds to capture the observer's attention, since this is the average time used on advertisements;
- Four pairs of bitrate/resolution that are typically used on web streaming would be an adequate choice and the video clips shown in a random way.

Then, the selection of observers was studied, and the decision was for a two-stage selection:

- An online survey for pre-selection and database creation;
- The verification of visual acuity at the session day.

Another key decision was on the classification method to use that would fit better with the type of test session and the way it could be implemented therefore, the decision was the following:

- The grading scale would be an eleven points scale, where 0 is the worst possible rating and 10 the best one;

- The assessment window would be a pop-up window and it would appear at the end of each video clip to enable observers to rate without pressure;
- The assessment would also be a one-click scale for an easier and intuitive use.

Finally, the type of architecture to use was chosen. The client-server architecture with dedicated connection was considered the best to fit for this situation.

After all these decisions taken, an online survey was sent to a potential test group and the data collected was analyzed. A database was created and all the selected respondents were contacted via email to schedule the sessions.

At the session day, the right execution of the tests was ensured and the necessary data to infer about how the degree of interest influences the content classification was obtained.

Chapter 6 – Data Analysis

Chapter index

- 6.1. Introduction, 52
- 6.2. Data Analysis from the Initial Survey, 52
- 6.3. Data Analysis from the Video Quality Assessment, 54
- 6.4. MOS Estimation, 59
- 6.5. Summary, 70

6.1. Introduction

By analyzing the data from the initial survey, it is possible to identify the sports that young people, in particular, like more/less and the differences in interests between women and men. Through this type of analysis, behavioral and cultural interests can be also estimated.

With data collected in subjective tests, it will be possible to verify whether the type of content influences the subjective rating. In case it does, it is expected that for sports with lower interest levels the rankings should also be lower than the rankings for sports with a high interest levels.

From the analysis of the results exposed in this chapter, the formulated hypothesis that the interest level in content type would positively influence subjective ratings could be confirmed. And, this influence happens because observers that have more empathy for a type of content (in this case, certain sports), are willing to watch them on almost all occasions. But, when observers do not feel engaged by the content, they will only watch it if the quality is very good. Otherwise they will have a very critical attitude and will rate it below average.

6.2. Data Analysis from the Initial Survey

From the initial survey, 268 responses were collected but only 260 were validated. The responses that were invalidated were all fake answers.

The participants in the survey were mostly students with ages between 18 and 25 years old. The graph in Figure 12 shows the distribution of ages by groups from the survey responses.

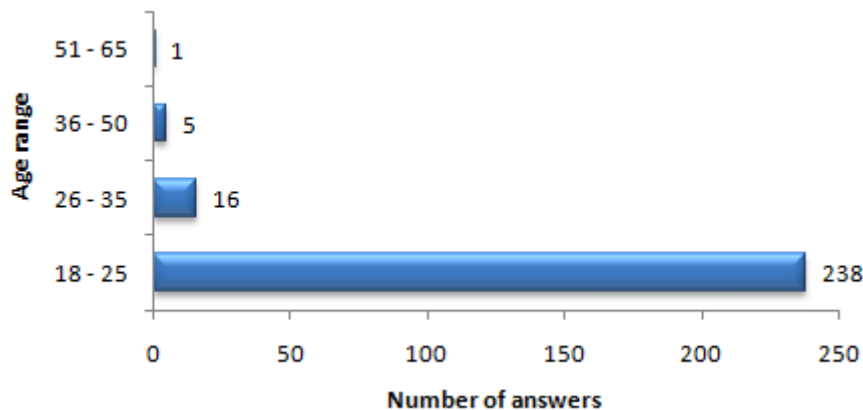


Figure 12 – Age distribution over all the age groups

As expected, the percentage of men's answers is much higher than the women's answers, as illustrated in Figure 13. Although this is an atypical distribution, the results are consistent for the study universe since students were from courses that have a lower percentage of women, (typically between 6% and 18%).

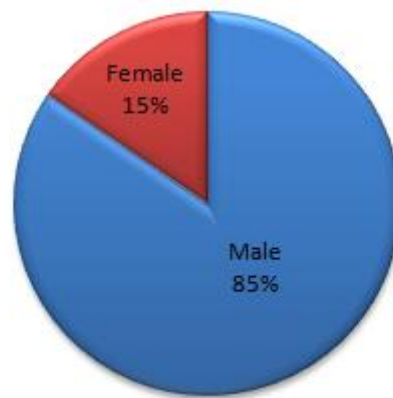


Figure 13 – Distribution by gender

From the collected answers, not surprisingly, the less popular sports were definitely Martial Arts, Weightlifting and Gymnastics. On the other hand, the sports with the highest interest level were Football, Volleyball and Tennis, as shown in Figure 14.

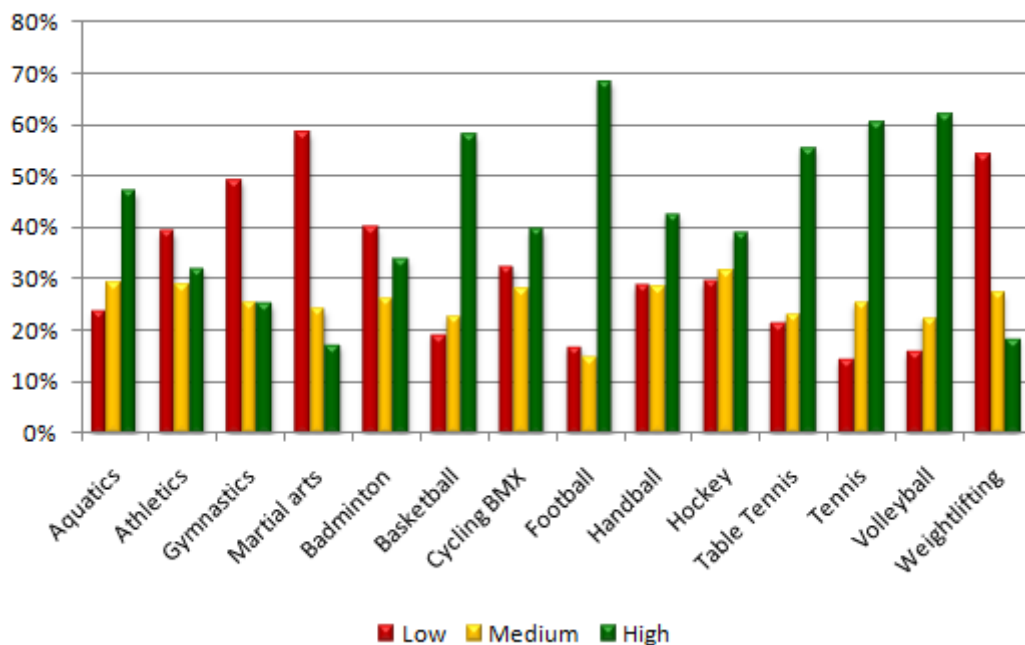


Figure 14 – Distribution of the sports interests by the levels low, medium and high

The interest level scale used, places “low interest” matching levels 1 and 2, “medium interest” matching level 3 and “high interest” matching levels 4 and 5. These five levels were labeled as follows:

- Level 1 – hate;
- Level 2 – don’t like;
- Level 3 – indifferent;
- Level 4 – like;
- Level 5 – love;

Comparing the interest level by gender, it appears that the sports with low interest level were similar to both genders. However, for sports with high interest level there are differences between genders, as evidenced in Table 5.

Table 6 – Woman/Man low and high sports interests

Woman's interests		Man's interests	
Low	High	Low	High
Weightlifting	Volleyball	Martial Arts	Football
Martial Arts	Aquatics	Gymnastics	Tennis
Athletics	Basketball	Weightlifting	Volleyball

Apart from the comparison of interest level by gender, it would have also been interesting to compare interest levels between cultures. The same survey developed in other European country is not expected to bring many changes, because football is sovereign throughout Europe. But the same survey realized in the USA would reverse the results, since football is undervalued compared to Martial Arts, which has many followers in that country, namely on boxing and wrestling modalities.

6.3. Data Analysis from the Video Quality Assessment

From the 260 respondents of the initial survey, only 25 were validated for the subjective tests. However, one of the candidates failed in the visual acuity test of Snellen and, as such, could not be included in the group of observers. A total of 24 observers have effectively participated and 144 videos were viewed. The percentage of videos shown, grouped by interest, is depicted in Figure 15.

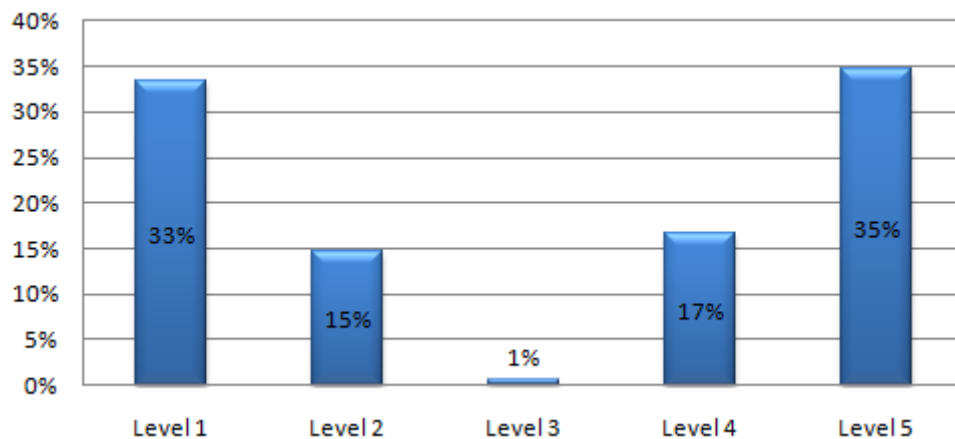


Figure 15 – Percentage of videos shown by Interest level

In these 144 videos, the sports corresponding to low interest levels were Boxing, Wrestling and Judo, as illustrated in Figure 16, and the most viewed sports, corresponding to high interest levels were Football, Swimming and Tennis, as shown in Figure 17.

Note that these results are not too far from those observed in the initial survey. However, it must be kept in mind that only 10% of the total universe participated in the tests, what may justify some of the differences found.

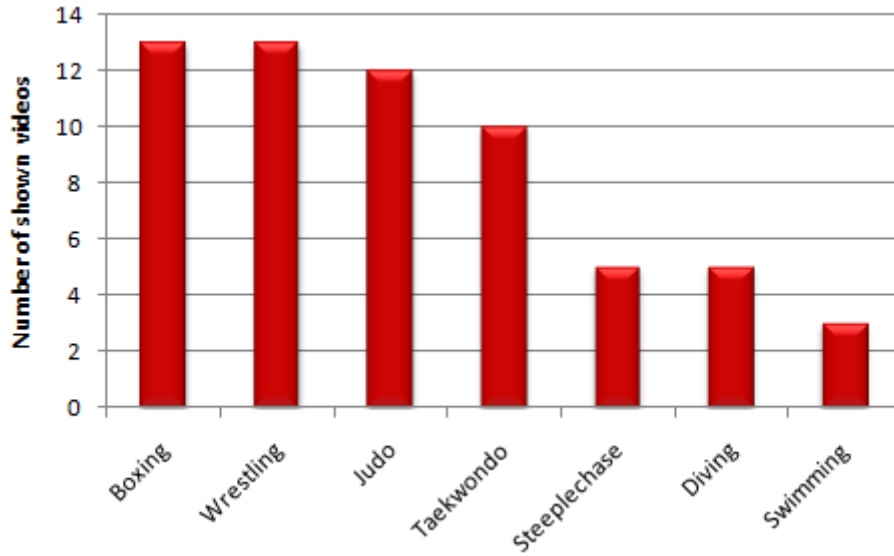


Figure 16 – Sports in the low interest levels

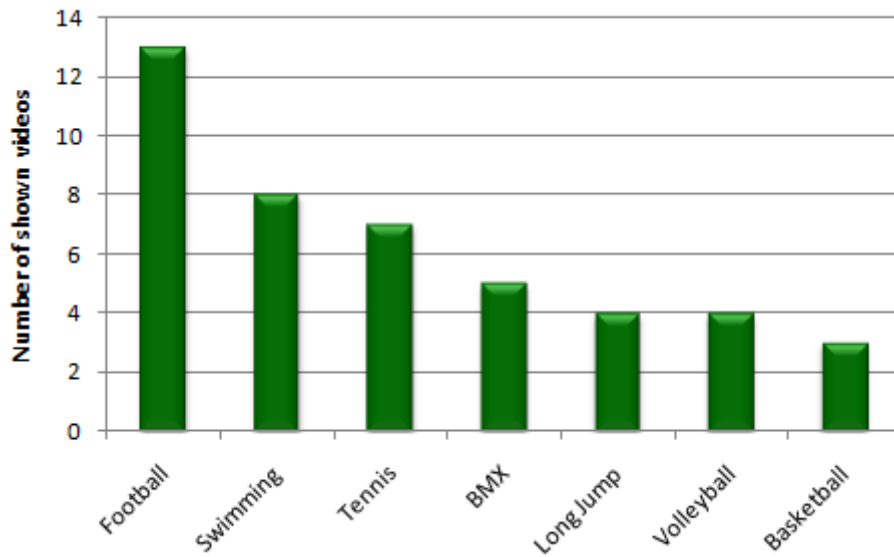


Figure 17 – Sports in the high interest levels

Figure 16 and Figure 17, only show the sports that were watched by more than 2 observers. With this condition, it is possible to derive conclusions and compare results. Note that, for the statistical analysis, only the sports that were watched by more than one third of the observers were considered.

Computing the average obtained for each interest level, regardless of the sport modality, produced the results plotted in Figure 18 turning evident that the observers tend to value more (around two values scale points higher) a video with the same bitrate, just because they have higher interest on it.

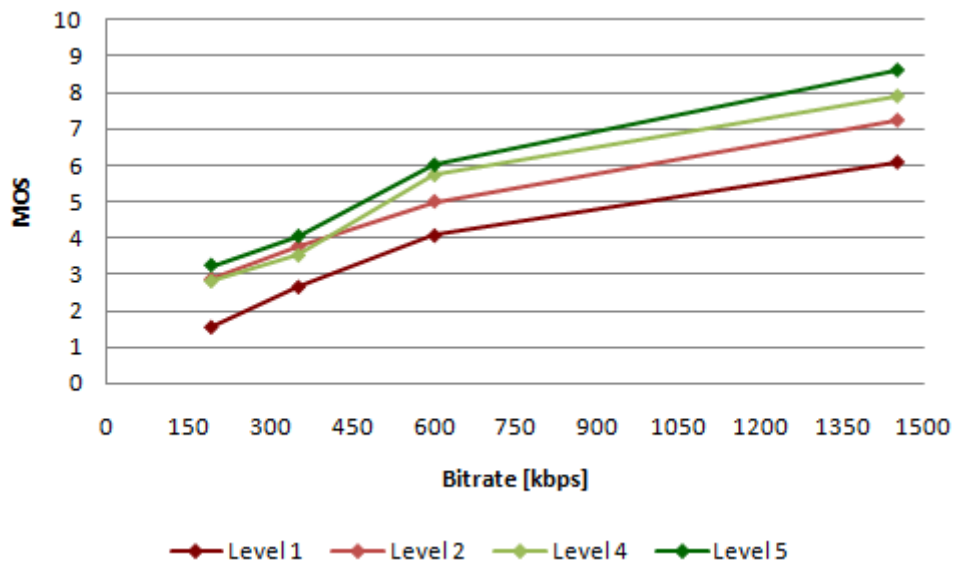


Figure 18 – Comparison between the average values assigned to each interest level

Notice also that, due to lack of sufficient data, level 3 was not considered in this analysis.

From the graph in Figure 18 it is possible to observe that:

- Within each interest level, as the bitrate increases the classification also increases and the curves have a logarithmic behavior (corresponding to what was expected);
- Keeping the bitrate constant, as the interest level increases the ranking also increases. This indicates that the interest level positively influences the rank (otherwise all curves would overlap at multiple points);
- For the lowest bitrate (190 kbps), with the exception for the interest level 1, the grading is independent of the interest level within an average of 3 point scale values;
- For the highest bitrate (1450 kbps), the classification is clearly different for each interest level. Notice that the difference between level 1 and level 5 is about 2.5 point scale values;
- Except for level 1, the difference between the two lowest bitrates (190 and 350 kbps) is lower than 1 point scale value. However, significant differences in ratings can be noticed for the others bitrates.

The graphs plotted in the Figure 19, show the standard deviation for the interest level curves previously analyzed in Figure 18. It is clear that for level 1, the worst bitrate has the lowest standard deviation. This shows that despite of the sport, observers tend to agree in the assessment. The best bitrate has, in contrast, the highest standard deviation. This quite interesting result confirms that observers were able to watch a sport they dislike but if it was shown in high quality. Otherwise they would not watch it. It is also important to observe that, as the bitrate increases the standard deviation also increases.

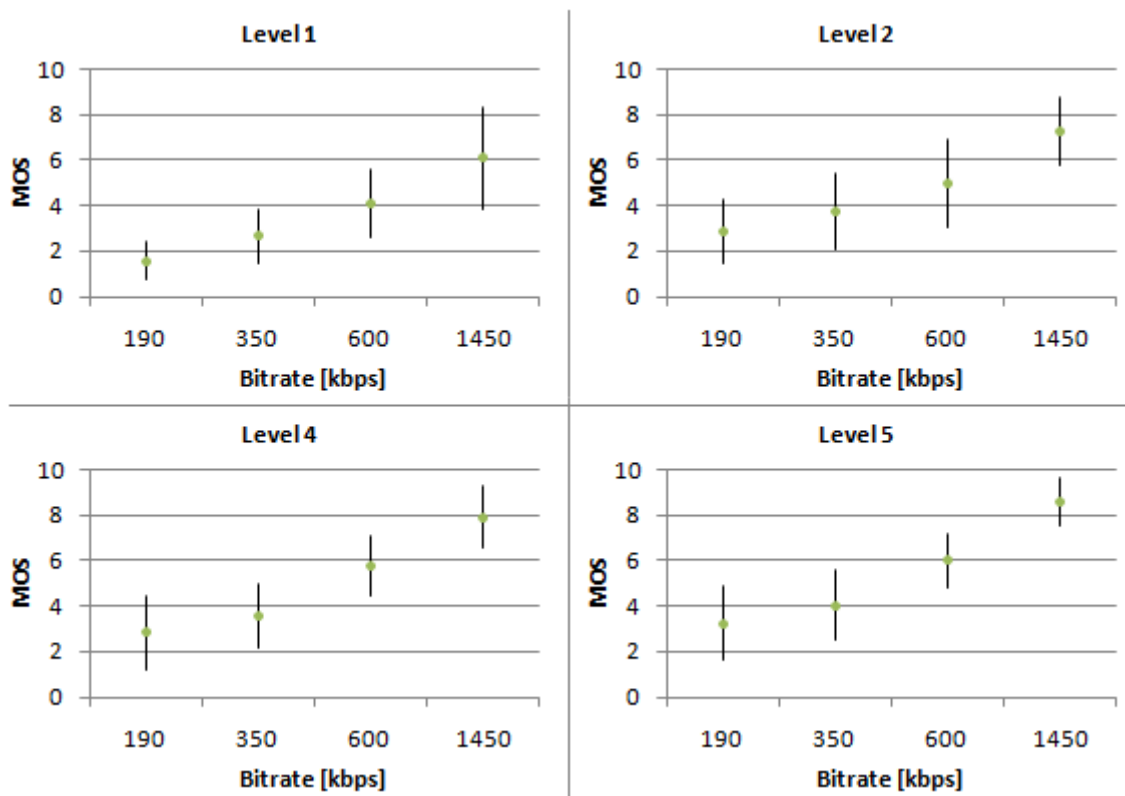


Figure 19 – Standard Deviation for each interest level curves

For level 5, the relation between the standard deviation and the bitrate are the opposite of what happens for level 1. For the highest bitrate, the standard deviation is the lowest and for the lowest bitrate the standard deviation is the highest. These results show that regardless of the sport type, observers tend to agree in the assessment for the best bitrate. On the other hand, for the lowest bitrate, observers can be radical on their assessment, confirming that, only for some sports the errors are tolerated. Although the standard deviation decreases as the bitrate increases, for the two lowest bitrates, the standard deviation is approximately the same.

For level 2 and level 4, the standard deviation is, somehow, independent of the bitrate because all bitrates have almost the same standard deviation. However, level 4 has a lowest standard deviation than level 2. Therefore, it may be concluded that for a higher interest level, observers are inclined to give more consistent ratings.

By comparing two different levels for the same sport it is possible to see the differences between rankings from the changes occurring in interest levels, as plotted in Figure 20 where the two lowest interest levels, level 1 and level 2, for boxing are confronted.

The difference between rankings is perfectly distinguished, with an average difference of 1 point scale value among all bitrates, except for the lowest one, where the difference is about 1.5. These values are consistent with the averages computed for each interest level (Figure 18).

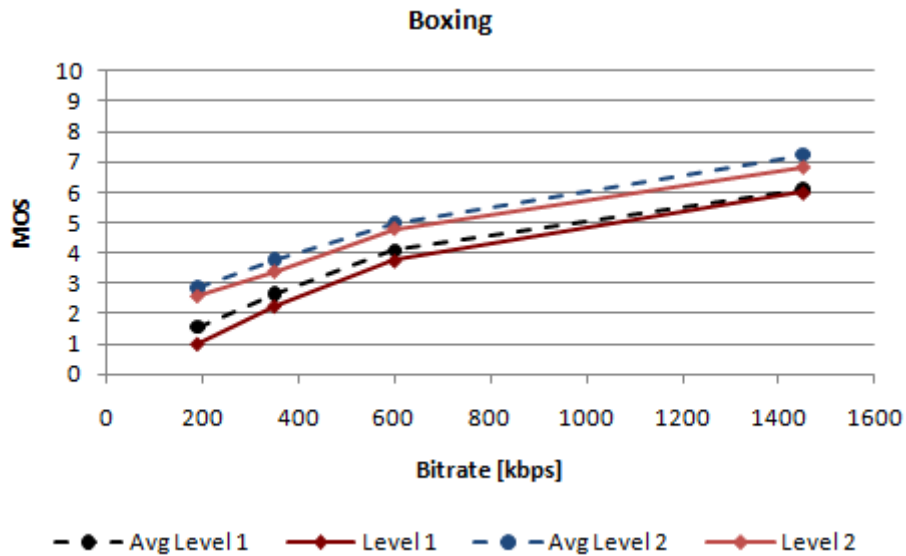


Figure 20 – Level 1 vs. Level 2 for Boxing

Figure 20 also plots the average curves for each interest level (the dashed lines) for a better comparison. Both averages curves provide a good approximation for each Boxing interest level. This very positive result shows that by knowing the averages curves, it is possible to estimate the MOS for a given sport, with a low margin of error.

Figure 21 shows, side by side, the three most watched sports with low interest. For the highest bitrate, 1450 kbps, all sports have a ranking almost coincident with the average for interest level 1. For 600 kbps, the ratings are also close to the average value. On the other hand, for the two lowest bitrates, the ratings are more dispersed, despite being within the limits. This result reinforces the conclusion previously obtained, that it is possible to estimate the MOS for a given sport by knowing the averages curves. However, for the two lowest bitrates, the estimation bears more errors because, for these bitrates, the observer tolerance has a high influence in the results.

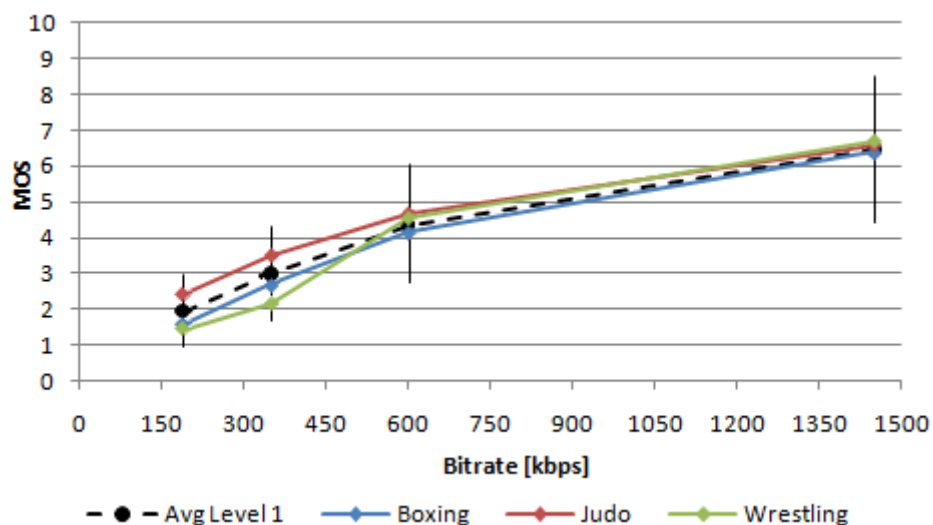


Figure 21 – Comparison between the three sports with low interest

In the graph plotted in Figure 22, the same comparison is made for the sports more watched but with high interest.

From in Figure 22, it may be observed that the rankings given to Football and Swimming are very similar between them for all bitrates and are in agreement with the average result. However, although Tennis ratings are within the limits for the highest bitrate, for the other bitrates, 190 and 350 kbps, the ratings are outside the limits. This phenomenon can be explained by the high movement that characterizes tennis as players of the game are constantly running from one side to another of the court, the ball is very small and can reach very high speeds.

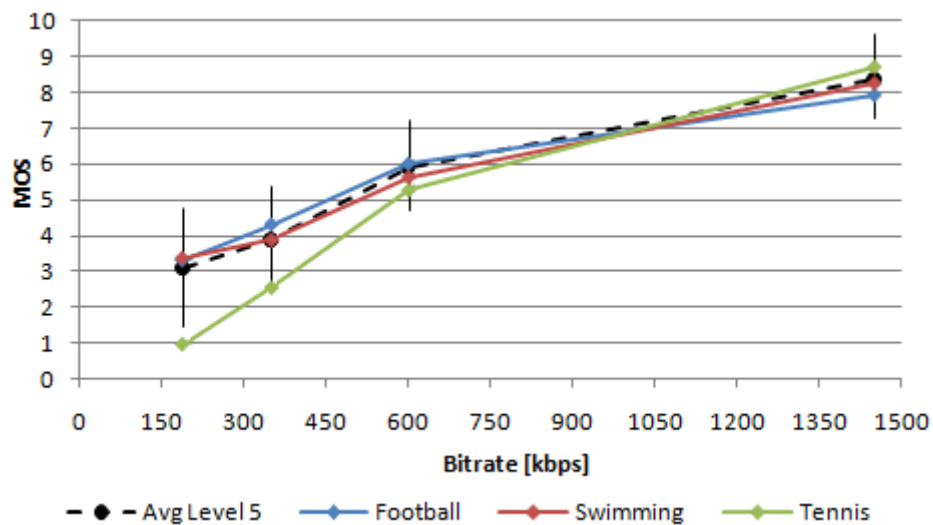


Figure 22 – Comparison between the three sports with high interest

The tennis ball has a diameter of 6.35 to 6.67 cm, a weight of 57.7 to 58.7 g and can reach speeds greater than 150 km/h. The current record is held by Ivo Karlovic in the 2011 Davis Cup, where his service reached the 251 km/h speed mark. The problem with this sport modality is that for the lowest bitrates, 190 and 350 kbps, observers hardly see the ball due to the loss of detail with the decrease of bitrate. Despite this adverse effect, due to the high speeds reached, the temporal and spatial displacements of the ball between video frames, is huge. This phenomenon makes the encoder share the few bits available to cover too many details, resulting almost impossible to encode the image of the ball with the available bits and, for these bitrates, the observers hardly see the ball.

In football, although the ball looks very small on a TV screen, it rarely reaches speeds above 100 km/h. Therefore, even for the lowest bitrates, observers can always see the ball. The same situation is valid for Swimming where the detail can be high and movement is very low. Even for the lowest bitrates, observers can always know who and where is their preferred swimmer.

6.4. MOS Estimation

With the collected data it is possible to express the MOS as a function of the bitrate (R) and the interest level (IL). The goal is to establish an additive formula, where the first term depends only on R and the second on IL, which can be expressed by the following equation:

$$MOS = f_1(R) + f_2(IL) \quad (12)$$

Since f_1 is a function of R , it has a logarithmic behavior because, from the moment that the observer thinks that the quality is maximum, the increase in the bitrate does not bring any perceived change in quality.

Plotting the trend lines that better fit each interest level allows to determine eventual relationships between them, as represented in Figure 23.

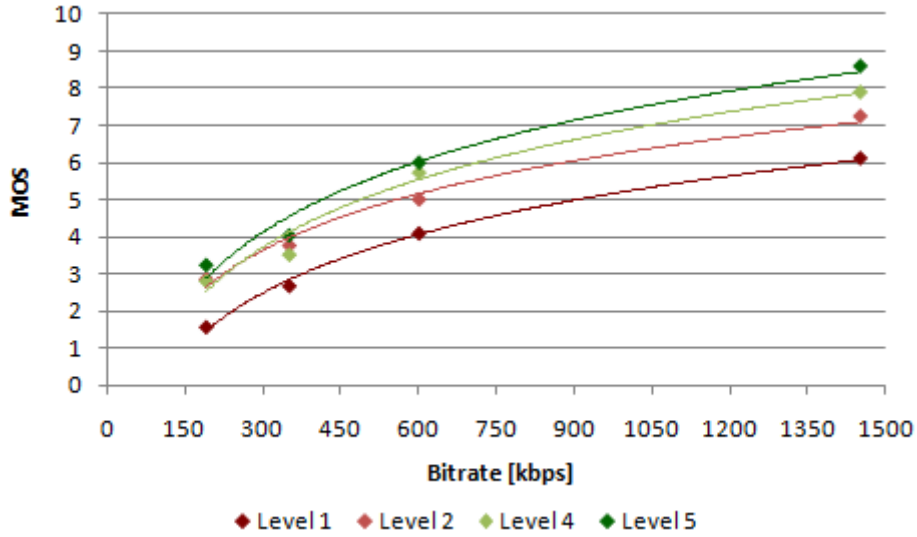


Figure 23 – Logarithmic trend lines for each interest level

Although the trend lines pass through all points for level 1 and level 2, for level 4 and level 5 the same is not true, the 350 kbps point falls far below the trend line. This situation may represent a discrepancy between the empiric function factors and the practical results for the highest interest levels.

The equations for each trend line are the following:

$$\text{Level 1: } y = 5.2264 \log_{10}(x) - 10.458 \quad (13)$$

$$\text{Level 2: } y = 5.0374 \log_{10}(x) - 8.8389 \quad (14)$$

$$\text{Level 4: } y = 6.0535 \log_{10}(x) - 11.277 \quad (15)$$

$$\text{Level 5: } y = 6.2987 \log_{10}(x) - 11.477 \quad (16)$$

From these, for each interest level, the MOS equation can be written as:

$$MOS = \text{scale_factor} \times \log_{10}(R) - \text{constant} \quad (17)$$

In this formula, despite the constant being negative, it represents a level function. The constant value in the function is used to drive up the mean curve as the level of interest increases. Therefore, the second term of the MOS equation must be a function of IL .

Observing the equations for each interest level, it appears that they are quite similar, especially between the two lowest and the two highest levels, despite low value for the constant factor to level 2. It can then be inferred that it is possible to obtain a function in function of R which approximates the behavior of each interest level equation.

Averaging the trend lines of those expressions, the first term for the general MOS function (18) is achieved:

$$f_1(R) = 5.6540 \log_{10}(R) - 10.513 \quad (18)$$

This new expression describes the MOS as a function of bitrate, regardless of the interest level. Figure 24 shows the comparison between (18) and the experimental curves. As can be seen, expresses the MOS only in function of bitrate is not enough to approximate the experimental results. Calculating the average standard deviation to $f_1(R)$ a value of 0.50 is obtained, this is a huge value. This reinforces the hypothesis that the interest level is an important factor and should be taken into account.

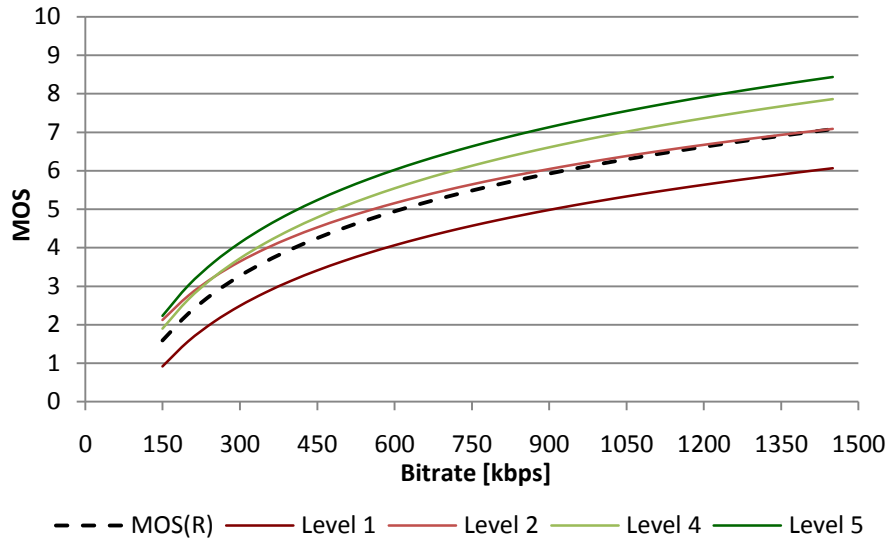


Figure 24 - Comparison between MOS(R) and experimental curves

The second term, that depends on IL, is use to level the MOS, in order to obtain a function close to what was obtained experimentally. Studying the MOS as function of interest levels for each available bitrate the graph of Figure 25 is obtained. This graph shows that MOS also has a logarithmic behavior for each interest level. However, this behavior is less pronounced in the two lowest levels, pointing out to a second term also with a logarithmic behavior.

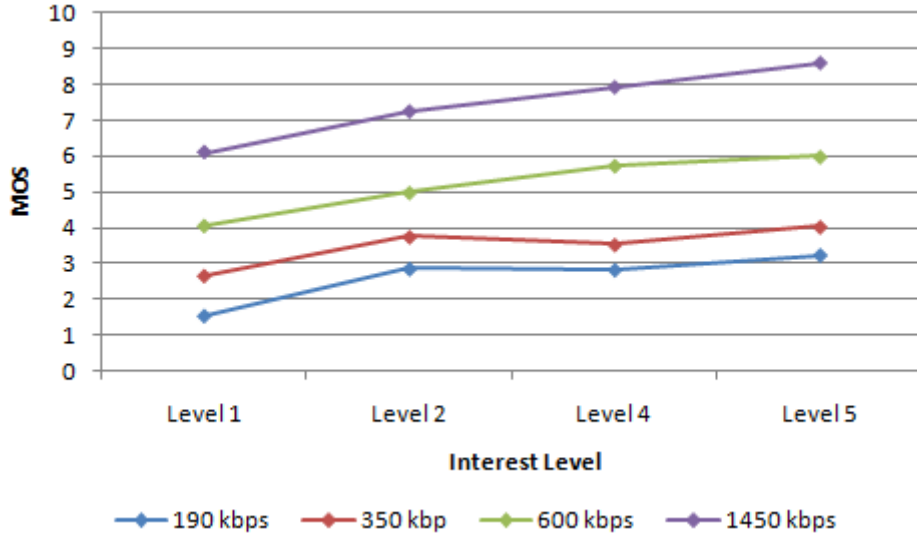


Figure 25 – MOS as function of interest level for each available bitrate

It may be then inferred that the final equation should be on the form of:

$$MOS = 5.6540 \log_{10}(R) - 10.513 + f_2(IL) \quad (19)$$

As the resulting curve should provide a good approximation for the average curves obtained for each interest level, it is possible then to estimate $f_2(IL)$, by keeping in mind that, it is also a logarithmic function:

$$f_2(IL) = 2.6318 \log_{10}(IL) - 1.041 \quad (20)$$

Combining the terms, the empiric MOS formulation can then be expressed as (21):

$$MOS = 5.6540 \log_{10}(R) + 2.6318 \log_{10}(IL) - 11.554 \quad (21)$$

Figures 26 to 29 show the average curve and the MOS computation for each interest level using (21). In these figures it is possible to verify that the new MOS formulation provides a really good estimative for each interest level. As illustrated in Figure 26, for interest level 1 both curves are almost overlap with a standard deviation of 0.10, the lowest obtained. For interest level 2 (Figure 27), the new MOS formulation gives a slightly lower estimative at the 190 kbps, in about 0.5 scale values, and an almost overlap for higher bitrates with a standard deviation of 0.13, note that this value is the higher standard deviation obtained. For interest level 4, Figure 28, the opposite happens, i.e., at the 190 kbps, the new MOS gives higher values but with a difference under 0.5 scale values, here the standard deviation is 0.12. In Figure 29, interest level 5, the new MOS comes slightly higher for 190 kbps and slightly lower for 1450 kbps, but with differences under 0.5 scale values. These standard deviation values show clearly that this new formulation that takes into account the interest levels, gives results very close to the experimental ones, being more than four times lower than the standard deviation obtained for the MOS expressed only as a function of the bitrate.

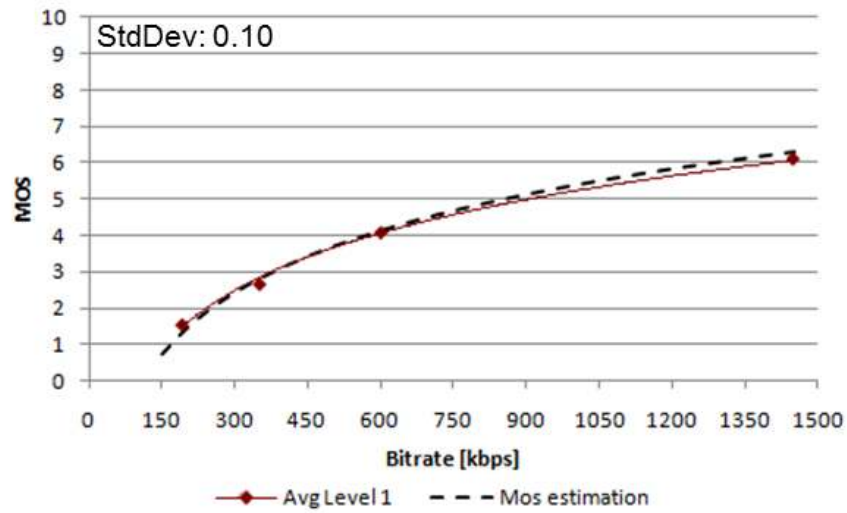


Figure 26 – Average curve vs. MOS estimation each for level 1

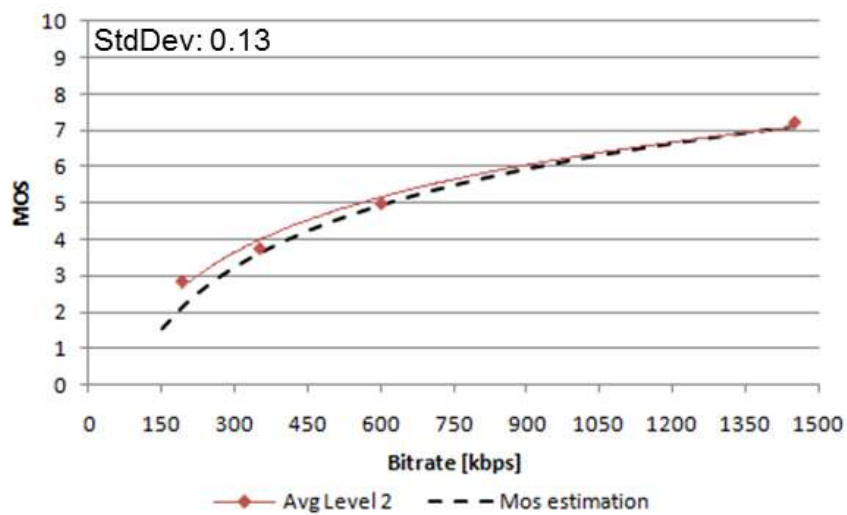


Figure 27 – Average curve vs. MOS estimation each for level 2

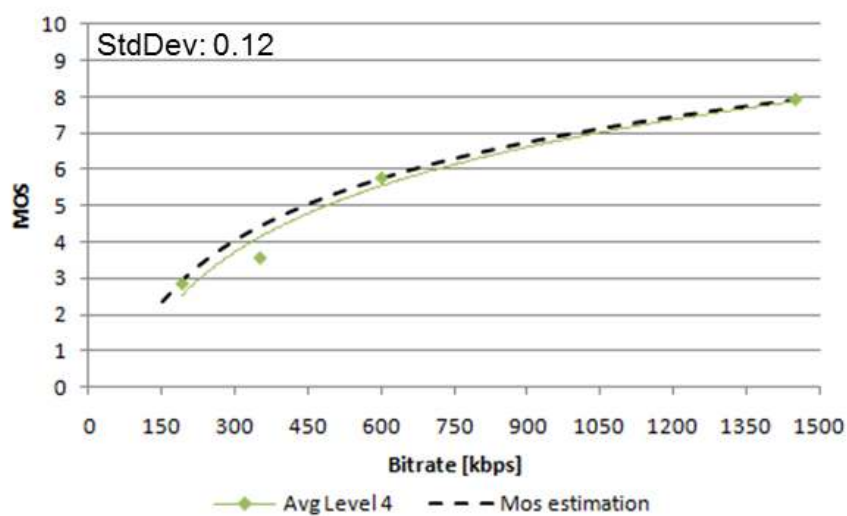


Figure 28 – Average curve vs. MOS estimation each for level 4

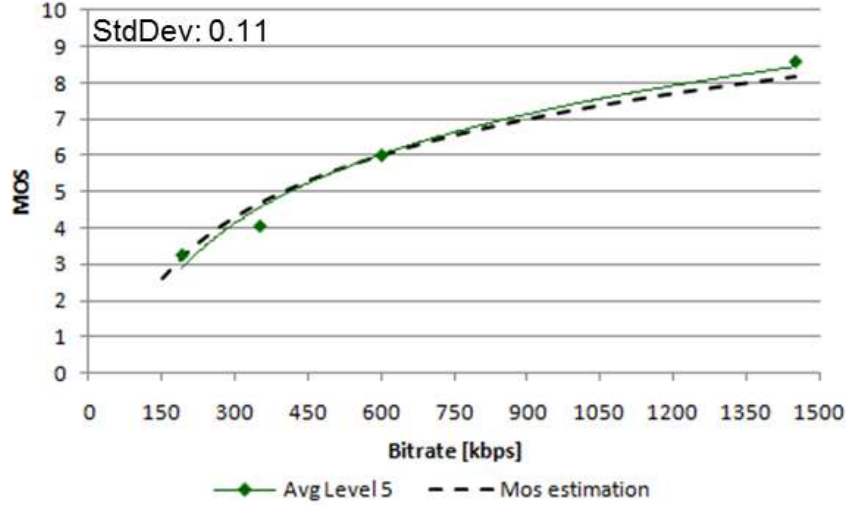


Figure 29 – Average curve vs. MOS estimation each for level 5

Despite these quite good results, there is still the need to introduce a new parameter in the estimation of MOS, related with sports with high temporal activity, such as Tennis. For this sport, as observed in Figure 22, the average curve does not represent a good approximation for the two lowest bitrates. Therefore, it is possible to conclude that, the new parameter must be a function of temporal activity (TA). The temporal activity can be estimated by making the difference, pixel by pixel, between two successive frames. ITU-T Recommendation [13] defines temporal activity as the maximum value of standard deviation found along the video frames, as expressed in equation (22):

$$TA = \max\{std[F_n(i, j) - F_{n-1}(i, j)]\} \quad (22)$$

In this equation, $F_n(i, j)$ is the pixel at the i th row and j th column of n th frame in time.

However, for sequences with changes of camera, the resulting temporal activity can have a high value even if the video has a low temporal activity. Since in sports capturing many changes of camera may occur (or even scene cuts, e.g., sports summary), in order to minimize and smooth this effect, the 99% percentile should be applied to the global temporal activity.

Figure 30 puts side by side the global temporal activity (g_act) calculated according to ITU-T Recommendation [13] and the 99% percentile (p_99) calculated over the global temporal activity according to [14]. By analyzing this graph, it is possible to see that for Triple Jump the global activity and the 99% percentile are very close. On the other hand, the Floor Exercises register the highest difference between the global activity and the 99% percentile. It is easy to understand that Floor Exercises have a lot of changes of camera, since the camera follows the athlete along the whole exercise.

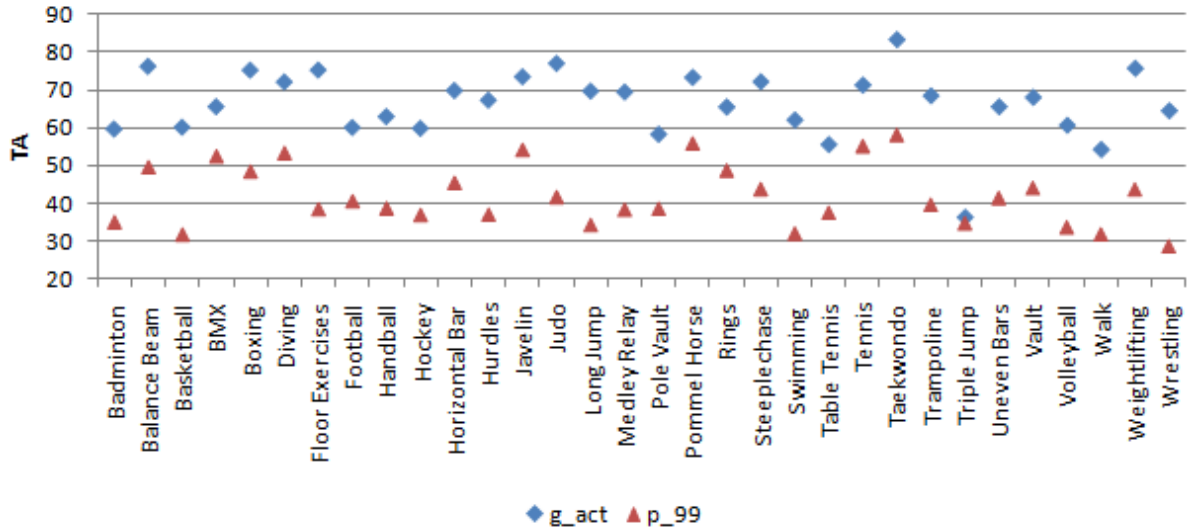


Figure 30 – Global temporal activity vs. 99% percentile

Figure 31 shows the graph obtained with the difference between consecutive frames for the Floor Exercises video capture. In this graph, it is possible to distinguish the peaks of the standard deviation values when a change of camera occurs. If these peaks were eliminated, the highest value for the standard deviation would occur approximately for the 650th frame, with values around 40. Checking this value in the graph of Figure 30, the standard deviation obtained is also around 40, demonstrating that the 99% percentile gives a good approximation to the temporal activity, if the change of camera is not accounted.

The global activity and the 99% percentile values of Figure 31 were obtained using MATLAB® tools according to [14]. Annex 4 provides the entire MATLAB® script code and the method used to calculate the 99% percentile.

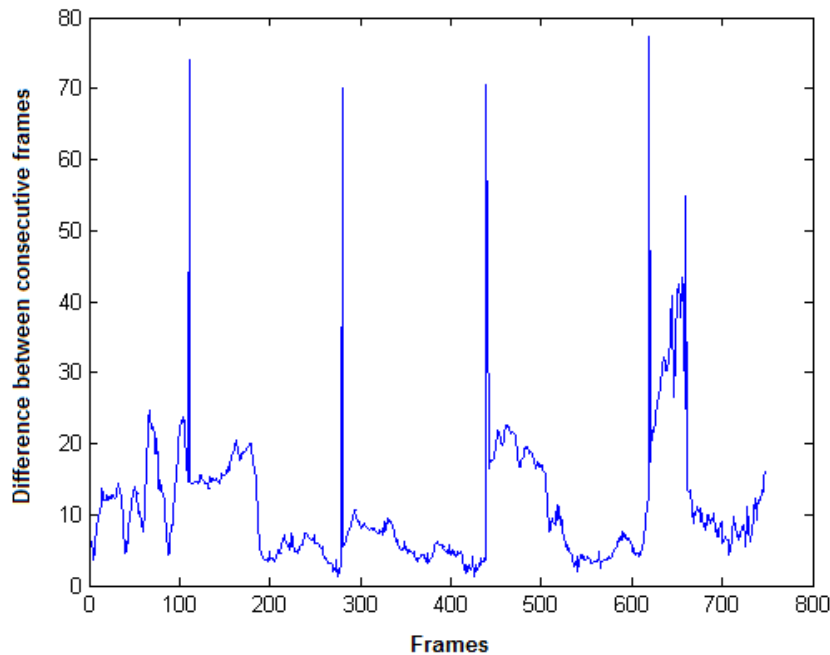


Figure 31 – Difference between consecutive frames to Floor Exercises

Since the 99% percentile gives a good approximation to the temporal activity estimation, the term “temporal activity” will always refer to the temporal activity given by the 99% percentile. Figure 32 shows the temporal activity for the 32 available sports.

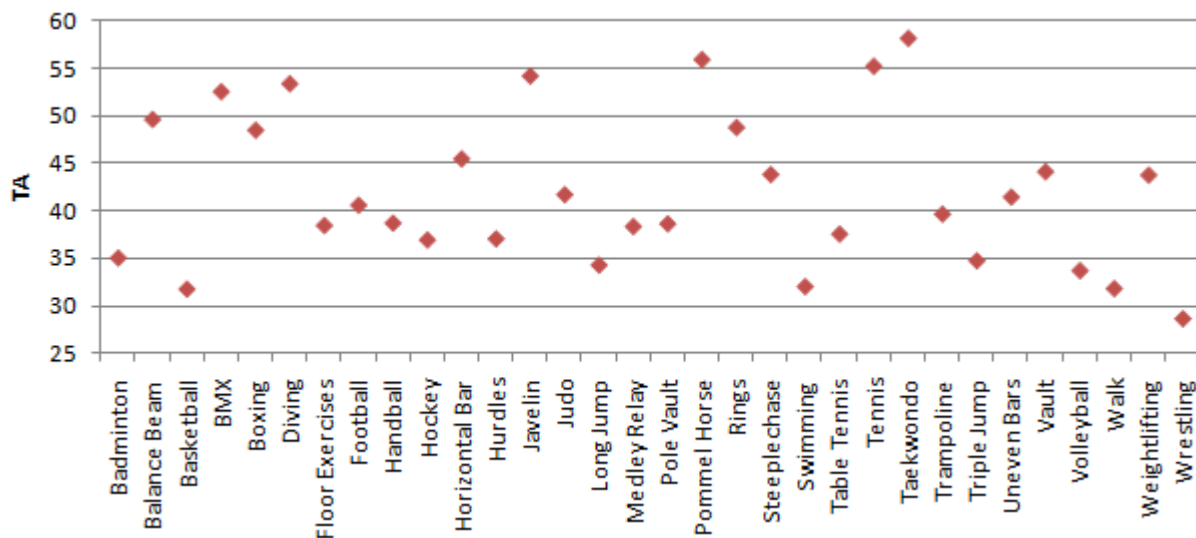


Figure 32 – Temporal activity for the 32 available sports

The problem now, relies in the identification of temporal activity characteristic for each sport. It is quite easy to conclude by observation that Taekwondo has the highest temporal activity, but what about Boxing? Does it have a high temporal activity too, or not?

Grouping sports at a high level, it is possible to establish the following three temporal activity stages:

- The low temporal activity have a $TA < 35$;
- The medium temporal activity have $35 < TA < 50$;
- The high temporal activity have a $TA > 50$.

With this approach, Tennis can be identified as a high temporal activity sport, confirming the experimental verification. In the experimental results, Javelin was also considered a high temporal activity sport, since for the lowest bitrates the javelin cannot be identify in the air, confirming again that the identified temporal activity stages provide a good characterization.

However, for BMX, Diving, Pommel Horse and Taekwondo, the high temporal activity stage does not apparently match, since the experimental results do not reveal such behavior. Although these sports are typically slow movement sports with one or two athletes, it is common to capture the event with several changes of camera. The cameraman is always looking for new plans, making zoom ins and zoom outs. Due to this non-intrinsic behavior of the sport, the difference between consecutive frames can be significant even when the 99% percentile is taken into account.

Figure 40 shows the graph obtained for Pommel Horse, which is full up of peaks. However, only seven of these peaks really represent camera changes. The other peaks are due to the camera

movement to follow up the exercise along the pommel horse. The peaks representing changes of camera are identified in the graph, with tag numbered 1 to 7. These tags correspond to the frozen frames illustrated in Figures 33 to 39:

1. Athlete's presentation to the juries;



Figure 33 – Athlete's presentation

2. A grand plan of the exercise;



Figure 34 – Exercise zoom

3. A detail of athlete's hands along the pommel horse;



Figure 35 – Zoom to athlete's hands along the pommel horse

4. The exercise finalization;



Figure 36 – Athlete overview

5. Athlete's coach's reaction;



Figure 37 – Athlete's coach

6. Athlete waiting for the scores;



Figure 38 – Athlete waiting for the scores

7. Preparation of other athlete to the exercise.



Figure 39 – New athlete presentation

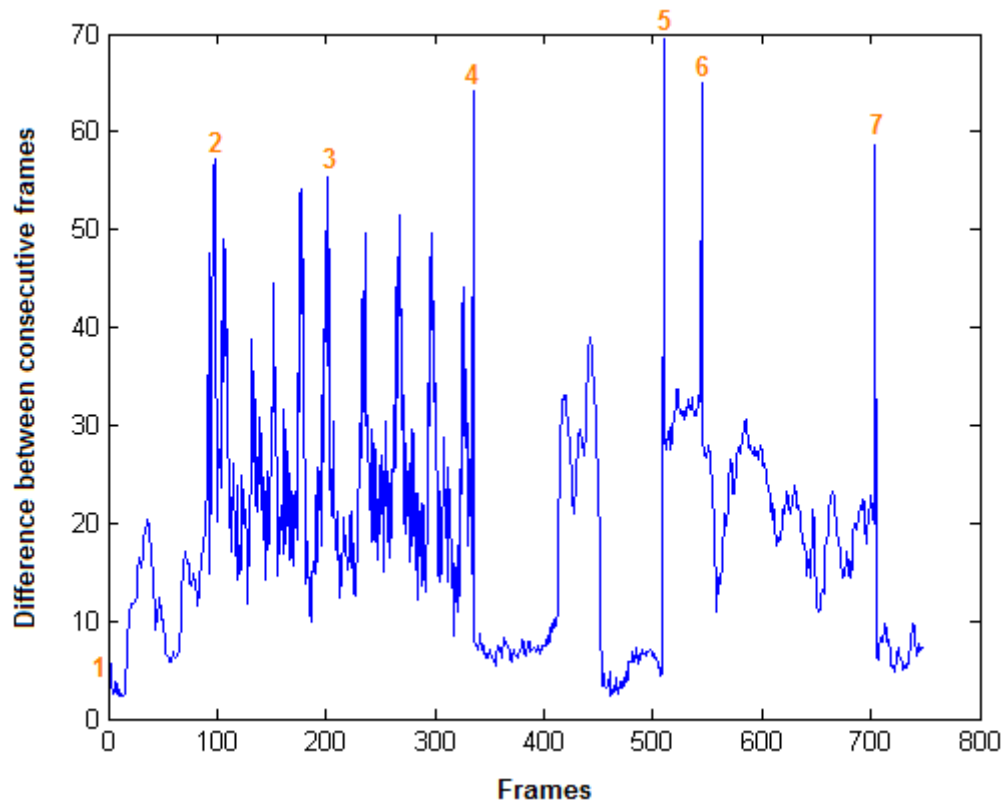


Figure 40 – Difference between consecutive frames to Pommel Horse

At the beginning of the video, the camera is stopped and the athlete's movements are slow and smooth. However, once the athlete climbs to the apparatus and starts his exercise, the camera cannot catch the entire athlete and his head is cut, as seen in Figure 34. Thus, while the athlete shows his exercise, the camera is gently rising, in order to catch the whole athlete. This camera movement is translated through several peaks (between tags 2 and 3).

At approximately the eighth second (frame 200) the camera is switched (Figure 35), showing in detail the movement of the athlete's hands along the apparatus, from one extreme to the other. As previously, the camera follows the athlete's movement, which leads to the existence of intermediate peaks between tags 3 and 4.

In both cases, the intermediate peaks are always lower than the peaks where the change of camera really occurs. This happens when the camera slowly follows the athlete's movement, and despite the difference between consecutive frames, there are always elements in the background that remain nearly at the same position.

The change between peaks identified with tags 4 and 5, are due to a slight change in the camera position, to follow the athlete on his projection to the ground.

As can be observed in Figure 37, after the athlete finishes the exercise, his trainer is focused, leading now to peak 5. The peak 6 is a new change of camera, showing the athlete waiting for the scores, as Figure 38 illustrates. Finally, the last change of camera occurs at the 700th frame, showing a new athlete ready to begin his exercise, as Figure 39 demonstrates. As the camera has remained stable over the last three peaks there are no intermediate peaks between them, unlike what happens between the peaks 2 and 3 and peaks 3 and 4.

This brief description shows that the high temporal activity stage must be analyzed for each case. Despite Pommel Horse having only six explicit changes of camera, the 99% percentile does not eliminate all the existing intermediate peaks, which are due to the camera movements in order to follow the athlete. This explains why Pommel Horse is classified as high temporal activity. The same reasoning is valid to BMX, Diving and Taekwondo. These sports have a significant amount of camera movements creates the intermediate peaks in temporal activity. Due to this phenomenon, the 99% percentile cannot smooth the effect of these peaks. For these sports, the 95% percentile may then be considered in order to reduce the intermediate peaks effect, avoiding these sports being classified in the high temporal activity stage.

Through the temporal activity analysis, only two sports, Tennis and Javelin, were identified at the high temporal activity stage. Since in the assessment tests no data were obtained for Javelin and Tennis only had data at the interest level 5, more test session must be performed to introduce the TA parameter in the MOS equation, due to the lack of data. Since till now only two sports have the designation of high temporal activity, the best solution is to introduce the TA parameter is to develop another formula for sports with high temporal activity. For sports with low and medium temporal activity the empiric formula (21) is used and for sports with high temporal activity a new empiric formula (23) should be used:

$$MOS = \begin{cases} 5.6540 \log_{10}(R) + 2.6318 \log_{10}(IL) - 11.554 & \text{low and medium TA} \\ \text{new equation to develop} & \text{high TA} \end{cases} \quad (23)$$

Due to the lack of data, the new formula for sports with high temporal activity could not be developed in due time for this dissertation, but for future work in this field.

6.5. Summary

In the initial survey a total of 268 responses were obtained, of which 260 were validated. From the analyzed sports Martial Arts, Weightlifting and Gymnastics in general, were the sports with the lowest interest. At the top of sports with high interest there were Football, Volleyball and Tennis.

Comparing the interest level by gender, it appears that the low interest sports, Weightlifting and Martial Arts, are shared for both genders. However, for the high interest, women and men only have in common the Volleyball.

From the 260 respondents of the initial survey, only 25 participated in the subjective tests. With the data collected it was concluded that the interest level has a positive influence in the subjective assessment. As the interest level increases, the subjective assessment also increases. Other interesting conclusions were also obtained, such as:

- Except for the lowest interest level, for 190 and 350 kbps, observers almost do not notice differences in quality, since, the rating assigned was less than 1 point scale value.
- Also for the highest bitrates, maintaining the exception, the classification is almost independent of the interest level, ranging from less than 0.5 values between the different interest levels. This shows that observers have a more critical opinion for the lowest interest level.
- The average curves obtained for each interest level, gives a good approximation for all the sports that are contained in the specified level. However, for sports with high interest and high movement, the obtained curve is up to 2 values below the average. For sports with low interest and low bitrates, the difference can be up to 1 value, although it is not very common and remained within the limits set by the standard deviation.

Also in this chapter, based on the experimental results, a new empiric formula to estimate the MOS as a function of the bitrate and the interest level was developed. As referred, although this formula gives a good approximation to almost all the sports, for the two lowest bitrates of Tennis, the experimental results are outside the limits of the standard deviation. So, the temporal activity was analyzed to identify the sports with low, medium and high temporal activity. Since sports have a lot of changes of camera, the temporal activity was computed based on the 99% percentile, to smooth the effect of the changes in camera's peaks. Even taken into account this effect, BMX, Diving, Pommel Horse and Taekwondo were classified in the high temporal activity stage, due to the camera's movements in order to follow the athlete. So, for these sports, the 95% percentile would be the adequate choice to reduce this effect.

Chapter 7 – Conclusions and Future Work

Chapter index

7.1. Conclusions, 72

7.2. Future Work, 72

7.1. Conclusions

The results obtained in Chapter 6, allowed concluding that the interest level has a positive influence on the subjective rating, as formulated in the hypothesis. For the same content, observers tend to increase the ratings for the same bitrate, only because they feel more interested about it. Between the lowest interest level and the highest one, the difference between ratings can achieve the 2.5 scale values in the MOS scale and this result is independent of the sport type.

It was also possible to conclude that, the average curves obtained give a good approximation for all the sports that are contained in the specified level. However, for sports with high interest and high temporal activity, the obtained curve is up to 2 values below the average. For sports with low interest and low bitrates, the difference can be up to 1 value from the average, although it is not very common, the result is within the limits set by the standard deviation. Because of that phenomenon, the empirical MOS formula developed, as a function of bitrate and interest level, still has to take into account, another parameter related with the temporal activity (TA), in order to have a more general MOS expression. However, since only two sports were identified in the high temporal activity stage, Tennis and Javelin, it makes no sense to introduce in this stage the TA parameter in the MOS expression, due to the lack of data collected for these two sports. For that purpose, the best solution would be to develop another formula related to sports with high temporal activity and integrate it in the general MOS formula. Once again, due to the lack of collected data, this development will remain for future work and not developed for this dissertation.

7.2. Future Work

Although only 24 observers participated in the test session, with 144 videos watched, this sample was just enough to allow us to derive valid and relevant results and conclusions from the data collected. But, additional research still needs to be done in this area, with a larger a more diversified group of observers, in order to collect data with statistical relevance to allow tuning the parameters of all dimensions, but essentially the temporal activity parameter, namely:

- Sports with high temporal activity, such as Tennis and Javelin. The performed subjective test only had enough data to evaluate the tennis behavior to the interest level 5. So, other interest levels must be also analyzed to Tennis and Javelin, to verify if the same phenomenon can be clearly identified.
- Sports with a medium interest level. Again, due to lack of sufficient data, level 3 was not considered in this analysis, since only one observer has watched one sport with this interest level.
- Test more bitrates between the 190 kbps and 1450 kbps, to obtain smoother curves. The test session performed only considered four bitrates, with a gap of information between the 600 kbps and the 1450 kbps.

References

- [1] My eDirector. My eDirector 2012 Website, 2009. URL <http://www.myedirector2012.eu>.
- [2] 7th Research Framework Programme, URL http://cordis.europa.eu/fp7/home_en.html/.
- [3] ITU-T. Definitions of terms related to quality of service. Recommendation E.800, International Telecommunication Union - Telecommunication Standardization Sector, 2008.
- [4] Marcio N. Zapater and Graça Bressan. A Proposed Approach for Quality of Experience Assurance of IPTV. In *Proceedings of the First International Conference on the Digital Society (ICDS'07)*, pages 25-25. IEEE, 2007. doi:10.1109/ICDS.2007.4.
- [5] ITU-T. Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100. Recommendation P.10/G.100, International Telecommunication Union - Telecommunication Standardization Sector, Jan. 2008.
- [6] ETSI. Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services. Technical Report TR 102 643, European Telecommunications Standards Institute, Jan. 2010.
- [7] Dialogic. Quality of Experience for Mobile Video Users. White Paper 11681-01, Dialogic Corporation, Dec. 2009. URL <http://www.dialogic.com/medialabs/>.
- [8] Philip Kortum and Marc Sullivan, "The Effect of Content Desirability on Subjective Video Quality Ratings," *Human Factors*, vol. 52, 2010, pp. 105-118.
- [9] Mario Vranješ and Snježana Rimac-drlje and Krešimir Grgiü, "Locally Averaged PSNR as a Simple Objective Video Quality Metric," *Symposium A Quarterly Journal In Modern Foreign Literatures*, vol. 10, 2008, pp. 10-12.
- [10] Alain Horé and Djemel Ziou. Image Quality Metrics: PSNR vs. SSIM. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 2366-2369. IEEE, 2010. doi:10.1109/ICPR.2010.579.
- [11] ITU-T. Objective perceptual multimedia video quality measurement in the presence of a full reference. Recommendation J.247, International Telecommunication Union - Telecommunication Standardization Sector, 2008.
- [12] ITU-R. Methodology for the subjective assessment of the quality of television pictures. Recommendation BT.500-12, International Telecommunication Union - Radiocommunication Sector, Sep. 2009.
- [13] ITU-T. Subjective video quality assessment methods for multimedia applications. Recommendation P.910, International Telecommunication Union - Telecommunication Standardization Sector, 2008.
- [14] Luís Miguel Roque, "Quality Evaluation of Coded Video," pp. 1-10.

Annex 1 – FFmpeg Comands

The videos used in this work were encoded with the FFmpeg version 0.6 tools. The videos were coded in four different pairs of rate/resolution, as described in Chapter 5, using the H.264 codec baseline profile. The coding was done in two steps to ensure better results with the second pass allowing to refining the video output to better fit the specifications. Table 7 describes the syntax/description of the encoding controls used in FFmpeg. The input commands are also shown.

Table 7 – FFmpeg command syntax and description

Syntax	Description
-i <i>filename</i>	Input file name
-y	Overwrite output files
-t <i>duration</i>	Restrict the output video sequence to the duration specified in seconds (-t <i>hh:mm:ss</i> syntax is also supported)
-b <i>bitrate</i>	Set the video bitrate in bps (default = 200 kbps)
-r <i>fps</i>	Set the video frame rate (default = 25 fps)
-s <i>size</i>	Set the video size in the format ' <i>w×h</i> ' (default = same as input)
-bt <i>tolerance</i>	Set the video bitrate tolerance in bits (default = 4000 kbits)
-vcodec <i>codec</i>	Force video codec to <i>codec</i>
-pass <i>n</i>	It is used to do two-pass video encoding, <i>n</i> can be 1 or 2
-g <i>gop</i>	Set the group of pictures size
-ss <i>position</i>	Find the given time position in seconds (-ss <i>hh:mm:ss</i> syntax is also supported)
-an	Disable audio recording
-fpre <i>filename</i>	Takes the filename of the preset as input and can be used for any kind of codec. The preset file contains a sequence of option=value pairs

Commands for the first pass:

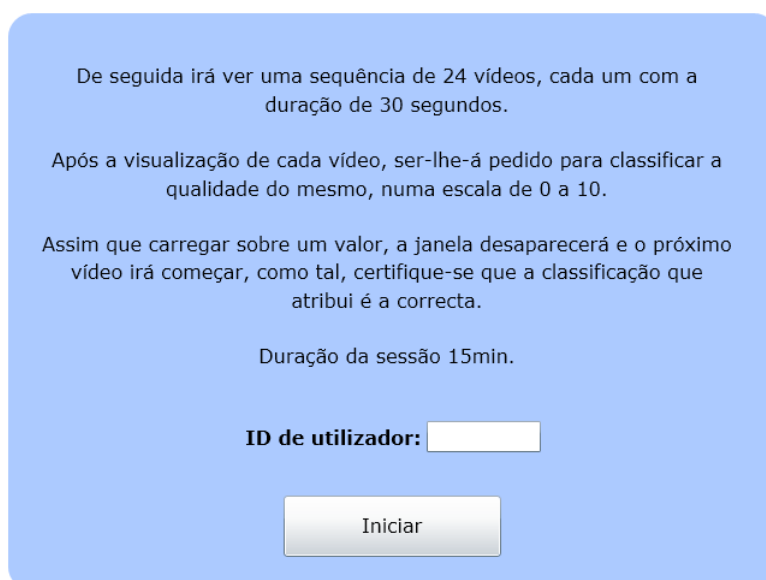
```
ffmpeg -i filename -an -pass 1 -vcodec libx264 -fpre libx264-  
slow_firstpass.ffpreset -fpre libx264-baseline.ffpreset -b bitrate -bt  
tolerance -threads 0 -g 25 -r 25 -s size -t 00:00:30 -ss position  
output_file
```

Commands for the second pass:

```
ffmpeg -i filename -an -pass 2 -vcodec libx264 -fpre libx264-  
slow.ffpreset -fpre libx264-baseline.ffpreset -b bitrate -bt tolerance -  
threads 0 -g 25 -r 25 -s size -t 00:00:30 -ss position -y output_file
```

Annex 2 – Application’s Appearance

As described in Chapter 5, the application for the video quality assessment had three stages, the login, the play-out and the logout. In the login, as shown in Figure 41, observers just needed to click over the box “ID de utilizador”, insert their ID and click on the start button “Iniciar”. The logout is automatic, done by the application, as soon as observers have completed rating the last video, as illustrated in Figure 42.



De seguida irá ver uma sequência de 24 vídeos, cada um com a duração de 30 segundos.

Após a visualização de cada vídeo, ser-lhe-á pedido para classificar a qualidade do mesmo, numa escala de 0 a 10.

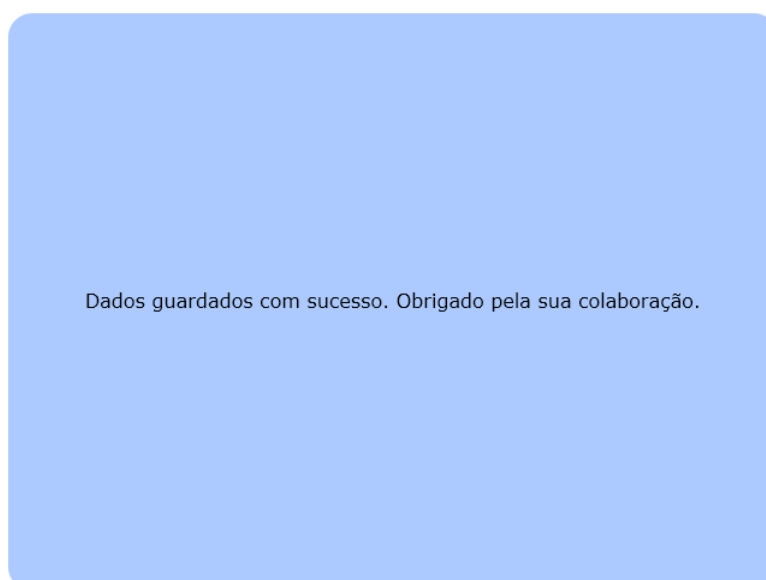
Assim que carregar sobre um valor, a janela desaparecerá e o próximo vídeo irá começar, como tal, certifique-se que a classificação que atribui é a correcta.

Duração da sessão 15min.

ID de utilizador:

Iniciar

Figure 41 – Application login screen



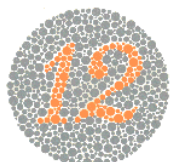
Dados guardados com sucesso. Obrigado pela sua colaboração.

Figure 42 – Application logout screen

Annex 3 – Ishihara Plates

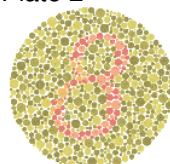
The Ishihara test consists of a set of 24 plates that observers have to watch in order to identify numbers, in plates 1 to 17, and one or two wiggly lines in plates 18 to 24. The plates and corresponding solutions are presented below, for what normal color vision people can see and what color-blind people can or cannot see.

Plate 1



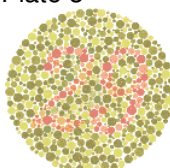
12 All people should see the number.

Plate 2



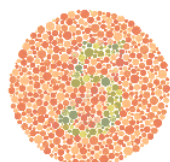
8 Those with normal color vision.
3 Those with red green color blindness.
Nothing Those with total color blindness see nothing.

Plate 3



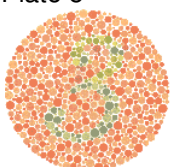
29 Those with normal color vision.
70 Those with red green color blindness.
Nothing Those with total color blindness see nothing.

Plate 4



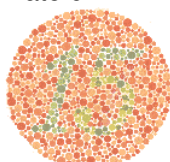
5 Those with normal color vision.
2 Those with red green color blindness.
Nothing Those with total color blindness see nothing.

Plate 5



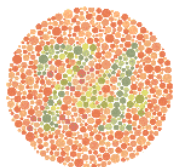
3 Those with normal color vision.
5 Those with red green color blindness.
Nothing Those with total color blindness see nothing.

Plate 6



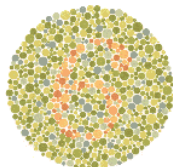
15 Those with normal color vision.
17 Those with red green color blindness.
Nothing Those with total color blindness see nothing.

Plate 7



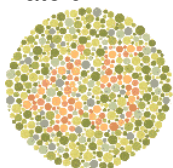
74	Those with normal color vision.
21	Those with red green color blindness.
Nothing	Those with total color blindness see nothing.

Plate 8



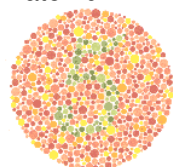
6	Those with normal color vision.
Nothing	The majority of color-blind people cannot see.

Plate 9



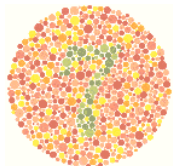
45	Those with normal color vision.
Nothing	The majority of color-blind people cannot see.

Plate 10



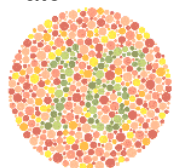
5	Those with normal color vision.
Nothing	The majority of color-blind people cannot see.

Plate 11



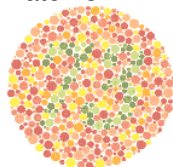
7	Those with normal color vision.
Nothing	The majority of color-blind people cannot see.

Plate 12



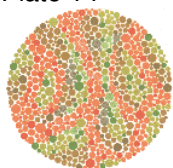
16	Those with normal color vision.
Nothing	The majority of color-blind people cannot see.

Plate 13



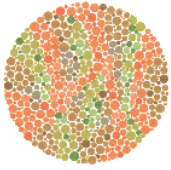
73	Those with normal color vision.
Nothing	The majority of color-blind people cannot see.

Plate 14



Nothing	People with normal vision or total color blindness should not be able to see any number.
5	Those with red green color blindness should see.

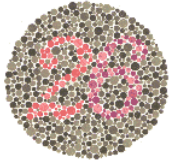
Plate 15



Nothing People with normal vision or total color blindness should not be able to see any number.

45 Those with red green color blindness should see.

Plate 16

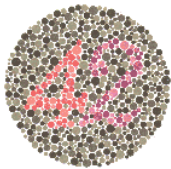


26 Those with normal color vision.

6 Red color-blind people will see a 6; mild red color-blind people will also faintly see a number 2.

2 Green color-blind people will see a 2; mild green color-blind people may also faintly see a number 6.

Plate 17

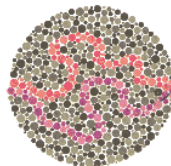


42 Those with normal color vision.

2 Red color-blind people will see a 2; mild red color-blind people will also faintly see a number 4.

4 Green color-blind people will see a 4; mild green color-blind people may also faintly see a number 2.

Plate 18

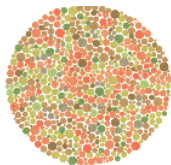


Those with normal color vision should be able to trace along both the purple and red lines.

Those with red colorblind should be able to trace the purple line; those with weak red vision may be able to trace the red line, with increased difficulty.

Those with green color-blind should be able to trace the red line; those with weak green vision may be able to trace the purple line, with increased difficulty.

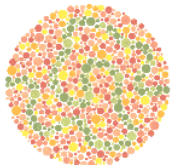
Plate 19



Those with normal color vision or total color blindness should be unable to trace the line.

Most people with red green color blindness can trace the wiggly line, depending on the severity of the condition.

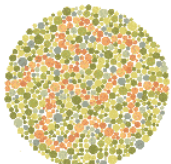
Plate 20



Those with normal color vision should be able to trace a green wiggly line.

Most people with any form of color blindness will be unable to trace the correct line.

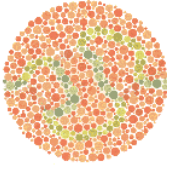
Plate 21



Those with normal color vision should be able to trace an orange wiggly line.

Most people with any form of color blindness will be unable to trace the correct line.

Plate 22

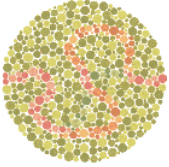


Those with normal color vision should be able to trace the blue-green/yellow-green wiggly line.

Red green color-blind people will trace the blue-green and red line.

People with total color blindness will be unable to trace any line.

Plate 23

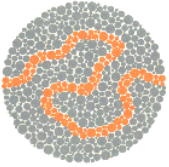


Those with normal color vision should be able to trace the red and orange wiggly line.

Red green color-blind people will trace the red and blue-green wiggly line.

People with total color blindness will be unable to trace any line.

Plate 24



Everyone should be able to trace this wiggly line.

Annex 4 – Temporal Activity MATLAB® Code

The temporal activity was calculated based on a MATLAB® program that was provided by Professor Paula Queluz and her student Tomás Brandão. This program allows calculating the temporal activity of a given video, by comparing two consecutive frames pixel by pixel. The program has as arguments the name of the video file, the vertical and horizontal resolutions and the number of frames that will be compared. Note that the video file must be in the YUV format and the number of frames must be the total number of frames minus one.

```
%TEMPORAL_ACTIVITY Computes temporal activity of YUV video sequence, as
%                      suggested by recommendation ITU-T P.910.
%
% [G_ACT, P_99, STD_ACT] = TEMPORAL_ACTIVITY(FILENAME, V_RES, H_RES, N_FRAMES)
% computes temporal activity values on the first N_FRAMES from the yuv
% file FILENAME, assuming that the dimensions of every frame are
% V_RES x H_RES.
%
% V_RES and H_RES are the vertical and horizontal resolutions, respectively.
%
% In case of success, TEMPORAL_ACTIVITY returns:
%
% G_ACT - global temporal activity value as specified in Rec. ITU-T P.910
%         (maximum value found along the video frames);
%
% P_99 - similar to previous, but the value at percentile 99 is returned,
%        instead of the maximum (useful if the video sequence contains
%        camera changes).
%
% STD_ACT - standard deviation of temporal activity along the video frames.
%
% On error (invalid number of arguments, file not found, incorrect
% video resolutions), TEMPORAL_ACTIVITY returns empty matrices.
%
% Tomas Brandao 2011/04/14

function [g_act, p_99, std_act] = temporal_activity(filename, v_res, h_res, n_frames)

g_act=[];
p_99=[];
std_act=[];

% Check number of arguments
if (nargin ~= 4)
    fprintf('Error: temporal_activity - invalid number of arguments.\n');
end

% Opens file, returning an empty matrix on error;
fp_vid = fopen(filename, 'rb');
if (fp_vid == -1)
    fprintf('Error: temporal_activity - could not open yuv file.\n');
    return;
end

% Check file size
fileinfo = dir(filename);
filesize = fileinfo.bytes;
n_bytes = v_res * h_res * 3 / 2 * n_frames;
if (n_bytes > filesize)
    fprintf('Error: temporal_activity - check resolutions / number of frames.\n');
    return;
end;

% Read first frame
img1 = fread(fp_vid, [h_res v_res], 'uint8');

% Initialize frame-wise temporal activity values
temp_act = ones(n_frames-1,1);
```

```

% Compute frame-wise temporal activity values
for k=2:n_frames

    fseek(fp_vid, (k-1)*3*h_res*v_res / 2, 'bof');
    img2 = fread(fp_vid, [h_res v_res], 'uint8');

    frame_diff = double(img2)-double(img1);

    temp_act(k-1) = std(frame_diff(:));

    img1 = img2;

end

% Compute global values
g_act = max(temp_act(:))

sorted_activity = sort(temp_act(:));
pos_percentil_99 = round(length(sorted_activity) * 0.99);
p_99 = sorted_activity(pos_percentil_99);

std_act = std(sorted_activity(1:pos_percentil_99));

% Close input file
fclose(fp_vid);

```

The 99% percentile is calculated according to the “Nearest Rank” method only in three steps.

1. The N values must be ordered, arranged from the lowest to the highest value:

```
sorted_activity = sort(temp_act(:));
```

2. The P th percentile rank of the N ordered values, must be calculated according to (24) and the result must be rounded to the nearest integer:

$$P_{n\%rank} = \frac{n}{100} \times N \quad (24)$$

```
pos_percentil_99 = round(length(sorted_activity) * 0.99);
```

3. Then, taking the value that corresponds to that rank, the P th percentile is calculated:

```
p_99 = sorted_activity(pos_percentil_99);
```