AGH University of Science and Technology
Faculty of Electrical Engineering, Automatics, Computer Science and Electronics

Ph.D. Thesis

Piotr Romaniak

Assessment of Perceptual Video Quality Affected by Acquisition and Bit-rate Reduction Artifacts

Supervisor:
Prof. dr hab. inż. Zdzisław Papir

AGH University of Science and Technology
Faculty of Electrical Engineering, Automatics, Computer Science and Electronics
Department of Telecommunications

Al. Mickiewicza 30, 30-059 Kraków, Poland
tel. +48 12 6173937
fax +48 12 6342372

www.agh.edu.pl
www.kt.agh.edu.pl

Printed in Poland

Acknowledgements

This dissertation would not have been possible without the support of many people. I would like to express my sincere gratitude to my supervisor, Prof. Zdzisław Papir who was abundantly helpful and offered invaluable assistance, support and guidance. His substantive comments on my research have significantly increased its value.

Deepest gratitude are also due to my collaborators and friends, without whose knowledge and assistance this study would not have been successful. My special thanks to Lucjan Janowski for his valuable advices in mathematics and statistics. To Mikołaj Leszczuk, for his bright ideas and significant support, especially in a conceptual work. To Michał Grega, for the offered support and for sharing his knowledge of digital photography. To Katarzyna Kosek-Szott and Szymon Szott, for answering my questions and their support in English.

I wish to express my love and gratitude to my beloved families; special thanks to my wife Justyna and to my son Adam for their support at home and making me forget about the work whenever I needed.

Abstract

This dissertation deals with the problem of the perceptual video quality assessment for video content. A huge increase in popularity of video-based services over recent years has raised unsolved quality assurance problems. Video service providers cannot afford the risk of going to market with a sub-standard video offer and need to manage their quality problems. Competition between providers is fierce and the key to success is to provide a service with the highest possible level of user satisfaction (QoE - Quality of Experience) .

Traditional approaches towards quality assessment of video delivery focus on the network aspects only. In order to perform a more reliable assessment of the quality experienced by the user, a more comprehensive quality assessment approach is needed. Additional quality aspects related to video acquisition, bit-rate reduction, service environment, and end users' preferences should be also addressed.

Therefore, the following thesis is formulated and proven:

> **It is possible to assess perceptual quality of video content affected by artifacts related to acquisition and bit-rate reduction, using no-reference metrics, in a real time.**

The thesis is proven by proposing a set of no-reference video quality metrics and demonstrating their high performance in terms of a correlation with end users' experience. Three stages of the end-to-end video delivery chain were addressed, namely video acquisition, compression, and service environment including end users' preferences. The metrics were verified using subjective experiments and objective models were derived upon the results.

The obtained results show that the reliable video quality assessment can be realized using no-reference metrics, for video-based services using H.264/AVC compression, and in a real-time for standard definition video. High performance in terms of a correlation with end users' experience was obtained for a diversified video content. Mentioned features meet all the requirements of a comprehensive video quality assessment system and can be utilized by video service providers

for a constant and reliable quality monitoring.

Keywords: video quality metrics, quality assessment, quality of experience, perceptual models, no-reference quality assessment

Contents

Nomenclature

Acronyms

ACR	Absolute Category Rating
ACR-HR	ACR with Hidden Reference
ALF	Asymmetric Logit Function
AM	Artifacts Measurement
ANOVA	Analysis of Variance
AVC	Advanced Video Coding
BER	Bit Error Rate
CCD	Charge Coupled Device
CMOS	Complementary Metal Oxide Semiconductors
CPU	Central Processing Unit
CSF	Contrast Sensitivity Function
DCT	Discrete Cosine Transform
DoF	Depth of Field
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
FPS	Frames Per Second
FR	Full-reference

GDA	Gabor Difference Analysis
GLZ	Generalized Linear Model
GoP	Group of Pictures
HDR	High Dynamic Range
HSDPA	High Speed Downlink Packet Access
HVS	Human Visual System
IP	Internet Protocol
IPTV	Internet Protocol Television
ITU	International Telecommunication Union
JND	Just-Noticeable Distortion
LCD	Liquid Crystal Display
LTE	Long Term Evolution
MAE	Mean Absolute Error
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MPQM	Moving Pictures Quality Metric
MSE	Mean Square Error
NR	No-reference
OS	Opinion Score
PC	Personal Computer
PDA	Personal Digital Assistant
PEVQ	Perceptual Evaluation of Video Quality
PLR	Packet Loss Rate
PSE	Peak Signal Error
PSNR	Peak Signal-to-Noise Ratio

QoD	Quality of Delivery
QoE	Quality of Experience
QoS	Quality of Service
RFP	Reverse Frame Prediction
RGB	Red Green Blue
RMSE	Root Mean Squared Error
RR	Reduced-reference
RTP	Real Time Protocol
SA	Spatial Activity
SAD	Sum of Absolute Differences
SC	Stimulus Comparison
SD	Standard Definition
SLR	Single-Lens Reflex
SS	Single Stimulus
SSIM	Structural Similarity Index
TA	Temporal Activity
UMTS	Universal Mobile Telecommunications System
VoD	Video on Demand
VQEG	Video Quality Experts Group

Variables

MOS(B)	MOS predicted from the blockiness model
MOS(B, F, IF)	MOS predicted from the integrated model for the H.264/AVC compression
MOS(B, SA, TA)	MOS predicted from the blockiness model including SA and TA
MOS(Bl)	MOS predicted from the blur model

MOS(Ex_o)	MOS predicted from the over-exposure model
MOS(Ex_u)	MOS predicted from the under-exposure model
MOS(F)	MOS predicted from the flickering model
MOS(Fr, d)	MOS predicted from the FPS model including d
MOS(Fr, SA)	MOS predicted from the FPS model including SA
MOS(N)	MOS predicted from the noise model
MOS(R, SA, TA)	MOS predicted from the frame resolution model including SA and TA
$\overline{x}$	Mean value of x
abs	Absolute value
B	Blockiness metric
$b(i, j)$	Magnitude of an image
Bl	Blur metric
$cutoff$	Value of the threshold used for an image edges filtering
d	Proposed video motion metric
Ex	Exposure metric
F	Flickering metric
Fr	Frame per second rate parameter
IF	I-frame flickering metric
L_b	Bright luminance bound of an image
L_d	Dark luminance bound of an image
N^{frame}	Frame level noise metric
N^{loc}	Local noise metric
R	Video frame resolution parameter
R_t^2	Pearson linear correlation coefficient obtained for the training set

R_v^2	Pearson linear correlation coefficient obtained for the verification set
R_{t+v}^2	Pearson linear correlation coefficient obtained for both sets
S^h	Horizontal Sobel filter
SA	Spatial activity of an image
SAD^{norm}	Normalized sum of absolute differences
$std(x)$	Standard deviation of x
TA	Temporal activity of an image
th	Value of the threshold used for detecting smooth image regions
x_{norm}	Normalized values of x

Chapter 1

Introduction

Media market watchers expect a turning point in the area of modern video-based services such as IPTV (Internet Protocol Television) and VoD (Video on Demand) . After several years of technical development of underlying network infrastructure, quality of service enforcing techniques, and broadband access to the Internet, the mentioned services have been nominated as "killer applications" once again. The reason for this comeback is that the current state of the multimedia delivery infrastructure finally allows for a seamless access and assures ubiquity of these services, to be used in three crucial areas of human existence (home, work, and travel). Competition between providers is fierce and the key to success is to provide a service with the highest possible level of user satisfaction (QoE - Quality of Experience).

1.1 Motivation

Development of broadband Internet technology as well as the high performance of modern personal computers and IPTV sets allow for the introduction of pilot video services over the IP networks. It became possible to provide all telecommunication services (telephony, video streaming, and data transmission) through one common network referred to as the Triple Play network. While the data transmission and telephony remains a stable offer for at least few years, video streaming has not found an appropriate technological support from telecom operators so far.

Most operators are moving cautiously with their video-based services rollouts because of service assurance concerns. Providers, of course, cannot afford the risk of going to market with a sub-standard video offer and need to resolve their quality problems.

While video-based services are becoming more and more popular in the fixed networks, their equivalents in mobile networks are celebrating their debut in the multimedia market. New technologies of data transmission in mobile networks (UMTS, HSDPA, LTE), and increased processing capacity of mobile terminals (smart phones, PDAs) allow users to receive video streams in the Mobile TV service (which is the mobile equivalent of the combined IPTV and VoD services in the IP networks).

Among the variety of advanced multimedia services, a new trend has began in the first decade of the twenty-first century, known as Web 2.0. The name suggests the next version of Web technology, however, it does not imply any changes in technology, but only sets the direction of the Internet evolution. The development of Web 2.0 technology significantly increased the amount of multimedia content generated by users, defined as user-generated content. Web 2.0 introduces also new challenges in video quality assessment, i.e. a necessity to address issues related to the source video quality (quality of the original video). A wide range of end users' video capturing devices and their limited capabilities often results in a significant quality degradation at the video acquisition phase. For example, a non-professional IP web-cam will produce such undesirable artifacts as noise, blur and improper exposure, depending on a scene conditions.

A common point for all the presented technologies and video services is a necessity to ensure high video quality from the end user's perspective. This requires the solution of certain problems associated with the scalability of the transmission, interaction between services, the need to maintain a high level of QoE for multimedia services and a large diversity of the underlaying networks. These problems do not exist in a dedicated homogeneous network and occurs only for the large scale heterogeneous networks with multiple services running.

In order to ensure a reliable and constant monitoring for video-based services, it is necessary to develop mechanisms for perceived video quality assessment. This is a key requirement in a scenario where a large number of video streams, varying in terms of video content and parameters, is transmitted in a heterogeneous environment. This means ensuring acceptable quality for both a user watching video on a plasma TV, and for a user connected to a wireless network using PDA. Derivation of a model assessing perceived video quality in a heterogeneous content distribution network is even more challenging, because of the need to reflect feelings and preferences of end-users.

1.2 Objectives and Challenges

Traditional approaches towards video quality assessment focus on the network aspects only. Metrics such as delay, jitter, packet loss and throughput are used to specify network performance and, then, to assess the service quality. Quality

of Service (QoS) is typically understood as a measure of performance from the network perspective at the packet level. However, in order to perform a more reliable assessment of the quality experienced by the user in case of a video delivery, a more comprehensive quality assessment approach is needed.

QoS parameters measurement does not represent an exhaustive set of metrics to enable an end-to-end quality management focusing on the user experience. The concept of Quality of Experience (QoE) has been recently introduced to address the issue concerning the assessment of how well a video service meets the customers' expectations. It describes the performance of a device, system, service, or application (or any combination thereof) from the user's point of view. In video content delivery systems, QoE is a measure of the end-to-end performance at the service level from the user perspective.

Video service providers are looking for reliable solutions for a constant QoE monitoring in an in-service mode. The implementation of such solutions is beneficial not only for the providers but also for the end-users. A key factor here is to ensure the perfect balance between cost and quality of the provided services.

The question arises whether it is possible to derive a reliable quality monitoring system operating in an in-service mode? The main challenges that have to be faced are:

- high performance expressed in terms of a correlation with end users' experience, for diverse video content is required,
- a multi-modal quality degradation nature in hybrid block-based motion compensated predictive video codecs needs to be addressed,
- source video quality aspects need be addressed,
- real time restrictions for quality assessment need to be met,
- time and cost consuming subjective experiments need to be carried out,
- ability to utilize in in-service applications is required, what implies a no-reference scenario.

1.3 Thesis

The following thesis is proposed:

> **It is possible to assess perceptual quality of video content affected by artifacts related to acquisition and bit-rate reduction, using no-reference metrics, in a real time.**

The thesis is proven by proposing a set of no-reference, real time video quality metrics and demonstrating their high performance in terms of a correlation with end users' experience expressed as Mean Opinion Score (MOS). The metrics were verified using subjective experiments in a controlled environment and objective models were derived upon the results obtained by statistical analysis.

1.4 Approach

The following assumptions were made in order to prove the thesis:

- video compression is realized using the H.264/AVC coding scheme,
- in-service application of the proposed metrics imposes no-reference approach (no access to the reference video content available),
- subjective experiments are used to in order to obtain "ground truth" regarding the perceived video quality,
- perceptual quality models are derived based on the results from subjective experiments,
- artifacts measurement approach is used to assess source video quality,
- artifacts measurement approach is used to assess compressed video quality,
- service environment factors and users' preferences are reflected in the results from subjective experiments,
- real time quality assessment for a standard resolution (SD) video using a standard PC computer.

The proposed metrics for QoE assessment encompass the first two stages of the video delivery chain, i.e. acquisition and compression (see section 7). Impairments caused by network transmission were out of the scope of this work, because this problem was extensively analyzed in the past (see section 3.2.3). Additionally, approaches based on a bit-stream analysis do not require to decompress video content, so do not constitute a challenge in a real time measurement systems. The last stage of the video delivery chain (end user's equipment and preferences) was addressed in the subjective experiments and is reflected in the gathered subjective scores.

Source video quality assessment is realized using the acquisition related metrics for over- and under-exposure, blur, and noise. Compressed video content quality is evaluated using metric for blockiness artifact, flickering artifact, and I-frame flickering artifact. Additionally, two other domains (expect compression)

of video bit-rate reduction are considered: 1) temporal – by changing frames per second FPS rate, and 2) spatial – by decreasing frame resolution.

1.5 Research Projects and Published Papers

The results presented in this dissertation were partially funded by the following research projects:

1. FP6 Network of Excellence CONTENT: Content Networks and Services for Home Users (grant no. 038423),
2. FP7 IP INDECT: Intelligent Information System Supporting Observation, Searching and Detection for Security Citizens in Urban Environment (grant no. FP7-218086),
3. Polish national project: Next Generation Services and Networks – Technical, Application and Market Aspects (grant no. PBZ-MNiSW-02/II/2007),
4. Polish national project: Impact of Quality of Service (QoS) on the User Quality of Experience (QoE) for Multicast Streaming Services in Heterogeneous IP Environment (grant no. NN517438833).

The results presented in this dissertation were partially published in the following papers:

Journal Papers

1. Janowski, L., Romaniak, P., Papir, Z., "Content Driven QoE Assessment for Video Frame Rate and Frame Resolution Reduction", preliminarily accepted for publication in Multimedia Tools and Applications, Springer, **IF**
2. Mu Mu, Romaniak, P., Mauthe, A., Leszczuk, M., Janowski, L., Cerqueira, E., "Framework for the Integrated Video Quality Assessment", preliminarily accepted for publication in Multimedia Tools and Applic., Springer, **IF**
3. Janowski, L., Leszczuk, M., Papir, Z., and Romaniak, P., "The Design of an Objective Metric and Construction of a Prototype System for Monitoring Perceived Quality (QoE) of Video Sequences", Journal of Telecommunications and Information Technology, no. 3, pages 87-94, 2011
4. Głowacz, A.,Grega, M., Gwiazda, P., Janowski, L., Leszczuk, M., Romaniak, P., Romano, S. P., Automated Qualitative Assessment of Multi-Modal Distortions in Digital Images Based on GLZ, Special Issue of Annals of Telecommunications on Quality of Experience and Socio-Economic Issues

of Network-Based Services, Springer, Volume 65, Numbers 1-2, Pages 3-17, February, 2010, **IF**

5. Janowski, L., Leszczuk, M., Papir, Z., Romaniak, P., Ocena postrzeganej jakości (Quality of Experience) usług strumieniowania wideo, w scenariuszu bez referencji (No-Reference), ze skalowaniem przepływności, Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne, ISSN 1230-3496, 8-9 2009, s. 800-804

6. Głowacz, A.,Grega, M., Janowski, L., Leszczuk, M., Romaniak, P., Zautomatyzowana ocena wielomodalnych zniekształceń w obrazach cyfrowych, Telekomunikacja Cyfrowa – Technologie i Usługi, Tom 9, strony 50-60, 2008-2009

7. Grega, M., Janowski, L., Leszczuk, M., Romaniak, P., Papir, Z., Quality of experience evaluation for multimedia services – Szacowanie postrzeganej jakości usług (QoE) komunikacji multimedialnej, Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne; ISSN 1230-3496. 2008 R. 81 nr 4 s. 142-153

Conference Papers

1. Romaniak, P., Janowski, L., Leszczuk, M., Papir, Z., "Perceptual Quality Assessment for H.264/AVC Compression", CCNC'2012 FMN, 4th International Workshop on Future Multimedia Networking, January 14, 2012, Las Vegas, NV, USA

2. Janowski, L., Romaniak, P., Papir, Z., "Assessing Quality of Experience for High Definition Video Streaming under Diverse Packet Loss Patterns", 4th International Conference on Multimedia Communications, Services and Security, 2-3 June 2011, Krakow, Poland

3. Romaniak, P., Janowski, L., Leszczuk, M., Papir, Z., "A No Reference Metric for the Quality Assessment of Videos Affected by Exposure Distortion", IEEE International Conference on Multimedia and Expo, July 11 to 15, 2011, Barcelona, Spain

4. Janowski, L., Romaniak, P., "How Do Video Frame Rate and Resolution Influence QoE", 3rd International Workshop on Future Multimedia Networking FMN'10, 2010 June 17-18, Krakow, Poland

5. Romaniak, P., Janowski, L., "How to Build an Objective Model for Packet Loss Effect on High Definition Content Using the SSIM and Subjective Experiment", 3rd International Workshop on Future Multimedia Networking FMN'10, 2010 June 17-18, Krakow, Poland

6. Janowski, L. , Romaniak, P., "Wpływ zmiany rozdzielczości i liczby klatek wyświetlanych na sekundę na jakość postrzeganą sekwencji wizyjnych", Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji KKRRiT 2010 Kraków, 16-18 czerwca 2010

7. Romaniak, P., Janowski, L., "Budowa obiektywnego modelu wpływu utraty pakietów na jakość postrzeganą dla telewizji wysokiej rozdzielczości (HDTV) z wykorzystaniem metryki SSIM oraz testów subiektywnych", Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji KKRRiT 2010 Kraków, 16-18 czerwca 2010

8. Cerqueira, E., Janowski, L., Leszczuk, M., Papir, Z., and Romaniak, P., Video Artifacts Assessment for Live Mobile Streaming Applications, DEMOFMN 2009 - Demonstrations on Future Multimedia Networking, June 2009, Coimbra, Portugal

9. Janowski, L., Leszczuk, M., Papir, Z., Romaniak, P., Ocena postrzeganej jakości (Quality of Experience) usług strumieniowania wideo, w scenariuszu bez referencji (No-Reference), ze skalowaniem przepływności, Krajowe Sympozjum Telekomunikacji i Teleinformatyki, 16-18 września 2009, Warszawa, Polska

10. Romaniak, P., Janowski, L., Leszczuk, M., Papir, Z., Ocena jakości sekwencji wizyjnych dla aplikacji strumieniowania na żywo w środowisku mobilnym, Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji KKRRiT 2009 Warszawa, 17-19 czerwca 2009

11. Bułat, J., Grega, M., Janowski, L., Leszczuk, M., Papir, Z., Romaniak, P., Zieliński, T., Quality of Experience for Image Searching and Video Streaming, Kierunki działalności i współpraca naukowa Wydziału Elektrotechniki, Automatyki, Informatyki i Elektroniki, 28-29 maj 2009, Kraków, Polska

12. Romaniak, P., Mu Mu, Mauthe, A., D'Antonio, S., Leszczuk, M., A Framework for Integrated Video Quality Assessment, 18th ITC Specialist Seminar on Quality of Experience, May 29 - 30, 2008, Blekinge Institute of Technology, Karlskrona, Sweden

13. Głowacz, A., Grega, M., Gwiazda, P., Leszczuk, M., Romaniak, P., Romano, S. P.: Automated Qualitative Assessment of Multi-Modal Distortions in Digital Images Based on GLZ, 18th ITC Specialist Seminar on Quality of Experience, May 29 - 30, 2008, Blekinge Institute of Technology, Karlskrona, Sweden

14. Boavida, F., Cerqueira, E., Chodorek, R., Grega, M., Guerrero, C., Leszczuk, M., Papir, Z., Romaniak, P., "Benchmarking the Quality of Experience for Video Streaming and Multimedia Search Services: the CONTENT Network of Excellence", XXIII Krajowe Sympozjum Telekomunikacji i Teleinformatyki, Bydgoszcz, 10-12 września, 2008

15. Papir, Z., Leszczuk, M., Janowski, L., Grega, M., Romaniak, P., "Quality of experience evaluation for multimedia services – Szacowanie postrzeganej jakości usług (QoE) komunikacji multimedialnej", Referat plenarny, Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji, Wrocław, 9-11 kwietnia 2008, S. 35-50

Workshop Papers

1. Romaniak, P., Towards Realization of a Framework for Integrated Video Quality of Experience Assessment, INFOCOM Student workshop 2009, Rio de Janeiro, Brazil, April 2009

2. Romaniak, P., "Quality of Experience Assessment of Video-Based Applications - Introduction to Research", Med-Hoc-Net-2007, Ionian University, Corfu, Greece - June 12-13, 2007

3. Romaniak, P., "Hybrid Solution for Quality of Experience Assessment of Video Streams Integrating Network Provider and User Approaches - Introduction to Research", CONTENT PhD Student, Spain, Madrid, February 2007 Workshop,

1.6 Dissertation Structure

The dissertation is structured as follows. Chapter 2 provides some background information on perceptual video quality assessment techniques and approaches. The state-of-the-art study is presented in Chapter 3. In the following chapters the original contributions are provided. Chapter 4 details derivation of perceptual video quality metrics. The subjective experiments performed and the analysis of results are given in Chapter 5 and Chapter 6 respectively. Chapter 7 concludes the dissertation.

Chapter 2

Introduction to Video Quality Assessment Techniques

This chapter contains some background information related to assessment techniques for video quality of experience (QoE). A classification of video quality metrics is proposed in order to present main approaches, advantages and disadvantages of different types of metrics. A discussion on parameters affecting perceived quality is given in order to realize a complexity of the overall QoE assessment task. The last section is devoted to subjective experiments, essential to derive perceptual models mapping measured parameters into QoE.

2.1 Classification of Video Quality Metrics

Perceived video quality assessment is a massive and challenging task. There are many different factors affecting the perceived video quality. The examples are screen size (mobile terminal vs. plasma TV screen), screen illumination (PDA's screen in a sunny day vs. cinema screen with lights turned off), movie content (talking heads vs. action movie), application (YouTube videos vs. video for the medical diagnose purpose), viewing distance (20 centimeters vs. several meters), user profile (amateur vs. professionalist), and many others. The one, permanently addressed quality factor, is video fidelity considering the distortion level introduced by the codec (lossy compression) and network during the transmission (packet loss ratio PLR).

In order to address mentioned factors, different types and classes of video

quality metrics have been proposed over last years [11], [63], [64], [72]. Performed efforts towards metrics classification resulted in at least one well-defined classification criterium: amount of required reference information. Other, more detailed types of classification, are presented by Winkler in [73] or by Eskicioglu in [10].

In this section, a classification being a super-set of existing ones and allowing for more accurate aggregation of existing metrics is proposed. Video quality metrics are classified using three orthogonal classifications: by the amount of the reference information required to assess the quality, by the measured features, and by the metric output (the way the quality is expressed). The proposed idea is presented in Fig. 2.1. All three mentioned classifications are discussed in the following sections.

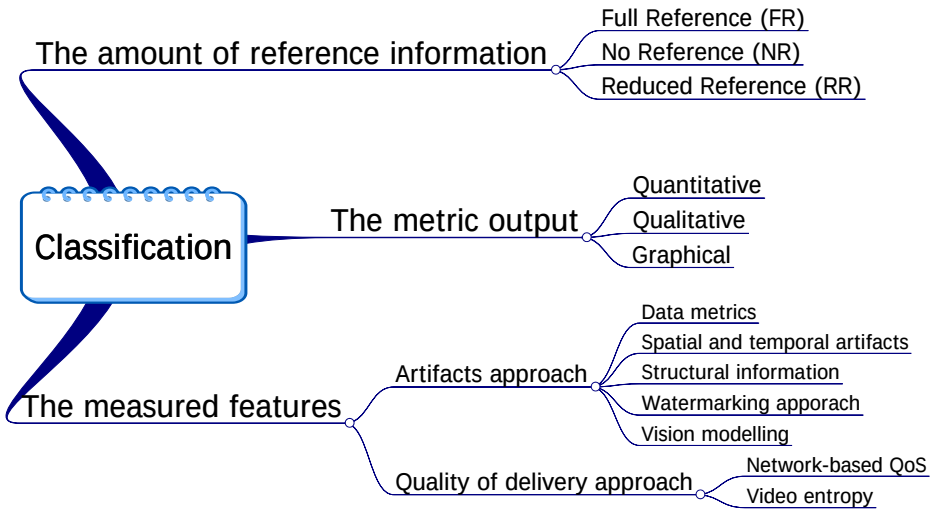


Figure 2.1: Classification of video quality metrics

2.1.1 Performance and Feasibility

Each class presented in Section 2.1.4 (the most detailed classification) is rated using two key parameters for comprehensive video quality metrics comparison and benchmark, namely "performance" and "feasibility" see Table 2.1. The first parameter, "performance", can be considered as the accuracy of the metric: degree of correlation with subjective rating. This usually applies to a well-defined test case under clearly specified conditions. In order to assess and compare the performance of different models in a quantitative manner, the test must be executed in a controlled environment under identical conditions for all target models.

The second parameter, "feasibility", considers such aspects as the flexible deployment in different network architectures, scalability to various user terminals, and efficient evaluation of a large number of concurrent video streams (even considering different end-systems). It is a crucial parameter since real-time quality of experience assessment becomes a key requirement for service providers.

Table 2.1: Performance and feasibility for different metric classes

Class	Performance	Feasibility
Data metrics	low – poor correlation with MOS	low – plenty of tools, restricted to the FR scenario
Spatial and temporal artifacts	medium-high – perceptual and spatio-temporal segmentation may assure high quality	medium-high – computational complexity distributed over user terminals, artifacts are measurable in the NR scenario
Structural information	medium-high – good cross-content and cross-distortion correlation	low-medium – restricted to the FR and RR scenario
Vision modeling	high – catch the degradation of key perceptual information	low – heavy computational complexity
Watermarking	medium – different susceptibility of the mark and the content	medium – additional amount of information in the video stream
QoS parameters measurement	low – pure correlation with MOS	high – instant assessment
Extended QoS	medium – analyze impact of network parameters on perceived quality	high-medium – almost instant assessment

2.1.2 Classification Based on Metric Output Type

Video quality metrics can be classified by the way the actual quality is expressed. This can be qualitative, quantitative or graphical ones [10], [21]. Quantitative criteria are usually expressed by a numerical score in some unlimited (e.g. logarithmic) or limited range (that is usually 1..5, 1..7, 1..10, −3..3). It is important

to note, that the quantitative measures can be calculated, but there are no inherent mappings on quality scales (like Mean Opinion Score, MOS [23]) and exact quality of user experience. On the other hand, qualitative criteria are considered with either textual or numerical measures. Textual criteria rely on a corresponding verbal description (e.g. MOS ranges from "bad" to "excellent"). Numerical criteria (like R-Value [27]) can be based on, e.g., the percentage of users who are satisfied with the quality. Graphical criteria rely on a set of measures, which reflect the most important image features. The examples are Hosaka plots [21] or Eskicioglu charts [11].

2.1.3 Classification Based on the Amount of the Required Reference Information

Classification based on the amount of the required reference information is the most popular classification criterium with three metric classes. The first one is called full-reference (FR) approach, assuming unlimited access to the original (reference) video (see Fig.2.2). Quality assessment is performed in a comparative way: *What is the quality of the distorted video compared to the original one?*

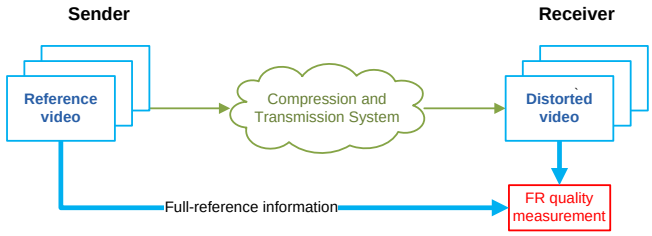


Figure 2.2: Diagram for Full-Reference approach

Advantages of this approach are a good correlation with MOS, relatively easy implementation as well as no CPU consumption and time limitations because of an off-line analysis. Disadvantages are the required amount of the reference, only off-line or laboratory applications, and need of a spatial and temporal alignment to ensure that adequate frames are compared. The area of possible applications is restricted to laboratory tests like codecs comparison and testing, encoder tuning or the quality acceptance level testing. Examples of the metrics are PEVQ [43], SSIM [68], and [48].

The second approach is commonly referred to as no-reference (NR) and stands for a blind quality evaluation (see Fig.2.3). Quality assessment is performed in an absolute way: *What is the quality of the video?*

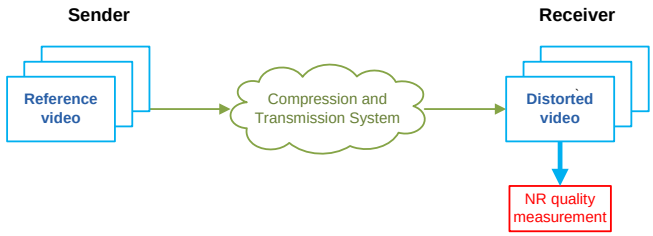


Figure 2.3: Diagram for No-Reference approach

This is an emerging and still not well defined approach enabling such desirable features as a real time in-service application for network performance monitoring, alarms generation or quality-based billing. In contrast to the FR scenario, there are no ITU recommendations related to NR video quality assessment. Disadvantages of the approach are lower correlation with MOS (refers to most of the currently existing metrics), complicated implementation and CPU load and time limitations. Preliminary work related to a design of a NR metric are described in [9], [13], [14], [19], [33], [52], [49].

The last class is referred to as reduced-reference (RR) approach that takes advantage (or disadvantage) of both previous approaches (see Fig.2.4).

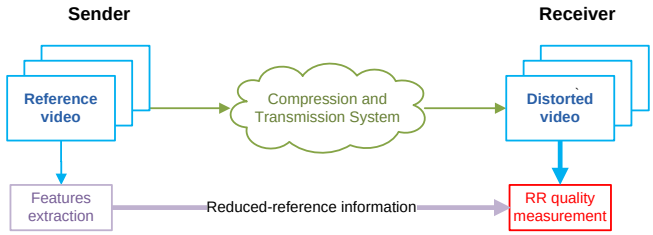


Figure 2.4: Diagram for Reduced-Reference approach

Only some certain features (like motion information or some spatial details) are extracted from the reference video stream and used to compare with the distorted one. This forces creation of an additional communication channel to send extracted information, and partial alignment. Amount of extracted information is still manageable and allows for more precise quality evaluation. The possible applications are both laboratory and in-service. There are few video quality metrics utilizing this approach [34], [74].

2.1.4 Classification Based on Measurement Type

There are two different approaches to the video quality assessment, based on different philosophies. The first one considers the whole end-to-end video delivery system as a black box and analyzes only decoded video quality at the receiver site (in a comparative or an absolute way) — it is commonly referred to as "artifacts measurement" (AM) approach. In the second approach, referred to as a "quality of delivery" (QoD) approach" all possible parameters of the delivery system are collected in order to predict output video quality. The first approach is well suited for source video quality and compression artifacts measurement. The second one is commonly used for network impairments assessment.

Artifacts Approach

If the artifacts approach is taken into account, video metrics can be divided into groups with the metric calculation complexity ranging from a simple pixel-to-pixel comparison algorithms (data metrics), through the separated artifacts measurement, up to the sophisticated HVS-based frame-level artifacts analysis [51]. The artifacts measurement approach tries to detect some certain distortions introduced by all stages of the end-to-end video delivery system (acquisition, compression, network transmission, decompression). Based on the level of detected distortion the overall quality grade is calculated (all detected distortions can contribute with different weights).

Data metrics look at the fidelity of the signal without considering its content [72]. In other words, every pixel error contributes to a decrease of quality, even if it is imperceptible for a human eye [57]. There are several simple methods intended for calculation of a scalar pixel-to-pixel measure to enable image comparison. Examples of such measures are: Peak Signal-to-Noise Ratio (PSNR), Mean Absolute Error (MAE), Mean Square Error (MSE), Peak Signal Error (PSE), or Root Mean Squared Error (RMSE). Some other have been analyzed in [11]. Data metrics were extremely popular over last decades and widely used in image and video quality assessment. The reason for this popularity was instant quality assessment, low computational complexity, and simple implementation. However, in their simplicity, they always operate on the whole frames and do not consider any other important factors strongly influencing the perceived quality - like HVS characteristics. For this reason, data metrics show low correlation with psychophysical experiments and are inadequate for precise quality assessment.

Fig. 2.5 illustrates a number of pictures with the same quality in terms of the MSE metric while the diversity in perceived quality is strong. Data metrics fail in cross-distortion and cross-content quality assessment [11]. The satisfactory performance appears only in case of some certain distortions measurement, e.g.,

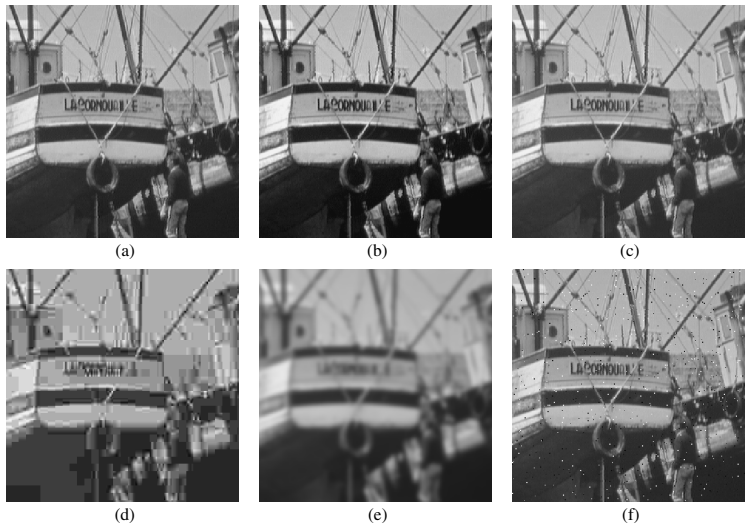


Figure 2.5: Images with different types of distortions, all with the MSE = 210 [67]

the MSE is accurate in additive noise but outperformed by other metrics for coding artifacts [2]. An excellent example for inverse-correlation is the PSNR metrics that will show a lower quality after dithering of an image with a reduced color depth (in fact, it is an improvement in quality) [73].

In order to overcome drawbacks attached to the data metrics, a number of more advanced approaches have been developed recently [14], [43]. Spatial artifacts measurement is performed on separated video frames. The principal idea assumes whole frames analysis. However, in order to assure higher performance, some features derived from vision modeling are employed. These are namely mechanisms for perceptual decomposition of video frames, weighting and pooling procedures (detailed in the further sections). Temporal artifacts refer to the video as a whole and focus on unnatural motion detection of video stream. This type of artifacts is mainly caused by severe network conditions, but sometimes also by video scaling in temporal domain.

Structural information approach is an innovative and promising idea, showing good performance for cross-distortion and cross-content video and image quality assessment. It is based on some certain features extraction (structure of the objects, spatial orientation and activity) and comparison between the original and

the distorted video sequences [68], [74]. The main drawback and limitation is a fact, that it cannot be applied in the NR scenario (it is restricted to the FR and RR only).

The vision modeling approach is the most sophisticated one, and one of the most popular at the same time [56], [71]. A video sequence or a single video frame is considered with respect to the visual information it contains. The idea of the approach is to reproduce human perception by modeling some components of HVS. According to [72] the most important characteristics of HVS are light adaptation, contrast sensitivity, spatial and temporal masking, and multi-channel model of human vision (all the properties are described in Section 2.2.5). Model for each component is built upon results from psychophysical experiments [73]. The limitations attached to this approach are related to high complexity of HVS-based metrics, especially for NR scenario.

In the watermarking (also referred to as data-hiding) approach an additional data is embed into a video stream; it is a mark [5], [13]. It is the NR approach in terms of architecture, and at the same time the FR comparative approach in terms of methodology. The mark is a well-known pattern, stored at the receiver side. The idea of quality assessment is based on the fidelity of the mark extracted from the distorted video measurement. The extracted mark is compared with the reference one. There are several issues influencing performance of the approach. The first one is different susceptibility to coding and transmission distortions of the mark and the video content. Other important issues are original video quality degradation caused by the embedded mark and additional data amount attached to the video stream.

Quality of Delivery Approach

All the artifacts-based metrics operate on the decompressed video frames level what implies a high computational complexity: first, the video stream has to be decompressed, then some artifacts have to be calculated. In order to overcome this problem, the "quality of delivery" approach can be applied. In this approach the video quality is predicted based on information gathered during video transmission. It is mainly based on network QoS parameters and video stream analysis. All the required information can be gathered on a video stream packet level or partially decompressed stream (video entropy). The more parameters measured the better video quality estimation.

This approach is preferable for in-service applications since computational complexity is significantly limited. In the simplest scenario, quality indicators are only some QoS parameters as packet loss ratio (PLR) or bit error rate (BER). This is an artificial approach that shows as poor correlation with MOS as data

metrics because a packet loss can have a drastically different impact on perceived quality depending on several factors. These factors are: compression algorithms (MPEG-2 vs. H.264), group of pictures (GoP) structure (I, P and B frames ratio), type of information lost (I, P, B frame), codec performance (coding, decoding), complexity and diversity of the video content (talking heads vs. action movie or cartoon).

An extension of the presented simple scenario was proposed recently in [35], [36], [39], [53], [57]. The idea is to assess an impact on the perceived quality of each packet loss based on information regarding video entropy. Estimation of network distortions based on partially decompressed stream analysis for H.264/AVC coded video is presented in [41].

2.2 Parameters for Video Quality Assessment

Video content delivery end-to-end system is illustrated in Fig. 2.6. The sources of potential quality problems are located in different parts of the end-to-end video delivery chain. The first group of distortions (1) can be introduced at the time of image acquisition. Other distortions (2) appear as a result of further compression and processing. Problems can also arise when scaling video sequences in the quantization, temporal and spatial domains. Then (3), for transmission over the network, there may be some artifacts caused by packet loss. At the end of the transmission chain (4), problems may relate to the equipment used to present video sequences.

After each stage, some reduction in the quality of the original video sequence may occur. The most common problems related to acquisition process are noise, lack of focus or improper exposure [51]. Lossy compression and network transmission will result in both spatial and temporal artifacts. Intra-frame compression results in a well known blockiness artefact; a flickering effect is associated with inter-frame compression [7]. Artifacts cause by packet losses will have spatial nature (missing frame slices), will propagate over successive frames (result of a predictive coding), and will cause playback discontinuities (missing frames) [48], [32]. All these effects will affect QoE for video-based applications [51].

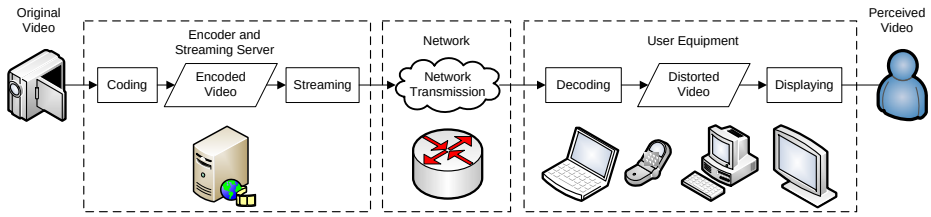


Figure 2.6: Video content delivery end-to-end system

In this section all mentioned factors affecting the visual quality of the video streams as well as video entropy and HVS-based characteristics are presented. The proposed parameters classification reflects the one proposed for video quality metrics.

2.2.1 Spatial Artifacts

Spatial artifacts refer to the frame-level of a video sequence and may be introduced by lossy compression as well as during an acquisition phase [29]. These artifacts have been the most frequently addressed video content quality parameters over past years. The commonly used examples are listed and described in

Table 2.2.

Table 2.2: Spatial artifacts

Name	Description
Contrast	The contrast of the distorted and the reference sequence
Blur	A distortion characterized by reduced sharpness of contour edges and spatial detail
Brightness	The brightness of the reference and the degraded signal
Blockiness	Often the result of a low bit rate coding that uses a block matching algorithm for the motion estimation and a coarse quantization for the image blocks
Global noise	Refers to the quality of the original video sequence, may be a result of analog-to-digital transformation
Color distortion	Low original video content quality and high compression may result in color distortion
Exposure level	Another important quality factor in this group, not analyzed in the literature. It is related to a video acquisition phase and described in detail in Section 4.1.1

2.2.2 Temporal Artifacts

Temporal artifacts refer to the video sequence as a whole and are mainly related to the playback continuity. This type of artifacts is caused mainly by severe network conditions but sometimes also by reduction in temporal domain. Examples presented in Table 2.3 are derived from existing metrics [43], [74].

2.2.3 Video Entropy Parameters

Video entropy information allows evaluation of an impact of network parameters and artifacts on quality of experience. In order to obtain desired information, video stream has to be partially decompressed. Parameters presented in Table 2.4 are used in existing video quality metrics [43], [53], [34].

Table 2.3: Temporal artifacts

Name	Description
Jerkiness/unnatural motion	Describes the smoothness of a video playback which is often impaired by down-sampling, coding processes and perturbed transmissions
Effective Frame Rate	Down-sampling of a video signal on a frame by frame basis often results in loss of information which often leads to the degradation of the video signal. The effective frame rate is an indicator quantifying the severeness of such a process
Frame Skips and Freezes	Temporal artifacts occurring in video transmissions caused by e.g. overloaded networks
Flickering effect	Another important quality factor in this group, not intensively analyzed in the literature. It is especially important for H.264/AVC video compression scheme and results from inter-frame compression. It is detailed in Section 4.2.2

Table 2.4: Video entropy

Name	Description
Compression ratio	Ratio of I frames to I+P+B frames
I frame count	Number of "I" (Intra) frames for the current sample period
P frame count	Number of "P" (Predictive) frames for the current sample period
B frame count	Number of "B" (Bidirectional) frames for the current sample period
Temporal and Spacial Activity	Temporal and spacial activity indicators quantify the amount of activity movement in the video content. These indicators can be derived from ITU-T recommendation P.910 [24]

2.2.4 Network QoS Parameters

Except typical QoS parameters as delay, delay jitter, available bandwidth or PLR, some other network parameters were successfully used in video metrics [53] (see Table 2.5).

Table 2.5: Network parameters

Name	Description
Program rate	Transport stream rate as observed
Program clock rate overall jitter	Jitter of synchronization stream
Jitter discards	Number of frames discarded due to jitter
Out of sequence	Number of misordered frames delivered
In sequence	Number of properly ordered frames delivered
Network loss probability	Statistically accurate predictive calculation of frame loss
Max loss length	Maximum number of consecutively lost frames
Max loss	Cumulative count of losses since first observation
Multicast join time	Actual time the stream way joined in "unix epoch" seconds
Multicast first time	Actual time the first data arrived in "unix epoch" seconds

2.2.5 Human Vision Modeling

Many studies to understand and model Human Visual System (HVS) have been performed recently [3], [20], [38]. Nevertheless, HVS is still not well defined because it is enormously complex. Only some approximate models have been proposed that account for limited number of HVS characteristics. In this section selected characteristics that influence video and image perception will be discussed.

Sensitivity to Light

The first characteristic that influence perceived visual quality is light adaptation. It was proved that we can adopt to almost unlimited range of the light intensities [72]. An immediate effect of these phenomena is perception of the relative rather than the absolute contrast. This is known as a Weber-Fencher law, where contrast is defined as a relative variance of the luminance:

$$C^W = \frac{\Delta L}{L} \tag{2.1}$$

The threshold contrast, that defines if a change in intensity is visible, is almost constant for intensities range considered in digital visual content. However, the value of the threshold is attached to the stimuli and depends mainly on color,

spatial and temporal frequency [72]. In order to account for this dependencies, Contrast Sensitivity Function (CSF) was designed. In the simplified approach CSF can be explained by the fact that human eye is less sensitive to the higher spatial frequencies than to the lower ones (see Fig. 2.7).

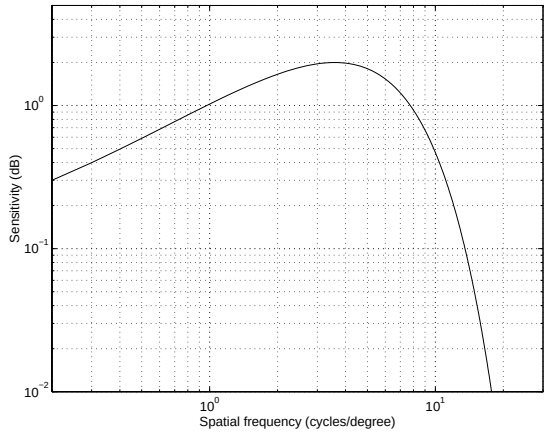


Figure 2.7: Sensitivity of the human eye as a function of spatial frequency [56]

In fact, the eye is the most sensitive to the bandpass stimuli, around 4 cycles per degree (cpd) [56]. It means that threshold contrast for higher spatial frequencies becomes very high (or even infinite). Thus, contrast sensitivity is close to the zero, since it is defined as an inverse detection threshold. Fig. 2.8 presents Campbell-Robson spatial CSF chart, which illustrates this phenomena. Inverted U-shape envelope is visible on the chart and the pick location depends on the viewing distance [72].

More advanced and complete approach defines CSF as a function of spatial and temporal frequency, and orientation. Contrast sensitivity is higher for vertical and horizontal directions, and lower for diagonal ones [56] (compare Fig. 2.8 with Fig. 2.9 or flip the page). Spatio-temporal CSF is presented in Fig 2.10.

In the most advanced approach CSF is also considered in relation to the achromatic, chromatic and colour stimuli; more detailed discussion is presented by Winkler [72].

Masking

As the CSF phenomen refers to the perception of a single wavelength, the second one called “masking” accounts for the interactions among coexisting stimuli. A

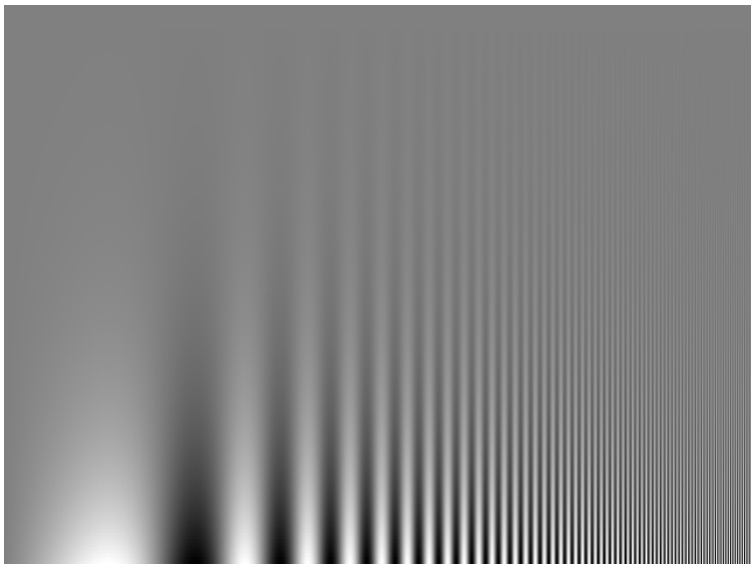


Figure 2.8: Campbell-Robson spatial CSF chart

typical effect of a spatial masking is illustrated in Fig. 2.11.

Perception of a foreground (target) is a function of the background (masker). Detection threshold of a foreground depends on the contrast and characteristics of the background [56], [72]. Typically, visibility threshold increase for the masker contrast higher than the target contrast (see Fig. 2.12). This happens when the masker and the target have different characteristics [72]. The other case is a local decrease of a detection threshold for the masker contrast around the target threshold; the target becomes more visible. This effect is called "facilitation" and refers to the case of a similar target and masker characteristics.

For modeling purpose the masking phenomena can be simplified and described by the following function [56]:

$$C_T = \begin{cases} C_{T0} & if\ C_M < C_{T0} \\ C_{T0}\left(\frac{C_M}{C_{T0}}^{\varepsilon}\right) & otherwise, \end{cases} \tag{2.2}$$

where C_T is detection threshold, C_{T0} is the absolute target detection threshold (without a masker), C_M is the masker contrast, and ε is slope of the masking function.

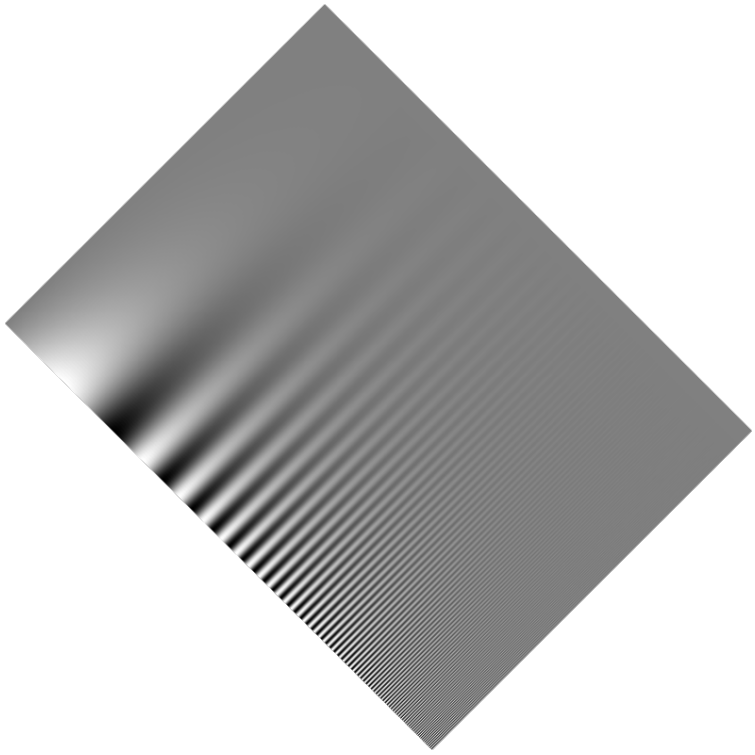


Figure 2.9: Campbell-Robson spatial CSF chart

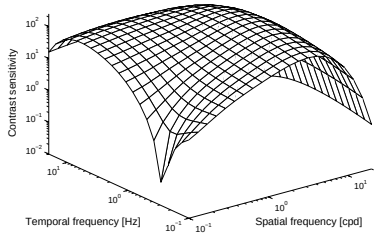


Figure 2.10: Spatio-temporal chromatic CSF [72]

Multi-Channel Model of Human Vision

Pattern sensitivity modeling is aided by the multi-channel human vision theory, proved by psychophysical experiments recently. The theory is based on a fact

Figure 2.11: Example of a spatial masking phenomena. Left: original image, right: ununiformly distorted image [73]

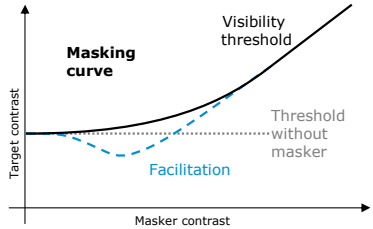


Figure 2.12: Spatial masking effect [73]

that human brain possesses a collection of a separate mechanisms, each of them tuned for a certain class of visual information. Based on this assumption, Gabor filter bank was successfully used in order to model this property [19], [56]. In this approach visual information is decompressed into separated channel, limited in orientation, spatial and temporal frequency. According to [56] for the spatial frequency domain a number of the channels ranges from four to eight. The same number of the channels were identified for the orientation. Winkler in [72] present an average bandwidth 1-2 octaves for spatial frequency and 20-60 degrees for orientation. For the temporal frequency only two or three channels exist. However, importance of the third one was not justified properly [56]. The frequency responses for the first channel (referred to as "sustained") and the second one (referred to as "transient") is presented in Fig. 2.13

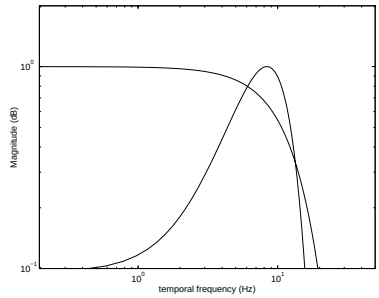


Figure 2.13: Temporal frequency response of two filter banks [56]

Pooling and Weighting

Human brain processes visual information in spatial and temporal channels as it was revealed in the previous section. Each channel is considered as an independent object and the quality evaluation should be performed for each separately. To account for a higher level of perception (video frame or a whole video sequence level) all the quality scores have to be combined together. This process is termed "pooling"; according to Winkler [72] it remains one of the most uncertain aspects of the vision. The pooling mechanism applied for the visual channels can be referred to as a "channel pooling".

Other type of the pooling is called "spatio-temporal pooling". It is based on an assumption, that a human observer never looks at the whole image at the same time. It happens because of the focus of attention and the viewing distance [56]. According to this, the global video quality metric should be computed over three-dimensional blocks (two spatial and one temporal dimension). The spatial dimensions should be adjusted according to the display size and resolution (e.g. to cover two degrees of visual angle), while the temporal dimension should fit the artifacts characteristics (i.e. should cover an average transmission error visibility stretch, that is around 1-2 seconds) [74].

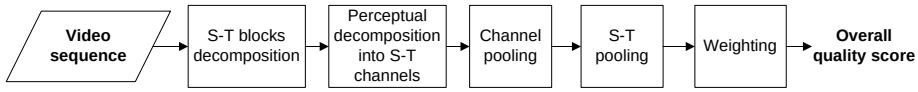


Figure 2.14: Diagram for pooling and weighting technique in case of the overall video quality evaluation

A common rule for a pooling mechanism implementation is a Minkowski summation (also known as Minkowski addition [69]). However, not every spatio-

temporal block should be considered with the same weight when the overall perceived quality is computed. This is because each discovered artifact (each block) may cover an area of different perceptual importance. This property is still not defined or fully understood. Some implementation attempts tend to use weighting according the the luminance level, spatial and temporal activity [68].

2.3 Subjective Experiments and Users' Responses Modeling

In order to enable video quality assessment in terms of user satisfaction level (qualitative scale), psychophysical experiments have to be performed. The whole process is a massive challenge and extremely time consuming. What is then the reason justifying such inconvenient process? The answer is simple: it is the only possible way of collecting subjective scores required for a model derivation. The model is in general a set of rules discovered using some statistical techniques, used for mapping quantitative value (or vector of values) returned by quality metric into a qualitative scale.

2.3.1 Subjective Quality Scales

The most popular qualitative scale is Mean Opinion Score (MOS) scale, described in ITU-T P.800 recommendation [23]. It was designed for encoded and transmitted multimedia content quality assessment. It is 5-grade scale, where 1 is the lowest perceived quality, and 5 is the highest perceived quality. MOS is constructed by averaging Opinion Scores (OSs) gathered during the subjective tests. As presented in Table 2.6, MOS scale can refer to the absolute content quality or to the impairment level.

Table 2.6: 5-grade quality and impairment scales

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible, but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

An extension of 5-grade MOS scale was presented in ITU-T p.910 [24] (see Table 2.7).

Another quality scale is R-Value, which is an example of quantitative scale for subjective quality expressing defined in ITU-T G.107 recommendation [27]. It was designed for VoIP service quality assessment. R-value can range from 1 (worst) to 100 (best) and defines a percentage of users satisfied with the service quality. In R-value scale 20 points are equivalent to 1 point in MOS scale, however, it is not very precise since the mapping is not linear. It is worth mentioning that the scale does not cover a case, when no users at all are satisfied with the quality. In

Table 2.7: 11-grade quality scale

MOS	Quality
10	The number 10 denotes a quality of reproduction that is perfectly faithful to the original. No further improvement is possible.
9	Excellent
8	
7	Good
6	
5	Fair
4	
3	Poor
2	
1	Bad
0	The number 0 denotes a quality of reproduction that has no similarity to the original. A worse quality cannot be imagined.

such case, R-value should be equal to 0, but this value is out of the scale.

2.3.2 Methodology for Subjective Experiments

General provision for the subjective assessment of the quality is presented in ITU-R BT.500-11 recommendation [22]. The methodology is designed for image or image sequence quality assessment, however, can be adopted for video sequences. According to the recommendation, subjective tests should be conducted on the diverse and numerous groups of experts (testers or subjects). For all the distorted video sequences a number of subjective opinion scores (OS) should be collected. In order to assess, how strongly few distorted parameters influence the perceived quality, each test session should include evaluation of both mono- and multi-modally distorted sequences. The number of scores collected for a single distortion has to be at least 15 [22]. What is more, a single user is able to answer correctly for about 20 minutes [17], [18], after this time the answers become less reliable and a short break is needed.

The test session consists of a series of assessment trials. Each trial refers to one distorted video sequence. The trials should be presented in a random order for each subject. The test sequences are presented only once in the test session: experiment normally ensures that the same video is not presented twice in succession with the same level of impairment [22].

Methodology for the different methods of the subjective tests is described in [22]. The first method, called Double Stimulus Impairment Scale — DSIS, operates on the five-level impairment grading (MOS). The reference object is always shown with the distorted one. Assessment of the quality refers to the distortion level, not to the absolute quality. The second method is called Double Stimulus Continuous Quality Scale — DSCQS. The quality is assessed on a continuous quality scale ranging from excellent to bad. Experts are not informed which object is the reference one, the absolute quality is assessed.

According to [22], in appropriate circumstances, also the single-stimulus (abbreviated to "SS") and the stimulus-comparison (SC) methods can be used. In the Single-Stimulus methods, a single image or image sequence is presented and the tester provides an index of the entire presentation. There are three types of SS methods: Adjectival Categorical Judgement Methods, Numerical Categorical Judgement Methods, and Non-categorical Judgement Methods [22]. An example of ITU quality and impairment scale (MOS) used in SS methods is presented in Table. 2.6. In stimulus-comparison methods, two images or sequences of images are displayed and the viewer provides an index of the relation between the two presentations. The ITU-R comparison scale is shown in Table 2.8.

Table 2.8: ITU-R comparison scale

Score	Comparison
-3	Much worse
-2	Worse
-1	Slightly worse
0	The same
+1	Slightly better
+2	Better
+3	Much better

Correct methodology for every subjective test can differ, depending on the specific test requirements. In order to fulfill the requirements, existing methods can be used directly as well as a custom composition can be applied [22].

2.3.3 Testbed

The testbed should be unique and restricted for one specific service (e.g., IPTV or mobile networks) and video content type (low quality H.263 for video conferencing vs. source quality High Definition MPEG-2). It should allow conducting subjective investigation upon prepared test material. Two main parts of each

testbed can distinguished: a distortion tool and an environment (infrastructure) for the subjective trials.

The distortion tool should allow applying such distortions to the test set as it was transmitted through the real network, with similar characteristics as considered service. The test set should consist of a number of video sequences, distorted due to the specific rules, according to the service requirements (single-parameter as well as a multi-parameters distortions should be applied). The total number of the distorted video sequences depends on the total number of the video quality parameters considered in the video quality metric.

The environment for the subjective trials should consist of at least three components, mutually connected using some network infrastructure. The network infrastructure is enforced by the service (e.g., IPTV or mobile networks). The first two components are the client and the video server applications, installed on the remote hosts. The last component is a proxy server, transparent for the video application. It should allow ports forwarding, IP addresses switching, and dropping any video packets on demand. The last functionality is assured by installing the mentioned above distortion tool.

2.3.4 User Responses Modelling

The process of the responses mapping and the model creation is presented in Fig. 2.15. The analysis of user responses consists of few steps [17]. The first two steps are common for all types of statistical models, and these are data cleaning and splitting. The first step in a mandatory one, since human answers always have some errors like missclicks or obviously incorrect answers. An example of an incorrect answer is giving much better response to much more distorted video [17]. The second step is splitting up the cleaned data in order to obtain two sets, a training set and a verification set. The model is computed on the basis of the training set only. The verification set is used to verify the model and prevent too strong correlation with the training set (the model has to perform well for general case not only for a particular case).

The next step is a model selection. The most feasible approach for mapping seems to be the linear regression algorithm (easy to understand, well defined and simple in implementation). Nevertheless, the basic assumption of the linear regression algorithm is that the response distribution can be approximated by the normal distribution. Unfortunately, subjective results do not have normal distribution and there are few reasons for this [17], [18].

One of the reason is that for the case where one of the extreme value (1 or 5) is chosen by the most testers the normal distribution reveals unrealistic symmetric error. What is more, in most of the psychophysical tests user can rate quality in numerical or verbal MOS scale (see Table 2.6). As the consequence the numbers

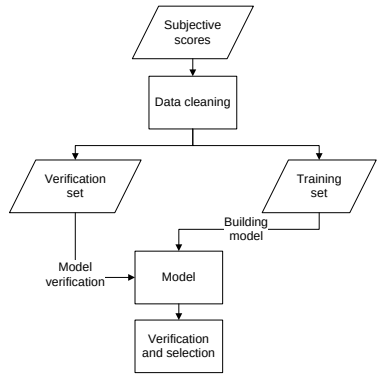


Figure 2.15: Block diagram for model creation and mapping

are only a convention and the analyzed variable (response) is of the ordinal type [1]. An ordinal variable defines only ordering but an distance between variable is not defined. Therefore, modeling the ordinal answers in the same way as strictly numerical data is a common mistake [1].

In order to overcome the problem with ordinal variables it is possible to use an 11-point quality scale described in ITU-T P.910 recommendation [24]. The assumption that residuals of obtained answers have a Gaussian distribution is correct in this case. The scale is designed in a way that the extreme values are select by testers occasionally and resulting MOS never reaches the edge of the scale.

2.4 Summary

Conclusions that emerge from the analysis presented in this chapter indicate that a comprehensive approach to quality assessment in the video delivery system requires:

- Measurement of multiple parameters at different stages (acquisition, compression, transmission). It is essential to define only key quality parameters that contribute to the overall QoE. Measurement of all possible parameters is not acceptable from the feasibility point of view.

- Selection of a measurement approach (AM or QoD), depending on the measured parameter. As already presented, the AM approach is better suited for source quality and compression artifacts, while QoD for network impairments.

- Assuring a balance between the performance and the feasibility. The AM approach aided with some basic HVS properties should provide satisfactory performance while being still computationally light. In case of the QoD approach additional inspection of a video bit-stream parameters allows to improve performance while preserving high feasibility.
- No-reference scenario for in-service applications.
- Mapping from quantitative parameters into qualitative scale.

Chapter 3

State-of-the-Art in Video Quality Metrics and Models

This chapter contains a comprehensive analysis of existing video quality metrics, classified according to the criteria proposed in the previous chapter. A study on no-reference metrics is more extensive than for full-reference metrics, what reflects current trends. Within the presented NR metrics there are two groups of a special interest: 1) Spatial Artifacts Approach and 2) Quality of Delivery Approach. The first group is important because it constitutes a background to the presented work. The second group is important because it constitutes a motivation to consider quality issues related to network losses as a mature and extensively analyzed research topic.

3.1 Full-Reference Video Quality Metrics Review

In this section, selected video quality metrics representing the Full-Reference approach are presented. This approach has been quite popular recently and a large diversity of metrics is available to the public (in the most cases only as a description but some implementations are available as well). The presented metrics are the best within this category and their high performance was proved by using a proper methodology.

3.1.1 Integrated Spatial and Temporal Artifacts Approach

Perceptual Evaluation of Video Quality (PEVQ)

An example of a quality metric offering the overall quality score in the MOS scale is the Perceptual Evaluation of Video Quality (PEVQ) based on ITU-T J-144 [26]. It is designed to estimate the video quality degradation occurring through a network and is based on spatial and temporal artifacts measurement aided by a model of HVS. PEVQ outputs MOS and additional indicators for a more detailed analysis of the perceptual level of distortion in the luminance, chrominance and temporal domains [46]. The algorithm can be divided into four independent blocks as presented in Fig. 3.1 [43].

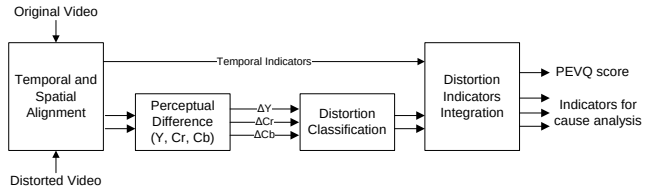


Figure 3.1: Block diagram of PEVQ [46]

The first block is a pre-processing block responsible for a spatial and temporal alignment of original and distorted video sequences in order to assure comparison of the corresponding frames. The second block is responsible for distortions (differences) calculation. Each type of distortion is weighted according to the influence on the perceived quality it provides and calculated separately for the luminance and two chrominance channels. Another indicator taken into account is motion of the reference video sequence, since perception of details is much higher for slow-motion videos. The third block classifies computed indicators and detects certain types of distortions. The last block calculates the final quality score based on indicators and distortions previously detected [43].

Except the overall MOS, the perceptual level of distortion in the luminance, chrominance and temporal domains are provided. The list of other indicators (quantitative values only) is presented in Table 3.1 [46]:

3.1.2 Structural Information

Structural Similarity Index Metric (SSIM)

In [67] an innovatory Full-Reference quality metric based on structural similarity was proposed. It was originally designed for still images quality assessment, however, an extension for video applications was presented in [68]. As presented by Wang in [65], [66] the human visual system (HVS) is very sensitive to the

Table 3.1: Additional indictors in PEVQ

Name	Description
Delay	The delay of each frame of the test signal related to the reference signal
Brightness	The brightness of the reference and degraded signal
Contrast	The contrast of the distorted and the reference sequence
PSNR	Allows for a coarse analysis of he distortions in different domains the $PSNR$ is provided for the Y, Cb and Cr components separately
Jerkiness	Describes the smoothness of a video playback which is often impaired by down-sampling, coding processes and perturbed transmissions
Blur	A distortion characterized by reduced sharpness of contour edges and spatial detail
Blockiness	Often the result of a low bit rate coding that uses a block matching algorithm for the motion estimation and a coarse quantization for the image blocks
Frame Skips and Freezes	Temporal artifacts occurring in video transmissions caused by e.g. overloaded networks
Effective Frame Rate	Down-sampling of a video signal on a frame by frame basis often results in loss of information which often leads to the degradation of the video signal. The effective frame rate is an indicator quantifying the severeness of such a process
Temporal and Spacial Activity	Temporal and spacial activity indicators quantify the amount of activity /movement in the video content. These indicators are derived from ITU-T P.910 [24]

structural information provided on an image in the viewing field. Based on this assumption, it was proved that metric for structural changes assessment can have good correlation with the perceptual quality [67]. The Structural Similarity Index (SSIM) is a top-down approach modeling functionality of the overall HVS. Suppose x and y are the reference and the distorted image signals, respectively. The overall similarity measure $S(x, y)$ is combined of three components: local luminance $l(x, y)$, local contrast $c(x, y)$, and structure $s(x, y)$ comparison between the original and the distorted images as presented in (3.1).

$$S(x, y) = f(l(x, y), cS(x, y), s(x, y)) \tag{3.1}$$

Block diagram of the SSIM measure system is presented in Fig. 3.2.

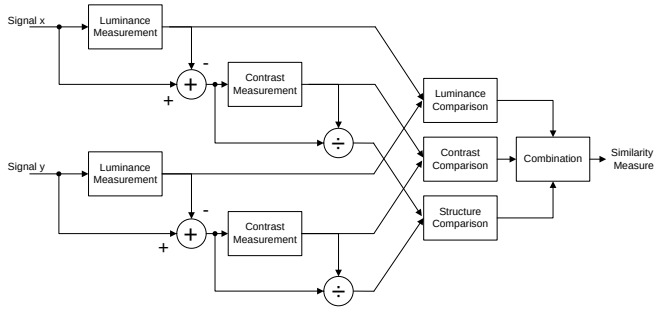


Figure 3.2: Diagram of the SSIM measure system [67]

Video quality assessment using SSIM is performed on the local region level, the frame level, and the video sequence level as presented in Fig. 3.3 [68]. First, random 8×8 pixels blocks are extracted from the original and the distorted video sequences. In the first level, SSIM index is calculated for each block for Y, Cb and Cr components separately. $SSIM_{ij}$ denotes similarity index for j-th block in i-th frame.

$$SSIM_{ij} = W_Y SSIM_{ij}^Y + W_{Cb} SSIM_{ij}^{Cb} + W_{Cr} SSIM_{ij}^{Cr} \tag{3.2}$$

In the second step, local quality values are combined into a frame-level quality. The quality of local regions is weighted according to the mean luminance level (dark regions are less sensitive to the quality degradation). Q_i denotes the quality index for i-th frame, and w_{ij} denotes weight for j-th block in i-th frame.

$$Q_i = \frac{\sum_{j=1}^{R_s} w_{ij} SSIM_{ij}}{\sum_{j=1}^{R_s} w_{ij}} \tag{3.3}$$

In the last step, the quality for the entire video sequence Q is computed. Frame-level quality is weighted using frame motion vectors, since some types of distortion (e.g., blur) do not affect the perceived quality for scenes, where large motion occurs. W_i denotes the weight for i-th frame.

$$Q = \frac{\sum_{i=1}^{F} W_i Q_i}{\sum_{i=1}^{F} W_i} \tag{3.4}$$

Assuming the original video sequence identical to the distorted one, then $w_{ij} = 1$ for all i, j and

$$W_i = \sum_{j=1}^{R_s} w_{ij} = R_s$$

for all i.

Cross-video and cross-distortion tests were performed using VQEG Phase I test data set [59]. In order to evaluate quantitative performance of video SSIM, four methods standardized by VQEG Phase I FR-TV were employed [61]. The results proved a good correlation with the subjective score and a strong potential of the structural similarity approach. Detailed results are presented in [68].

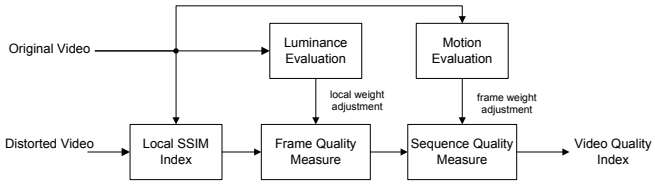


Figure 3.3: SSIM based video quality assessment system [68]

3.1.3 Vision Modelling Approach

Moving Pictures Quality Metric (MPQM)

Moving Pictures Quality Metric (MPQM) is an objective quality metric for moving picture using the vision modeling approach [56]. The model accounts for the spatio-temporal aspects of HVS, namely the contrast sensitivity and masking. Based on the assumption that HVS processes visual information in separated spatial and temporal channels, the authors proposed the filter bank approach to vision modeling, which decomposes the original and coded video sequences into perceptual channels as presented in Fig. 3.4.

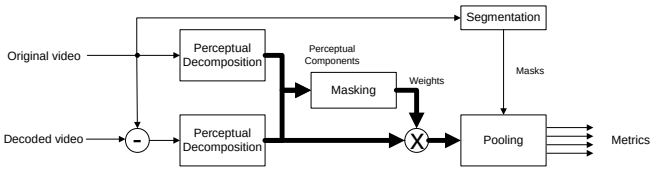


Figure 3.4: Block diagram of MPQM. Thick lines represent a set of perceptual components, thin lines represent video sequences [56]

At the same time, the original video sequence is being segmented using uniform areas, textures, and the contours classification block by block. In the next step, contrast sensitivity and masking are considered for each perceptual channel in detection threshold calculation. The assumption is that the original video sequence is a masker for the distortion. Afterwards, the filtered error signal is divided by the detection threshold. In the last step, data from the channels is gathered together in order to account for a higher level quality; this process is called pooling [56] and the distortion level is measured. MPQM provides the overall quality metric and three detailed metrics for uniform areas, textures, and contours. In case of the detailed metrics, three different masks created in the segmentation block are used. In case of the overall metric, the distortion level is computed by each three-dimensional block. The first dimension is temporal, the other as spatial, selected to cover 2 degrees of visual angle. MPQM for a given block is computed according to (3.5):

$$E = \left(\frac{1}{N} \sum_{c=1}^{N} \left(\frac{1}{N_x N_y N_t} \sum_{t=1}^{N_t} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} |e[x,y,t,c]| \right)^{\beta} \right)^{\frac{1}{\beta}} \tag{3.5}$$

where $e[x, y, t, c]$ is the masked error signal at position (x, y) and time t in the channel c; N_x, N_y, and N_t are dimensions of the blocks; N is the number of the channels.

In order to present distortion E in a some well-know scale, the authors proposed two alternatives. The first one is the logarithmic scale referred to as "visual decibels"; the metric can be named Masked PSNR (MPSNR):

$$MPSNR = 10 \log_{10} \frac{255^2}{E^2} \tag{3.6}$$

The second idea is to map distortion E onto the five-grades MOS scale by using the following equation:

$$Q = \frac{5}{1 + NE} \tag{3.7}$$

where N is empirical constant.

High performance and good correlation MOS was proved by using video sequences coded with MPEG-2 and H.263 video codecs.

3.2 No-Reference Video Quality Metrics Review

In this section selected video quality metrics representing the No-Reference approach are presented. Since it is an emerging approach, only one metric is well defined and commercialized – V-Factor [8].

3.2.1 Vision Modeling Approach

Gabor Difference Analysis (GDA) and Reverse Frame Prediction (RFP)

Two mutually related quality metrics are presented in [19]. The first one, the FR Gabor Difference Analysis (GDA), is directly utilized in the second one, i.e., in NR Reverse Frame Prediction (RFP). The GDA metric is based on HVS characteristics. Both, the original and distorted video sequences and transformed into the SCT color space (commonly used in medical imaging as being reasonably insensitive to variations of illumination [19]). Then the Gabor filter bank is used to transform video frames to different scale (spatial frequency) and orientation channels. The GDA metric is well correlated with subjective experiment results. However, it does not include spatial and temporal registration (it can be done in the future work). The second quality metric, RFP, assumes only few frames in the distorted video sequence are corrupted (the rest of the frames are of acceptable quality). The metric grades video quality by comparing two frames of the distorted video sequence using the GDA metric. A flowchart of the RFP metric is presented in Fig. 3.5. Every time a scene cut is detected in the video sequence, the quality analysis process is restarted. Cuts are detected using cross correlation coefficients $C(0,0)$ of the successive frame.

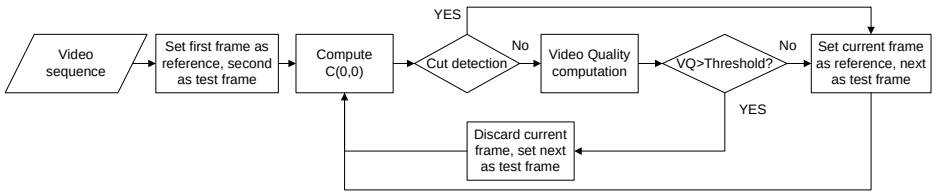


Figure 3.5: Flowchart of RFP method [19]

Just-Noticeable Distortion (JND)

An interesting approach to digital video quality assessment is presented in [33]. The approach is not capable of direct perceived quality assessment, however it can be useful in more effective bandwidth utilization, that may result in a higher video quality. In [33] the authors presented a model for the Just-Noticeable Distortion (JND) estimation, which refers to the maximum distortion

that the Human Visual System (HVS) cannot perceive. The model can be applied for still images and video sequences as well. It incorporates both spatial and temporal HVS properties and operates in the Discrete Cosine Transform (DCT) domain. The factors affecting JND are spatio-temporal Contrast Sensitivity Function (CSF), eye movement, luminance adaptation and intra- and inter-band contrast masking. The presented model outperforms relevant existing models because it is capable of predicting more aggressive JND values without any noticeable distortions. According to [33], the model can be useful in image/video compression, watermarking, perceptual quality evaluation and visual signal enhancement.

3.2.2 Spatial Artifacts Approach

Metric Based on Artifacts

In [14] the authors presented NR quality metrics based on image artifact measurements: blockiness, blurriness and noisiness. The metric for the first artifact, blockiness, is based on the method proposed by Vlachos [58]. The blockiness signal strength is estimated by comparing the cross correlation of pixels inside (intra) and outside (inter) a frame. The blurriness metric uses the Canny Edge Detection algorithm in order to obtain the magnitude of the edge pixels $M(i, j)$, and their orientation $O(i, j)$. The blurriness level is computed by averaging width of the strong edges for all video frames. The noisiness metric is based on the work presented by Lee [37] that estimates the noise level by a local variance of flat areas. The overall quality level is expressed as a combination of three single metrics using both Minkowski metric and the linear model. Test preformed on results obtained in a subjective experiment proved good performance of the metric (linear Pearson correlation $r = 0.86$). Weakness points of the proposed approach are three quality parameters only and performance of the metrics proved only upon MPEG-2 video sequences.

MPEG Blocking and Random Digital Error Metrics

In [9] the authors presented a prototype NR video quality system based on real time analog television and digital video stream quality measurement. Digital quality measurement consists of two metrics: Digital MPEG Blocking Metric and Random Digital Error Metric. The fist one is based on a movement of 8×8 pixels block boundaries along the whole video sequence. The second one is based on an assumption that corrupted pixels are routinely very bright green and, therefore, easy to detect. Few precautionary actions were taken in order to prevent misclassification of the corrupted pixels. The perceptual impairment is computed by rescaling metrics output by a Level of Detail Metric, that implement the spatial masking theory (artifacts are more visible on smooth regions).

Flickering Artefact for H.264/AVC Video

According to the work presented by Pandel in [44], flickering is the most annoying temporal artefact inherent for predictive inter-coded video sequences (in particular for H.264 encoded sequences). Video sequences with slow camera movements or zoom are especially exposed to flickering artefact. In temporal predictive coding macro-blocks are not updated (encoded) until the difference between corresponding macro-block of successive frames is above a specific threshold. The stronger the compression, the higher the threshold. This suggests that each macro-block remains in one of two possible states: 1) no-update – differences between successive frames are below the threshold, and 2) update – the opposite case [44]. Frequent changes between the two states denote a severe flickering artefact. The metric is calculated as a normalized number of transitions between states. The two state model including a hysteresis is detailed in [44].

Metrics for Degraded and Enhanced MPEG Video

An interesting work presenting an integrated no-reference quality metric for degraded and enhanced video is presented in [6]. Several image and video artifacts were combined into an integrated formula for the overall quality assessment. Considered artifacts were related to the source quality and video compression. Subjective experiments were carried out in order to train and verify the proposed model. Obtained correlation was satisfactory and significantly higher than for PSNR but only a limited number of tests sequences was used. Additionally, no discussion on the sequences complexity and diversity was provided. The results have been elaborated for previous coding standards (when compared to H.264/AVC) where except blockiness other compression artifatcs like ringing or clipping were essential [6]. For the current hybrid block-based motion compensated predictive coding schemes (H.264/AVC) it is more important to account for temporal artifacts like flickering [44]

3.2.3 Quality of Delivery Approach

V-Factor

V-Factor based on Moving Picture Quality Metrics (MPQM) is the only well standardized approach towards No-Reference video quality assessment [8]. The block diagram of the V-Factor quality metric is presented in Fig. 3.6.

V-Factor is an implementation of MPQM and provides not only the user-perceived video quality score but also some additional information [53], [54]:

Extra information listed in Table 3.2 allows monitoring and diagnosing the root of problems. The parameters are transport stream key parameters defined in ETSI TR 101 29 [12] and those related to the network layer, defined in ITU-T Y.1540/1541 [25] or IETF RFC2330 [8].

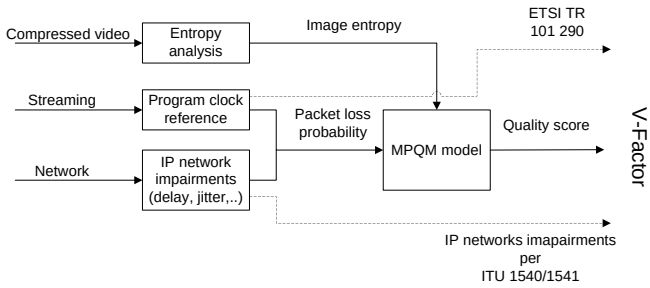


Figure 3.6: Block diagram of V-Factor [8]

The MPQM model should be fed with several parameters, namely the Packet Loss Rate (PLR) probability and the image entropy as presented in Fig. 3.7.

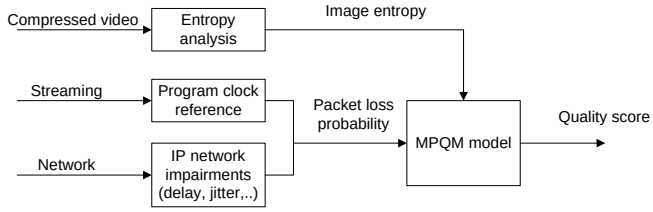


Figure 3.7: MPQM model [8]

The image entropy consists of such features as relative size of frames, size of Group of Pictures (GoP) since loss of an I, B and P frames can have different impact on the perceived quality [8].

V-Factor is calculated according to (3.8):

$$VFactor = Q_{er}(qs)\left(1 - e^{\frac{P_{lr}}{\sqrt{\Gamma_r}}}\right) \tag{3.8}$$

where $Q_{er}(qs)$ is a function that measures impairment coming from a given codec, P_{lr} is a PLR probability, and Γ_r refers to video complexity (entropy).

Customer Oriented Metric for H.264/AVC

An application of the customer oriented measurements for H.264/AVC video streaming standard is presented in [39]. The process of delivering and decoding is described as a ďż˝black boxďż˝ where some noncontrollable parameters like the packet loss or delay jitter interact with controllable parameters like the video bit rate, see Fig. 3.8.

Table 3.2: Additional information in V-Factor

Name	Description
Program rate	Transport stream rate as observed
Program clock rate overall jitter	Jitter of synchronization stream
Jitter discards	Number of frames discarded due to jitter
Out of sequence	Number of misordered frames delivered
In sequence	Number of properly ordered frames delivered
Network loss probability	Statistically accurate predictive calculation of frame loss
Max loss episode length	Maximum number of lost frames per episode
Max loss episodes	Cumulative count of loss episodes since first observation
Multicast join time	Current time the stream way joined in "unix epoch" seconds
Multicast first time	Current time the first data arrived in "unix epoch" seconds
Compression ratio	Ratio of I frames to I+P+B frames
I frame count	Number of "I" (Intra) frames for the current sample period
P frame count	Number of "P" (Predictive) frames for the current sample period
B frame count	Number of "B" (Bidirectional) frames for the current sample period

Response consist of the amount of detected video and audio distortions [39]. Analysis includes the influence of every factor in the response output (Pareto analysis) and contribution of these factors to the variability of the entire system (ANOVA analysis) [39]. Subjective tests with a single-stimulus methodology were conducted, the amount of distortions was counted. Different network conditions were emulated with *Netem* software. Results show the most prominent factors affecting the observed response in both Pareto QoE Audio/Video and ANOVA QoE Audio/Video analysis. The list of the considered parameters is presented in Table 3.3.

According to the results, the authors proposed an adaptive multimedia streaming over IP technique, allowing to preserve the sustained quality of experience. For those interested please refer to [39].

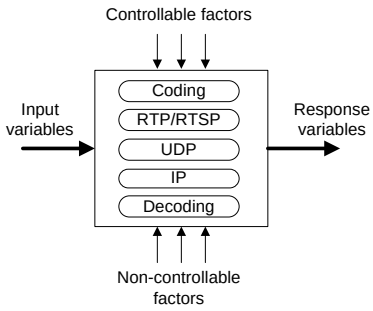


Figure 3.8: Black box modeling of H.264 delivered over RTP [39]

Table 3.3: Considered network parameters

Packet Loss Rate
Delay
Delay Variation
Video coding rate (CODVideo)
Delay CODVideo
CODVideo for Real Time Protocol

MPEG-2 Quality Prediction and Control

Verscheure in [57] explains why a common belief, that increasing video stream bit rate enhances the perceived quality, is misleading. The paper address the problem of quality prediction and control of an MPEG-2 video stream transmitted through the lossy (in terms of a packet loss) network. The MPEG-2 video standard is analyzed and the impact on the visual quality of packet loss is discussed. Two quality factors are considered: a video bit rate and a packet loss ratio. Impact of each factor is considered separately as well as joint impact of both factors is analyzed. The ground truth regarding video quality typically obtained from the subjective experiments is replaced using MPQM metric [56]. As presented in Fig. 3.9, video quality may decrease when the bit rate increases.

The explanation for this is simple. The higher bit rate the more packets lost. Hence, the probability of a packet lost causing a significant quality degradation is higher. The most important conclusion is that for each PLR the optimal bit rate have to be found (different across diverse video sequences).

Metric Based on Network Parameters

In [52] the authors presented the NR quality metric based on two network parameters related to packet loss: MeanDistance (MD) that reflects the mean

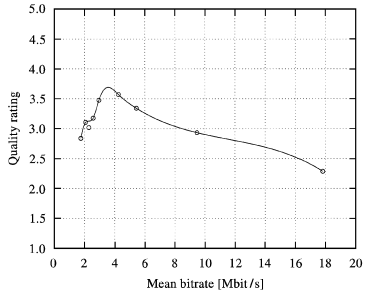


Figure 3.9: MPQM quality versus encoding bit rate for $PLR = 5 \times 10^{-3}$ [57]

distance (in frames) between packet losses, VarDistance (VD) that reflects variance of the distance. The ground truth for the perceived video quality is expressed as a MeanQuality (MQ) that is a mean value of a blockiness parameters for all the frames in the video sequence. To build up the relation between mentioned network parameters and the video quality MQ, the linear regression method was used. Results proved that MD and VD parameters show quite good correlation with the video quality defined by the authors (correlation coefficient $r^2 = 0.8367$). The main drawback of this solution is the fact that it is restricted to an MPEG-2 video stream and does not consider subjective tests results as a reference video quality.

Quality Monitoring for H.264/AVC Video Recent work devoted to network impairments assessment for H.264/ACV video is presented in [41]. The NORM (NO-Reference video quality Monitoring) algorithm estimates quality degradation due to channel errors (network losses). Quality assessment is based on lack of motion vectors, prediction of residuals, and temporal error propagation. The performance of the algorithm was verified against a FR metric – SSIM.

Metric for Compression

Although QoD approach is mainly suited for network impairments, there are some emerging efforts utilizing this technique to assess compression artifacts. In [4] the authors present a no-reference quality assessment metric for H.264/AVC compression. The proposed model represents a parametric approach, i.e. quality is estimated using parameters extracted from a bit-stream (H.264 bit-stream) without video decompression. It comprises two main steps: coding error estimation and perceptual weighting of this error. The authors show that the metric outperforms PSNR and achieves the Pearson correlation coefficient $R^2 \approx 0.9$. Another work devoted to H.264/AVC compression quality assessment is presented in [55].

This time it is a hybrid approach where parameters extracted from the H.264 bit-stream are aided with video content characteristics calculated on decompressed video frames. According to the results presented, the model estimates the subjective quality at a correlation coefficient $R^2 \approx 0.9$ and outperforms PSNR. In both papers discussed, quality estimation is based on quantizer information from the H.264 bit-stream. Such an approach is restricted to a given bit-stream model, video codec, and even codec profile; however, it cannot be easily adopted to other compression schemes. Moreover, it is common for video content to be compressed prior to transmission and trans-coded then. In such a scenario quality estimation based on quantizer information will reflect only artifacts that result from the final trans-coding. In contrast, a metric based on image analysis detects a cumulative effect, i.e. all the artifacts actually seen by the user.

3.2.4 Watermarking (Data Hiding) Approach

Data Hiding

An innovative approach towards the NR video quality assessment based on data hiding was presented in [13]. Video quality assessment is based on calculation of the impairment of the well known mark embed in a video stream using the Total Square Error (TSE) pixel-to-pixel metric. The authors presented two models. The first allows to estimate the "best" strength of the mark embedded in the video stream, based on its visibility and the data hiding capacities of the host video. The second model allows to predict TSE of the transmitted video that is mapped into MOS. Performance of the presented metric is quite good (correlation coefficient $r^2 = 0.8833$ with the MOS values) and was proved by subjective experiments.

Data Hiding In Perceptually Important Areas

The work presented in [5] can be perceived as a improvement of the method presented in [13] based on data hiding. The authors presented the NR quality metric using a data hiding technique with a mark embedded into perceptually important areas of the video sequence only. In order to calculate the important areas, the authors used overall importance maps consisting of three separated maps for motion, contrast and color. The conducted tests proved that the presented metric outperforms the simple PSNR metric and correlates well with MOS (no exact correlation value specified). However, since the metric operates on DCT coefficients it is codec dependent.

3.3 Summary

The presented SotA analysis reveals some points that require improvement and further studies. An aspect of source video quality (artifacts related to the acqui-

sition phase) is underestimated and almost never has been taken into account. A comprehensive approach towards perceptual assessment of video compression schemes requires addressing not only artifacts related to intra- and inter-frame compression but also an influence of frame resolution and frames per second rate. An impact of video compression for diverse video content cannot be estimated properly using a metric measuring only one artefact. Furthermore, any metric suited for source video quality and video compression should represent the no-reference approach because in a real system it is very likely that the original sequence is not available [73].

In contrast to the QoD approach, metrics based on image analysis detect *a cumulative effect* of the entire video delivery chain, i.e. all the artifacts actually seen by the user. Artifacts measurement approach is not aware of transmission scheme, bit-stream standard and coded parameters, therefore it can be generalized over different acquisition and compression schemes based on DCT and involving both intra- and inter-frame compression.

Chapter 4

Derivation of Objective Video Quality Metrics

This chapter presents derivation of objective video quality metrics devoted to quantitative assessment of video artifacts. A metric should be understood as a definition of a particular artifact measurement method. It should not be confused with a perceptual model, responsible for mapping quantitative metric into qualitative MOS scale.

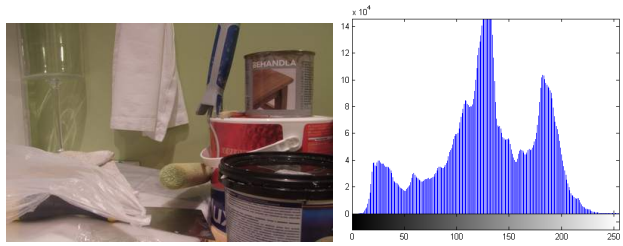
4.1 Acquisition Related Quality Metrics

This section includes details on derivation of metrics for three artifacts which are essential to a video acquisition phase. The metric for exposure is a pioneer approach, while the metrics for blur and noise were based on some well-known image properties inherent for each artifact. All the metrics were developed by the author and implemented in Matlab environment.

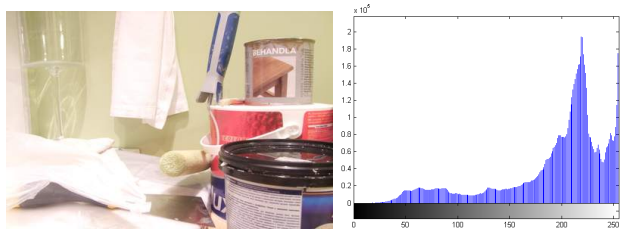
4.1.1 No-reference Exposure Metric

Exposure is the total amount of light allowed to fall on a photographic medium during the process of taking a photograph [70]. In the digital domain the photographic medium is represented by a matrix of light sensors. A correct exposure means that light sensors were exposed to such an amount of light that the image histogram covers a desired range of the luminance. This definition refers to both still and moving digital images. Any distortion in the exposure domain may occur only during the image acquisition process. The same picture taken at different

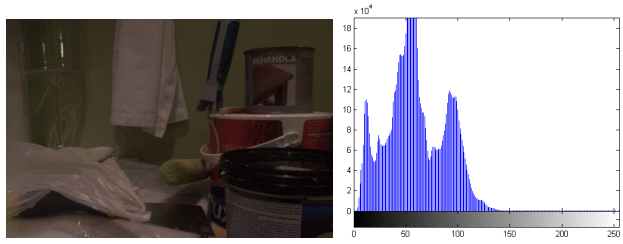
exposure levels is presented in Fig. 4.1. Associated histograms (the distribution of image pixels luminance) illustrate correct-, over-, and under-exposure. Picture in Fig. 4.1(a) shows a correctly exposed image. Its histogram is spread over the whole range of the luminance. Histograms of over- and under-exposed images are shifted to the bright and the dark side, respectively (see Fig. 4.1(b) and Fig. 4.1(c)).



(a) Proper exposure



(b) Over-exposure



(c) Under-exposure

Figure 4.1: Pictures and histograms for different exposure levels

The most popular research area dealing with image exposure is referred to as High Dynamic Range (HDR) imaging. The idea of this technique is to enhance a dynamic range of an image (greater range of luminance between the lightest and the darkest area) beyond capture device capabilities. It is obtained by merging a

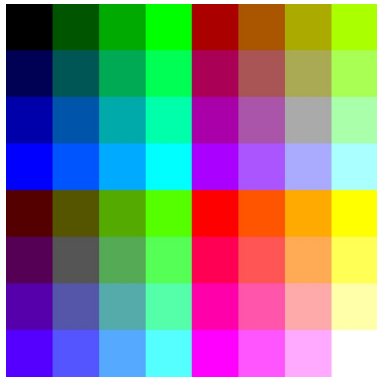
number of pictures (usually 2 or 3) taken at different exposure times [42]. Other research taking advantage of exposure modifications is presented in [45]. In the experiment optimized exposure and ISO values are verified with respect to image quality reproduction. In the result, a dynamic range of an image is extended and visual noise is minimized. In both of the presented research areas image exposure is considered as a controllable parameter that can be described using acquisition parameters (aperture, ISO, exposure time). The open questions are how to quantify the overall image exposure and how it relates to the quality perceived by people (QoE).

The purpose of the presented research was to develop a No Reference (NR) metric assessing QoE for video affected by exposure distortion. The presented exposure metric is unique and no similar research has been found in the literature [50]. Since the research is devoted to video sequences, it was necessary to develop an exposure generation model. Such a model should allow for an introduction of a desired level of over- and under-exposure into a video sequence. The reason is that in contrast to still images, it is not possible to produce exactly the same video sequences with different exposure levels using a single camera. What is more, the research was performed on existing sequences, shared by the video quality community [59].

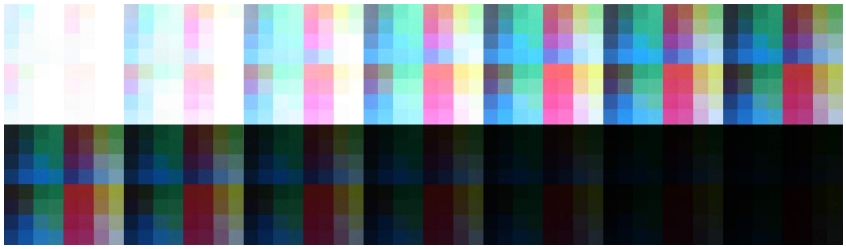
Exposure Generation Model

In contrast to still images, it is not possible to produce exactly the same video sequences with different exposure levels using a single camera [50]. Therefore, it was necessary to develop an exposure generation model capable of introducing the desired level of over- and under-exposure into a video sequence. The model is required to modify the luminance of the original video sequences in such a manner as they were captured with different exposure parameters.

There are three methods to control exposure in digital cameras. Firstly, by opening or closing the aperture. The larger the hole of the iris the more light is reaching image sensors in a given time. Secondly, by changing the light signal boost what is commonly referred to as ISO sensitivity. The last way to control image exposure is to change the exposure time. It is the most straightforward way and assures that other image parameters remain unchanged. In the case of the aperture adjustment the depth of the field DoF is affected additionally. It is not a desired effect since a large hole of the iris yields significant reduction of DoF. Another undesirable effect is associated with ISO adjustment. The higher the ISO the more sensitive image sensors are and therefore the possibility to take pictures in low-light situations. The disadvantage of this process is associated noise level increase. The above consideration became the cue to select the exposure time adjustment as the only way to control exposure in the performed experiment.



(a) Test chart



(b) Different exposure times

Figure 4.2: Test chart used for the experiment

In order to build as generic an exposure generation model as possible a custom test chart was created (see Fig. 4.2(a)). It includes 64 uniform regions. For each color component (R, G, B) four values of color saturation (0, 85, 170, and 255) were selected. The chart includes all possible combinations (i.e. $4 * 4 * 4 = 64$). For the purpose of the experiment it was printed on a paper sheet. Afterwards, pictures of the chart were taken (see one of each two in Fig. 4.2(b)) using a digital Single-Lens Reflex (SLR) camera with a fixed aperture (F8), fixed ISO (400) and 28 different exposure times (typical values from 0.0005 to 0.25 seconds).

In order to propose a model capable of exposure generation it was necessary to analyze the pixels color saturation for the 28 produced pictures. It was decided to use the RGB color space instead of the YCbCr one because both the Charge Coupled Device (CCD) and the Complementary Metal Oxide Semiconductors (CMOS) image sensors separate light into the RGB components during the light acquisition process.

For each of the 64 uniform regions the average color saturation was calculated

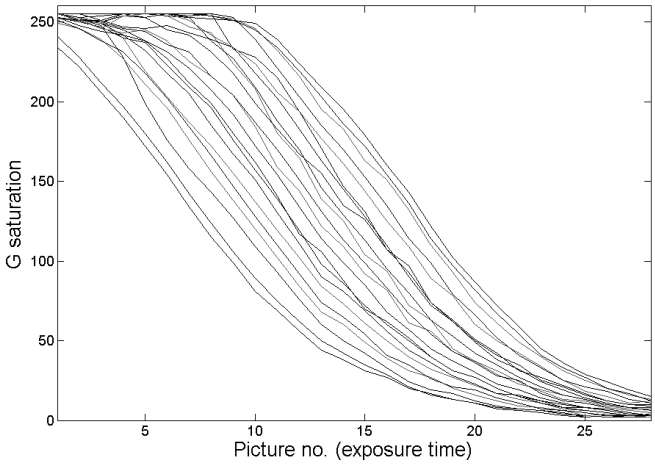
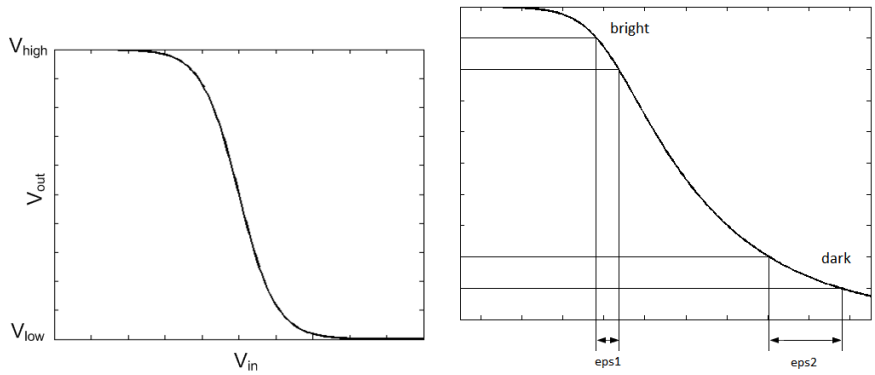


Figure 4.3: Results obtained for the G color component

on each picture. The color reproduction was obviously not perfect when compared to the original chart. However, for the experiment purpose the only requirement was to be able to analyze a complete representation of the color space. The results for the G color component (the most representative one from R, G, and B) are presented in Fig. 4.3. Each curve represents the color saturation of the G component of the same uniform region read from 28 pictures taken at different exposure times. 64 lines correspond to 64 uniform regions.

Exposure time, which changes exponentially, has been replaced on the x-axis by a picture number (linear scale). This is a good approximation of the resultant change in the exposure (observed in the picture) due to a nonlinear light to the brightness (color saturation) conversion used in photography. Replacing a picture number with an exposure time on the x-axis causes a significant increase in the observed asymmetry.

Obtained plots present a characteristic that is in line with the intuition. It is similar to the physical characteristics of a typical image sensor presented in Fig. 4.4(a). For the simplicity input voltage V_{in} represents the amount of light reaching an image sensor (exposure) and output voltage V_{out} represents the resulting color saturation. The voltage transfer characteristic can not be directly adopted because there are different image sensors and different characteristics of the remaining part of the image acquisition path (e.g. lens or postprocessing). The transition from $V_{\mathrm{out}} = V_{\mathrm{high}}$ to $V_{\mathrm{out}} = V_{\mathrm{low}}$ is symmetric and centered around $(V_{\mathrm{high}} - V_{\mathrm{low}})/2$. In digital photography the voltage transfer characteristic is modified asymmetrically in order to encode the bright light with a higher resolution



(a) Characteristic of CMOS inverter

(b) Asymmetric characteristic of an image sensor used in digital photography

Figure 4.4: Voltage transfer characteristics

than the dark one. It is presented in Fig. 4.4(b) where a much greater exposure change ($eps2 >> eps1$) is necessary to produce the same luminance change in the dark than in the bright light.

The same properties can be observed in the obtained plots (compare Fig. 4.3 versus Fig. 4.4(b)). It is clearly visible that regardless of the initial color saturation value all curves follow nearly the same function.

It was necessary to propose a function that can be used for mapping between exposure level x and pixel saturation values y for all RGB components. The following properties were desirable: limited values, saturation around the extremes, and asymmetry in order to account for any non-linear processing. The selected function was an asymmetric logit function (ALF) [16] given by Eq. 4.1.

$$y = \left(\frac{e^{ax+b}}{1+e^{ax+b}} \right)^{c} \tag{4.1}$$

where a, b and c are estimated parameters.

An example of the ALF function is shown in Fig. 4.5. Note that all presented functions have similar behavior for $x > 3$ and different for lower values (i.e. the desired asymmetric behavior).

The ALF function is a non-linear function, so the parameter estimation is not a trivial task. For the research purpose a nonlinear least squares method was used. Note, that for this method it is important to choose proper starting points for all estimated parameters. In the presented research the exposure level

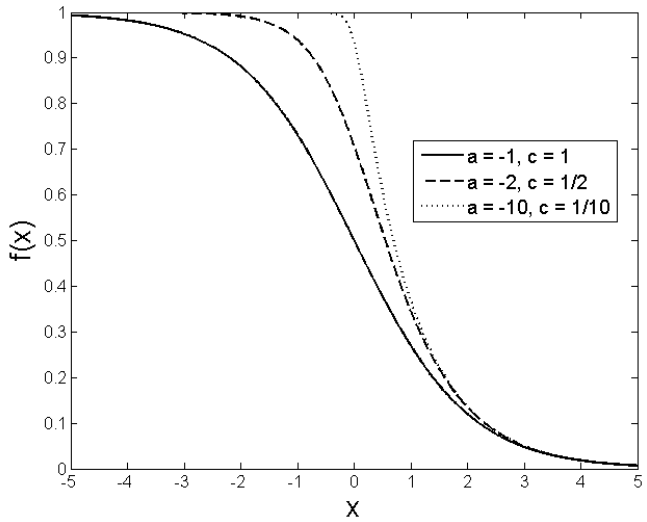


Figure 4.5: An example of different asymmetric logit functions, $b = 0$

was described by a virtual variable, i.e. experiment number from 1 to 28. It was necessary to normalize it prior to the estimation using the following formula:

$$x_{norm} = \frac{x - \overline{x}}{std(x)} \tag{4.2}$$

where $std(x)$ is .

The analysis shows that the best results (averaged R^2 values higher than 0.995 for R, G and B color components) have been obtained for the following values (with 95% confidence bounds): $a = -3.2$ $(-3.71, -2.69)$, $b = -1.3$ $(-1.71, -0.88)$, $c = 0.4$ $(0.30, 0.49)$.

Obtained results show that the proposed model fits the test chart data. Hence, the model is expected to produce a realistic exposure modification effect when applied to the video content. The effect is obtained by the modification of the color components' saturation (for each single frame of a video sequence) according to the derived mapping function.

Verification of the Proposed Model for Exposure Generation

The next step after the finalization of the model for exposure generation was to validate it against results obtained for real pictures. It was necessary to verify whether a model built upon a given image and digital camera can be generalized

over different images and cameras. For this purpose 3 different scenes were captured using two different digital cameras, with numerous exposure times. Each scene varies in terms of content, lightning, amount of details, and colors.

Figure 4.6: Illustration of the observed problem with increasing difference between color components. From the left to right: original image, under-exposed image with visible effect, and under-exposed image with compensated effect.

During a visual inspection of the exposure generation results one problem was discovered. It is illustrated in Fig. 4.6. The problem was identified on image regions were two out of free color components were saturated (value 0 or 255). In case of under-exposure generation three glaring colors were likely to appear on the bright regions, namely Aqua, Fuchsia, and Yellow. The common feature of these colors is that one color component has a much lower value that the two others. It was a natural consequence of the function used for the mapping between exposure level and pixel values. For two saturated color components the same exposure modification resulted in a much lower value change than compared with the third component (not fully saturated on the original image). The stronger exposure modification the higher difference between color components. This was not an intended feature but also an inevitable one due to the use of a limited range of values, i.e. the value of the pixels that were already saturated does not change even if stronger exposure is applied.

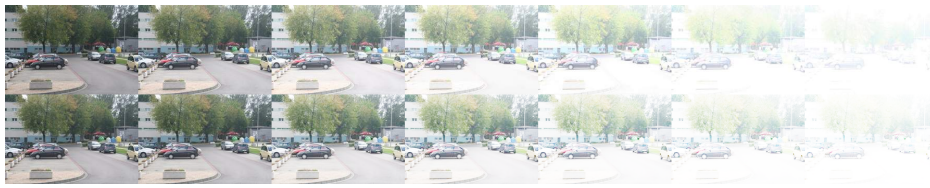
To eliminate this adverse effect a compensation method was proposed. This method involves a slight reduction in the value of the saturated pixels. In result, a difference between the color components does not grow as much as in the described example. Results of the improved exposure generation model are presented in Fig. 4.7[1]. Top pictures are the original ones captured by digital cameras with different exposure times. Bottom pictures are obtained from the original picture using the proposed exposure generation model. Generated exposure level is closely adjusted to the corresponding original images but more important is proper color handling. Some colors fade much faster than others (see green tree in Fig. 4.7(b) or bright cars in Fig. 4.7(d)).

The generation of a desired exposure level is illustrated in Fig. 4.8. The first

[1]In order to get a full impression it is necessary to view this figure on the source PDF file.

(a) Test pictures

(b) Landscape over-exposure example



(c) Outdoor over-exposure example

(d) Outdoor under-exposure example

(e) Indoor under-exposure example

Figure 4.7: Visualization of the exposure generation model performance for over- and under-exposure

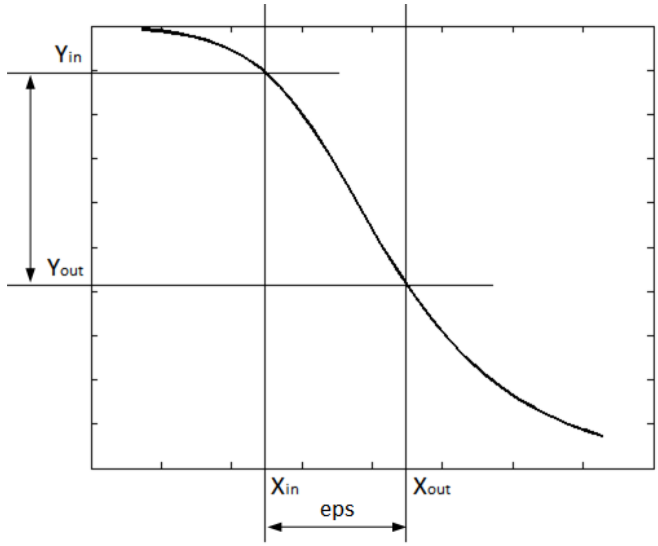


Figure 4.8: Generation of a desired exposure level based on pixel luminance/color saturation modification

step is to take the original picture, read pixel saturation values y_{in} for all RGB components and calculate the original exposure level x_{in} for each pixel, using the following equation (inverse function for Eq. 4.1):

$$x_{in} = \frac{ln\left(\frac{y_{in}^{1/c}}{1-y_{in}^{1/c}}\right) - b}{a} \tag{4.3}$$

where $a = -3.2$, $b = -1.3$, $c = 0.4$.

Afterwards, shift exposure value by a given offset *eps* obtaining a desired exposure level $x_{out} = x_{in} + eps$. In the last step a new pixel saturation value y_{out} is calculated:

$$y_{out} = \left(\frac{e^{a*x_{out}+b}}{1+e^{a*x_{out}+b}}\right)^{c} \tag{4.4}$$

where $a = -3.2$, $b = -1.3$, $c = 0.4$.

Exposure Metric

Another important step was to develop a methodology for the exposure distortion measurement. An exposure distortion is understood as the overall quality degradation caused by an improper exposure. In the exposure generation experiment its value was represented by a virtual variable (shot number from 1 to 28). The proposed metric should quantify exposure in a more meaningful manner. It means that the same value should preferably correspond to the same quality degradation level across different scenes. In contrast, fixed exposure time would yield a completely different impact on QoE depending on scene lightning (correct exposure time for low-light scene would result in over-exposure distortion for bright scene).

The metric was inspired by a shape of the histograms presented in Fig. 4.1. A histogram of a correctly exposed image spreads over the whole range of luminance. Histograms of over- and under-exposed images are shifted to the bright and the dark side respectively. The higher the exposure distortion the more significant the shift. In other words, there are no completely black and white regions on over- and under-exposed images, respectively. Consequently, the exposure metric is based on histogram range inspection.

The metric is calculated locally for each video frame. In the first step mean luminance is calculated for each macro-block of a given video frame. The average of the three macro-blocks with the lowest and the highest luminance represent luminance (histogram) bounds. The exposure metric for a single frame is calculated using the following formula:

$$Ex = \frac{L_b + L_d}{2} \tag{4.5}$$

where L_b and L_d are bright and dark luminance bounds.

Video level metric is calculated by averaging frame metric over one scene. The proposed methodology assumes that each natural video sequences has at least some bright and dark regions. It is much more accurate approach than a simple histogram average luminance calculation. For instance, it eliminates the problem when images showing black objects with very few bright regions would be classified as under-exposed.

4.1.2 No-reference Blur Metric

The most common approach towards image blur estimation utilizes the fact that blur makes image edges less sharp. Recent works representing this approach are described in [9] and [14]. The proposed metric is based on an average width of sharp edges only [47]. It is critical in terms of prediction accuracy since eliminates strong content dependency. Metric based on an average width of all edges

may introduce misleading results in case of an images with an inherent regions of smoothly changing pattern. The average edges width for such image is relatively high, suggesting strong image blur. The proposed metric in based on an assumption that every image has at least few sharp edges and consists of two phases: 1) edge detection and 2) mean width calculation.

$$S^h = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

$$S^{mod} = \begin{bmatrix} 0.125 & 0 & -0.125 \\ 0.25 & 0 & -0.25 \\ 0.125 & 0 & -0.125 \end{bmatrix}$$

$$b(i,j) = x^{fil}(i,j)^2 \tag{4.6}$$

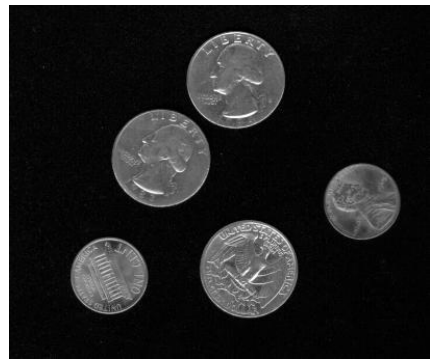
$$cutoff = \frac{16}{m*n} \sum_{i=1}^{m} \sum_{j=1}^{n} b(i,j) \tag{4.7}$$

The metric is calculated over image luminance (see Fig. 4.9 b). In the first step of the edge detection phase image gradient in horizontal direction (see Fig. 4.9 c) and its magnitude (see Fig. 4.9 d) is calculated using 'sobel' filter S^h. For the purpose of metric calculation a matrix containing pixel luminance values was filtrated using modified 'sobel' filter S^{mod}. For the resulting filtrated image x^{fil} a magnitude was calculated (see equation 4.6). The mean of the magnitude squared image multiplied by an appropriate scale is used for an automatic cutoff threshold calculation $cutoff$. The $cutoff$ value is used for sharp edges filtering. Empirically, the selected scale was set to 16 during the experiment (see equation 4.7).

In the last step of the edge detection phase, neighboring edges are suppressed during a thinning procedure. It is an algorithm presented using a Matlab source code (see Fig. 4.10). The resulting edge map is illustrated in Fig. 4.9 e.

In the second phase average edges width is calculated. Each pixel marked as an edge (white pixels in Fig. 4.9 e)) is presumed to be localized in the middle of the corresponding edge. Edge width is measured as a number of neighboring pixels (localized on the left and right in the same horizontal line) that fulfill the following criteria: a) right-localized pixels intensity values is monotonically increasing/decreasing for rising/falling edges, b) analogically for left-localized pixels, and c) edge slope value doesn't fall below a certain level. The slope value threshold th

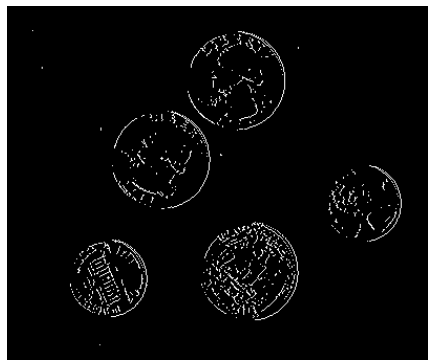
(a) Original image



(b) Image luminance x

(c) Sobel filtrated image x^{fil}

(d) Magnitude image b



(e) Thinning results e

Figure 4.9: Image preprocessing in Matlab prior to the edges width calculation

```
% Description of variables:
%     m - number of image rows
%     n - number of image columns
%     r - current row
%     c - current column
%     b1...b4 - tested conditions
%     e(r,c) - resulting edge map

for r=1:m,
  for c=1:n,
        if (r < 1) || (r > m) || ((c-1) < 1)
          b1 = true;
        else
          b1 = (b(r,c-1) <= b(r,c));
        end

        if (r < 1) || (r > m) || ((c+1) > n)
          b2 = true;
        else
          b2 = (b(r,c) > b(r,c+1));
        end

        if (c < 1) || (c > n) || ((r-1) < 1)
          b3 = true;
        else
          b3 = (b(r-1,c) <= b(r,c));
        end

        if (c < 1) || (c > n) || ((r+1) > m)
          b4 = true;
        else
          b4 = (b(r,c) > b(r+1,c));
        end
        e(r,c) = (b(r,c)>cutoff) & ...
                 (( b1 & b2) | ...
                  ( b3 & b4 ));
  end
end
```

Figure 4.10: Thinning code in Matlab

```
% detect a rising edge
if ((a(r,c2 - 1) > a(r,c2)) || (a(r,c2) > a(r,c2+1)))

    % increase a number of detected edges
    edgesNo = edgesNo + 1;

    % calculate an edge width
    while ((a(r,c2) - a(r,c2+1) > th) && (c2 < n - 1) )
        vlow = a(r,c2+1);
        c2 = c2 + 1;
    end
    width = width + c2-c;
    c2 = c;
    while ((a(r,c2 - 1) - a(r,c2) > th) && (c2 > 2) )
        vhigh = a(r,c2 - 1);
        c2 = c2 - 1;
    end
    width = width + c-c2;
```

(a) Raising edge code

```
% detect a falling edge
elseif ((a(r,c2 - 1) < a(r,c2)) || (a(r,c2)  < a(r,c2+1)))

    % increase a number of detected edges
    edgesNo = edgesNo + 1;

    % calculate an edge width
    while ((a(r,c2+1) - a(r,c2) > th) && (c2 < n - 1) )
        vhigh = a(r,c2+1);
        c2 = c2 + 1;
    end
    width = width + c2-c;
    c2 = c;
    while ((a(r,c2) - a(r,c2 - 1) > th) && (c2 > 2) )
        vlow = a(r,c2 - 1);
        c2 = c2 - 1;
    end
    width = width + c-c2;
end
```

(b) Falling edge code

```
% Description of variables:
%     m - number of image rows
%     n - number of image columns
%     r - current row
%     c - current column
%     e(r,c) - resulting edge map
%     edgesNo - number of detected edges
%     width - cumulative width of all detected edges
%     th - edge slope threshold
%     c2, vlow, vhigh - temporar variables

for r=1:m,
  for c=1:n,
      if (e(r,c))
          c2 = c;
          vlow = a(r,c2);
          vhigh = a(r,c2);

          % detect a rising edge
          ...
          % detect a falling edge
          ...

      end
  end
end
```

(c) Edges width code

Figure 4.11: Calculation of edges width in Matlab

is proportional to the standard deviation of surrounding pixels' intensity. The Matlab code illustrating the calculation steps in detail and variables description is given in Fig. 4.11. Mean edges width for the entire image corresponds to the frame level blur value. Video level metric Bl is calculated by averaging frame metric over one scene.

4.1.3 No-reference Noise Metric

The idea behind the proposed noise metric is based on work presented by Lee [37]. According to [37], the most convenient method for noise estimation for remotely sensed images is to identify homogenous image areas and use them to calculate noise statistics. The statistics are limited to a local mean and a variance of pixels luminance. The more recent work utilizing this approach is presented by Dosselmann in [9] and Farias in [14].

Block diagram of the proposed noise metric is presented in Fig. 4.12. The metric is calculated locally for each video frame. In the first step, standard deviation $std(x)$ of pixels luminance for each coding block is calculated:

$$std(x) = \left(\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2 \right)^{\frac{1}{2}} \tag{4.8}$$

where $\overline{x}$ is the mean over x:

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$$

In the next step, the minimal standard deviation $std(x)^{min}$ for each image slice is calculated. The mean value for all slices is considered the threshold th value for the given video frame:

$$th_i = \frac{1}{n} \sum_{j=1}^{n} std(x)_j^{min} \tag{4.9}$$

where n is the number of slices on the i-th frame.

All the coding block having standard deviation of pixels luminance lower or equal to the th are classified as smooth regions and considered in the further steps. Presented approach towards identifications of smooth regions assures selection of comparable number of blocks for images ranging from low to high spatial complexity. It outperforms approach presented in [9] based on a fixed threshold in terms of a visual annoyance prediction performance.

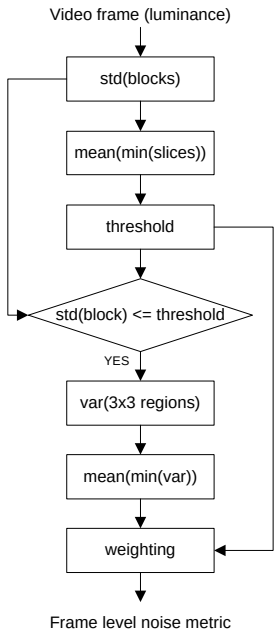


Figure 4.12: Block diagram for noise metric

On each coding block classified as smooth region further operations are performed. First, calculation of pixels variance ($v = std(x)^2$) is applied to 9 overlapping regions, each $3x3$ pixel size. Next, the mean value over 3 regions having the lowest variance is calculated. It corresponds to the local-level noise (for one smooth region) and is denoted as N^{loc}. The mean value over all smooth regions stands for a frame-level noise value N^{frame}.

The presented metric shows a good performance but a noticeable video content dependence was identified. In order to overcome this drawback, spatial masking phenomena was addressed by weighting frame-level noise value according to the threshold th_i, calculated in the first steps.

$$N_i^{frame} = \frac{1}{n * th_i} \sum_{j=1}^{n} (n_j^{loc}) \tag{4.10}$$

where n is the number of smooth regions detected on the i-th frame.

It compensates a well-known property that images having low spatial complexity are more exposed to the visual distortion caused by noise. In order to improve metric performance in terms of computational complexity, number of

coding blocks investigated in the first step may be reduced by a certain margin, without meaningful impact on the prediction accuracy. Video level metric is calculated by averaging frame metric over one scene.

4.2 H.264/AVC Compression Related Metrics

This section includes details on derivation of metrics for the three artifacts related to video compression assured using the H.264/AVC scheme. All the metrics were developed by the author and implemented in Matlab environment. Metrics for blockiness and flickering were inspired by the existing work, while the I frame flickering metric is a pioneer approach.

4.2.1 No-reference Blockiness Metric

Blockiness artifact measurement is based on a well-known property of a DCT-based coding. Each blocking artifact has a least one visible corner. Recent works utilizing this fact are described in [9] and [14].

Blockiness artifact is calculated locally, for each coding block. Fig. 4.13 presents the methodology applied for one block. Absolute difference of pixels' luminance is calculated separately for: 1) intra-pairs, represented by pixels connected with 'up' and 'left' arrows, and 2) inter-pairs, represented by 'down' and 'right' arrows.

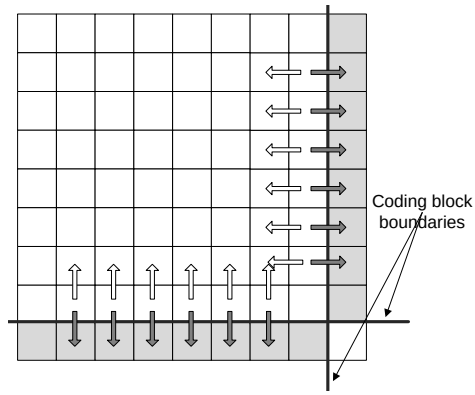


Figure 4.13: Calculation of blockiness artifact for one coding block

A cumulative sum of the intra- and inter-difference is calculated. The ratio between these values averaged over the whole image (all coding blocks) represents the blockiness metric for a given frame. For a real time application the metric

should be calculated over a time window (the number of video frames). Mean value for the window represents a blockiness level B.

4.2.2 No-reference Flickering Metric

The flickering metric was inspired by work presented by Pandel [44]. According to [44], flickering is the most annoying temporal artifact inherent for predictive inter-coded video sequences (in particular for H.264 encoded sequences). Video sequences with slow camera movements or zoom are especially exposed to flickering artifact. In temporal predictive coding macro-blocks are not updated (not encoded) until the difference between the corresponding macro-block of the successive frames is above a threshold. The stronger compression the higher is the threshold. It is illustrated in Fig. 4.14 where two successive video frames are shown for the same video sequences but encoded with different bit-rates. It is clearly visible that the stronger compression (lower bit-rate) the more coding blocks remain exactly the same as on the previous frame. These blocks are marked with black and represent regions with the lowest difference calculated frame to frame. It suggests that each macro-block remains in one of the two possible states (*no-update* or *update*) [44]. Frequent changes between states denote severe flickering artifact. The metric is calculated as a normalized number of transitions between the states.

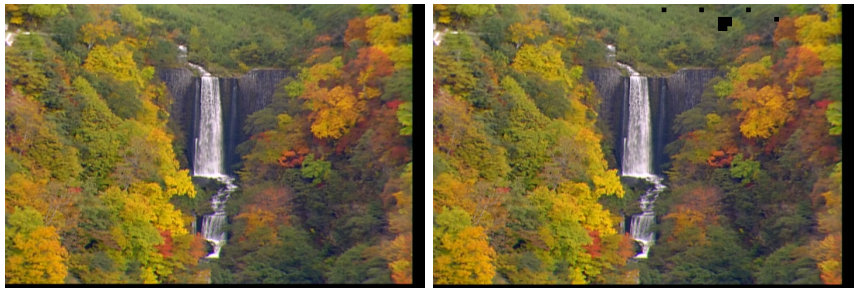
The state model used for flickering calculation including hysteresis is illustrated in Fig. 4.15. The task in the flickering metric implementation was threefold. First, to define threshold used to decide if a given macro-block remains in the *no-update* state. In [44], the threshold was defined as a mean squared difference between pixels of the current and corresponding macro-block, but the exact value was not revealed. The threshold was calculated as an average of absolute differences in pixels' luminance SAD for each 16×16 macro-block. In the proposed case, the normalized value (averaged for a single pixel) is given by equation:

$$SAD^{norm} = \frac{1}{256} \sum_{i=a}^{a+15} \sum_{j=b}^{b+15} abs\left(x_n(i,j) - x_{n+1}(i,j)\right) \tag{4.11}$$

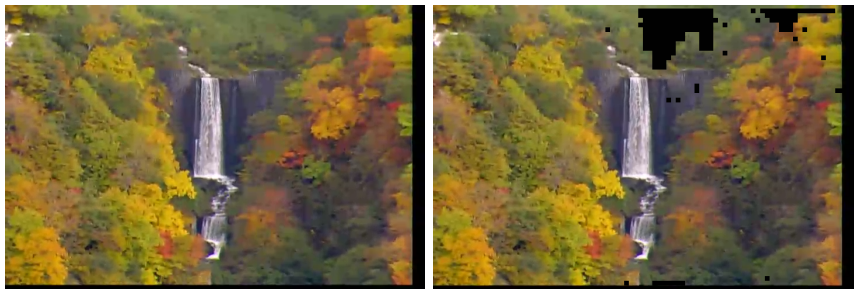
where $x_n(a,b)$ and $x_{n+1}(a,b)$ are the coordinates of the top left corner of a given block on the current and the next frame respectively.

Second, a different method for spatial pooling was proposed, i.e. calculate the frame-level flickering measure as a mean value over a small number of macro-blocks with the highest values (number of transitions between states). Third, two previous parameters were adjusted in order to optimize prediction performance defined as a correlation with subjective scores.

In order to maximize the correlation of the flickering metric F with MOS

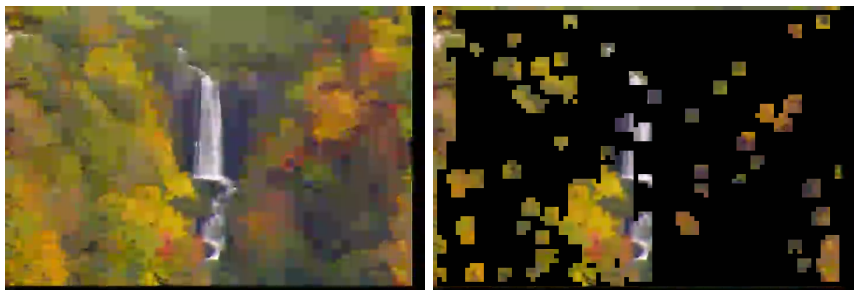


(a) 4000 kbit/s



(b) 500 kbit/s

(c) 200 kbit/s



(d) 100 kbit/s

Figure 4.14: Two successive frames from "Autumn" video sequence, 720x486, 30 fps, encoded with different bitrates. Left images represent previous frames while right images represent current frame with non-updated blocks marked in black.

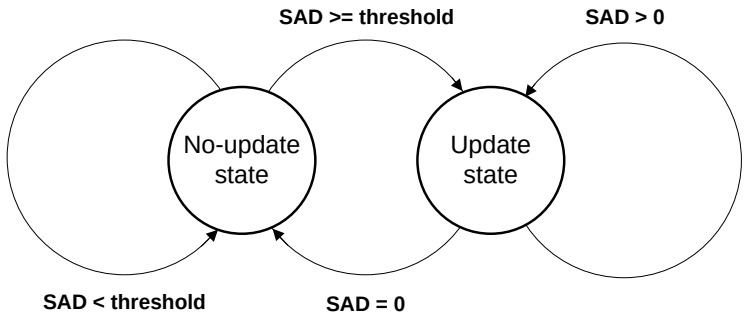


Figure 4.15: Proposed two-state flickering model

several threshold values (between 0.5% and 2% of luminance change) and several numbers of macro-blocks with the highest number of transitions between states (between 0.5% and 10% of macro-blocks) were considered. The highest correlation with MOS was achieved for the threshold equal to 1% and frame-level flickering averaging over 3% of the total number of macro-blocks.

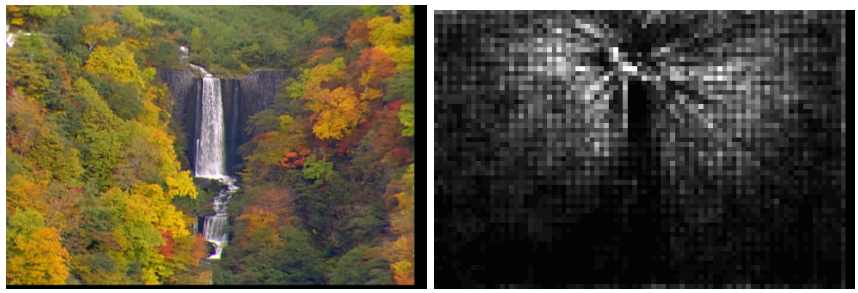


Figure 4.16: Visualization of flickering measurement over a time window

The metric is calculated over a time window (number of video frames). In Fig. 4.16 a total number of state changes for the whole time window for each coding block is represented by the luminance level. The brighter coding block the more state changes were observed. In the given example 3% of the brightest blocks will represent frame level flickering.

4.2.3 No Reference I Frame Flickering Metric

During a visual inspection of video sequences encoded with the H.264 codec another temporal artefact associated with H.264/AVC compression was identified.

It can be defined as a flickering of the entire video frame whenever an I frame is decoded . In our case this means one flicker per second (FPS = 30 and GoP = 30). Sequences with a slow global motion and high spatial activity are especially vulnerable to this artefact [75]. For such sequences, strong compression imposes that the majority of coding blocks remain in the no-update state during the entire GoP structure. This results in a significant flickering whenever I frame arrives (suddenly all coding block are updated). For lower compression most of the coding blocks are updated (even several times) during the GoP structure and the effect disappears. It should be noted that this effect is inherent for GoP structures starting with one I frame, and should not be generalized over different schemes.

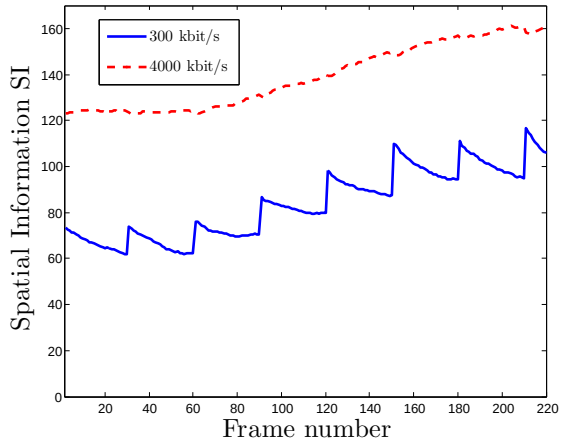


Figure 4.17: Time plot explaining I frame flickering root for "Autumn" video sequence and two different bit-rates.

Fig. 4.17 presents time plots for SRC 2 "Autumn" (slow camera zoom) and two different bit-rates (300 and 4000 kbit/s). Spatial activity SA (see Section 5.2 for calculation details) is calculated for each video frame. Values obtained for 4000 kbit/s are obviously higher than for 300 kbit/s (stronger compression yields details decrease). Significant peaks on 300 kbit/s plot represent I frames.

By analogy with the two-states model for flickering metrics described in Section 4.2.2, each decoded I frame activates the *update state* and causes visible global flickering. In contrast to P and B frames, all macro blocks are updated (intra-coded) on I frames even when strong compression is applied. We propose

the following formula to calculate the I-frame flickering IF effect:

$$IF = \text{mean}\left[\frac{SA_I(n)}{SA(n-1)}\right] \tag{4.12}$$

where $SA_I(n)$ and $SA(n-1)$ are spatial activities calculated for the I frame and the preceding one respectively (see Section 5.2 for details). IF for the entire video sequence is calculated as a mean value over all pairs (I frame and the preceding one). As with both previous metrics, averaging over a time window is required, and for the purpose of the experiment the window size was equal to the sequence length. It was verified that this artifact is periodic and does not change significantly over time within the same video scene.

Chapter 5

Subjective Experiments

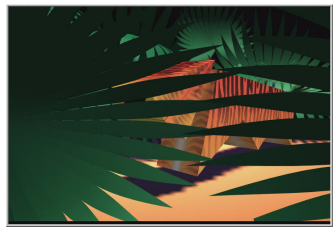
In order to verify the derived video quality metrics and propose perceptual models, two subjective experiments were carried out. The purpose of the experiments was to calculate a mapping function between the metrics and quality perceived by users expressed in MOS scale. The first experiment was devoted to acquisition related artifacts, while the second one to video bit-rate reduction. Video bit-rate reduction can be realized in three domains: 1) compression – by increasing quantization parameter QP, 2) temporal – by changing frames per second FPS rate, and 3) spatial – by decreasing frame resolution. In contrast to the other considered quality factors, FPS rate and video resolution are explicit parameters and no dedicated metrics are required to measure them. On the other hand, both factors may significantly influence perceived video quality and for this reason are covered by the subjective experiments.

Human testers rated video sequences affected by different degradation levels of all the considered artifacts. The degradation level was adjusted to cover the whole quality scale (from very slight to very annoying distortions).

5.1 Acquisition Related Artifacts

Four source sequences from standard VQEG content [60], namely SRC 16, 18, 19, and 21 were used (see Fig. 5.1). The same SRC (source) numbers as those presented in [60] was used in this work. The video sequences reflect two different content characteristics (i.e. spatial and temporal complexity).

The experiment was conducted by approximately 100 students. The applied subjective test methodology was the ITU's ACR-HR (Absolute Category Rating with Hidden Reference) described in ITU-T P.910 [24]. It represents a Single-



(a) SRC 16 – “Betes pas betes”

(b) SRC 18 – “Autumn leaves”

(c) SRC 19 – “Football”

(d) SRC 21 – “Susie”

Figure 5.1: VQEG Test Sequences

Stimulus (SS) approach, i.e. all video sequences contained in a test set are presented one by one without a possibility to compare with the reference video. The reference sequences are also included in the test set and rated according to the same procedure. This approach is commonly referred to as ACR Hidden Reference (ACR-HR). According to the newest VQEG’s guidelines for ACR methods, the eleven-grade numerical quality scale was used [24]. This is a quite innovative approach since five-grade MOS scale was a default scale so far [23].

For both over- and under-exposure, six degradation levels were introduced into each test sequence. In case of blur and noise, 10 and 7 degradation levels were used respectively.

5.2 Video Bit-rate Reduction

The second subjective experiment was devoted to video bit-rate reduction by means of frame resolution FR, frame per second FPS rate, and H.264/AVC compression. This experiment was much more extensive in terms of a number of test sequences. This is because video bit-rate reduction is content dependent and a full spectrum of different content characteristics had to be included. Furthermore, artifacts associated with H.264/AVC compression occur together and the

Table 5.1: Characteristics of the selected video sequences

SRC	SA	TA	O	Train.	Ver.
21	54.59	7.27	5.96		√
18	76.14	5.58	6.06	√	
20	189.82	2.30	6.09	√	
14	69.55	16.49	7.08		√
16	77.30	17.50	7.30	√	
13	85.57	19.98	7.44	√	
19	65.73	25.81	7.48		√
5	75.90	25.23	7.57	√	
2	138.45	16.62	7.70	√	
9	77.90	31.79	7.82		√
7	83.08	36.79	8.00	√	
3	128.69	23.37	8.01		√
10	171.00	27.14	8.45	√	

resultant impact on the quality should be considered. This is in contrast to the acquisition related artifacts which are independent from each other.

The first step in the experiment design process was to select a pool of test sequences. The key parameters describing any video sequence characteristics are spatial activity SA and temporal activity TA, i.e. the number of details and motion respectively. In order to make the selection task easier we used a *scene complexity* o measure, which is a combination of SA and TA [15].

A method presented in [15] was used, where *scene complexity* o is defined as

$$o = \log_{10}\left(\underset{n}{\text{mean}}\left[SI(n)\cdot TI(n)\right]\right) \tag{5.1}$$

where $SI(n)$ is spatial information computed for the nth frame and given by

$$SI(n) = \underset{space}{\text{rms}}\left[\text{Sobel}\left(F(n)\right)\right] \tag{5.2}$$

and $TI(n)$ is temporal information computed on the base of nth and $n-1$th frames given by

$$TI(n) = \underset{space}{\text{rms}}\left[F(n) - F(n-1)\right] \tag{5.3}$$

In both equations (5.2) and (5.3) $F(n)$ denotes the nth video frame luminance channel. Sobel is the Sobel filter [15] and rms_{space} is the root mean square function over an entire video frame.

The scene complexity analysis resulted in choosing 13 source sequences from

standard VQEG content [60] (see Fig. 5.2), namely SRC 2, 3, 5, 7, 9, 10, 13, 14, 16, 18, 19, 20, and 21, with a scene complexity o varying from 5.96 (SRC 21) to 8.45 (SRC 10). The same SRC (source) numbers as those presented in [60] was used in this work. Results ordered according to the scene complexity o are detailed in Table 5.1.

Subjects were pre-screened for the following features:

- normal (20/30) visual acuity with or without corrective glasses (self-declaration based on Snellen test chart),
- normal color blindness vision (per Ishihara test),
- contrast sensitivity,
- familiarity with the Polish language sufficient to comprehend the instruction and to provide valid responses using the semantic judgment terms expressed in that language.

Since the ultimate goal was to build a no reference models therefore ACR-HR (Absolute Category Rating with Hidden Reference) methodology was used [24], i.e. a no reference subjective test. In the ACR-HR tests, the original version of each video sequence is presented for rating somewhere in the test, without identifying it as the original. Subjects rate the original as they rate any other degraded video sequence. Additionally, each test condition is presented only once for the subjective assessment. The subjects were instructed to watch the entire video sequence before voting. The interface instructed when to vote.

The original sequences were 10 seconds long with the SD resolution and 30 FPS rate. Six different frame rates (from 5 to 30), five different resolutions (from SD to SQCIF), and six different bit-rates (100, 200, 300, 500, 1000, and 4000 kbit/s) with a constant Group of Pictures (GoP) length equal to 30. We used x.264 coded, main profile, and the constant bit-rate mode for the rate-control.

The videos were displayed in the center of a 17' LCD monitor with a native resolution of 1280x1024 pixels. All considered resolutions were up-scaled to SD because an assumption was that the tested service has a constant display size.

The test was run using a web browser and the content was downloaded from a server. Each video sequences was downloaded first and then viewed by a subject. The sequences were played out in a random order which was different for each subject. After a subject saw a sequence he or she scored it using an eleven point discrete scale [24].

The experiment started with a training phase in order to familiarize subjects with the specificity of the test. The phase consisted sequences (selected from the main pool) covering the entire range of considered distortions. The answers obtained were not considered in the further work.

(a) SRC 2 – “Barcelona”

(b) SRC 3 – “Harp”

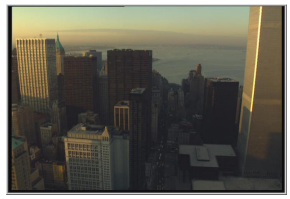
(c) SRC 5 – “Canoa Valsesia”

(d) SRC 7 – “Fries”

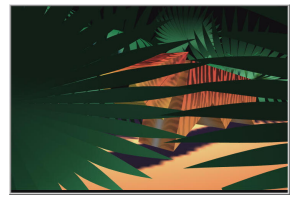
(e) SRC 9 – “Rugby”

(f) SRC 10 – “Mobile & Calendar”

(g) SRC 13 – “Baloon-pops”



(h) SRC 14 – “New York 2”



(i) SRC 16 – “Betes pas betes”

(j) SRC 18 – “Autumn leaves”

(k) SRC 19 – “Football”

(l) SRC 20 – “Sailboat”

(m) SRC 21 – “Susie”

Figure 5.2: VQEG Test Sequences

A post-experiment inspection of the subjective results was necessary in order to discard viewers who were suspected to give random answers. The rejection criteria verified the level of consistency of the scores of one viewer according to the mean score of all the subjects over the entire experiment. Example criteria to discard subjective data sets were the following: (i) the same rating was used for all or most of the video sequences or (ii) the subject's ratings correlate poorly with the average ratings from other subjects (correlation coefficient R^2 lower than 0.75).

It was assumed that exactly 24 valid subjects are required, however, 29 subjects took part in the experiment (in order to allow for the removal of some of the subjects in the post-screening procedure). A valid viewer meant a viewer whose ratings were accepted after post-experiment results screening. Scores from valid subjects only are discussed in the remaining part of this paper.

The following procedure [62] was used to obtain ratings for 24 valid subjects:

1. Conduct the experiment with 24 subjects.
2. Apply post-experiment screening to eventually discard viewers who are suspected to have voted randomly.
3. If n viewers are rejected, run n additional subjects.
4. Go back to step 2 and step 3 until valid results for 24 subjects are obtained.

Chapter 6

Derivation of Perceptual Models Mapping Objective Metrics into MOS Scale

This chapter presents derivation of perceptual models mapping objective metrics into the qualitative scale, i.e. MOS scale. The mapping is proposed using regression analysis and two different fitting functions. The first one is called an asymmetric logit function (ALF) and was used only in case of single parameter models (all acquisition related models). For other models, including more than one parameter, a linear combination of polynomials was used as a fitting function instead. The ground truth regarding quality is represented by MOSs obtained in the subjective experiments.

The first section details three models built upon the acquisition related metrics for over- and under-exposure, blur, and noise. The mapping functions are derived using subjective data gathered in the first subjective experiment (see Section 5.1). All the artifacts in this group are independent and represented by separate models.

The second section is devoted to video bit-rate reduction artifacts. It is verified using the H.264/AVC compression scheme. Video bit-rate reduction can be realized in three domains: 1) compression – by increasing quantization parameter QP, 2) temporal – by changing frames per second rate, and 3) spatial – by decreasing frame resolution. The mapping functions are derived using subjective data gathered in the second subjective experiment (see Section 5.2).

At the end of this chapter a summary of derived QoE estimation models is

given. Usability of the proposed metrics is presented on an example of a real-time video QoE assessment.

6.1 Video Acquisition Related Models

The most important task was to find a function mapping a distortion level on the user experience (MOS). Since in the test a dense eleven-grade quality scale was used it could be assumed that the testers' answers are an interval variable. Therefore, the obtained results were approximated by a continuous function. The important problem was to predetermine a shape of the function to be used. The shape cannot be predicted so numerous different functions could be considered but some basic properties of these functions can be specified.

The argument of the mapping function is an objective metric and its value predictes user experience in MOS scale. The user answers are scored on a limited scale. On the other hand, a metric value cannot be easily limited (it is not known what is the maximum possible distortion range). Therefore, using linear regression is not feasible because linear functions are not limited. A function that satisfies all mentioned requirements is monotonic and has two horizontal asymptotes.

A general class of functions having the above features was proposed. The function is an asymmetric logit function (ALF) [16], given by:

$$f(x) = \left(\frac{e^{ax+b}}{1 + e^{ax+b}} \right)^c \tag{6.1}$$

where a, b and c are estimated parameters.

An example of an ALF function is shown in Figure 6.1. A special case of ALF is a symmetric logit function (SLF) (parameter $c = 1$ in Fig. 6.1).

The ALF function is a non-linear function, so the parameter estimation is not a trivial task. For the research purpose a nonlinear least squares method was used. Note, that for this method it is important to choose proper starting points for all estimated parameters.

6.1.1 Model Based on the Exposure Metric

It was presumed that both over- and under-exposure degrade QoE and that the task is to find a proper mapping function between the metric and QoE. The function was derived from the results obtained in a subjective experiment. Based on the analysis performed, symmetric logit function SLF was proposed:

$$f(x) = 10 \left(\frac{e^{ax+b}}{1 + e^{ax+b}} \right) \tag{6.2}$$

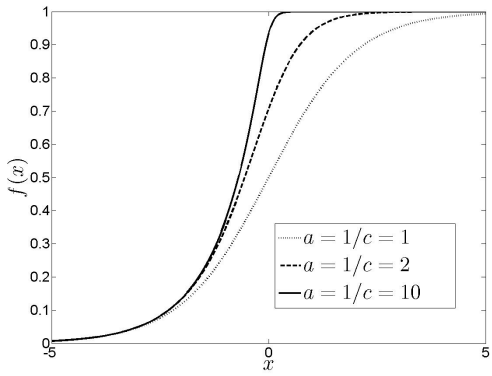
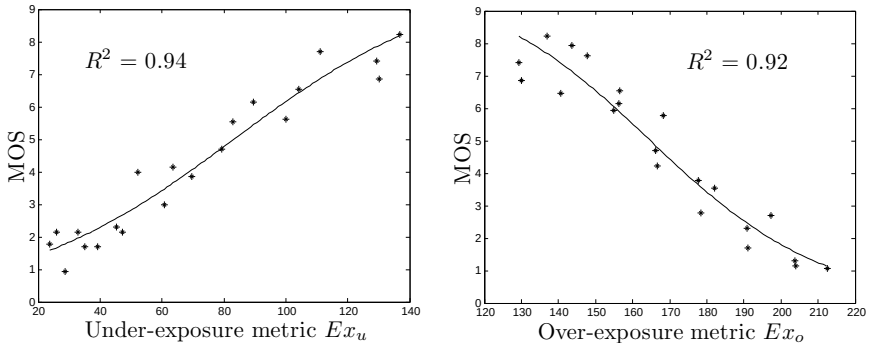


Figure 6.1: An example of different asymmetric logit functions



(a) Under-exposure vs. MOS correlation (b) Over-exposure vs. MOS correlation

Figure 6.2: The correlation results for the exposure metric with the estimated SLF function

The initial values range of SLF function values $[0; 1]$ was multiplied by 10 in order to cover the entire 11-point MOS scale (from 0 to 10). It was also decided to eliminate asymmetry ($c = 1$) because too large confidence interval were obtained for parameter c.

It was necessary to estimate two different function for over- and under-exposure. Estimation results are presented in Fig. 6.2. The obtained correlation coefficient for the over-exposure model MOS(Ex_o) was $R^2 = 0.92$ for the following parameters: $a = -4.31$ and $b = 7.10$. For under-exposure MOS(Ex_u) the results are: $R^2 = 0.94$, $a = 2.81$ and $b = 2.34$.

6.1.2 Model Based on the Blur Metric

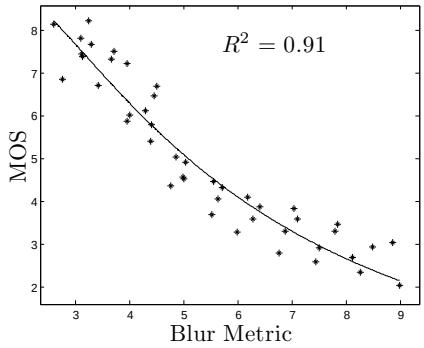


Figure 6.3: The correlation results for the blur metric with the estimated ALF function

Based on the analysis performed an asymmetric logit function ALF was selected, given by:

$$f(x) = 10\left(\frac{e^{ax+b}}{1+e^{ax+b}}\right)^{c} \tag{6.3}$$

Estimation results are presented in Fig. 6.3. The obtained correlation coefficient for the blur model MOS(Bl) was $R^2 = 0.91$ for the following parameters: $a = -1.50$, $b = 2.87$, and $c = 0.14$.

6.1.3 Model Based on the Noise Metric

Based on the analysis performed an asymmetric logit function ALF was selected, given by:

$$f(x) = 10\left(\frac{e^{ax+b}}{1+e^{ax+b}}\right)^{c} \tag{6.4}$$

Estimation results are presented in Fig. 6.4. The obtained correlation coefficient for the noise model MOS(N) was $R^2 = 0.91$ for the following parameters: $a = -3.46$, $b = -8.82$, and $c = 0.02$.

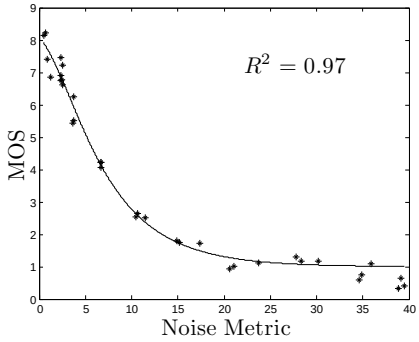


Figure 6.4: The correlation results for the noise metric with the estimated ALF function

6.2 Video Bit-rate Reduction Related Models

As already mentioned in this chapter, video bit-rate reduction can be realized in three domain: compression, spatial, and temporal. In case of the compression domain, a comprehensive approach towards perceptual quality assessment requires addressing artifacts related to intra- and inter-frame compression. An impact of video compression for diverse video content cannot be estimated properly using a metric measuring only one artefact. Intra-frame compression is addressed by a typical metric measuring the blockiness artefact B. In order to account for inter-frame compression artifacts we used an improved flickering metric operating on a macro-block level F. The overall QoE assessment system is supplemented by a custom metric dedicated to I-frames flickering IF.

Three QoE models are presented: 1) a model based on the blockiness metric B, 2) a model based on the flickering metric F, and 3) an integrated model for H.264/AVC compression including all three metrics (B, F, and IF). Because the I-frame flickering effect is restricted to low bit-rates only (strong compression) we consider it as an additional parameter of the integrated model rather than a stand-alone compression metric. Two video content characteristics were considered in order to improve model performance in terms of correlation with Mean Opinion Scores (MOS).

In case of two other domains (spatial and temporal) it was necessary to derive models mapping frame per second FPS rate and frame resolution FR into MOS [31]. The FPS rate and FR are video stream parameters which can be easily measured and no dedicated metric is required. Therefore one could think that providing a QoE metric based on them is easy. However, for different video content the influence of FPS rate or FR is different [28]. For example, if a video is in fact

a still image the quality remains the same for any frame rate. The consequence of this fact is that a model needs to take into account the video content, too.

6.2.1 Methodology for Models Derivation

The sequences pool in the bit-rate reduction related subjective experiment was much more extensive when compared to the acquisition related experiment (13 vs. 4 test sequences). This is because video bit-rate reduction is content dependent and a full spectrum of different content characteristics had to be included. Furthermore, artifacts associated with H.264/AVC compression occur together and the resultant impact on the quality should be considered. This is in contrast to the acquisition related artifacts which are independent from each other.

A more sophisticated methodology for model derivation was necessary. The proposed one consists of several steps that are reflected in the remaining part of work, namely: 1) divide subjective date into the training and the verification sets, 2) propose a model type and calculation methodology, 3) define model parameters (explanatory variables), 4) calculate model coefficients using the training set, 5) analyze statistical significance of the proposed parameters, 6) remove statistically not significant parameters, and finally 7) verify a model using the verification set.

Once a satisfactory model derived upon the training set is obtained, its performance can be validated using the verification set. Such a test allows to determine whether the model has a general value or if it is only over-fitted to the training data set.

The whole sequences set was divided into two groups; *training set* (src 2, 7, 10, 13, 16, 18, 19, and 20) and *verification set* (src 3, 5, 14, and 21). Both of these sets contain sequences covering a similar range of scene complexity o values but the training set is larger than the verification set.

Since an eleven point quality scale [24] was used the assumption that residuals of obtained answers have a Gaussian distribution is likely. This allows to use linear regression for the modeling purpose. This is an advantage when compared to five point MOS scales [24] where the GLZ (Generalized Linear Model) should be used instead [30], [40].

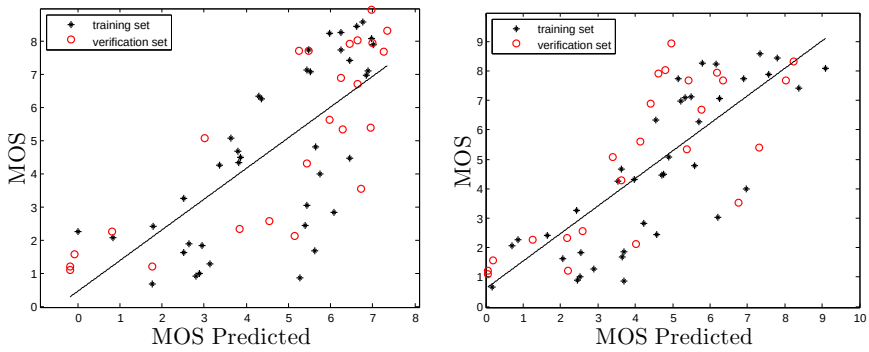
Any model considered in this part of the dissertation is a linear combination of explanatory variables such as

$$\mathrm{MOS}(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2 + ... \tag{6.5}$$

where x_i is an explanatory variable (for example the sequence frame rate) and a_i is a coefficient of x_i. When estimating a model like (6.5), a_i coefficients are computed and their statistical significance is assessed (a probability that the coefficient is different from 0). Any non statistically significant parameter has to

be removed from the model and an the estimation procedure has to be repeated. The final models contain only statistically significant coefficients estimated for specific explanatory variables.

6.2.2 Model Based on the Bockiness Metric



(a) $MOS(B)$ (b) $MOS(B, SA, TA)$

Figure 6.5: Performance of two models involving the blockiness metric for different sets

Based on the preliminary discussion of results it was assumed that for the blockiness metric spatial and temporal activity will improve performance in terms of correlation with MOS. In order to verify this assumption two models were derived: 1) a basic model denoted as $MOS(B)$ based only on a single explanatory variable – the blockiness metric B, and 2) a complex model denoted as denoted as $MOS(B, SA, TA)$ including two additional explanatory variables, i.e. spatial activity SA and temporal activity TA. The derived models are given by the following equations:

$$\text{MOS}(B) = -10.38 + 17.86B \tag{6.6}$$

$$\text{MOS}(B, SA, TA) = -10.88 + 14.68B + 0.02SA + 0.08TA \tag{6.7}$$

Figure 6.5 presents correlation of both models with MOS. The following correlation coefficients were obtained for the $MOS(B)$ model: 1) $R_t^2 = 0.50$ for the training set, 2) $R_v^2 = 0.65$ for the verification set, and 3) $R_{t+v}^2 = 0.55$ for both sets at the same time. The complex model $MOS(B, SA, TA)$ with both SA and TA being significant achieved higher correlation: 1) $R_t^2 = 0.66$, 2) $R_v^2 = 0.61$,

and 3) $R^2_{t+v} = 0.62$. The most meaningful correlation results are obtained using both sets at the same time while the verification set allows to determine whether a model is a general one. In case of a significant correlation drop for the verification set, a model should be considered as over-fitted to the training set. This is not the case for the blockiness metric, in particular if the most outstanding points were included in the verification set. Nevertheless, correlation obtained for the complex model is not high enough to represent the overall quality for H.264/AVC compression.

6.2.3 Model Based on the Flickering Metric

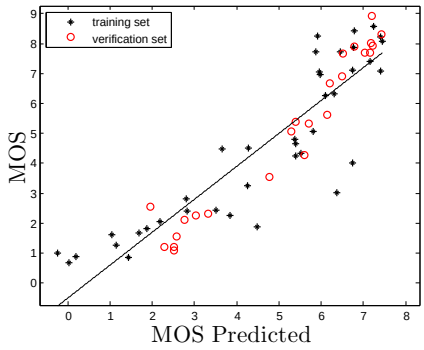


Figure 6.6: Performance of the flickering model for different sets

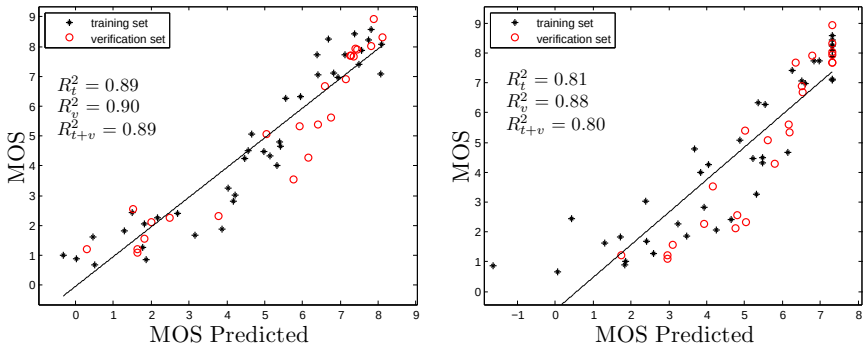
The preliminary analysis of results did not significantly help in understanding of the relation between content characteristics and the flickering metric. As a result, the approach from the blockiness metric case was repeated. Two models were derived: 1) a basic model denoted as $MOS(F)$ based only on a single explanatory variable – the flickering metric F, and 2) a complex model denoted as $MOS(F, SA, TA)$. Statistical analysis revealed that for the complex model neither was the correlation coefficient R^2 improved nor were SA and TA statistically significant. Therefore the final model consists of the flickering metric only. It suggests that either the flickering artefact in H.264/AVC compression or the flickering metric calculation methodology are content insensitive. The model is given by the linear equation:

$$\mathrm{MOS}(F) = 7.68 - 33.61F \tag{6.8}$$

Figure 6.6 presents correlation of $MOS(F)$ model. The following correlation coefficients were obtained for the $MOS(F)$ model: 1) $R^2_t = 0.78$ for training set,

2) $R_v^2 = 0.94$ for verification set, and 3) $R_{t+v}^2 = 0.83$ for both sets. As presented in Fig. 6.6, outstanding sequences were not included in the verification set. This explains why the correlation obtained for the verification set is so high. Correlation obtained for all sets is significantly higher than for the blockiness models and has a stronger potential in representing the overall quality for H.264/AVC compression.

6.2.4 Integrated Model for H.264/AVC Compression



(a) $MOS(B, F, IF)$ (b) $MOS(SSIM)$

Figure 6.7: Correlation with MOS of the integrated vs. SSIM models

In order to verify the assumption that H.264/AVC compression yields blockiness artefact, flickering artefact, and the I-frame flickering effect at the same time, an integrated model $MOS(B, F, IF)$ was derived, including all three parameters:

$$\mathrm{MOS}(B, F, IF) = -14.55 + 6.33B - 26.22F + 16.72IF \tag{6.9}$$

The results are presented in Fig. 6.7, and for all sets they demonstrate that the model is very accurate ($R_t^2 = 0.89$, $R_v^2 = 0.90$, and $R_{t+v}^2 = 0.89$). All three model parameters are statistically significant and constitute the general model of perceptual evaluation of H.264/AVC compression.

Its performance was compared with the Structural Similarity Index Metric (SSIM) [68], a well-known quality metric operating in a full-reference mode. The motivation for the choice was SSIM's availability, simplicity and good correlation with human perception. As presented by Wang [65], the human visual system (HVS) is very sensitive to the structural information provided on an image in a viewing field. Based on this assumption, SSIM can have good correlation with the

perceptual quality in our case, since artifacts caused by H.264/AVC do destroy structural information.

Comparison of results between SSIM and the proposed model is presented in Fig. 6.7. The integrated model outperforms the SSIM metric in terms of correlation with MOS, for all sets. Correlation obtained for SSIM model was $R_t^2 = 0.81$, $R_v^2 = 0.88$, and $R_{t+v}^2 = 0.80$. This can be explained by realizing that the proposed model is devoted to this specific kind of artifacts. From the other hand, SSIM is a more general metric and even the full-reference approach does not guarantee so high correlation for every scenario.

Another advantage of the proposed model is no reference approach, which significantly improves the application potential [73].

6.2.5 Models Based on FPS Rate and Frame Resolution

The simplest possible model consists of a single explanatory variable – FPS rate denoted by Fr or frame resolution FR denoted by R. In Figures 6.8 and 6.9 MOS is shown as a function of Fr and R, respectively.

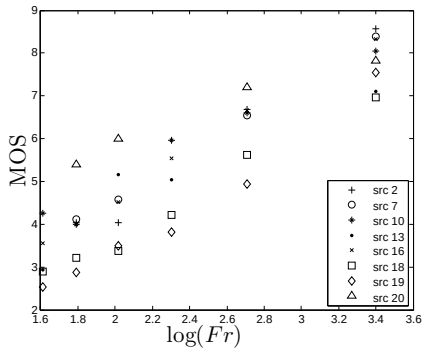


Figure 6.8: MOS as a function of a logarithm of frame rate. Only points from the training set are shown.

For both plots the obtained MOS scores are changing monotonically along with the explanatory variable but for any given FPS rate or FR obtained results are strongly scattered. More careful inspection of these figures reveals that sequence is one of the most important parameters. Note that for the frame rate case sequence 19 (marked by a diamond) was rated the lowest for most of the cases and sequence 20 (marked by a triangle) was rated the highest for most of the cases (see Figure 6.8). Note also that only for the original sequences (frame rate 30) the obtained MOS order is different. It means that the MOS results obtained for lower frame rates were driven mainly by content complexity while

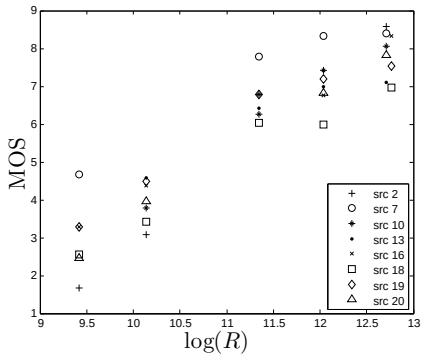


Figure 6.9: MOS as a function of a logarithm of total frame pixels. Only points from the training set are shown.

for the original sequences a decisive factor was the content type instead (we tend to score higher content that we are familiar or sympathize with).

The above consideration shows that sequence complexity has to be taken into account. Because spatial and temporal information seem to play a crucial role in determining differences between sequences, it was decided to use these parameters as additional explanatory variables.

The Final Models

From the previous section it turns out that temporal information TA and spatial information SA should improve a performance of both models in terms of their correlation with MOS. The first approach was to combine the FPS rate with temporal information. As one may expect, FPS rate reduction should affect perceived quality more for sequences with fast motion. On the other hand, it is expected that resolution should be combined with spatial information because for lower resolutions we loose some details of the original sequence. Obviously, for a sequence with more details more information will be lost than for a plain one.

However, the results obtained for both proposed models i.e. MOS(Fr, TA) and MOS(R, SA) were strongly scattered and not considerably better than the basic ones. It was decided to analyze new models containing such explanatory variables as TA, SA, and either Fr or R and all possible cross-relations between them. A detailed analysis revealed that models MOS(Fr, SA) and MOS(R, SA, TA) outperformed all others in terms of the correlation with MOS. In case of the FR model, both SA and TA are statistically significant but TA is much more important. It means that both SA and TA determine how resolution change

influences QoE but TA does it to the higher degree.

In the case of the frame rate model SA is the single parameter being statistically significant! This suggests that for sequences with lots of details (high SA) frame rate change is much more critical than for sequences with lots of motion (high temporal information). This result is surprising for us and is in contrast to the common understanding of the FPS reduction problem. If this property is successfully validated using the verification data set it will be an important and interesting discovery for the research community. In case if the assumption is wrong and it does not represent a real property of human perception it was decided then to build another model. According to the discussion from Section 6.2.5, the other model has to be proposed prior to the verification of the first one.

Note that if a sequence has high SA then TA value can be higher than for a sequence with low SA, even if the second one has more motion. It is a consequence of the way the TA value is calculated (see [60]). It means that the same TA and different SA indicate different actual motions. Therefore, it was decided to use a different motion metric expressed by $d = TA/SA$. Using d, another frame rate metric expressed by $\mathrm{MOS}(Fr, d)$ was proposed.

The three obtained models are given by equations

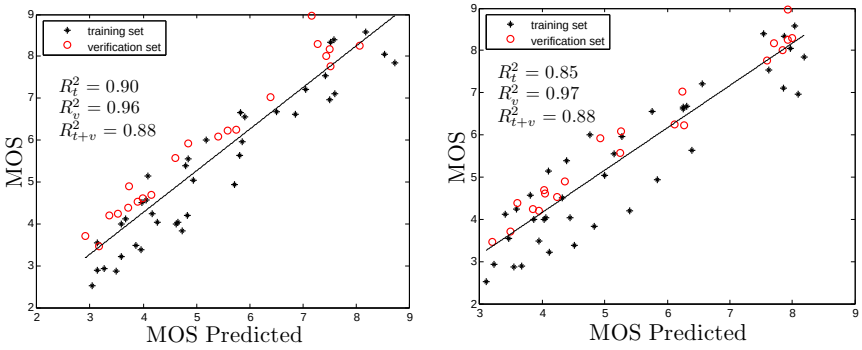
$$\mathrm{MOS}(Fr, SA) = -1.56 + 1.09\frac{SA}{100} + 2.43 \log Fr \tag{6.10}$$

$$\mathrm{MOS}(Fr, d) = -1.49d + 2.34 \log Fr \tag{6.11}$$

$$\begin{aligned}\mathrm{MOS}(R, SA, TA) = &- 12.8 + 0.62\frac{SA}{100} + \\ &5.66\frac{TA}{100} + 1,51 \log R\end{aligned} \tag{6.12}$$

The correlation results obtained for both data sets are shown in Figures 6.10 and 6.11. Both FPS models present similar performance in terms of a correlation R^2 and a smooth distribution of the points. Note that the results obtained for the first FPS model (equation (6.10)) are slightly better than those obtained by the second one (equation (6.11)). Therefore, the simpler one is recommended for use, i.e. $\mathrm{MOS}(Fr, SA)$.

The following correlation coefficients were obtained for the $MOS(Fr, SA)$ model: 1) $R_t^2 = 0.90$ for training set, 2) $R_v^2 = 0.96$ for verification set, and 3) $R_{t+v}^2 = 0.88$ for both sets. For the resolution model $MOS(R, SA, TA)$, correlation coefficients are the following: 1) $R_t^2 = 0.94$ for training set, 2) $R_v^2 = 0.91$ for verification set, and 3) $R_{t+v}^2 = 0.91$ for both sets.



(a) MOS(Fr, SA) (b) MOS(Fr, d)

Figure 6.10: Correlation of Fr based models obtained for both data sets

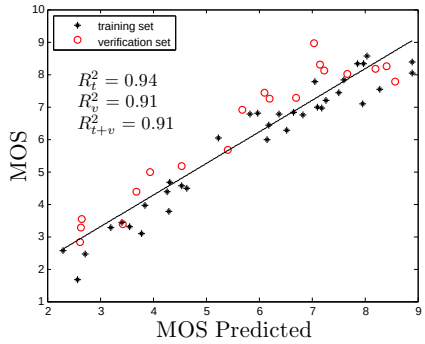


Figure 6.11: MOS(R, SA, TA) correlation obtained for both data sets

Both FPS models predict the verification set better than the training set. In this particular case, perceived quality of video sequences from the verification set turned out to be much easier to predict than for the training one. This happened by a chance and for any other quality metric the same sets may reveal quite opposite properties.

6.3 Summary of Video Quality Metrics and Models

According to the classification proposed in section 2.1, the proposed video quality metrics are qualitative, no-reference and represent artifacts measurement approach. The following factors affecting perceived video quality were addressed:

- Exposure Ex: spatial artifact related to the acquisition phase,
- Blur Bl: spatial artifact related to the acquisition phase,
- Noise N: spatial artifact related to the acquisition phase,
- Blockiness B: spatial artifact related to video compression,
- Flickering F: temporal artifact related to video compression,
- I-Frame Flickering IF: temporal artifact related to video compression,
- Frame per second rate Fr: video parameter resulting in a temporal artifact,
- Frame resolution R: video parameter resulting in a spatial artifact,
- Spatial Activity SA: video content parameter describing the amount of image details,
- Temporal Activity TA: video content parameter describing the amount of motion.

In order to improve the performance of the proposed quality metric some of them were aided with selected human visual system properties. The following properties were addressed:

- Spatial masking addressed in the noise metric,
- Spatial pooling addressed in the flickering metric,
- Weighting according to the local contrast in the flickering and blur metrics,
- Asymmetric and non-linear processing of visual stimuli reflected in the mapping functions.

The proposed metrics for QoE assessment encompass the first two stages of the video delivery chain, i.e. acquisition and bit-rate reduction. Impairments caused by network transmission were out of the scope of this work, because this problem was extensively analyzed in the past (see section 3.2.3). Additionally, approaches based on a bit-stream analysis do not require to decompress video content, so do

not constitute a challenge in a real time measurement systems. The last stage of the video delivery chain (end user equipment and preferences) was addressed in the subjective experiments and is reflected in the gathered MOSs. The summary of the derived models for perceptual video quality assessment in no-reference scenario is provided in Table 6.3.

Table 6.1: Summary of QoE Assessment Models

Parameter	Description
MOS(Bl)	MOS obtained for the QoE assessment model for blur artifacts
MOS(Ex)	MOS obtained for the QoE assessment model for exposure artifacts
MOS(N)	MOS obtained for the QoE assessment model for noise artifact
MOS(Fr, SA)	MOS obtained for the integrated QoE assessment model for video bit-rate scaling in temporal domain, including FSP rate Fr and video spatial activity SA
MOS(R, SA, TA)	MOS obtained for the integrated QoE assessment model for video bit-rate scaling in spatial domain, including video resolution R, video spatial activity SA, and temporal activity TA
MOS(B, F, IF)	MOS obtained for the integrated QoE assessment model for video bit-rate scaling in quantization domain, including blockiness B, flickering F, and I-frame flickering IF metrics

6.4 Real Time Verification

Applicability is another important aspects that needs considering when designing video quality metrics, although it is frequently overlooked in scientific papers. Applicability is understood here as the ability to utilize a metric in real life applications. Full-reference metrics are restricted to laboratory applications, while no-reference metrics offer an option to assess live video-based applications under real-time restrictions.

It was proved using a simulation that not only is the proposed models well correlated with MOS, but also that the metrics are computationally light enough to fulfil the real-time requirements. Fig. 6.12 shows an overview of the graphical interface for live video quality assessment. Gauges located at both sides of the preview screen represent single metrics and video content characteristics. A

Figure 6.12: Overview of the interface for live video quality assessment.

gauge located at the top in the middle represents video quality derived from the integrated model for H.264/AVC compression. All metrics and content characteristics (SA and TA) are calculated simultaneously using the same CPU core. Performance tests were carried out using 25 FPS videos in SD resolution, captured live from web cam or streamed from files. For both scenarios the overall calculation speed was higher than necessary including a significant overhead due to the graphical interface for live video quality assessment. Metrics were implemented in a C/C++ environment using open video libraries available for Linux. The graphical interface was implemented using QT libraries.

Chapter 7

Conclusions and Future Work

The problem of a reliable and comprehensive video quality assessment was discussed in this dissertation. The research was motivated by a huge increase in popularity of video-based services over recent years. Video quality assurance concerns became important again in a scenario where a large number of video streams, varying in terms of video content and parameters, is transmitted in a heterogeneous environment.

Video service providers are looking for reliable solutions for a constant QoE monitoring in an in-service mode. They need to manage service quality problems in order to become more competitive and attract more customers. However, traditional approaches towards video quality assessment does not satisfy the requirements of a comprehensive quality monitoring system (Chapter 3).

Because of the presented reasons, the following thesis was formulated and proven:

> **It is possible to assess perceptual quality of video content affected by artifacts related to acquisition and bit-rate reduction, using no-reference metrics, in a real time.**

Three out of four stages of the end-to-end video delivery chain (Section) were addressed, namely video acquisition (representing source quality), compression, and service environment including end users' preferences. The main original contribution of this dissertation is a set of no-reference perceptual video quality metrics (Chapter 4) derived in order to address the most important factors affecting the perceived video quality. For the source quality assessment, three

acquisition related metrics were developed, namely for over- and under-exposure, blur, and noise. Compressed video content quality was evaluated using metrics for blockiness artifact, flickering artifact, and I-frame flickering artifact.

Two subjective experiments were conducted and MOS scores were collected. The methodology for subjective experiments conformed to ITU-T P.910 [24], ITU-R BT.500 [22], and VQEG [59] recommendations. The first subjective experiment (Section 5.1) was devoted to acquisition related artifacts, while the second one to the video bit-rate reduction.

The proposed metrics were verified using the subjective data and objective models were derived (Chapter 6). Each model allows for mapping between a quantitative metric and MOS scale. Two models were proposed for video bit-rate reduction by means of spatial and temporal scaling (Section 6.2.5). Additionally, an integrated model for H.264/AVC compression was proposed and verified (Section 6.2.4). It was proven that not only is the proposed models well correlated with MOS, but also that the metrics are computationally light enough to fulfil the real-time requirements.

The obtained results show that the reliable video quality assessment can be performed using no-reference metrics, for video-based services using H.264/AVC compression, and in a real-time for standard definition video. High performance in terms of a correlation with end users' experience was obtained for a diverse video content. Mentioned features meet all the requirements of a comprehensive video quality assessment system and can be utilized by video service providers for a constant and reliable quality monitoring.

Future work will focus on realization of the end-to-end quality monitoring system for video streaming services, including additionally an impact of network transmission. For this purpose, it is possible to utilize some existing metrics representing the quality of delivery QoD approach. The research problem to be solved is a design of an integrated model predicting the resultant impact of all video processing stages; from video acquisition, through the bit-rate reduction and network transmission, to the service environment and end users' preferences.

According to the results presented in [17], a resultant impact on QoE of different coexisting distortions can be accurately modeled using a minimum measure. The underlying assumption is that the overall QoE cannot be higher than QoE measured for the worst single quality aspect. Another important assumption, not trivial this time, is that coexisting image and video artifacts do not reveal an additive nature. The assumption was proven in [17] where a cross influence of seven different image quality aspects was analyzed. The results presented in [17] will be verified upon a subjective experiment including video sequences with all possible degradation types.

Bibliography

[1] Agresti, A. *Categorical Data Analysis, 2nd Edition.* John Wiley & Sons, Ltd, 2002.

[2] Avcibas, I., Sankur, B., and Sayood, K. Statistical Evaluation of Image Quality Measures. *J. Electronic Imaging 11*, 2 (2002), 206–223.

[3] Brainard, D. H. *Handbook of Optics: Fundamentals, Techniques, and Design.* 1995.

[4] Brandao, T., and Queluz, M. P. No-reference quality assessment of h.264/avc encoded video. *IEEE Transactions on Circuits and Systems for Video Technology 20*, 11 (November 2010), 1437 – 1447.

[5] Carli, M., Farais, M., Gelasca, E. D., Tedesco, R., and Neri, A. Quality Assessment Using Data Hiding on Perceptually Important Areas. *IEEE International Conference on Image Processing, ICIP 2005 3* (September 2005), III – 1200–3.

[6] Caviedes, J., and Oberti, F. No-reference quality metric for degraded and enhanced video. In *Digital Video Image Quality and Perceptual Coding* (2006), H. Wu and K. Rao, Eds., CRC Press, pp. 305–324.

[7] Cerqueira, E., Janowski, L., Leszczuk, M., Papir, Z., and Romaniak, P. Video Artifacts Assessment for Live Mobile Streaming Applications. In *FMN '09: Proceedings of the 2nd International Workshop on Future Multimedia Networking* (Berlin, Heidelberg, 2009), Springer-Verlag, pp. 242–247.

[8] CMP Media LLC. *Measuring IPTV QoS Performance at the Box.* `http://www.digitaltvdesignline.com/`.

[9] Dosselmann, R., and Yang, X. D. A Prototype No-Reference Video Quality System. *Fourth Canadian Conference on Computer and Robot Vision, CRV '07 2007*, 411–417.

[10] Eskicioglu, A. M. Quality Measurement for Monochrome Compressed Images in the Past 25 Years. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Istanbul, Turkey, 2000), IEEE, pp. 1907–1910.

[11] Eskicioglu, A. M., and Fisher, P. S. Image Quality Measures and Their Performance. *IEEE Transactions on Communications 43*, 12 (December 1995), 2959–2965.

[12] ETSI. *Recommendation ETSI TR 101 290, Digital Video Broadcasting (DVB);Measurement Guidelines for DVB Systems.* ETSI, May 2001.

[13] Farias, M., M.Carli, and S.K.Mitra. Objective Video Quality Metric Based on Data Hiding. *IEEE Transactions on Consumer Electronics 51*, 3 (August 2005), 983–99.

[14] Farias, M., and S.K.Mitra. No-reference Video Quality Metric Based on Artifact Measurements. *IEEE International Conference on Image Processing, ICIP 2005 3* (September 2005), III – 141–4.

[15] Fenimore, C., Libert, J., and Wolf, S. Perceptual effects of noise in digital video compression. In *140th SMPTE Technical Conference* (Pasadena, CA, Oct. 1998), pp. 28–31.

[16] Fukumoto, K. Taking bounded variables seriously: Extended beta binomial, asymmetric logit, and time series. In *Research Workshop in Applied Statistics* (2004).

[17] Głowacz, A., Grega, M., Gwiazda, P., Janowski, L., Leszczuk, M., Romaniak, P., and Romano, S. P. Automated qualitative assessment of multi-modal distortions in digital images based on GLZ. *Special Issue of Annals of Telecommunications on Quality of Experience and Socio-Economic Issues of Network-Based Services 65*, 1-2 (February 2010), 3–17.

[18] Głowacz, A., Grega, M., Romaniak, P., Leszczuk, M., Papir, Z., and Pardyka, I. Compression and distribution of panoramic images concatenated using mpeg-7 descriptors. *Multimedia Tools and Applications 40*, 3 (December 2008), 321–339.

[19] Guo, J., Dyke-Lewis, M. V., and Myler, H. Gabor Difference Analysis of Digital Video Quality. *IEEE Transactions on Broadcasting 50*, 3 (September 2004), 302–311.

[20] Hood, D. C., and Wilkestein, M. A. *Handbook of Perception and Human Performance.* John Wiley, 1986.

[21] Hosaka, K. A New Picture Quality Evaluation Method. In *Proc. International Picture Coding Symposium* (Tokyo, Japan, 1986), pp. 17–18.

[22] ITU-R. *Recommendation ITU-R BT.500-11, Methodology for the Subjective Assessment of the Quality of Television Pictures.* Geneva, Switzerland, August 1998.

[23] ITU-T. *Recommendation ITU-T P.800, Methods for subjective determination of transmission quality.* ITU-T, Geneva, Switzerland, 1996.

[24] ITU-T. *Recommendation ITU-T P.910, Subjective Video Quality Assessment Methods for Multimedia Applications.* ITU-T, 1999.

[25] ITU-T. *Recommendation ITU-T Y.1541, Network Performance Objectives for IP-Based Services.* ITU-T, Geneva, Switzerland, February 2003.

[26] ITU-T. *Recommendation ITU-T J.144, Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference.* Geneva, Switzerland, 2004.

[27] ITU-T. *Recommendation ITU-T G.107, The E-model, a Computational Model for Use in Transmission Planning.* Geneva, Switzerland, March 2005.

[28] Janowski, L., Leszczuk, M., Papir, Z., and Romaniak, P. Ocena postrzeganej jakości usług strumieniowania wideo w scenariuszu bez referencji ze skalowaniem przepływności. *Przegląd Telekomunikacyjny, Wiadomości Telekomunikacyjne 82*, 8-9 (2009), 800–804.

[29] Janowski, L., Leszczuk, M., Papir, Z., and Romaniak, P. The design of an objective metric and construction of a prototype system for monitoring perceived quality (qoe) of video sequences. *Journal of Telecommunications and Information Technology*, 3 (2011), 87–94.

[30] Janowski, L., and Papir, Z. Modeling subjective tests of quality of experience with a generalized linear model. In *QoMEX 2009, First International Workshop on Quality of Multimedia Experience* (California, San Diego, July 2009).

[31] Janowski, L., and Romaniak, P. How do video frame rate and resolution influence qoe. In *3rd International Workshop on Future Multimedia Networking FMN'10* (Krakow, Poland, June 17-18 2010).

[32] Janowski, L., Romaniak, P., and Papir, Z. Assessing quality of experience for high definition video streaming under diverse packet loss patterns. In *4th International Conference on Multimedia Communications, Services and Security* (Krakow, Poland, 2-3 June 2011).

[33] Jia, Y., Lin, W., and Kassim, A. A. Estimating Just-Noticeable Distortion for Video. *IEEE Transactions on Circuits and Systems for Video Technology 16*, 7 (July 2006), 820–829.

[34] Kanumuri, S., Cosman, P., and Reibman, A. A Generalized Linear Model for MPEG-2 Packet Loss Visibility. *Packet Video Workshop, PV2004* (December 2004).

[35] Kanumuri, S., Cosman, P. C., Reibman, A. R., and Vaishampayan, V. A. Modeling Packet-loss Visibility in MPEG-2 Video. *IEEE Trans. Multimedia 8*, 2 (2006), 341–355.

[36] Kanumuri, S., Subramanian, S., Cosman, P., and A.R.Reibman. Predicting H.264 Packet Loss Visibility using a Generalized Linear Model. *Image Processing, 2006 IEEE International Conference on* (8-11 October 2006), 2245 – 2248.

[37] Lee, J., and Hoppel, K. Noise Modeling and Estimation of Remotely-sensed Images. *in Proc. International Geoscience and Remote Sensing, Vancouver, Canada 2* (1989), 1005–1008.

[38] Legge, G. E., and Foley, J. M. Contrast Masking in Human Vision. *Journal of the Optical Society of the America 70*, 12 (1980), 1458 – 1471.

[39] Lopez, D., Gonzalez, F., Bellido, L., and Alonso, A. Adaptive Multimedia Streaming over IP Based on Customer-Oriented Metrics. *ISCN'06 Bogazici University, Bebek Campus, Istanbul* (June 16 2006).

[40] McCullagh, P., and Nelders, J. *Generalized Linear Models*, 2nd ed. Chapman & Hall, 1991.

[41] Naccari, M., Tagliasacchi, M., and Tubaro, S. No-reference video quality monitoring for h.264/avc coded video. *IEEE Transactions on Multimedia 11*, 5 (August 2009), 932–946.

[42] Nayar, S., and Mitsunaga, T. High dynamic range imaging: spatially varying pixel exposures. vol. 1, pp. 472 –479 vol.1.

[43] OPTICOM. *PEVQ Advanced Perceptual Evaluation of Video Quality*, 2007. `http://www.opticom.de/download/PEVQ-WP-v07-A4.pdf`.

[44] Pandel, J. Measuring of flickering artifacts in predictive coded video sequences. In *WIAMIS '08: Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services* (Washington, DC, USA, 2008), IEEE Computer Society, pp. 231–234.

[45] Park, H. J., and Har, D. H. The optimum exposure decision for the enhanced image performance using a digital camera. pp. 333 –336.

[46] PEVQ. *Perceptual Evaluation of Video Quality*, 2007. `http://www.opticom.de/technology/pevq.html`.

[47] Romaniak, P. Towards realization of a framework for integrated video quality of experience assessment. In *INFOCOM Student Workshop 2009* (Rio de Janeiro, Brazil, April 2009).

[48] Romaniak, P., and Janowski, L. How to build an objective model for packet loss effect on high definition content using the ssim and subjective experiment. In *3rd International Workshop on Future Multimedia Networking FMN'10* (Krakow, Poland, June 17-18 2010).

[49] Romaniak, P., Janowski, L., Leszczuk, M., and Papir, Z. Ocena jakości sekwencji wizyjnych dla aplikacji strumieniowania na żywo w środowisku mobilnym. In *Krajowa Konferencja Radiokomunikacji, Radiofonii i Telewizji KKRRiT 2009* (Warszawa, Polska, 17-19 czerwca 2009).

[50] Romaniak, P., Janowski, L., Leszczuk, M., and Papir, Z. A no reference metric for the quality assessment of videos affected by exposure distortion. In *IEEE International Conference on Multimedia and Expo* (Barcelona, Spain, July 11 to 15 2011).

[51] Romaniak, P., Mu, M., Mauthe, A., D'Antonio, S., and Leszczuk, M. A Framework for Integrated Video Quality Assessment. *18th ITC Specialist Seminar on Quality of Experience* (May 2008).

[52] Shengke, Q., Huaxia, R., and Le, Z. No-reference Perceptual Quality Assessment for Streaming Video Based on Simple End-to-end Network Measures. *International conference on Networking and Services, ICNS '06* (2006), 53–53.

[53] Spirent Communications, Inc. *IPTV REAL-TIME VIDEO QUALITY TESTING*, 2007. `http://www.spirentcom.com/documents/3942.pdf`.

[54] Spirent Communications, Inc. *TESTING IPTV VIDEO QUALITY IN A PORT DENSE ENVIRONMENT*, 2007. `http://www.spirentcom.com/documents/4095.pdf`.

[55] Sugimoto, O., Naito, S., Sakazawa, S., and Koike, A. Objective perceptual video quality measurement method based on hybrid no reference framework. *IEEE International Conference on Image Processing (ICIP)* (Nov. 2009), 2237 – 2240.

[56] van den Branden Lambrecht, C. J., and Verscheure, O. Perceptual Quality Measure Using a Spatio-temporal Model of the Human Visual System. *in Proc. SPIE 2668* (1996), 450–461.

[57] Verscheure, O., Frossard, P., and Hamdi, M. User-oriented QoS Analysis in MPEG-2 Delivery. *Journal of Real-Time Imaging (special issue on Real-Time Digital Video over Multimedia Networks) 5*, 5 (October 1999), 305–314.

[58] Vlachos, T. Detection of Blocking Artifacts in Compressed Video. *Electronics Letters 36*, 13 (2000), 1106–1108.

[59] VQEG. *The Video Quality Experts Group.* `http://www.vqeg.org/`.

[60] VQEG. *The VQEG sequence description.* `ftp://vqeg.its.bldrdoc.gov/SDTV/VQEG_PhaseI/TestSequences/Reference/ThumbNails/TestSequencesIndex.pdf`.

[61] VQEG. *Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment*, March 2000. `http://www.vqeg.org/`.

[62] VQEG. *Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content*, 2009.

[63] Wang, Y. Survey of Objective Video Quality Measurements.

[64] Wang, Z. Objective Image/Video Quality Measurement – A Literature Survey. *EE 381K: Multidimensional Digital Signal Processing*.

[65] Wang, Z. *Rate Scalable Foveated Image and Video Communications.* PhD thesis, Dept. Elect. Comput. Eng. Univ. Texas at Austin, Austin, TX, December 2001.

[66] Wang, Z., Bovik, A. C., and Lu, L. Why is Image Quality Assessment so Difficult. *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing 4* (May 2002), 3313–3316.

[67] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing 13*, 4 (April 2004), 600–612.

[68] Wang, Z., Lu, L., and Bovik, A. C. Video Quality Assessment Based on Structural Distortion Measurement. *Signal Processing: Image Communication 19*, 2 (2004), 121–13.

[69] Wikipedia. *Minkowski addition*. Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Minkowski_addition`.

[70] Wikipedia. *Wikipedia, The Free Encyklopedia*. `http://www.wikipedia.org/`.

[71] Winkler, S. A Perceptual Distortion Metric for Digital Color Video. *in Proc. SPIE 3644* (1999), 175–184.

[72] Winkler, S. *Digital Video Quality - Vision Models and Metrics*. John Wiley & Sons, Ltd, 2005.

[73] Winkler, S. Video Quality and Beyond. *in Proc. European Signal Processing Conference, Poznan, Poland* (September 3-7 2007).

[74] Wolf, S., and Pinson, M. H. Spatial-temporal Distortion Metrics for In-service Quality Monitoring of any Digital Video System. *in Proc. SPIE 3845* (1999), 266–277.

[75] Yang, J. X., and Wu, H. R. Robust filtering technique for reduction of temporal fluctuation in h.264 video sequences. *Circuits and Systems for Video Technology, IEEE Transactions on 20*, 3 (march 2010), 458 –462.