An Objective and Subjective Quality Assessment for Passive Gaming Video Streaming

Kingston University London

Nabajeet Barman

Faculty of Science, Engineering and Computing

Kingston University London, United Kingdom

Supervisor: Prof Maria G. Martini

A thesis submitted for the degree of

Doctor of Philosophy

August 2019

"*Find something to believe in and find it yourself. When you do, pass it on to the future.*" – Metal Gear Solid 2: Sons of Liberty (Video Game 2001)

To Mom, Dad, and my brother, Debjeet

Declaration of Authorship

I, Nabajeet Barman, declare that this thesis titled, *"An Objective and Subjective Quality Assessment for Passive Gaming Video Streaming"* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my work.
- I have acknowledged all main sources of help.

NBarman

Nabajeet BARMAN
August 16, 2019

Acknowledgements

Personal

I owe my deepest gratitude to my supervisor and line manager, Professor Maria G. Martini, for her belief in me and giving me this opportunity to pursue this PhD. Her encouragement, support, and guidance throughout these past three years, not only limited to scientific research have helped me to finish this marathon while fulfilling many other goals and objectives of mine. Her personality and kindness will continue to inspire me beyond my professional career.

Also, my heartiest thanks to my colleagues and friends, Saman Zadtootaghaj and Steven Schmidt without whose collaboration, many of these works would not have been possible. Special thanks to Professor Sebastian Möller from the Berlin Institute of Technology for his guidance and support during my secondment at Technische Universität Berlin, Germany.

Thanks to the team at Multidimensional Insight Laboratory, Yonsei University, especially to Prof Sanghoon Lee, Woojae Kim, Duc Nguyen and Sewoong Ahn for hosting me at their institute for my secondment, leading to the completion of one of my earliest works.

Special thanks to all QoENet colleagues for all the training and network-wide events, online courses, scientific discussions and their continuous support during this period. Special thanks to Arslan Ahmad, Werner Robitza and Alcardo Barakabitze.

Thanks to Dr Manzoor Razaak and Dr Ognen Ognenoski for their guidance and tremendous help, especially during the initial few months, to help understand and overcome different challenges, both personal and professional.

Also my heartiest thanks to the previous and current team members at our WMN lab: Prof. Christos Politis, Dr Nabeel Khan, Dr Deepak G C, Alexandrous Ladas, Nikolaos, Solomon, Obinna, Bola, Silvia, Hadi for the friendly and supportive environment in the lab. Special thanks to my friends Sana, Bakhtiyar, Ioannis and Olga for the wonderful time together. My sincere thanks to the Tim and Tom from SEC-IT team and Nick, Lina and Karen Ingyon at Kingston University who during the course of the studies helped me with many of the project as well as personal requirements. My deepest gratitude to Rosalind Percival for being very kind, helpful and accommodating with the various requirements during the course of this PhD.

No story is complete without the support of a family behind. I would not be here without the support, encouragement, and love of my parents, G C Barman and Mrinalinee Barman. Thanks to my brother, Debjeet Barman for all his help and love throughout my life. Special thanks to my sister-in-law Jucy and my beautiful, recently born nephew, Nishkarsh.

Institutional

This work is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 643072.

Abstract

Gaming video streaming has become increasingly popular in recent times. Along with the rise and popularity of cloud gaming services and e-sports, passive gaming video streaming services such as Twitch.tv, YouTubeGaming, etc. where viewers watch the gameplay of other gamers, have seen increasing acceptance. Twitch.tv alone has over 2.2 million monthly streamers and 15 million daily active users with almost a million average concurrent users, making Twitch.tv the 4th biggest internet traffic generator, just after Netflix, YouTube and Apple. Despite the increasing importance and popularity of such live gaming video streaming services, they have until recently not caught the attention of the quality assessment research community. For the continued success of such services, it is imperative to maintain and satisfy the end user Quality of Experience (QoE), which can be measured using various Video Quality Assessment (VQA) methods. Gaming videos are synthetic and artificial in nature and have different streaming requirements as compared to traditional non-gaming content. While there exist a lot of subjective and objective studies in the field of quality assessment of Video-on-demand (VOD) streaming services, such as Netflix and YouTube, along with the design of many VQA metrics, no work has been done previously towards quality assessment of live passive gaming video streaming applications.

The research work in this thesis tries to address this gap by using various subjective and objective quality assessment studies. A codec comparison using the three most popular and widely used compression standards is performed to determine their compression efficiency. Furthermore, a subjective and objective comparative study is carried out to find out the difference between gaming and non-gaming videos in terms of the trade-off between quality and data-rate after compression. This is followed by the creation of an open source gaming video dataset, which is then used for a performance evaluation study of the eight most popular VQA metrics. Different temporal pooling strategies and content based classification approaches are evaluated to assess their effect on the VQA metrics. Finally, due to the low performance of existing No-Reference (NR) VQA metrics on gaming video content, two machine learning based NR models are designed using NR features and existing NR metrics, which are shown to outperform existing NR metrics while performing on par with state-of-the-art Full-Reference (FR) VQA metrics.

Preface

This doctoral thesis presents my research work in the field of gaming video streaming applications, which I performed at the School of Computer Science and Mathematics at Faculty of Science, Engineering and Technology, Kingston University. The results presented in this thesis have been published in the following publications:

- N. Barman, E. Jammeh, S. A. Ghorashi and M. G. Martini, "No-reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications", *IEEE Access*, vol. 7, pp. 74511-74527, 2019.

- N. Barman and M. G. Martini, "A Survey of QoE Modeling for HTTP Adaptive Video Streaming", *IEEE Access*, vol. 7, pp. 30831-30859, 2019.

- N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini and S. Möller, "An Objective and Subjective Quality Assessment Study of Passive Gaming Video Streaming", *International Journal of Network Management.* 2018. https://doi.org/10.1002/nem.2054.

- N. Barman, S. Schmidt, S. Zadtootaghaj, M. G. Martini, and S. Möller, "An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming", *In Proceedings of 23rd Packet Video Workshop 2018 (PV' 18)*, ACM, Amsterdam, Netherlands, June 2018, pp. 7-12.

- N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "GamingVideoSET: A Dataset for Gaming Video Streaming Applications", *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, Amsterdam, Netherlands, June 2018, pp. 1-6.

- N. Barman, S. Zadtootaghaj, M. G. Martini, S. Möller, and S. Lee, "A Comparative Quality Assessment Study for Gaming and Non-Gaming Videos", *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Sardinia, Italy, May 2018, pp. 1-6.

- N. Barman and M. G. Martini, "H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 Codec Comparison for Live Gaming Video Streaming", in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX 2017)*, Erfurt, Germany, May 2017, pp. 1-6.

Additionally, during the course of my PhD studies, I have also been involved in different collaborative works mentioned below which are not an integral part of this thesis.

- A. A. Barakabitze, N. Barman, A. Ahmad, S. Zadtootaghaj, L. Sun, M. G. Martini, L. Atzori, "QoE Management of Multimedia Services using SDN and NFV: State-of-the-Art and Future Challenges", in *IEEE Communications Surveys & Tutorials.* Revision.

- S. Zadtootaghaj, N. Barman, S. Schmidt, M. Martini, and S. Möller, "NR-GVQM: A No Reference Gaming Video Quality Metric", in *20th IEEE International Symposium on Multimedia (IEEE ISM 2018)*, Taichung, Taiwan, 2018, pp. 131-134.

- S. Zadtootaghaj, S. Schmidt, N. Barman, S. Möller, and M. G. Martini, "A Classification of Video Games based on Game Characteristics linked to Video Coding Complexity", in 2018 16th Annual Workshop on Network and Systems Support for Games (NetGames 2018), (Amsterdam, Netherlands), 2018. **Best Paper Award**.

- N. Barman, S. Valentin, and M. G. Martini, "Predicting Link Quality of Wireless Channel of Vehicular Users Using Street and Coverage Maps", in *27th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2016)*, (Valencia, Spain), Sept 2016.

ITU-T Contribution

- N. Barman, M. Martini, and S. Zadtootaghaj, "Impact of video codecs on Gaming Quality of Experience", *ITU-T Study Group 12, ITU-T Contribution C.096 (2017)* Available: https://www.itu.int/md/T17-SG12-C-0096/en.

- S. Schmidt, S. Zadtootaghaj, S. Möller, F. Metzger, M. Hirth, M. Sužnjević, N. Barman, and M. G. Martini, "Requirement Specification and Possible Structure for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Contribution C.200-R1 (2018). Available: https://www.itu.int/md/T17-SG12-C-0200/en

- S. Schmidt, S. Zadtootaghaj, F. Schiffner, S. Möller, S. S. Sabet, C. Griwodz, N. Barman, and M. G. Martini, "Data Assessment for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Contribution C.293 (2018). Available: https://www.itu.int/md/T17-SG12-C-0293/en

- S. Zadtootaghaj, S. Schmidt, M. Utke, S. Möller, N. Barman, and M. G. Martini, S. S. Sabet, and C. Griwodz, "Evaluation of Standardized Video Quality Models for Assessing Video Quality of Cloud Gaming Services", ITU-T Contribution C.294 (2018). Available: https://www.itu.int/md/T17-SG12-C-0294/en

- S. Schmidt, S. Zadtootaghaj, M. Utke, S. Möller, N. Barman, M. G. Martini, S. S. Sabet, and C. Griwodz, "First Draft for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Contribution C.387 (2019). Available: https://www.itu.int/md/T17-SG12-C-0387/en

Contents

List of Figures

List of Tables

List of Acronyms

ACR	Absolute Category Rating
ACR-HR	ACR with Hidden Reference
AOM	Alliance for Open Media
AVC	Advanced Video Coding
BR	bitrate
BD-BR	Bjontegaard-Delta Bitrate
BIQI	Blind Image Quality Index
BP-ANN	Back-Propagation Artificial Neural Network
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
CABAC	Context-Adaptive Binary Arithmetic Coding
CAVLC	Context-Adaptive Variable-Length Coding
CBR	Constant Bitrate
CDN	Content Delivery Network
CRF	Constant Rate Factor
CSGO	Counter Strike: Global Offensive
CTU	Coding Transform Unit
CUQI	Cardiac Ultrasound Video Quality Index
DASH	Dynamic Adaptive Streaming over HTTP
DCR	Degradation Category Rating
Diablo	Diablo III
Dota2	Defense of the Ancients 2

DCT	Discrete Cosine Transform
DSCQS	Double Stimulus Continuous Quality Scale
DLM	Detail Loss Metric
DSIS	Double Stimulus Impairment Scale
FIFA	FIFA
FPS	First Person Shooter
fps	frames per second
FHD	Full HD
FR	Full-Reference
GOP	Group of Pictures
GP	Gaussian Process
GTA	Grand Theft Auto
H1Z1	H1Z1: Just Survive
HAS	HTTP Adaptive Streaming
HD	High Definition
HDR	High Dynamic Range
HDS	Adobe HTTP Dynamic Streaming
HEVC	High Efficiency Video Coding
HLS	HTTP Live Streaming
HLS	HTTP Live Streaming
HS	Heathstone
HVS	Human Visual System
IQA	Image Quality Assessment
JCT-VC	Joint Collaborative Team on Video Coding
JND	Just Noticeable Difference

LF	Light Field
LoL	League of Legends
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
MOS	Mean Opinion Scores
MPEG	Moving Pictures Expert Group
MSCN	Mean Subtracted Contrast Normalized
MSE	Mean Square Error
MSS	Microsoft Smooth Streaming
NFS	Need for Speed
NIQE	Natural Image Quality Evaluator
NN	Neural Network
NR	No-Reference
NR-GVSQE	No-Reference Gaming Video Streaming Quality Estimator
NR-GVSQI	No-Reference Gaming Video Streaming Quality Index
NSS	Natural Scene Statistics
OTT	Over The Top
OR	Outlier Ratio
PC	Project Cars
PCA	Principal Component Analysis
PLCC	Pearson Linear Correlation Coefficient
PLR	Packet Loss Ratio
PSNR	Peak Signal to Noise Ratio
PPSNR	Partial PSNR
QUIC	Quick UDP Internet Connections

QoE	Quality of Experience
QoS	Quality of Service
QP	Quantization Parameter
RD	Rate-Distortion
RES	Resolution
RF	Random Forest
RMSE	Root Mean Square Error
ROI	Region of Interest
RR	Reduced-Reference
SA	Spatial Activity
SAO	Sample Adaptive Offset
SC	Starcraft 2
SD	Standard Definition
SI	Spatial Information
SpEEDQA	Spatial Efficient Entropic Differencing for Quality Assessment
SROCC	Spearman's Rank Correlation Coefficient
SSIM	Structural Similarity
SSCQE	Single Stimulus Continuous Quality Rating
ST-RRED	Spatio-Temporal-Reduced Reference Entropic Differences
SVM	Support Vector Machine
SVR	Support Vector Regression
TA	Temporal Activity
TCP	Transmission Control Protocol
TI	Temporal Information
UHD	Ultra High Definition

VMAF	Video Multimethod Assessment Fusion
VOD	Video-on-demand
VQA	Video Quality Assessment
VQEG	Video Quality Experts Group
VBR	Variable Bit Rate
VIF	Visual Information Fidelity
VIFP	Visual Information Fidelity - Pixel Domain
VR	Virtual Reality
WEKA	Waikato Environment for Knowledge Analysis
WoW	World of Warcraft

Video games are bad for you? That's what they said about rock 'n' roll.

— Shigeru Miyamoto *Representative Director/Fellow, Nintendo*

1 Introduction

Contents

Humans like to be entertained and visual information is one of the most preferred forms of entertainment, courtesy of the fact that 70-90% of the neurons in the human brain is dedicated to vision [1]. The advancements in the field of video streaming have recently resulted in the rise of both Video-on-demand (VOD) (YouTube, Netflix, Amazon Video, Hulu, etc.) and Live (Twitch.tv, YouTubeGaming, Facebook Live Stream, etc.) streaming services. As evident, video streaming is not a niche market anymore, and there exist a wide range of options for the consumers to choose from. Hence, as a service provider, it is no more sufficient just to provide a service, but it is equally important to make sure that the needs and expectations of the end user of the offered services are met.

One of the reasons behind the rising popularity of such Over The Top (OTT)[1] services is the constant improvement of the streaming technologies. There has been a considerable amount of work on video delivery over the Internet to meet this increased demand. With the deployment of new wireless technologies such as 4.5G LTE-Advanced, the available end-user bandwidth has increased considerably over the recent years, and it will further increase with 5G systems. OTT services are usually of two types: VOD and Live. VOD refers to the services which allow users to select and watch content as per their own choice. The content is usually encoded and put in a suitable container on the service providers' server or a Content Delivery Network (CDN). The user then can select and demand which videos to watch and when. Services offering VOD have in recent years grown in much popularity such as the likes of YouTube, Netflix, Amazon Prime Video, Hulu etc. as they give the users more freedom on the choice of any content, any place and any time. Since the content can be prepared and stored (cached) without strict time constraints, such services can benefit from optimized encoding settings (e.g., Variable Bit Rate (VBR), multiple pass, per-title encoding [2], dynamic optimizer [3]) and client side optimization techniques (buffering, etc.). In VBR mode of encoding, the encoded audio or video data rate can vary over the duration of the audio or video depending on the content complexity. Multiple pass encoding allows the encoder to estimate the content complexity and hence achieve a better rate-distortion trade-off, resulting in higher quality encodes as compared to single pass encoding. Buffering refers to the downloading and storing of data ahead of time so that playback can continue without interruption in case of changing network throughput. Hence, such approaches can help mitigate some factors to improve the end user Quality of Experience (QoE).

[1] OTT media services refers to the services which provides delivery of media content via the internet.

In contrast, live streaming services are streamed in real-time and hence have strict encoding requirements. Due to the nature of such services, users generally have less tolerance towards playback issues such as stalling/rebuffering, low picture quality, etc., as that may result in missing out on an essential part of the stream (e.g., a player being bowled out during a cricket match). Traditionally used methods to overcome network outages, such as buffering, have limited applicability, as the media to be streamed is not yet created. Therefore, live streaming services have a different set of constraints and requirements than that of VOD streaming services, which need to be taken into account in order to improve the end user QoE.

The growth and success of existing and new services, in the end, depends on the acceptance by the end user. With the proliferation of many streaming services as well as the introduction of the latest video formats and services, the user has many options to choose from. Hence, it is no more sufficient for a service provider to just provide a service but also meet user expectations and satisfaction. Traditionally, the performance and hence success of a service has been based on Quality of Service (QoS) estimation which involves the measurement of network parameters such as Packet Loss Ratio (PLR), throughput, delay, jitter, etc. Such an approach does not necessarily correlate well with user expectations and experience. QoE, on the other hand, takes into account user expectations and experience and hence is considered a better indicator of user level satisfaction. This has led to a shift from traditional technical QoS based assessment (see, e.g., [4]) to QoE based assessment (see, e.g., [5], [6]).

1.1 Gaming Entertainment

Gaming has always been a prevalent and widely accepted form of entertainment, especially for the younger generation (usually 16-34 year olds). Over the years, it has evolved and come a long way from small abstract games, such as Mario and Pacman, to very complex and realistic games such as Battlefield, Grand Theft Auto (GTA), etc. A recent survey shows that over 87% of the internet users reportedly have played games on at least one device [7]. Gaming video streaming, in general, can be divided into two major applications: interactive (also called cloud gaming) and passive (also called spectator gaming).

Cloud gaming refers to applications such as PlayStation Now[2] and GeForce Now[3], where the gameplay is performed by the user using the cloud services offered

[2] https://www.playstation.com/en-gb/explore/playstation-now/

[3] https://www.nvidia.com/en-us/geforce/products/geforce-now/

by the client. Such applications eliminated the need for high performance hardware at the client (user) side while also addressing security and piracy concerns. Recent years have seen increasing acceptance of such services by the gaming community.

Another form of gaming which has gained attention in recent years is mobile gaming, where the users primarily use their mobile devices such as tablets and smartphones to play games. Mobile gaming continues to grow in popularity due to the huge potential in monetization and improvements in the new generation smartphones. Mobile gamers nowadays spend approx. 3 hours and 20 minutes per day online via their smartphone [7]. With the rise of gaming and the growth of the gaming community worldwide, a new form of gaming termed as passive gaming video streaming (also referred to as *Spectator Gaming*) has gained popularity. A recent survey in [7] shows that 1 in 4 gamers have watched a live gaming stream in the past month with 18% having watched an eSports tournament. The increasing popularity of such genre of entertainment has lead to the rise of OTT streaming services such as Twitch.tv and YouTube gaming, with Twitch.tv alone consisting of over 2 million monthly streamers and over 15 million daily active visitors with almost a million concurrent users making it the 4th largest peak traffic generator in the US, just after top three OTT on-demand streaming services, Netflix, Google and Apple. Figure 1.1 shows the number of average concurrent viewers and streamers for various live streaming platforms.

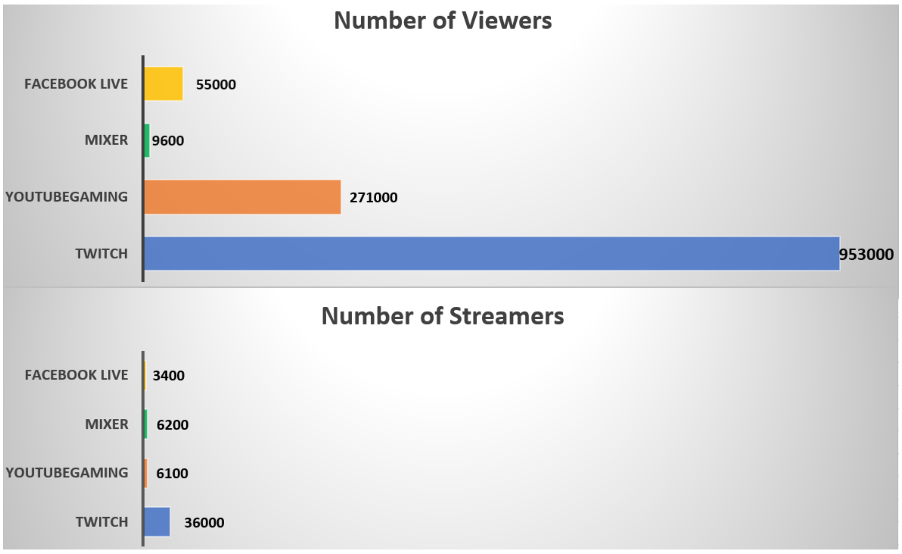
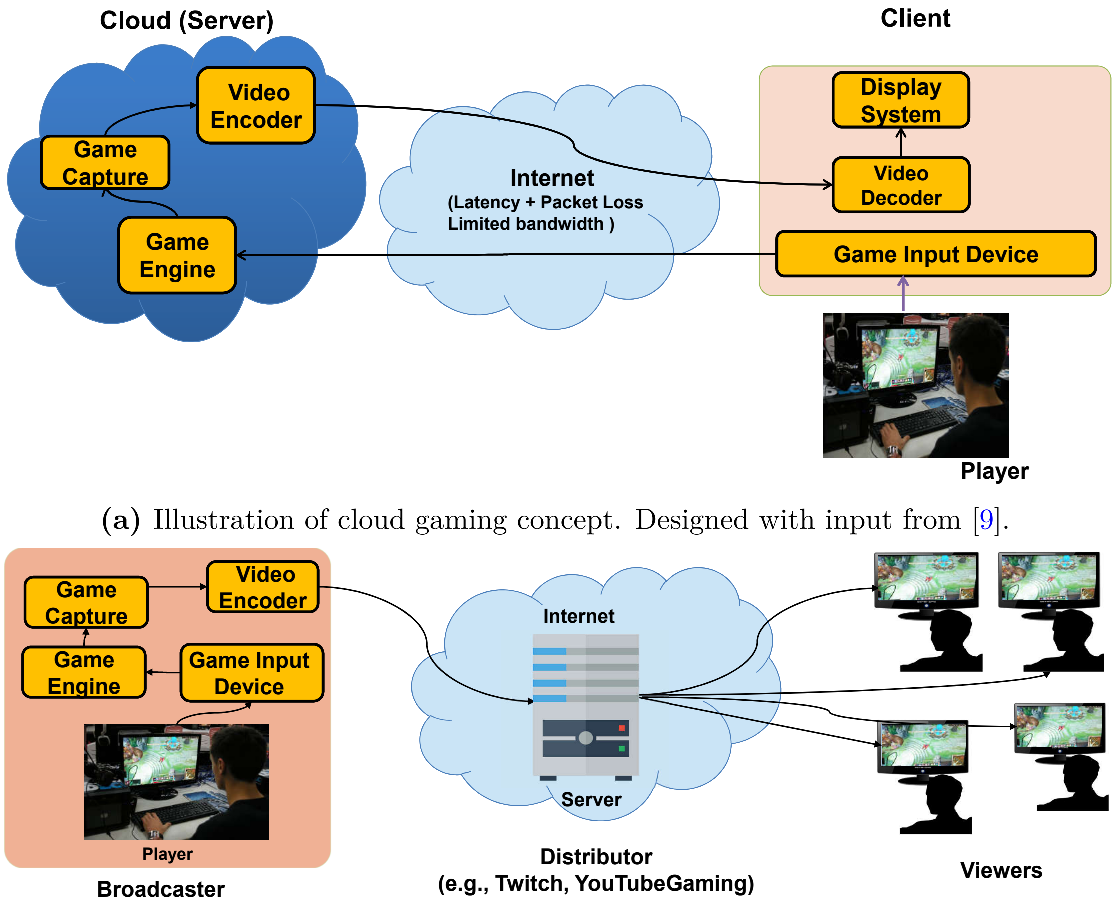


Figure 1.1: Average concurrent streamers and viewers by platform. Adapted based on input from [8].

Figure 1.2a describes the major components and process of cloud gaming. The game engine processes the game related input commands which are then encoded



(a) Illustration of cloud gaming concept. Designed with input from [9].

(b) Illustration of passive gaming video streaming applications (Spectator Gaming).

Figure 1.2: Illustration of cloud gaming and passive gaming video streaming.

in a suitable format by the video encoders to be displayed at the end user after decoding using the video decoder. The encoding is challenging because, during transmission through the internet, the video could be impaired by many factors such as packet loss, delay, bandwidth limitation etc., hence affecting the end user received gaming video quality. Figure 1.2b shows the process of passive online gaming video streaming services over the internet as provided by Twitch.tv and YouTubeGaming. In such applications, the gameplay is usually performed at the client end. The gameplay video and/or audio is then encoded using a video encoder and sent over the internet to the respective OTT server which then transcodes it into different representations and then transmits them to the end user.

1.2 Objectives

This research aims to investigate gaming video quality using subjective and objective quality assessment methods considering passive live gaming video streaming applications such as Twitch.tv, YouTubeGaming, etc. The major objectives of this work are:

(a) *To evaluate the compression efficiency of the three most popular compression standards: H.264/AVC, H.265/HEVC and VP9 considering gaming video content.*

One of the solutions to reduce the increasing demand for bandwidth is by achieving higher compression efficiency during the source encoding process without loss of visual quality, thus preserving the end user QoE. Our initial investigation showed that so far all major OTT gaming video streaming providers are using the H.264 encoder. The introduction of newer encoders such as VP9 and H.265/MPEG-HEVC results in much higher compression efficiency when considering non-gaming videos but at the cost of higher computational complexity. Many previous studies, depending on the configuration and videos considered, have reported different performance results for different codecs. So far, such a performance evaluation considering gaming videos and real-time streaming scenarios is missing. Therefore, to address this research gap, an objective evaluation of eight most popular games encoded using H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 encoders for live game video streaming applications as currently used by Twitch.tv and YouTubeGaming is performed. The results are reported in terms of three objective video quality metrics (Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), Visual Information Fidelity - Pixel Domain (VIFP)), Bjontegaard-Delta Bitrate (BD-BR) analysis, and encoding duration.

(b) *To study and evaluate if there exist any objective and subjective difference in terms of the trade-off between quality and data-rate after compression between gaming and non-gaming videos.*

Gaming videos are artificial and synthetic, and hence the user may perceive the quality differently and have a different set of expectations from such services. Additionally, they have different streaming requirements which can lead to a difference in QoE of the end user as compared to services streaming non-gaming videos. Therefore, it is worth investigating the specificity of gaming videos in relation to compression and the resulting end-user QoE. Towards this end, an objective and subjective quality comparison study for non-gaming videos and gaming videos, with 30 video sequences (15 per type), encoded using the High Efficiency Video Coding (HEVC) compression standard is carried out. The study to evaluate potential similarity and dissimilarity between the two video types is carried out considering various objective and subjective aspects.

(c) *To create a dataset of gaming videos to support reproducible research in the field of gaming video streaming.*

One of the reasons behind the growth of the research in Video Quality Assessment (VQA) was the availability of publicly available databases. Such open source datasets have often been used by researchers as a baseline for performance evaluation of various existing and newly proposed quality estimation methods, hence helping in the advancement of the field of quality assessment. Therefore, an openly available dataset for gaming content will allow researchers to gain comparable and more generalizable results for gaming content, e.g., for VQA, QoE prediction modeling, and selection of optimized encoding settings. Towards this end, the GamingVideoSET dataset, consisting of twenty-four uncompressed raw gaming videos of 30 seconds duration, 1080p resolution, and 30 fps, is designed. The dataset includes subjective quality assessment results for 90 video sequences. In addition to the reference videos, the objective quality assessment results (average and per-frame) using three video quality assessment metrics is available for 576 video sequences obtained by encoding the reference videos in 24 different resolution-bitrate pairs.

(d) *To evaluate the performance of the state-of-the-art VQA metrics for gaming videos considering real-time passive gaming video streaming applications.*

Video quality assessment is imperative to estimate and hence manage the QoE in video streaming applications to the end-user. Recent years have seen tremendous advancement in the field of objective VQA metrics, with the development of models that can predict the quality of the videos streamed over the Internet. However, no work so far has attempted to study the performance of such quality assessment metrics on gaming videos, which are artificial and synthetic and have different streaming requirements than traditionally streamed videos. To address this research gap, a study of the performance of objective quality assessment metrics for gaming videos considering passive streaming applications is carried out. Additionally, various temporal pooling methods are evaluated to see their effect on the performance of the evaluated VQA metrics.

(e) *To design lightweight, machine learning based No-Reference (NR) VQA metrics for passive gaming video streaming applications.*

Previous gaming video quality assessment studies indicated a not so satisfactory performance of existing NR VQA metrics. Also, due to the inherent nature and different requirements of gaming video streaming applications, as well as the fact that gaming videos are perceived differently from non-gaming content (as they are usually computer generated and contain artificial/synthetic content), there is a need for application specific light-weight, no-reference gaming video quality prediction models. To address this requirement, two NR machine learning based quality estimation models for gaming video streaming, NR-GVSQI, and NR-GVSQE, using NR features such as bitrate, resolution, blockiness, etc. are designed.

1.3 Contributions to Knowledge

The contributions to knowledge as an outcome of my PhD research are listed below.

(a) A performance evaluation of three widely used codecs, H.264/MPEG-AVC, H.265/MPEG-HEVC, and VP9, for performance efficiency in terms of BD-BR savings and encoding times considering gaming videos is carried out. For the encoding settings and the encoders used, in terms of BD-BR analysis, H.265/MPEG-HEVC is found to provide the best compression efficiency but is 2.6 times slower than H.264/MPEG-AVC. The magnitude of bitrate savings for VP9 compared to H.264/MPEG-AVC is found to be highly dependent on the content type, with H.264/MPEG-AVC resulting in higher average bitrate savings with an encoding speed four times faster than VP9.

(b) A comparative objective and subjective study of the difference between gaming and non-gaming content in terms of the trade-off between quality and datarate after compression is carried out. For the same encoding settings, in terms of the correlation between objective and subjective scores, it was found that, in general, non-gaming videos achieved a higher correlation score when compared to that of gaming videos, a possible reason for which is attributed to subject bias.

(c) An open source gaming video dataset called GamingVideoSET consisting of 24 reference videos and 576 distorted videos is created to encourage reproducible research in the field of gaming video streaming. The data set includes subjective quality assessment results for 90 video sequences obtained by encoding six different gaming videos using the H.264/MPEG-AVC codec standard in 15 different resolution-bitrate pairs (three resolution, five bitrates each). In addition to the reference videos, the dataset offers a total of 576 distorted videos in

MP4 format, obtained by encoding the videos in 24 different resolution-bitrate pairs, and their objective quality assessment results (average and per-frame) using three video quality assessment metrics. The dataset is freely available for research use at https://kingston.box.com/v/GamingVideoSET.

(d) A performance evaluation study of the eight most popular VQA metrics is carried out. The performance of existing NR metrics is found to be not so satisfactory. A performance evaluation of the metrics using different temporal pooling methods is performed. The results indicate that while depending on the metric, there exist a particular pooling strategy other than normal mean, no single pooling strategy increases the performance of all metrics as compared to the simple mean temporal pooling. Also, an analysis of these metrics as per video content complexity class and different pooling methods is carried out.

(e) Two machine learning based NR metric for passive gaming video streaming applications are proposed. Different features such as bitrate, resolution, contrast, blur, etc. are extracted from the videos at frame level which are then averaged for the whole duration of the videos and then used as input to the machine learning algorithms. The models are trained using subjective and objective measurement as the target, respectively. For the purpose of testing, the existing dataset, another open source dataset, KUGVD, which consists of both subjective (Mean Opinion Scores (MOS)) ratings and objective analysis considering six gaming videos is designed using encoding settings similar to that of GamingVideoSET. The performance evaluation of the two proposed metrics on the two gaming video datasets shows that the proposed models outperform the current state-of-the-art NR metrics, while also reaching a prediction accuracy comparable to the best known Full-Reference (FR) metric.

1.4 Ethics Consideration

This research involved conducting several subjective quality assessment tests that required human subjects to watch (gaming) videos encoded using different conditions and provide their opinions. The test participants were recruited on a volunteer basis and were briefed about the tests, and their questions and concerns were addressed. Taking into account that some games might include violent scenes, special consideration was taken to remove/minimize the violent content and the test participants were allowed to leave the test anytime they wanted. Any personal information collected was stored securely and was not used in any of the results presented in this thesis. The necessary ethics approval for the research was obtained from the Faculty Research Ethics Committee, Kingston University within the scope of Horizon 2020 Marie Skłodowska-Curie Initial Training Network QoENet[4].

1.5 Thesis Structure

Chapter 2 introduces the field of QoE starting with an introduction of QoE - its definition, influence factors and modeling. Different VQA methodologies are discussed followed by a brief literature review of content and context aware video encoding approaches.

Chapter 3 reports the work addressing our first objective where three most widely used codecs are compared for their compression efficiency and encoding run times considering gaming video streaming application scenarios.

Chapter 4 presents a comparative quality assessment study for gaming and non-gaming videos addressing our second objective. Using subjective and objective quality assessment methods, fifteen different gaming and non-gaming videos encoded using HEVC compression standard are analyzed to investigate the possible difference between gaming and non-gaming videos.

Chapter 5 first introduces the open-source dataset, GamingVideoSET which is designed in line with our third objective. The performance evaluation study of eight most popular VQA metrics addressing our fourth objective is also presented in this chapter along with a discussion on the effect of different temporal pooling strategies and content complexity on the performance of different VQA metrics.

Chapter 6 presents our work on the design of two machine learning based NR metrics for quality evaluation of gaming video streaming applications which addresses our fifth objective. An additional open source gaming video streaming dataset KUGVD is also introduced in this chapter.

Chapter 7 concludes this thesis with a discussion of the research findings and potential future work.

[4] http://www.qoenet-itn.eu/

I'm not young enough to know everything.

— J. M. Barrie *The Admirable Crichton, Act I (1903)*

2 Background

Contents

Part of the content of this chapter has been published as: N. Barman and M. G. Martini, *QoE Modeling for HTTP Adaptive Video Streaming–A Survey and Open Challenges*, in *IEEE Access*, vol. 7, pp. 30831-30859, 2019.

2.1 Introduction

The Cisco Visual Networking Index forecasts an increase in Internet traffic, with video alone being 82% of the net consumer Internet traffic by 2021 [10]. With the emerging video formats (e.g., Ultra High Definition (UHD), High Dynamic Range (HDR), Light Field (LF)) and new services such as Virtual Reality (VR), Social-TV, cloud gaming, the available network technology will not be able to meet the increased demand for high bandwidth for all the users and to satisfy users' expectations for any content, any place, any time. Depending on the streaming supply chain stage, there exist different strategies that can be used to optimize the available resources, such as encoding (e.g., VBR, multiple pass, per-title encoding [2], dynamic optimizer [3])), network (resource allocation, load balancing, scheduling, caching etc.) and client (buffering, media representation adaptation etc.). This thesis mainly deals with the encoding process and takes into account HTTP Adaptive Streaming (HAS) based applications (see Section 2.2) considering a live gaming video streaming application. Network and client side optimization methods are not investigated in this thesis.

Towards this end, we present in this chapter an overview of the various technologies and methods considered in this thesis. We start with an introduction of the HAS technology in Section 2.2 followed by a discussion on QoE, its definition, objective and subjective quality assessment methodologies and classification of various factors which might affect the end user QoE in Section 2.3. Section 2.5 discusses the importance of QoE modeling from various stakeholder's perspective. In Section 2.4, various subjective and objective quality assessment methodologies are presented along with a discussion of their advantages and disadvantages. Section 2.6 presents a brief review of the different content and context aware approaches for video encoding followed by a discussion of how requirements of a quality assessment model vary depending on the application scenario under consideration. Chapter 2.7 concludes this chapter along with a discussion of the major learnings and the identified research gap in the field of passive gaming video streaming applications.

2.2 HTTP Adaptive Video Streaming

In this thesis, we focus exclusively on HAS applications using reliable delivery mechanisms such as Transmission Control Protocol (TCP) and Quick UDP Internet Connections (QUIC). Reliable transport protocols such as TCP make sure that all data will be delivered correctly to the destination process without any errors. This

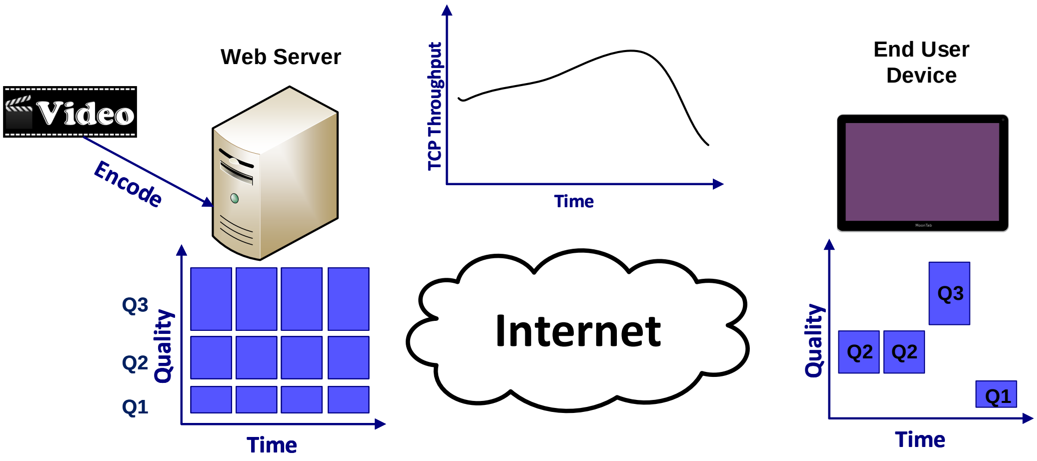


Figure 2.1: HAS Schematic (Q3, Q2 and Q1 denotes high, medium and low quality level respectively).

is usually achieved by a connection oriented approach between the sender and the receiver with the receiver acknowledging the receipt of packets and retransmission of lost or erroneous packets. HAS is one of the most popular streaming technologies for video delivery over the Internet, currently used by the primary OTT providers such as Netflix, YouTube and Twitch.tv. Some of the most widely used implementations of HAS include: Adobe HTTP Dynamic Streaming (HDS) [11], HTTP Live Streaming (HLS) [12], Microsoft Smooth Streaming (MSS) [13] and Dynamic Adaptive Streaming over HTTP (DASH) [14]. The first three are proprietary and vendor specific HAS implementations while DASH, also commonly known as MPEG-DASH, is an open source international standard developed by Moving Pictures Expert Group (MPEG) [15]. The underlying logic is common in all these implementations with some differences in the manifest file, recommended segment size, etc.

2.2.1 Concept Overview

Figure 2.1 illustrates the basic concept behind HAS applications. The video file is encoded at different representation levels (see Section 2.2.2) and then divided into chunks (also referred to as segments) of equal duration (often 2, 4 or 10 seconds, but depends on the standard) which are then stored on a server. The reverse process of first segmenting and then encoding can also be used, as currently done by most of the OTT providers to speed up the encoding process. When a first request for the video file is made by the client, the server sends the corresponding manifest file (e.g., *.mpd* for DASH, *.m3u8* for HLS) which consists of the details about the video file such as video duration, segment size, available representation levels, codec, etc. The client then requests for video chunks based on its rate adaptation logic. The client's rate adaptation logic can be broadly categorized into the throughput-based, buffer-based and hybrid approach. A comprehensive survey of the rate adaptation methods for HAS can be found in [16]. Figure 2.1 illustrates the concept of streaming assuming a throughput-based rate adaptation method. It can be observed that the client, based on its network condition, adapts the quality of the video to provide a smooth streaming experience to the end user.

2.2.2 Quality Switching Dimensions

Videos can be encoded at different bitrates (quality levels) by adjusting one/two/all of the following parameters: spatial resolution, frame rate and Quantization Parameter (QP). A bitrate decrease usually indicates lower quality but the reverse does not necessarily holds true, i.e., increasing the bitrate after a certain threshold (which depends on the video content type) does not necessarily result in higher (perceived) quality videos. Figure 2.2 illustrates the adaptation dimensions for video encoding.

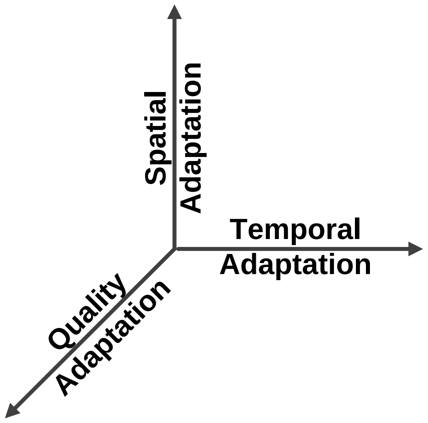


Figure 2.2: Video quality switching dimensions.

1. *Spatial Adaptation*: The videos are encoded at different resolutions, hence decreasing the number of pixels in the vertical and/or horizontal dimensions.
2. *Temporal Adaptation*: The temporal resolution of the video is decreased by dropping some of the frames, i.e., encoding a lower number of frames per second, hence reducing the encoded bitrate.

3. *Quality Switching*: Increasing (decreasing) QP values results in an allocation of less (more) bits per pixel, hence resulting in lower (higher) bitrate values.

The actual dimensions of adaptation depend on the application type and also on the content type. For most content types, compression based quality is considered the most important dimension. For similar bitrate values, spatial resolution reduction is perceived better than frame rate reduction (the actual impact of *upscaling* depends on the specific player used for video playback at the end user device), hence resolution is one of the most widely used adaptation dimensions [17]. For smaller screen sized devices such as mobile, tablets, etc., spatial resolution plays an important role in QoE. In general, in HAS, adaptation in multiple dimensions is perceived better than a single dimension adaptation [18] and hence is widely used by major OTT providers.

2.3 QoE: Definition and Assessment Methodologies

2.3.1 QoE Definition

The EU Qualinet community (COST Action IC1003: "European Network on Quality of Experience in Multimedia Systems and Services") defines QoE as "QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state", which has since then also been adopted by ITU-T as the definition of QoE [19, 20]. QoE takes into account the end user's experience and level of satisfaction and is of much interest to both academic and industrial players in the field of multimedia. Understanding the end users' expectations and experience is paramount to the development of future services as well as improvement of the existing technologies and services. While traditionally QoS has been used to measure the effectiveness of a service, it fails to take into account end user related factors (user expectation, environmental factors, etc.). Also, QoS is limited to telecommunication services and relies only on technical measurements. QoE, on the other hand, covers domains beyond telecommunications and is multidisciplinary in nature, including domains such as psychology, business, technical, environmental, etc. Figure 2.3 illustrates the encapsulation of QoS and QoE.

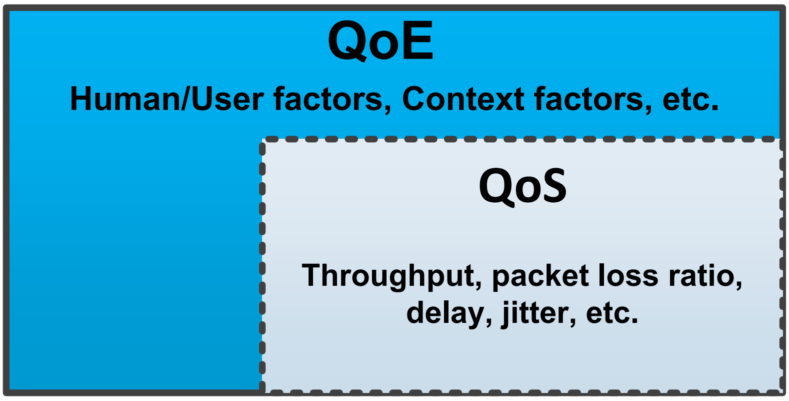

QoE
Human/User factors, Context factors, etc.
QoS
Throughput, packet loss ratio, delay, jitter, etc.

Figure 2.3: QoS and QoE encapsulation.

2.3.2 QoE Influence Factors

A QoE influence factor is "any characteristic of a user, system, service, application, or context whose actual state or setting may have an influence on the QoE for the user" [19]. As defined in ITU-T Rec. P.10/G.100 Amendment 5, QoE influence factors include the type and characteristics of the application or service, context of use, the user's expectations with respect to the application or service and their fulfillment, the user's cultural background, socio-economic issues, psychological profiles, emotional state of the user, and other factors whose number will likely expand with further research [20]. Influence factors on QoE can be grouped into the following four categories as described by the authors in [21].

System IFs

System IFs mostly consist of the technical aspects of quality, for example, the ones which can be measured using QoS based measurement approaches. They cover a wide range of aspects such as media related (quality switching events), network related (wired/wireless/mobile, bandwidth, delay, jitter, packet loss, etc., resulting in impairments such as temporal interruptions/pauses) or end-user device related (display resolution, playback capabilities such as supported codecs, formats, etc.) which affect the QoE of the end user. For example, frequent quality changes to adapt to the network bandwidth can be annoying to the end user. Hence, knowledge of such IFs can help in the design of proper strategies such as how frequently to adapt the quality, etc.

Human IFs

Human or User IFs include aspects which refer to the information about the end-user and related aspects. These include individual characteristics of a user such as expectations from the service, memory and recency effects, usage history of the application (e.g., browsing history, frequently played video), demographic and socio-economic background, physical and mental constitution (users' emotional state), memory, categorization, and attention among many others.

Context IFs

Context IFs deal with factors such as location, end user environment (viewing environment, acoustic conditions, etc.), time of the day, type of usage (e.g., just casual browsing, newly released episode of the favorite TV show), time of service consumption (peak time, offload time, etc.).

Content IFs

One of the most important is the content IFs which addresses the characteristics of the content. The aspects in this category include information about the content being offered by the service/application under consideration. For example, for video, the content level IFs are duration, video type and content complexity (spatial and temporal complexity).

2.3.3 QoE Assessment

ITU-T Recommendation P.10/G.100 Amendment 5 defines QoE assessment as the process of measuring or estimating the QoE for a set of users of an application or a service with a dedicated procedure and considering the influencing factors (possibly controlled, measured, or simply collected and reported) [20]. The main objective of QoE assessment is the design of a system which can identify the various factors and their influence on the end user QoE. Such information can then be used by the various stakeholders for optimization along the process of service delivery (encoding pipeline, load balancing, resource allocation, etc.) to provide a reasonable QoE to the end user while making optimized usage of the available resources. Lossy compression is usually required for multimedia data which need to be transported over the Internet, to decrease the required bandwidth and transport costs. During lossy compression, information is lost, with higher compression ratios resulting in a higher amount of information loss. Also, in traditional streaming technologies, transmission errors such as jitter, delay, packet loss, etc., lead to further artifacts which are annoying to the end user. Since it is almost impossible for most practical applications to provide a service without any artifact, a proper QoE model/metric can help quantifying the amount and kind of distortions and the magnitude of their effect on the end user QoE, which can then lead to the design of proper strategies to help minimize such artifacts.

2.4 VQA Methodologies

VQA approaches can be categorized into two main categories: objective and subjective. Objective VQA methods are mathematical models that aim at providing a quality score which closely resembles the perceived image/video quality. Subjective VQA, on the other hand, tries to take into account the user feedback in the form of ratings and targets to estimate the video quality as perceived by the end user.

2.4.1 Subjective Quality Assessment

Subjective assessment scores are typically reported as MOS which is the average of the opinion scores collected from the assessors for a given stimulus calculated as

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \tag{2.1}$$

where N is the total number of subjects and R is the individual rating provided by a user. For repeatability and validation purpose, common guidelines for conducting subjective tests are issued in ITU-T Rec. BT.500 and ITU-T Rec. P.910 [22, 23]. These recommendations include a detailed description of the test settings, methodology and procedures that need to be followed, including data processing guidelines, such as outlier detection, etc. Some of the standardized testing methods for conducting subjective tests are briefly summarized below:

1. Absolute Category Rating (ACR): Stimulus are shown in a random order, and the test participants rate the stimulus as 1, 2, 3, 4 or 5 where the numbers correspond to Bad, Poor, Fair, Good and Excellent respectively.

2. ACR with Hidden Reference (ACR-HR): This method is similar to ACR, but here reference sequences are also added but without informing the test subject. Using the reference videos scores as the baseline, differential rating for the impaired sequence is calculated.

3. Single Stimulus Continuous Quality Rating (SSCQE): Usually suitable for quality evaluation of longer sequences, here the test participants rate the video stimulus on a continuous basis which is sampled at regular intervals.

4. Double Stimulus Continuous Quality Scale (DSCQS): In this method, the reference and impaired sequences are shown in a random order. The test participants are allowed playback of sequences until they are sure and then rate the sequences.

5. Double Stimulus Impairment Scale (DSIS)/Degradation Category Rating (DCR): The reference sequence is shown first followed by the impaired sequence and subjects then rate the stimulus based on the perception of the impairments (perceptible, annoying, imperceptible etc.).

6. Pair Comparison: In this method, the two impaired sequences are compared and then rated for their quality.

ACR test methodology is easy and fast to implement which can help gather large number of ratings in a brief period of time which makes it one of the most widely used subjective test methodology in the research community including many publicly available datasets. ACR ratings confound the impact of the impairment with the influence of the content upon the subject (e.g., whether the subject likes or dislikes the production quality of the stimulus) [22]. Hence, taking into account the above-mentioned facts as well as the fact that no specific advantage offered by the other test methodologies was required in our studies, we decided to use ACR as the choice of our test methodology in this thesis. Figure 2.4 describes the subjective test procedure used in this thesis. We have used a software-based interface which displays the stimulus (videos) of duration T_V followed by the interface where the test subject can rate the stimulus. The test participant can take as much time, T_R, as required to rate the stimulus, and is illustrated by variable lengths of “Rate” in Figure 2.4. An option of break or “pause” is offered to the test participant if the test is of longer duration ($>$ 30 minutes) in line with the ITU-T Rec. P.910 and ITU-R Rec. BT.500.

2.4.2 Objective Quality Assessment

Based on the relationship between the input and output of the system, i.e., depending on the amount of source (reference) information required, VQA metrics can be classified as FR, Reduced-Reference (RR) and NR.

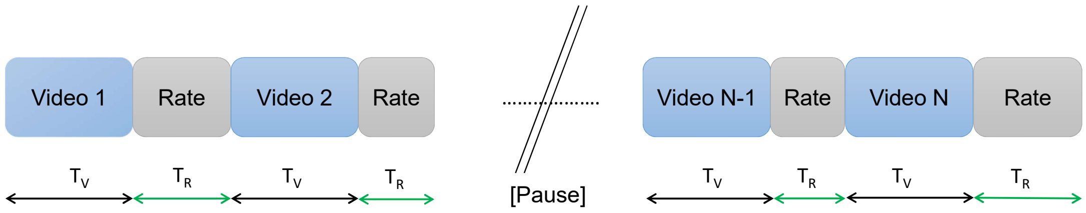


Figure 2.4: Subjective test procedure used in this thesis.

FR Metrics

As the name suggests, FR metrics require the availability of full information of the source video. They are computed based on a frame-by-frame comparison between the reference and the distorted image/video. The source video should be available in pristine quality (unimpaired and uncompressed) so that there can be a direct comparison (e.g., pixel by pixel) between the reference and distorted image/video. Due to the availability of full source information, these metrics are usually more accurate than their counterpart (RR or NR metrics) but as such are not suitable for most real-world applications. Some of the most widely used quality metrics in the field of image and VQA are FR metrics such as PSNR, SSIM [24], VIF [25] and Video Multimethod Assessment Fusion (VMAF) [26].

PSNR is one of the most commonly used metrics for both image and video quality assessment. Even though PSNR does not correlate well with subjective scores, due to its simplicity and ease of computation, it still finds application in many image and video quality assessment fields such as rate-distortion optimization in codecs, codec comparison etc. [27]. Given a reference image (frame in case of video) I_R and the corresponding distorted image, I_D of size MxN, Mean Square Error (MSE) is defined as:

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[I_R(i,j) - I_D(i,j)]^2 \tag{2.2}$$

PSNR in terms of decibels (dB) is then defined as:

$$PSNR = 10log_{10}\left(\frac{MAX_{I_R}^2}{MSE}\right) \tag{2.3}$$

SSIM, which computes the structural similarity between the two images, was shown to correlate better with subjective judgment and hence is also widely used for both image as well as video quality assessment. SSIM which is calculated on various windows of an image is calculated as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\mu_x\sigma_y + c_2)}{(2\mu_x^2 + 2\mu_y^2 + c_1)(2\sigma_x^2 + 2\sigma_y^2 + c_2)} \tag{2.4}$$

Here the measure is considered between two windows x and y of common size NxN and μ_x, μ_y, σ_x^2, σ_y^2 are the average of x and y and variance of x and y respectively. σ_{xy} is the covariance of x and y. c_1 and c_2 are two variables calculated as $(k_1L)^2$ and $(k_2L)^2$ respectively where k_1 and k_2 by default is equal to 0.01 and 0.03 respectively and L is the dynamic range of the pixel-values.

Visual Information Fidelity (VIF) is an image based quality assessment metric which predicts the quality of an image based on the amount of information present in a reference image and the amount of information that can be extracted from the respective distorted image [25]. Taking into account the Human Visual System (HVS), it quantifies the amount of information that the brain can extract from a reference image which is then quantified to the loss of the information corresponding to the distortion using the Natural Scene Statistics (NSS), HVS, and an image distortion model. While originally the proposed model was based on the wavelet domain, a multi-scale pixel domain implementation of VIF, VIFP, was proposed later. Despite comparatively lower performance than the originally proposed VIF, it has found wide spread acceptance and use by the research community due to its simplicity and ease of computation.

VMAF is a fusion based metric proposed by Netflix in 2016 which “fuses” elementary video quality metrics VIF and Detail Loss Metric (DLM) [28] along with Motion into an Support Vector Regression (SVR). VMAF while designed as a video quality metric, computes scores at frame-level which are then pooled temporally (averaged) to obtain a final score between 0 to 100, with a higher score denoting a higher quality. Figure 2.5 describes the steps and working of VMAF. The elementary metrics used are:

- VIF: VIF metric discussed above in its original form combines scores from four scales to calculate the final VIF score. In VMAF, the loss of fidelity score of each of the four scales is considered as individual features.

- DLM is another image based metric which takes into account the loss of useful visual information affecting the content visibility and attention distraction due to the presence of redundant information in the image. In VMAF, only loss of detail is taken into account with special consideration for cases such as black frames where the original metric fails.

- Temporal information or “motion” is considered only for the luminance plane by taking into account the pixel difference between the consecutive frames.

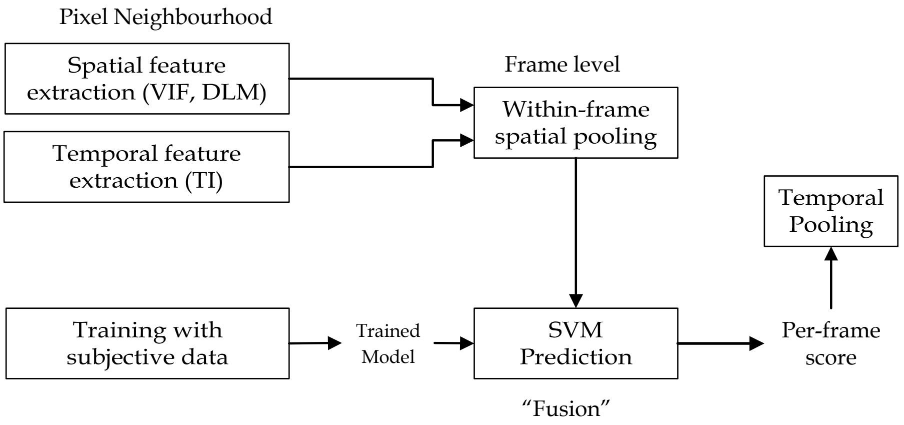

Pixel Neighbourhood
Spatial feature extraction (VIF, DLM)
Temporal feature extraction (TI)
Frame level
Within-frame spatial pooling
Temporal Pooling
Training with subjective data
Trained Model
SVM Prediction
Per-frame score
"Fusion"

Figure 2.5: VMAF system diagram. Reproduced from [29].

The six features (VIF scores over four different scales, DLM and motion) are then used as input to a SVR model which is trained using subjective scores as the ground truth. VMAF is primarily designed taking into account primarily two types of distortion which are the major distortion in HAS based applications: compression and scaling.

RR Metrics

RR metrics are used when only partial information about the reference video is available. Thus, they are usually less accurate than FR metrics but are useful in applications where there is limited source information available due to scenarios such as limited bandwidth transmissions. Taking into account the open source availability of the proposed metrics, as well as considering existing works comparing the performance of the existing RR metrics, we selected Spatio-Temporal-Reduced Reference Entropic Differences (ST-RRED) [30] as one of the RR metrics used in this thesis [31]. ST-RRED which measures the amount of spatial and temporal information differences in terms of wavelet coefficients of the frames and frame differences between the distorted and received videos is one of the most widely used RR metric with very good performance on various VQA databases. Multi-scale multi-orientation wavelet decomposition into different subbands using steerable pyramids is performed on which the ST-RRED algorithm is operated, which is very time consuming (due to the high complexity of the required operations and the number of subbands). In this thesis, we, therefore, used the recently developed optimized version of ST-RRED, known as STRREDOpt, which calculates only the desired sub-band while resulting in a similar performance as ST-RRED but is about ten times computationally faster [32]. In addition, we also use the recently proposed

Spatial Efficient Entropic Differencing for Quality Assessment (SpEEDQA) model, which is almost 70 times faster than the original implementation of ST-RRED and seven times faster than STRREDOpt as it considers only the spatial domain for its computation [33].

NR Metrics

NR quality metrics do no use any source/reference information and try to predict the quality based on the received signal. In the absence of source information, such metrics are usually less accurate than their counterparts, FR, and RR metrics. Since for gaming applications, a high-quality reference video is typically not available, the availability of good performing no-reference metrics is of very high importance. For this thesis, we selected three NR metrics: Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [34], Natural Image Quality Evaluator (NIQE) [35] and Blind Image Quality Index (BIQI) [36] which are based on NSS and are some of the most commonly used NR metrics. NSS is based on the observation that the distribution of pixel intensities of natural images differs from that of distorted images, which is more pronounced when the pixel intensities are normalized. The distribution computed over these normalized pixel intensities follows a Gaussian distribution, which is not true if the image is distorted or unnatural. Hence a deviation of the distribution from the ideal Gaussian distribution shape can be used as a measure of distortion in an image. The most commonly used method to compute the normalized pixel intensities is Mean Subtracted Contrast Normalized (MSCN) where the local mean pixel intensity ($\mu(i,j)$) is subtracted from the pixel intensities ($I(i,j)$) and then divided by the local variance field ($\sigma(i,j)$) as

$$\hat{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C} \tag{2.5}$$

where $\hat{I}(i,j)$ is the MSCN coefficient and C is a constant to make the denominator non-zero. We briefly discuss next the three NR metrics here. A more detailed discussion of the various other NR metrics is provided in Chapter 6.2.

BIQI is a NR metric which is based on a 2-stage modular framework - classification and prediction. In the first stage, it uses NSS to find the type of distortion in the image and classify the image into a specific distortion type among the five distortion classes. In the second stage, it uses distortion specific Image Quality Assessment (IQA) algorithms for quality assessment, which in the default case are specifically trained Support Vector Machine (SVM) models for each distortion type.

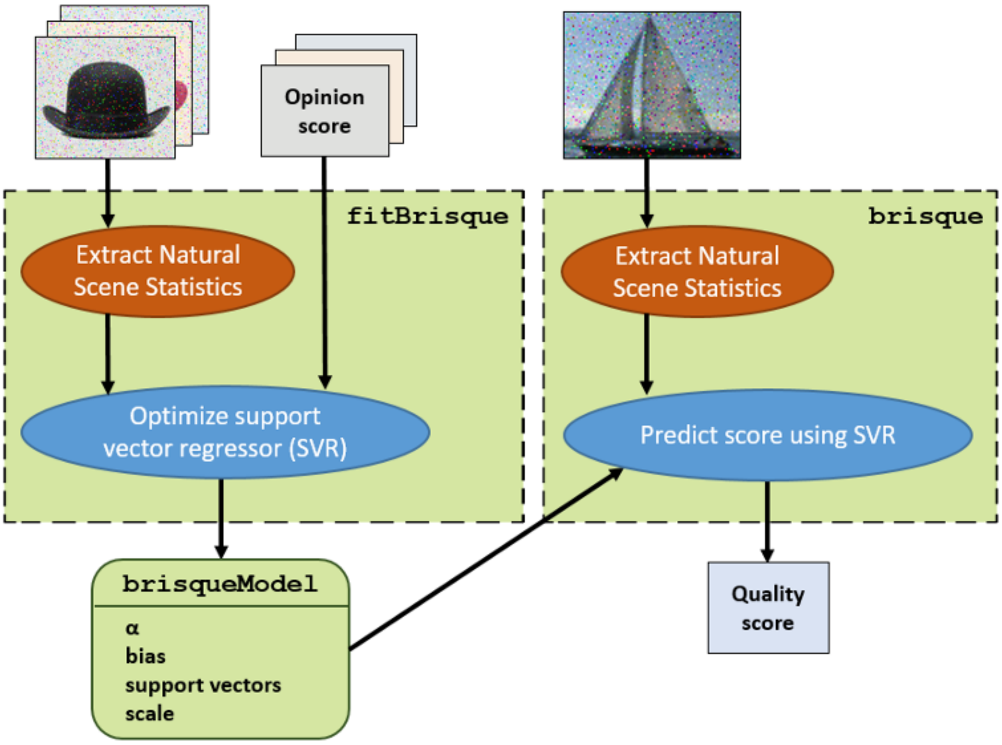

Opinion score
fitBrisque
brisque
Extract Natural Scene Statistics
Extract Natural Scene Statistics
Optimize support vector regressor (SVR)
Predict score using SVR
brisqueModel
α
bias
support vectors
scale
Quality score

Figure 2.6: Full BRISQUE workflow [37].

BRISQUE is a class of opinion-aware IQA which also uses NSS to predict the image quality without using any reference information. Figure 2.6 shows the workflow of the BRISQUE prediction model as used in this work [37]. The model uses training images from which NSS features are extracted and along with the opinion scores of the respective images they are used to train and optimize a SVR model. For estimation of the quality of an unknown image, the same NSS features are extracted which are then used as an input to the trained SVR model to predict the quality.

While BRISQUE is from an opinion-aware class of IQA models in that it requires opinion scores of the images to be used as an input to the training algorithm, NIQE belongs to the class of opinion-unaware IQA algorithms as it does not require any opinion scores for its model training. Figure 2.7 shows the workflow of the NIQE model as used in this work. During the training phase, various NSS features are extracted from the training images to which a generalized Gaussian model is fitted. For image quality prediction, the same NSS features are extracted and then a Gaussian model is fitted to the extracted coefficients. Based on the distance between the fitted Gaussian model of the test image and the generalized trained Gaussian model, a quality estimate score is provided.

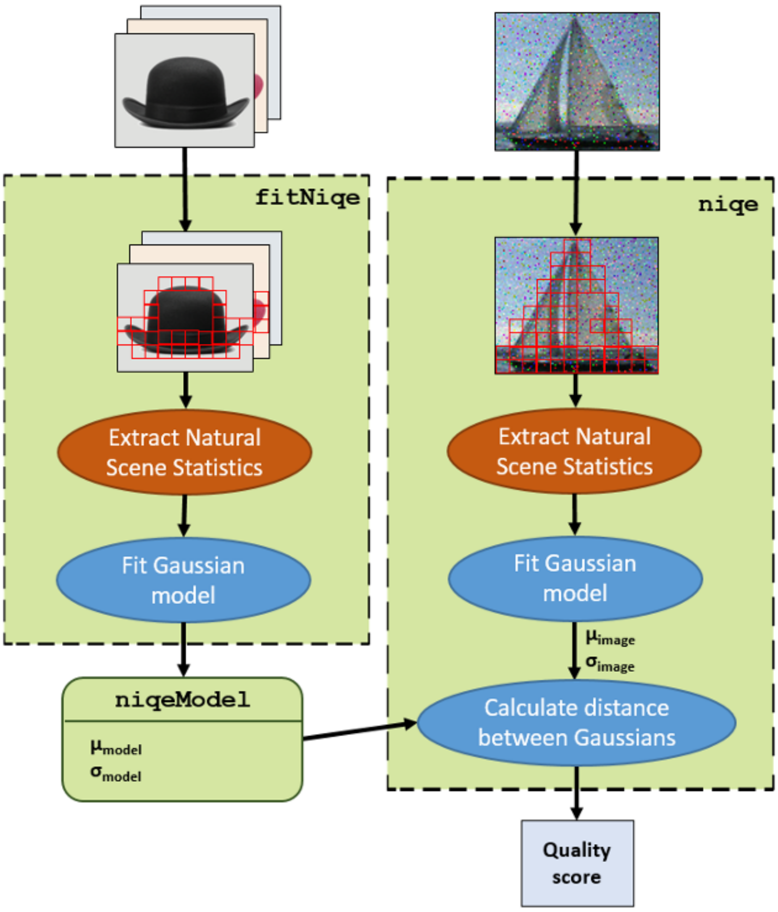


Figure 2.7: Full NIQE workflow [37].

2.4.3 Discussion

Both objective and subjective VQA approaches have inherent drawbacks. While subjective VQA provides information on the actual quality experienced by the users, it is not suitable for real-world applications. Also, conducting subjective tests incurs costs and time, and only a small number of influence factors can be evaluated due to the constraints such as limited test duration and a limited number of assessors. Objective VQA using metrics such as PSNR and SSIM, while fast and comparatively easier to implement, do not always correlate well with the end user quality [24, 38]. For two videos of different (perceivable) quality, the objective metric may provide a similar score and hence does not necessarily reflect the end user's perceived quality. Also, many objective metrics require source sequences, which is not practical in most of the real-world quality estimation scenarios. The quality metrics such as PSNR and SSIM were initially developed and used for IQA. For VQA, they are calculated on a frame-by-frame basis and then the final

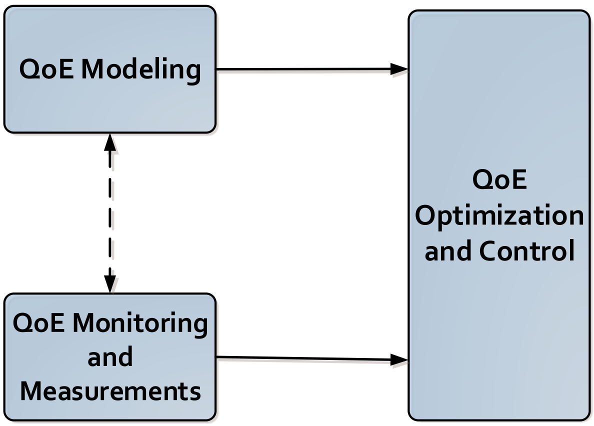


Figure 2.8: QoE Management Process.

score is reported as the average of the individual scores over the full duration of the video sequence. In this thesis, for computation of PSNR, SSIM and VIFP, VQMT tool provided by the researchers in [39] has been used. For VMAF, the implementation provided by Netflix in [40] was used. For RR metrics, SpEEDQA and STRREDOpt and NR metric BIQI, the implementation and default settings as provided by the respective authors has been used while for NIQE and BRISQUE, the implementation provided in the Matlab was used [41].

2.5 Importance of QoE Modeling and Performance Evaluation of QoE Models

Managing QoE in a communication system is a complex task, primarily consisting of three steps, as shown in Figure 2.8 and discussed in [42] and [43]. A key step in QoE management is the design of QoE models. ITU-T Recommendation P.1201 defines a QoE model as “An algorithm with the purpose of estimating the subjective (perceived) quality of a media sequence” [5]. QoE models take into account various influence factors and try to estimate the end user QoE. QoE monitoring and measurement(s) can be done by any stakeholder and the parameters measured will depend on the application and the interests of the stakeholder [44] [45]. The final step in QoE management includes QoE optimization and control, typically performed based on models or measurements. Again, the optimization process and the parameters controlled will depend on the stakeholder and the application type. In this chapter, we limit our discussion to the first step, focusing on QoE Modeling for HAS applications using reliable transport protocols such as TCP or QUIC [46].

QoE modeling is one of the critical steps in the QoE management process chain, as the performance of the QoE model will decide the reliability and accuracy of the next steps along QoE based management. We discuss next the importance of QoE modeling from the point of view of various stakeholders in the multimedia streaming process chain.

2.5.1 Importance of QoE Modeling for Different Stakeholders

Network Provider

With increasing demand for OTT services, both VOD and live, there is a tremendous pressure on the network operators to provide seamless connectivity and high QoE to the end users. QoE models can help network operators identify the various IFs, and their respective impact on the end user QoE and hence allow the network operators to take necessary actions (resource allocation such as network throttling, load balancing, caching and network provisioning) to prevent user churn.

Service Provider

In today's highly competitive environment with almost similar pricing schemes, the service provider cannot rely on profit generation based solely on the provision of a service, but should also take into account different factors which may shift the user base to the competitors. For example, for a service provider, measurable QoE factors such as viewing duration are of huge interest [47]. For advertisement based services, longer viewing duration implies more advertisement. On the other hand, for subscription based services, a shift of even a smaller percentage of viewer base can result in a significant effect on revenues. One of the disadvantages of HAS services is the requirement of additional storage space, as multiple copies of the same file are stored in the server. In such cases, optimized encoding bitrates can lead to huge storage space savings for the OTT provider while also reducing the demand for the required bandwidth. Hence, proper QoE models can provide an insight into the IFs and their impact on the service, and in turn, allow the service provider to take appropriate decisions/measures to ensure high end user QoE.

Device Manufacturer

Nowadays, most of the device manufacturers, such as Samsung, LG, Sony, etc., are involved in the manufacturing of both small screen devices (mobiles, tablets) and big screen devices (PC/TV). Different devices have different capabilities, and the perceived quality depends on various factors, one of which is the device screen size. Also, small screen devices have different processing capabilities compared to large screen devices. Hence, good QoE models can provide insight to the device manufacturers, considering the device features (display size, display resolution, RAM, etc.), on what settings to use such that the QoE of the end user can be maximized. The models, depending on their design can also be used for codec comparison and hence allow device manufacturers to provide optimized encoding and decoding capabilities so as to support the latest codecs in the shortest possible time. Many device manufacturers are also interested in QoE modeling for production of QoE monitoring solutions such as probes, QoE estimation modules etc.

End User

In the end, the user is the king or queen. The success of a service will depend on the acceptance of the same by users. As mentioned in [43], successful QoE management will lead to satisfied end users as their requirements and/or expectations will be met and hence they may be further open to adopting new and complex services, leading to the growth of more advanced technologies.

To summarize, QoE modeling can help us identify the various influencing factors which might have an impact on the end-user QoE. The actual applicability and performance of the model will vary depending on the stakeholder as different actors involved will focus on different aspects (mostly the ones they can control). For example, in the case of HAS, a network provider may be interested in rebuffering, quality switches, etc. and their corresponding effect on QoE as they are directly or indirectly related to the network QoS parameters such as delay, jitter, packet loss, etc. A content provider may be interested more in the effect of average bitrate, segment size, video popularity, etc., for example, to save storage costs, optimized video caching, etc. At the application layer, the service provider may be interested in IFs such as adaptation frequency, adaptation magnitude, etc. to take these into account for the design of the client's adaptation algorithm.

2.5.2 QoE Model Performance Evaluation

The aim of a VQA metric is to use objective measurements such as signal fidelity measurements to predict visual quality as perceived by human observers. The criteria for the evaluation of the performance of the objective QoE model, as mentioned initially in Video Quality Experts Group (VQEG) FRTV Phase I and later in VQEG FRTV Phase II [48, 49], are:

- *Prediction Accuracy*: It refers to the ability of a model to predict the subjective rating scores with low error.
- *Prediction Monotonicity*: It refers to the degree of model's prediction agreement with the relative magnitudes of the subjective rating scores.
- *Prediction Consistency*: It refers to the ability of a model to maintain prediction accuracy over a wide range of test sequences with a variety of video impairments.

The prediction accuracy of a model can be evaluated by using the Pearson Linear Correlation Coefficient (PLCC) between the predicted and actual subjective rating scores which determine the strength and direction of the linear relationship between them. Similarly, the prediction monotonicity of a model can be evaluated using the Spearman's Rank Correlation Coefficient (SROCC) between the predicted and actual subjective rating scores which indicates the strength and direction of the monotonic relationship between both scores. Finally, the prediction consistency of the model can be evaluated using measurements such as the Outlier Ratio (OR). A low OR value indicates a high consistency of prediction, with OR = 0 implying that the model will be stable to predict the QoE.

2.6 QoE Awareness for Video Coding

The new video formats such as 4K and HDR result in files of enormous size and hence call for modern video compression standards. The effort in this direction resulted in the recently introduced new video compression standard H.265/MPEG-HEVC, which on an average, for the tested sequences, is shown to achieve 50% higher compression efficiency than its predecessor H.264/MPEG-AVC [50–52]. VP9, a royalty-free encoder developed by Google as a competitor of the H.265/HEVC encoder, has gained much popularity and is supported by almost all browsers except for Safari. Licensing issues with H.265/HEVC and the aim to develop a more futuristic royalty-free video codec led to the creation of a consortium of industry partners called Alliance for Open Media (AOM)[1]. The joint efforts of the members of AOM have since then drove to the development of the AV1 codec[2] with the final bitstream specification frozen in early 2018. Recent studies comparing the performance of AV1 with x265, x264 and libvpx considering on-demand adaptive streaming applications have found it to result in the highest bitrate savings but at the cost of huge encoding times [53, 54]. The applicability of such encoders for live streaming applications remains an open question.

[1] http://aomedia.org/

[2] https://aomedia.googlesource.com/aom/

2.6.1 Compression

Video compression typically consists of three steps: signal decorrelation, discarding information (lossy compression) and entropy coding [1]. The three steps include methods for removal of statistical redundancies, such as motion compensation, entropy coding, intra frame predictions, etc. as well as exploiting perceptual redundancies such as removal of high frequency coefficients [55]. Many of the existing encoders already use the content information to achieve higher compression ratio, mostly by removal of statistical redundancy. Some aspects of QoE are considered in existing video codecs. One of the most common examples is chroma sub-sampling, i.e. representing image/video in the YUV colour space putting more emphasis on luminance and less on chrominance by taking into account the HVS. While there is some work which already exploits HVS to reduce perceptual redundancy to achieve higher compression efficiency, there still exist many redundancies considering HVS which can be exploited to achieve increased compression efficiency without much loss of perceived visual quality.

2.6.2 State of the Art Encoder: HEVC

The latest video compression standard, HEVC/H.265 achieve a compression efficiency which is twice the compression efficiency of H.264/AVC, thanks to the exploitation of content information such as motion vectors, uniform spatial regions, etc. during various steps of compression at the expense of higher computational complexity. Figure 2.9 shows the block diagram of a hybrid video encoder for HEVC. Video codec performance can be considered as a trade-off between three variables: quality, bitrate, and computation cost, with HEVC achieving a higher compression ratio at the cost of increased computation. The major improvements over the previous standards which result in improved compression efficiency are [50]:

1. **Larger and flexible coding block size**: Larger coding blocks called Coding Transform Unit (CTU) with a max size of 64x64 are allowed which decreases the signaling overhead. Each CTU can consist of a combination of both intra predicted blocks as well as inter predicted blocks. Taking into account video content for better adaptation, the CTU can further be divided into a coding unit and prediction unit, which can have different motion data or prediction modes.

2. **Larger Interpolation filter**: HEVC uses an 8/7-tap filter for luma component and 4-tap for chroma component (compared to 6-tap and bilinear filter for luma and chroma component respectively in Advanced Video Coding (AVC)). This allows for up to 1/4 pixel accuracy in motion vectors. The interpolation is achieved using integer pixels.

3. **High throughput Context-Adaptive Binary Arithmetic Coding (CABAC) and Advance Motion Vector**: CABAC helps achieve 10-15% higher coding efficiency compared to Context-Adaptive Variable-Length Coding (CAVLC). While the basic structure of CABAC is same as that in AVC, the improvements include reduced number of context bins and increasing the “fast” bypass bins thus achieving almost three times reduction in context memory and 20 times reduction in line buffer for context selection.

4. **Larger transforms and more sizes**: At the cost of increased complexity and memory requirements, HEVC supports 4x4 up to 32x32 integer transforms which helps in achieving a 5-10% increase in coding efficiency. Besides, for certain types of contents such as screen content, graphics etc., and the transform can be skipped to achieve higher compression.

5. **More prediction modes**: The number of intra-picture prediction modes supported in HEVC is increased to 35 (33 angular, one planar and one DC mode) compared to just ten modes supported in H.264/AVC.

6. **In-loop filtering (Deblocking filter and Sample Adaptive Offset (SAO) filter)**: One or two of the filtering stages can be applied. The deblocking filter is mostly similar to that of AVC but has been modified to simpler version and parallel processing friendly. The SAO is a non-linear amplitude mapping filter to address local discontinuities by adding an offset (band offset or edge offset) based on the values of the neighbour in one of the four directions (0, 45, 90, 135 degrees).

2.6.3 Content Aware Coding Strategies

Psychological studies have proven the fact that humans perceive various regions of an image/video differently. As reviewed in [55], human visual perception mechanisms such as contrast sensitivity, masking, fovea, and visual attention can be used to understand and design perceptual models using techniques such as a priori manual Region of Interest (ROI) selection, user input-based attended region selection, visual

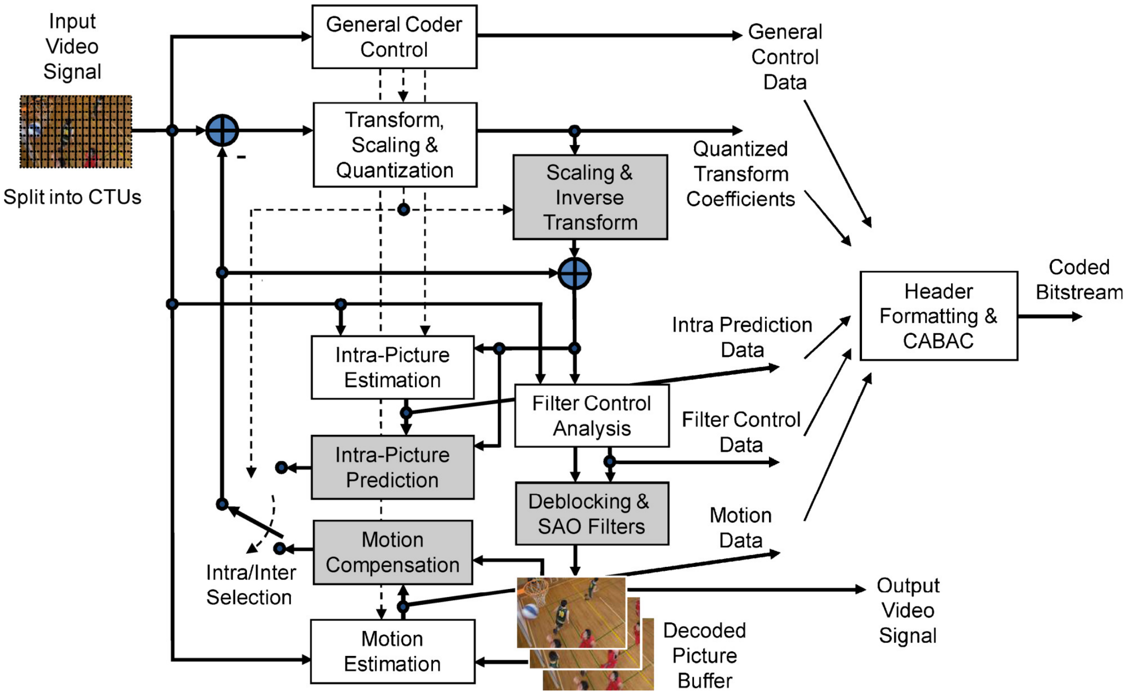


Figure 2.9: HEVC video encoder (with decoder modeling elements shaded in light gray) Courtesy of [50].

attention modelling and visual sensitivity modelling. Hence, exploiting the HVS, inattentive and attentive regions of an image/frame may be treated unequally, such as by providing more importance to the attentive regions. Such regions are termed as ROI which can be detected using either feature based or object based approaches. ROI based scalability allows the encoder to encode the important regions at higher quality while lowering the quality at not so important regions. ROI varies with context and is hence difficult to detect/predict. Examples for ROI include the face of a speaker during a video call or a region of the image/video consisting of a tumour in case of medical applications (e.g. e-Health scenarios). ROI detection has received huge interest for both image and videos, and much work has been done in the detection of such regions as well as for methods that exploit the ROI for improved performance gains. A ROI analysis may be used either to achieve improved error resilience and/or to achieve higher compression efficiency. One such work presented in [56] proposes a method to automatically detect the ROI based on visual content analysis and another based on eye tracking methods.

Content information can be used to improve existing technologies such as MPEG-DASH. Over the past few years, adaptive bitrate streaming technologies such as MPEG-DASH has gained tremendous attention due to its flexibility of quality adaptation, guaranteed video delivery, and firewall friendliness amongst other advantages. As mentioned in [57], subjective perception of video quality adaptation is influenced by the content type which can be exploited for improved optimization of video delivery using HAS. During the encoding process, the content information such as scene cuts, high motion contents, ROI etc. can be used for optimized selection of factors such as frame rate, resolution, quantization factor etc. The content information may also be used for adaptation strategies such as segment duration, adaptation/switching points, Group of Pictures (GOP) length selection etc. Authors in [58] proposes a content-aware adaptation scheme by classifying high motion content as important content since they appeal more to the users. The adaptation is performed according to the principles below:

- Selecting the same video representation for segments belonging to a scene to avoid quality oscillation.
- Buffering high motion scenes during low motion scene playback to avoid stalling.

In [59], the authors propose a method for HLS using a trellis representation and Just Noticeable Difference (JND) metric. Using the concept of JND, the authors propose an adaptation method which provides a smooth playback with quality switches limited to 3 JND (perceptually smooth transition) while maintaining a small buffer size. Similar work by Netflix referred to as "Per-Title Encode Optimization" demonstrates that content classification based on genres such as comedy, action, drama, sports etc. alone is not enough [2]. Each video is different and needs smarter processing for content analysis so as to achieve the best quality at a given bit-rate. Using perceptually spaced QP to obtain the next bit-rates, they produce bit-rate quality curves, ensuring smooth quality transitions.

Compound videos which refer to videos where there is a simultaneous occurrence of natural video, text, and graphics within a single frame can exploit content awareness to achieve better compression efficiency. Examples of compound videos include applications such as mobile gaming, conference calls etc. Traditional transform based coding is not suitable for such videos as they often miss out on uniform parts where there could have been more compression as well as poor performance, especially at the edge boundaries of the text. In such cases, content awareness can be used to achieve higher compression efficiency by distinguishing between natural video content and screen content. One such work in [60] divides the compound frames into sub-frames to different layers for compression using a combination of block-level classification, object-level natural video detection, and layer-level video compression. Based on the observation that for screen content encoding consisting of sharp edges on object boundaries, direct encoding of residual signals in spatial domain may not be efficient enough, authors in [61] propose edge mode scheme for HEVC which selects the best edge mode after identifying a set of edge modes based on intra prediction directions. In comparison to the unmodified HM encoder, the proposed scheme achieves up to 18% bitrate savings for normal videos and average 10.4% for screen content video.

2.6.4 Context Aware Coding Strategies

Context awareness has recently gathered much attention due to the advantages it provides for many technologies such as mobile networks, wearable gadgets, context (e.g. location) based services, etc. Objectives of the context aware approaches, in general, include maximization of the QoE of the user, optimizing the usage of distributed hardware resources and optimizing the usage of radio resources for video transmissions. Context information such as computational power, memory, battery life, communication medium (wireless/fixed), etc. may be used to decide on the encoding process. For example, a low end device such as older versions of iPhone may not be capable of decoding HEVC Main Profile Level 10. Hence, services incorporating real time encoding can use the context information (here device capabilities) to avoid bandwidth wastage by transmitting a video at a higher bitrate than the capability of the device (e.g. decoding and display capabilities). In addition, viewing context such as environment (lighting conditions etc.) effects the QoE of the end user as they have an impact on the users' perception of the video. Context awareness has much lately found significant application in e-Health/mobile health scenarios. As described in [62], possible use for context awareness includes a patient health monitoring system where patient data and status (critical/non-critical) are considered for optimized compression of video data. Higher compression can also be achieved by taking into account the context such as ROI detection, followed by lower and higher compression of the regions depending on the priority.

With the penetration of electronic devices, there exist numerous kinds of devices with varying capabilities. Hence, content adaptation is required depending on context information such as the capabilities and specification of the end device for optimum usage of available network resources as well as to provide improved QoE to the end user. For example, a video may be watched by a user on his/her laptop and by another user on a 55″ 4K TV. Hence, context information such as the device resolution, software/hardware (decoder) capabilities, and network type (fixed/wireless) is required to optimize the quality of the content provided by the network. This calls for context aware solutions which adapt the videos taking into consideration the context.

Context awareness in the case of adaptive streaming technologies such as MPEG-DASH is widely used to optimize many parameters such as for video segmentation, resolution/bit-rate adaptation etc. In general, the adaptation algorithm asks for the representation depending on the bandwidth information and/or buffer status. In [63], the authors identify and illustrate the impact of three important context factors (namely display size and viewing distance, surrounding luminance and body movement) on video coding and adaptation. The authors propose an environment aware video adaptation strategy based on the observation of environment based human eye sensitivity to distortion. The proposed scheme is proven to achieve the same perceptual quality at almost 30% reduced bandwidth requirement, hence advocating for the feasibility of context aware coding approaches in MPEG-DASH. Video decoding is a computationally intensive process with high quality videos requiring more power than lower quality videos. High quality videos may not always result in increased QoE. One such example is when the user device has limited battery left and wants to maximize the viewing duration rather than watching it at the highest quality. In such a context, device information such as battery status may be used for energy savings by streaming lower quality videos as illustrated in [64].

Many recent works have been proposed recently which uses location information to optimize video streaming especially in the field of adaptive streaming. Authors in [65] introduces a tunnel case study for HAS which uses location awareness to perform optimized video streaming by buffering video segments before entering a “coverage hole” with no signal. Such predictive approaches result in reduced or no stalling events at the cost of lower quality video playback, hence increasing the QoE (as stalling is one of the significant factors for decreased QoE). Authors in [66] propose a novel method using context information for channel prediction on a long time scale and discuss how such predictions can be used for predictive video streaming applications.

2.6.5 Discussion

Based on the brief review of existing works on content and context aware approaches to improve the end user QoE, it is evident that using different context and content information, the end user QoE can be improved. We saw how the content and context information used depends on the type of application and the type of content. One such domain which in recent years has increasingly seen the application of quality assessment models and methodologies in specific domains is e-Health/telemedicine.

Telemedicine also referred to as e-Health applications refer to the use of electronic and communication technologies such as medical images/videos for delivery of healthcare. Quality of the transmitted information in such applications is of paramount importance. For example, introducing distortions to the part of the image or video carrying sensitive information (tumor in an ultrasound image or location of the fracture in an X-ray image) can potentially affect the diagnosis by the doctor, hence risking patient's health. Quality assessment metrics and models can help to preserve the diagnostic quality[3] in such applications. Therefore, recent years have seen lots of work in quality assessment in the field of telemedicine research with the development of various technologies such as image processing, video compression, and transmission technologies [68]. Since the already proposed quality metrics are designed taking into account general requirements and hence does not necessarily takes into account the diagnostic quality requirements of medical videos, many different approaches such as receiver operating characteristics analysis [67], context-aware ultrasound video transmission [69], medical image quality index using grey relational coefficient calculation approach [70], Cardiac Ultrasound Video Quality Index (CUQI) using motion and edge information of the cardiac ultrasound video [71] among many others have been proposed in recent years.

2.7 Conclusion

In this chapter, we presented the basic concepts of QoE, different methods to measure QoE and the need for QoE measurement and how such QoE measurements can be used by various stakeholders for QoE based control and management of their services and products. While there has been much work towards content and context aware approaches to improve the QoE of traditional video streaming services such as Netflix, YouTube etc., live passive gaming video streaming services has received almost no attention from the research community. Gaming video streaming in recent years have seen a tremendous rise in terms of both popularity and acceptance by the end users but a quality assessment study of passive gaming video streaming applications has been missing so far. Similar to eHealth/Telemedicine, such a new but emerging field of gaming video streaming can be considered as an application with different requirements and properties as compared to the traditional 2D video streaming services which provide us with both new research opportunities as well as challenges. One of the starting research objective in this direction will be

[3] Diagnostic quality is defined as the acceptability measure of a medical image/video for diagnosis process [67].

to establish if and how passive gaming video streaming applications differ from traditional non-gaming videos streaming applications. Other challenges will be to study the effectiveness of existing VQA metrics for gaming video streaming applications in order to design a NR VQA metric with high performance on gaming videos considering passive gaming video streaming applications.

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe

3

H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 Codec Comparison for Live Gaming Video Streaming

Contents

The work presented in this chapter has been published as (with minor modifications): N. Barman and M. G. Martini, "H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 Codec Comparison for Live Gaming Video Streaming", in *9th International Conference on Quality of Multimedia Experience (QoMEX 2017)*, Erfurt, Germany 2017.

3.1 Introduction

As discussed previously in Chapter 2, one of the solutions to reduce this increasing demand for bandwidth is by achieving higher compression efficiency during the source encoding process without loss of visual quality, thus preserving the end user QoE. H.264/MPEG-AVC, jointly developed by ITU-T Video Coding Experts Group and ISO/IEC JTC1 MPEG and standardized in 2003, is one of the most widely used video codecs in many applications for compression, recording and video content transmission including real-time services such as Twitch.tv. Building on its predecessor, H.264/MPEG-AVC included many new features such as new Discrete Cosine Transform (DCT) design features, increased inter-picture prediction modes (quarter-pixel precision for motion compensation, multiple motion vectors per macroblock, etc.), lossless macroblock coding features, an in-loop deblocking filter, entropy coding design (CABAC and CAVLC), among many others.

With the increased resolution of the videos, increasing amount of video traffic, etc., there was an increasing demand for a newer standard to improve upon the compression efficiency with respect to that offered by H.264/AVC. Efforts in this direction resulted in the H.265/MPEG-HEVC codec which for the same quality achieves almost 50% bitrate reduction compared to its predecessor H.264/MPEG-AVC (based on the test conditions and video sequences used by the Joint Collaborative Team on Video Coding (JCT-VC)) [51, 52]. H.265/MPEG-HEVC, standardized in 2013, is the newest video compression standard developed by the JCT-VC which was presented and discussed in detail in Section 2.6.2. The increased compression by H.265/MPEG-HEVC is achieved at the cost of increased computational complexity. In the rest of this chapter, we interchangeably use H.264 or AVC to refer to H.264/MPEG-AVC and H.265 or HEVC to refer to H.265/MPEG-HEVC.

As a successor to the earlier VP8 standard and competitor of H.265/HEVC, Google released an open and royalty-free video codec called VP9 [72]. VP9 video in WebM container is currently used in YouTube and supported on most browsers. Based on the traditional block-based transform coding format, VP9 offers higher ($64x64$) coding unit blocks, increased intra-frame prediction methods, up to eight-pixel precision motion vectors, asymmetric DCT, etc. Similar to profiles and levels defined by H.264 and HEVC, VP9 support four different profiles (0, 1, 2 and 3) with varied bit-depth and chroma subsampling support.

In the field of gaming, while much effort has been put into design and optimization of games, cloud gaming platforms, etc., very few studies have focused on streaming of gaming videos. The importance of optimization of gaming videos streaming is evident from the fact that Twitch.tv is currently ranked 4th in terms of total traffic during peak hours in the U.S.A. [73]. The increased demand for such OTT services leading to increased bandwidth consumption calls for more efficient encoders. Gaming videos, unlike traditional streaming videos, consist of synthetic and artificial content and may have very high spatial and temporal complexity (depending on the genre), with some scenes of very high motion content followed by scenes with very low motion content. Also, they need to be encoded and streamed in real-time. To this end, we present in this chapter a video codec comparison of the three most popular codecs, taking into consideration the following factors:

1. The study is performed using gaming videos of some of the most popular games streamed on Twitch.tv, one of the most famous and widely used platform for live and on-demand gaming video streaming.

2. FFmpeg and its open source codec implementations for the three codec standards are used to encode the videos to make the results reproducible in an easy and cost effective manner.

3. Real-time encoding settings for live streaming of the games as recommended by Twitch.tv and FFmpeg are used [74, 75].

The main objective of this work is to study the performance of the three most widely used codecs for applications which include real-time encoding and transmission of gaming videos such as Twitch.tv. Earlier performance studies of the three encoders have mostly focused on non real-time applications such as VOD and TV Broadcast scenarios.

The rest of this chapter is organized as follows. Section 3.2 describes similar works involving codec comparison. Section 3.3 describes the source videos, encoder settings and encoding parameters used in this work. Results are discussed in Section 3.4 and Section 3.5 finally concludes this chapter.

3.2 Related Work

The authors of [76] performed a large scale analysis of x264, x265 and libvpx encoders for VOD application scenarios for Standard Definition (SD), High Definition (HD) and Full HD (FHD) resolutions. The videos were encoded content adaptively using 3-pass setting (video was encoded at eight different quality levels and then the encoded bitrates were used for 2-pass VBR encoding). Since there are no encoding time constraints, a *placebo* preset was used which lets the encoder achieve the best compression performance. Results are reported in terms of four objective metrics and BD-BR results with x265, on an average, providing the best compression efficiency compared to libvpx and x264.

In [77], the authors provide an objective and subjective assessment to measure the coding efficiency of the three codecs using the reference encoders but limited to UHD resolution and TV broadcast scenarios. HEVC was found to perform better than AVC and VP9. When compared subjectively, AVC performed better than VP9 while objectively VP9 performance was better than AVC.

In [78], the authors performed a codec comparison of the three encoders based on video content complexity for FHD resolution videos. It was observed that with increase in video content complexity the objective quality decreases, while the encoding time increases. Among the three codecs, X.265 was found to be superior to X.264 and VP9.

The authors in [79] performed a codec comparison for H.264, H.265, VP8 and VP9 for UHD, FHD and HD resolutions and reported the results in terms of three full reference objective quality metrics and BD-BR analysis. H.265/MPEG-HEVC and VP9 performance was found to be almost equal with both performing better than the earlier encoders. VP8 was found to be superior to the H.264/MPEG-AVC standard.

In 2013, Grois *et al.* did a performance comparison of H.264, H.265 and VP9 [80] and more recently a coding efficiency comparison of H.264, H.265 and AV1/VP9 encoders [81]. In both works, H.265 was found to be superior to H.264 and VP9 (AV1/VP9) with VP9 (AV1/VP9) performing worse than both H.264 and H.265 in terms of coding efficiency. In [82], they performed a comparison study of the three codecs using a low delay configuration. In terms of compression efficiency, the performance of H.265 was found superior to VP9 and H.264. In terms of encoding duration, VP9 was found faster than H.265 but much slower than H.264. Since the work in [82] was limited to objective evaluation, the authors in [83] performed a subjective evaluation of HEVC and VP9 assuming real-time applications using a low-delay configuration of the HM reference software. HEVC was found to be superior to VP9 and AVC.

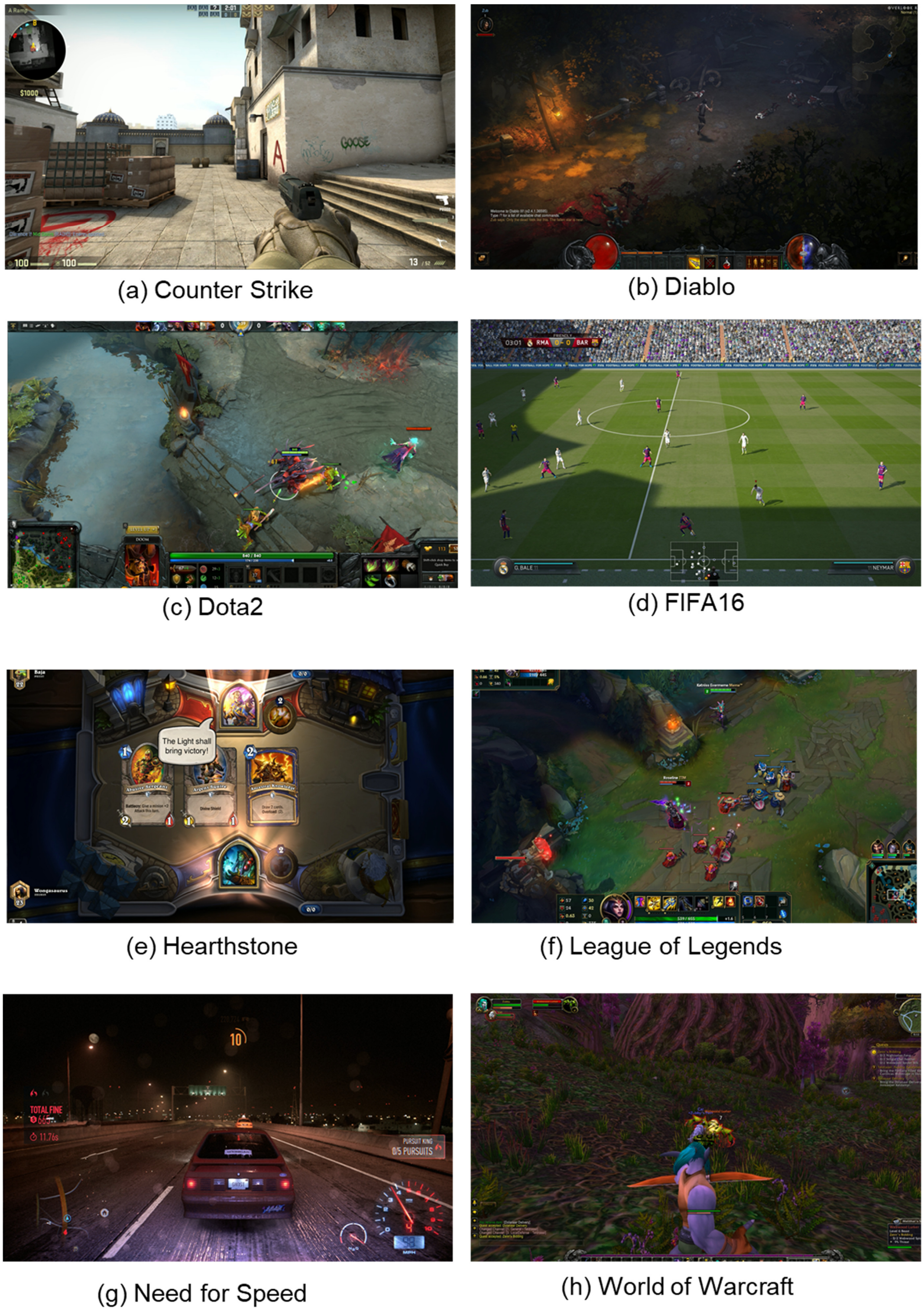
Most of the works discussed above were limited to non real-time applications such as VOD streaming with no constraint on the total encoding duration. Such studies, while indicative of the compression efficiency of the encoder for VOD scenarios, did not provide information on the relative performance of the three codecs for real-time scenarios. The adopted encoding settings, such as high encoding times (e.g., *placebo* preset which focus on maximizing the encoded video quality without any time constraint), 3-pass/2-pass VBR, etc., are not necessarily valid for live streaming applications. Few of these works ([82, 83]) evaluate the relative performance of the three codecs using low-delay configuration of the reference software. While the low-delay configuration emulates real-time encoding, the encoder as such cannot be used for real-time streaming applications (reference software are as such being developed to provide complete features and produce compliant bitstreams and not necessarily to optimize the encoding process which is left as an open question for the developers). Also, as discussed in [76], both of these works ([82, 83]) use a mix of reference software and practical encoders, which makes it unclear whether the assessment was of the compression efficiency achievable by the bitstream syntax or on what can be achieved by the codec implementations. In contrast, in this work we do not perform an assessment of the coding standards, but an evaluation of the performance of currently available practical encoders for H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 under real-time constraints such as fast encoding (*veryfast* preset, *realtime* deadline), strict Constant Bitrate (CBR) settings as used currently by live gaming videos streaming applications such as Twitch.tv.

3.3 Evaluation Methodology

3.3.1 Reference Videos

Figure 3.1 presents the screenshots of the gaming videos used in this work. The gaming videos used consist of some of the most popular games, namely Counter Strike: Global Offensive (CSGO), Diablo III (Diablo), Defense of the Ancients 2 (Dota2), FIFA (FIFA), Heathstone (HS), League of Legends (LoL), Need for Speed (NFS) and World of Warcraft (WoW). The games were played locally on a high end gaming desktop PC and were captured losslessly in RGB format at 1920x1080 resolution. The gaming videos cover a wide range of genres such as strategy, role play, racing etc. and are of 10 seconds duration each, with a framerate of 24 fps.

The complexity of encoding a video sequence and hence the compression difficulty depends on the content complexity which can be characterized by using Spatial Information (SI) and Temporal Information (TI) which define the amount of spatial details and the amount of temporal changes in a video sequence respectively. ITU-T Recommendation P.910 defines the SI value for a video sequence as the maximum value of individual SI values obtained for each frame (only luminance plane) at each time instant n [22]. Similarly, TI is defined as the maximum value of individual values obtained for the duration of the video sequence as the difference between the corresponding pixel values (of the luminance plane) of successive frames. Figure 3.2 shows the spatial and temporal content of the videos which confirms the fact that



(a) Counter Strike

(b) Diablo

(c) Dota2

(d) FIFA16

(e) Hearthstone

(f) League of Legends

(g) Need for Speed

(h) World of Warcraft

Figure 3.1: Sample videos used in this work.

the sample videos used in this work span a wide range of video complexity and are representative of most of the gaming videos streamed online.

3.3.2 Encoding Parameters

Table 3.1 describes the resolution-bitrate pairs used in this work. Since high definition (720p and 1080p) videos makes up almost 95% of the total live stream of Twitch.tv, we decided to include more bitrate values for these two resolutions compared to the lower resolutions (360p and 480p) [84]. The bitrate values were decided based on the recommended settings by Twitch.tv for the H.264 encoder [75]. Lower bitrate values were included considering latest encoders such as HEVC. For this work we used the FFmpeg library *libx264* which is the x264 H.264/MPEG-4 AVC encoder wrapper and *libx265* which is the x265 H.265/HEVC encoder wrapper. For VP9 we used the FFmpeg's *libvpx-vp9* library which is the open-source implementation of VP9 codec. FFmpeg is an easy to use free software capable of performing a wide range of multimedia operations including but not limited to record, convert and stream audio and video [85]. FFmpeg is widely used in applications such as YouTube, VLC Media Player, Blender among many others. In this work, FFmpeg version 2.8.10, built with gcc 5.4.0 was used for encoding the

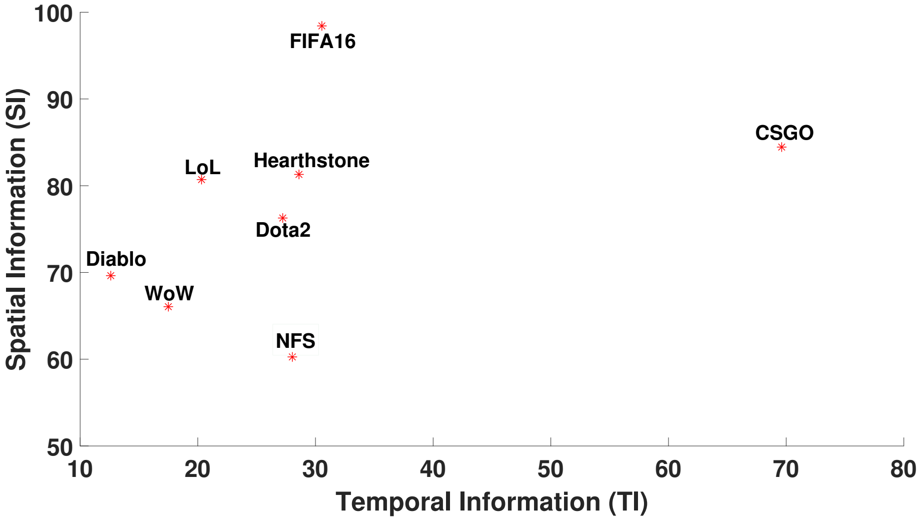


Figure 3.2: Spatial-Temporal plot for the videos used in this work.

Table 3.1: Resolution-Bitrate pairs for all three codecs.

Resolution	Bitrate (kbps)
1920x1080 (1080p)	1000, 1500, 2000, 3000, 4000
1280x720 (720p)	750, 1000, 1500, 1800, 2500
640x480 (480p)	400, 600, 900, 1200
480x360 (360p)	300, 400, 600, 800

videos. Encoding was performed on a Lenovo ThinkPad P50 laptop with 16 GB RAM, Intel Core *i7-6820HQ@2.70GHz x 8* and Nvidia Quadro 1000M running 64-bit Ubuntu 16.04 LTS *(4.4.0-59 generic)*. The encoder settings for the three encoders are described in Table 3.2. The settings are chosen to reflect real-time encoding and live streaming applications and are based on the general guidelines used for real-time streaming [74] and [75]. Comparison between the encoders is tricky, specifically between vp9 and x264/x265 as each have different parameter settings. While some other settings may have been included or made more exhaustive (e.g., *ultrafast* preset instead of *veryfast*), the selected settings, to the best of our knowledge, provide a balance between the encoding speed and quality for the application scenario under consideration and also lead to a fair comparison between the encoders.

Table 3.2: Encoder settings summary.

Encoder	Settings
libx264, libx265	preset=veryfast, profile=main, level=4.0
libvpx-vp9	deadline=realtime, quality=realtime
libx264, libx265 libvpx-vp9	single pass, buffer=bitrate closed gop=48(2s), CBR

3.4 Results and Discussions

In this section we discuss the results for the three codecs in terms of three FR objective quality metrics, BD-BR analysis and total encoding duration.

3.4.1 Objective Quality Assessment

We evaluate the performance of the metrics using three widely used objective quality metrics, PSNR, SSIM and VIFP. The three quality metric values are calculated using the Video Quality Measurement Tool [39]. Since the three quality metrics were originally designed for image quality assessment, for video quality assessment the individual score for each frame is calculated and the final score is reported as the average of the individual scores over all the frames of the video sequence. For the quality metric calculation, the videos encoded at lower resolutions were up-sampled to the source resolution (1080p) using *bilinear* interpolation.

Figure 3.3 compares the three codecs in terms of quality-bitrate curves for the three FR quality metrics and for four different resolutions for a sample video (CSGO). For all three quality metrics, a similar behavior can be observed among the three codecs for all four resolutions. x265 performs better than both x264 and vp9 for all resolutions. At a given resolution, for x265 vs. vp9, the difference in quality decreases with increase in the bitrate as also reported in [76]. Similar behavior is observed for x265 vs. x264 but with lower rate of decrease in quality gap with increasing bitrate at a given resolution.

Figure 3.4 presents the results of the quality-bitrate curves for the three FR quality metrics for all four resolutions considering the average of the quality scores of all the eight video sequences. It can be observed that, while for all resolutions and all quality metrics the performance of the x265 encoder is better than that of x264 and vp9, the performance of vp9 and x264 is comparatively better or worse depending on the resolution and choice of the quality metrics. Similar to the discussion in the previous paragraph, we observe that the difference in performance between vp9 and h265 decreases at higher bitrates. For a better understanding of the performance of the three encoders for each resolution and gaming video, we present next the BD-BR analysis of the three codecs.

Note: The actual bitrate of the encoded videos obtained with the three encoders is slightly different from the specified bitrate. For simplicity we report the nominal specified bitrates in Figure 3.3 and Figure 3.4; however, for BD-BR calculations the actual bitrate of the encoded videos is used.

3.4.2 BD-BR Analysis of the Codecs

BD-BR calculations are used to compute the average gain in PSNR or the average percent savings in bitrate between two Rate-Distortion (RD) curves [86]. Table 3.3 summarizes the percentage bitrate savings for PSNR for the RD curves for three combinations of the codecs for all videos and resolutions. A negative value represents a bitrate savings for the first codec compared to the second one and vice-versa.

x265 vs. x264

For all videos and for all resolutions, x265 outperforms x264 in terms of bitrate savings. The amount of savings depends on the resolution and also on the type of gaming video. For all the eight videos considered in this work, compared to x264, x265 results in an average bitrate savings of 20.64% at the same quality level. On an average, x265 results in higher bitrate savings at higher resolutions compared to the bitrate savings for same videos encoded at lower resolutions. This observation is in line with the observations reported in earlier works (e.g. in [76] considering the VOD scenario).

Table 3.3: BD-BR Results (PSNR-based) for all videos.

	x265 vs. x264				**vp9 vs. x264**				**x265 vs. vp9**			
Sequence/ Resolution	**1080p**	**720p**	**480p**	**360p**	**1080p**	**720p**	**480p**	**360p**	**1080p**	**720p**	**480p**	**360p**
CSGO	-30.63	-20.75	-12.02	-10.66	-13.31	-6.40	-0.10	-1.27	-18.33	-14.24	-11.67	-9.18
Diablo	-36.21	-31.67	-21.36	-21.70	-14.22	-10.18	-1.76	1.49	-25.90	-23.48	-20.41	-23.38
DOTA2	-16.69	-21.38	-15.40	-13.50	**21.95**	**19.08**	**19.64**	**19.78**	-31.56	-33.49	-29.57	-28.02
FIFA16	-43.11	-40.47	-27.95	-19.26	-9.06	-10.02	-0.39	**8.76**	-36.58	-28.06	-23.96	-23.45
Hearthstone	-33.54	-33.55	-21.64	-21.94	**14.97**	**9.35**	**17.82**	**18.42**	-42.17	-39.28	-33.36	-34.22
LOL	-8.29	-11.07	-0.82	2.60	**35.01**	**24.23**	**33.69**	**38.85**	-32.87	-27.79	-25.82	-25.90
NFS	-30.90	-23.39	-16.78	-12.99	-9.22	-5.33	**0.74**	**0.20**	-20.04	-16.75	-16.70	-12.42
WoW	-14.17	-23.67	-15.92	-11.70	**28.86**	**25.16**	**33.58**	**35.65**	-34.94	-39.71	-37.82	-35.26
Average	**-26.69**	**-25.74**	**-16.48**	**-13.64**	**6.87**	**5.73**	**12.90**	**15.23**	**-30.29**	**-27.85**	**-24.91**	**-23.97**
Total Average	**-20.64**				**10.18**				**-26.76**			

x265 vs. vp9

Similar to the trend observed in x265 vs. x264, x265 outperforms vp9 in all cases with an average bitrate savings of approximately 26.76%. This observation is also in line with many earlier works discussed in Section 3.2.

vp9 vs. x264

The bitrate savings comparison for x264 vs. vp9 provides interesting insights into the coding efficiency of vp9 and x264 for the considered live gaming video streaming scenario. As highlighted in red color in Table 3.3, for almost half of the videos (average complexity, medium SI and TI), x264 outperforms vp9. The savings with x264 are higher at lower resolutions as compared to those at higher resolutions. On an average, considering all videos and all resolutions, x264 outperforms vp9 by approximately 10.18%. Results of superior performance in terms of coding efficiency of x.264 compared to vp9 have been also reported in earlier works by [77, 80, 81].

As expected, the performance of x265 is superior compared to both x264 and vp9. The difference in bitrate savings of x264 against vp9 was found to be highly dependent on the content type, encoding bitrate and resolution. For low and medium complexity videos, at lower resolutions, x264 outperforms vp9.

Encoding Duration

Table 3.4 describes the total encoding run time of the sample videos for 1080p resolution performed on a Lenovo ThinkPad P50 laptop with 16 GB RAM, Intel Core *i7-6820HQ@2.70GHz x 8* and Nvidia Quadro 1000M running 64-bit Ubuntu 16.04 LTS *(4.4.0-59 generic)*. A similar behavior (decreasing encoding times with decrease in resolution) is observed for other resolution encodes. Since the total encoding duration depends on the machine and even varies for the same encoding process, the encoding was repeated twice for all bitrates for all videos and the

Table 3.4: Encoding Run Times (in seconds) for 1080p resolution videos.

	x264					**x265**					**vp9**				
Sequence/ Bitrate (kbps)	**1000**	**1500**	**2000**	**3000**	**4000**	**1000**	**1500**	**2000**	**3000**	**4000**	**1000**	**1500**	**2000**	**3000**	**4000**
CSGO	2.65	2.75	2.83	2.96	3.04	8.43	9.09	9.59	10.92	11.47	12.72	14.39	15.51	17.09	18.47
Diablo	3.05	3.05	3.20	3.36	3.33	6.95	7.60	7.18	8.82	9.46	9.85	10.48	11.38	12.64	13.34
DOTA2	4.09	4.14	4.19	4.11	4.37	6.12	6.53	6.68	7.50	7.87	9.22	10.19	10.79	11.41	12.21
FIFA16	2.90	3.03	3.18	3.32	3.09	5.76	6.26	6.42	7.20	7.98	10.49	10.99	11.43	11.63	12.35
Hearthstone	2.86	2.88	2.95	2.72	2.91	5.23	5.10	5.73	5.8	6.05	6.97	8.02	7.97	8.83	9.21
LOL	1.79	1.84	1.89	1.91	1.93	5.15	5.25	5.63	6.11	6.38	8.41	8.96	9.09	9.69	9.97
NFS	2.74	2.86	2.90	3.04	3.16	9.07	10.37	10.83	11.99	12.63	11.06	12.67	13.48	15.34	16.85
WoW	1.94	2.05	2.14	2.15	2.24	5.72	5.92	6.24	6.49	6.96	9.83	10.48	10.83	11.60	11.71
Average	**2.75**	**2.82**	**2.91**	**2.95**	**3.01**	**6.56**	**7.02**	**7.29**	**8.10**	**8.60**	**9.82**	**10.77**	**11.31**	**12.28**	**13.02**
Total Average	**2.89**					**7.51**					**11.44**				

average values are reported. As expected, the encoding times for x265 and vp9 encoders are very high compared to that of x264 encoder. On an average, x264 is approximately 2.6 and 4 times faster than x265 and vp9 encoders respectively. Similarly, compared to vp9, x265 is on an average 1.52 times faster.

3.5 Conclusion

In this work we performed a performance comparison of H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 video encoders using gaming videos for live streaming applications. Compared to VOD streaming scenarios which can benefit from more optimized (e.g. 3-pass/2-pass) and longer encoding durations (e.g. slow/placebo preset), live streaming scenarios have more constraints and hence have stricter and limited set of encoding choices (e.g. superfast/ultrafast preset settings, 1 pass, constant bitrate). As observed in previous studies for VOD streaming scenarios such as in [76], also for live streaming scenarios, in terms of bitrate gains at fixed quality, H.265/MPEG-HEVC outperforms H.264/MPEG-AVC and VP9. Also, for the settings and encoders used in this work, H.264/MPEG-AVC on an average offers superior performance gains compared to VP9.

The encoding duration for the three encoders was analyzed. While during "real" live streaming applications the actual encoding duration may vary with respect to the one obtained with our simulation, we believe that the orders of magnitude of the average encoding duration are still well representing the actual duration. With respect to encoding duration, the x264 encoder was found to be the fastest and vp9 the slowest. The BD-BR savings for vp9 compared to x264 at the cost of increased encoding duration makes vp9 less attractive for live gaming video streaming applications. In contrast, the BD-BR savings for x265 compared to that of x264 is almost 20% but at the cost of increased encoding duration. Such high encoding duration raises questions about the applicability of such encoders for applications such as live gaming video streaming where increased delays may result in huge broadcast latency issues resulting in decreased end user QoE [87]. While the present H.265/MPEG-HEVC practical encoders can still be used for obtaining different representations at different bitrates for on-demand streaming, for real-time encoding and streaming applications such as gaming videos streaming, there is a need for development of faster and more optimized practical encoders.

Based on the review of existing works and results obtained from this work, we conclude that, in general, the actual performance of the video codecs is highly dependent on a lot of factors such as the encoders (reference/practical), encoding duration constraint (real-time/offline), content type (natural/animation/synthetic videos), video complexity (low/medium/high SI and TI), video resolution (UHD/FHD/HD/SD), encoding settings (GOP size, single or multiple pass, VBR/CBR) and quality evaluation metric (objective/subjective).

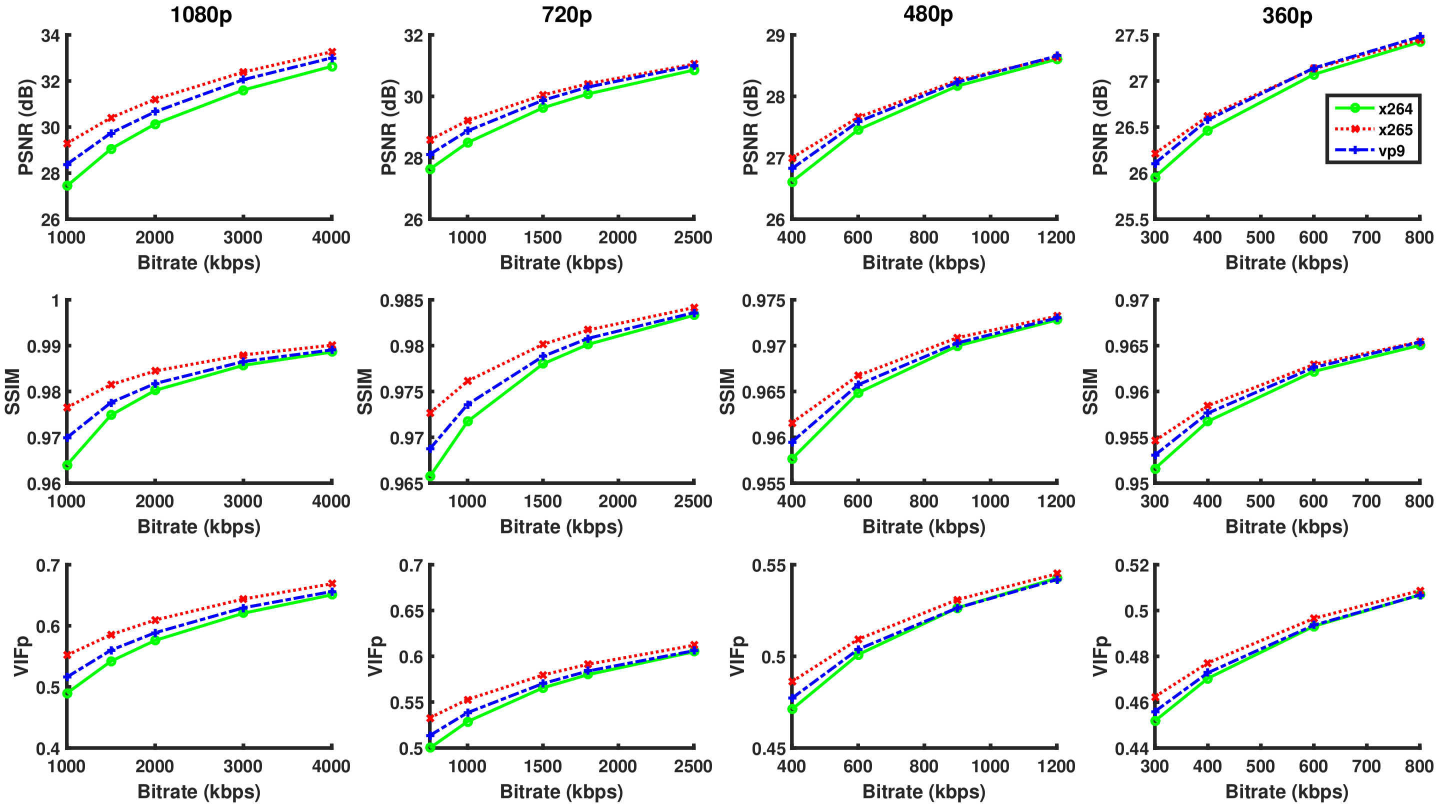


Figure 3.3: Objective quality vs. Bitrate for three codecs for one of the sample videos (CSGO).

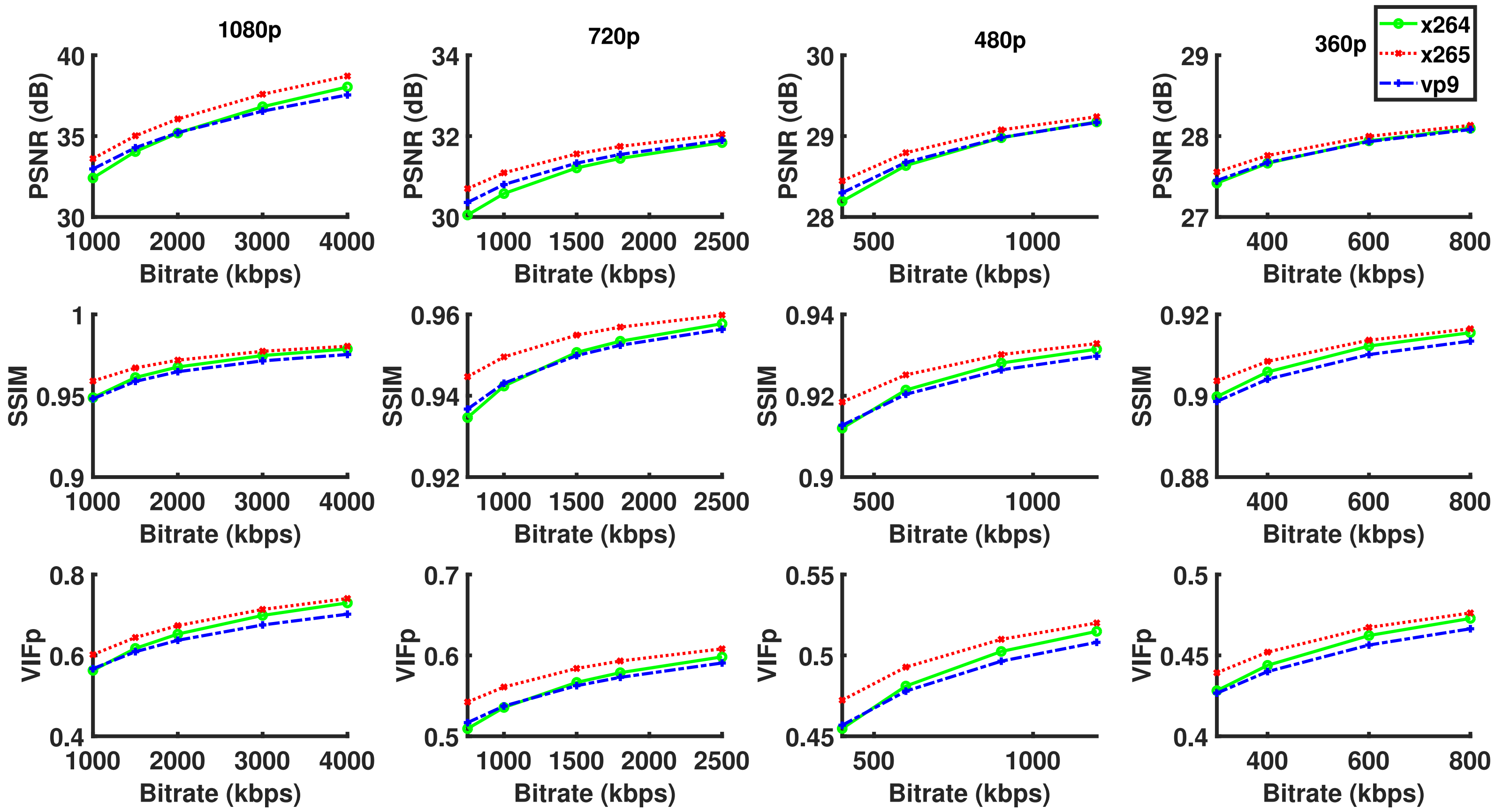


Figure 3.4: Objective quality vs. Bitrate for three codecs considering the average value of the respective score of the eight video sequences.

Untersuchen was ist, und nicht was behagt

Investigate what is, and not what pleases.

— JDer Versuch als Vermittler von Objekt und Subjekt (The Attempt as Mediator of Object and Subject) (1792)

4
A Comparative Quality Assessment Study for Gaming and Non-Gaming Videos

Contents

The work presented in this chapter has been published as (with minor modifications): N. Barman, S. Zadtootaghaj, M. G. Martini, S. Möller, and S. Lee, “A Comparative Quality Assessment Study for Gaming and Non-Gaming Videos,” in *Tenth International Conference on Quality of Multimedia Experience (QoMEX 2018)*, Cagliari, Italy 2018.

4.1 Introduction

Video gaming is one of the fastest growing fields in multimedia as evident with the recent rise of services such as Twitch.tv and YouTubeGaming, which provide videos (Live and VOD) of users playing games. We refer to videos being streamed on such sites which show the footage of user gameplay as *gaming videos* while by *non-gaming videos*, we refer to the commonly streamed videos in applications such as YouTube and Netflix or on broadcast TV (anything other than the gaming videos is considered here as non-gaming videos ranging from different genres such as Music, movies, drama, animation, etc.). Providing adequate quality for such services is essential and also requires studying the usability of existing objective metrics and models for monitoring the quality of gaming videos at the end-user side.

So far, most of the works related to objective and subjective quality analysis have been limited to traditional non-gaming videos and more importantly to VOD scenarios. Gaming videos, unlike non-gaming videos, consist of artificial and graphical components. Our earlier work on codec compression performance presented in Chapter 3) indicated that the performance of the codecs, especially when comparing VP9 and H.264, depends to a great extent on the type of the game, a behavior absent in similar studies considering non-gaming videos. This indicates a possible difference between gaming and non-gaming videos. Towards this end, we perform a comparative objective and subjective quality assessment study for gaming and non-gaming videos to investigate if there exists any significant difference between the gaming and non-gaming videos which can the be exploited further to improve the QoE of the end user.

The rest of this chapter is organized as follows. Section 4.2 describes the source videos and the encoding settings used in this work. Section 4.3 presents a content analysis of the reference videos in terms of spatial and temporal complexity. In Section 4.4 we report the objective analysis of the videos and in Section 4.5 the subjective analysis for a subset of those videos. We finally conclude in Section 4.6 with a summary of the key observations and possible future work.

4.2 Test Methodology

4.2.1 Source Videos

The source videos used in this work are selected based on diversity of genre, variety in degree of content complexity (both spatial and temporal) and for games, in addition, we considered the popularity of the game on streaming services such as

Twitch. Given that very few recorded gaming videos are available in raw format, we decided to capture ourselves some of the most popular games such as CSGO, FIFA, LoL, HS, Dota2, NFS, Diablo and WoW. The games were captured losslessly at 24 fps. Other gaming and non-gaming videos used in this work are obtained from [88]. In this work we used a total of 30 videos, at 1920x1080p, YUV format, consisting of 15 non-gaming videos and 15 gaming videos spanning a wide range of genres (non-gaming videos: animation, drama, etc.; gaming videos: First Person Shooter (FPS), racing, role-play, etc.) and are representative of most of the content types currently streamed over the Internet. Figure 4.1 shows the screenshots of some of the sample gaming and non-gaming videos.

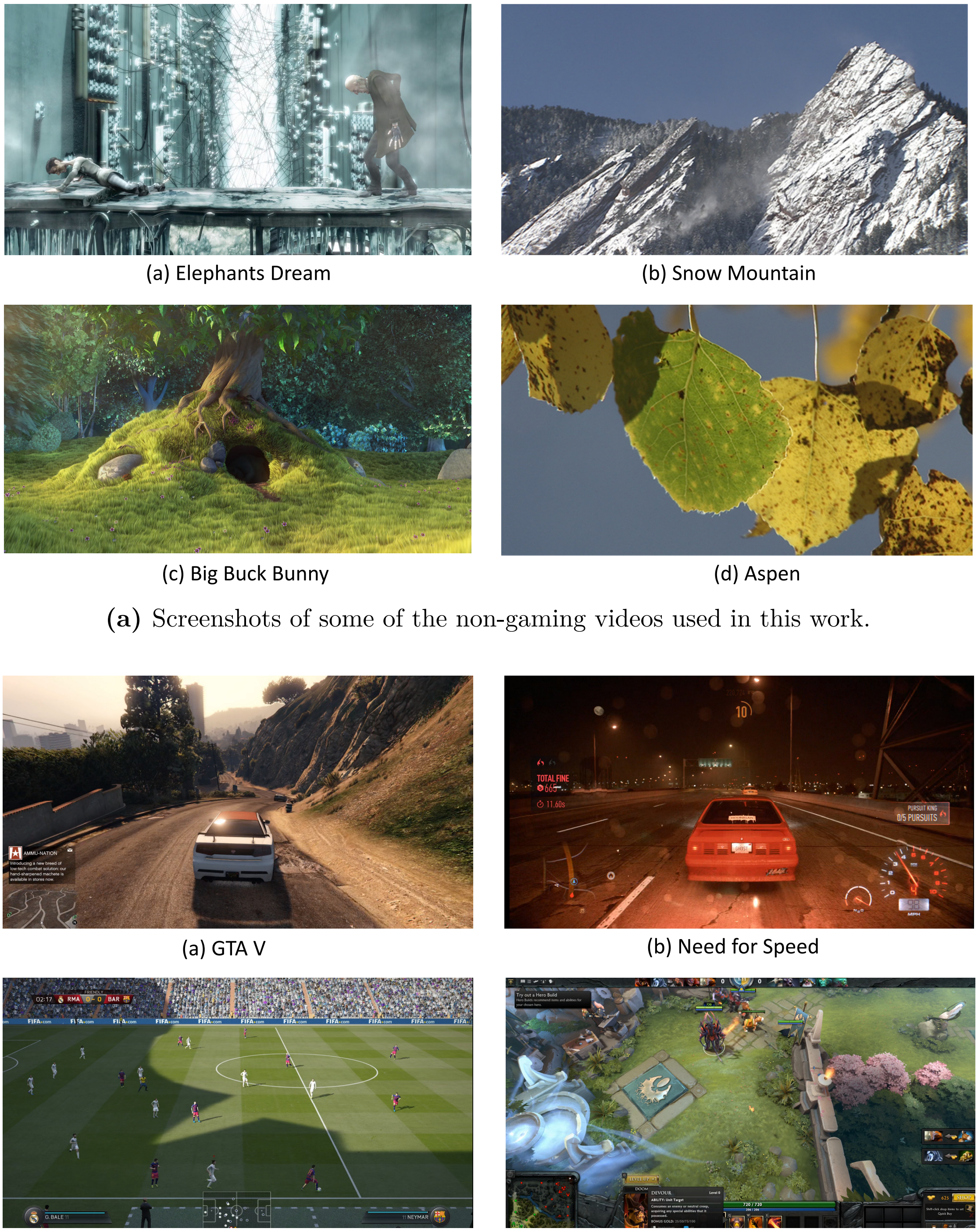
4.2.2 Encoding Parameters

The videos were encoded via HEVC at the native resolution of 1080p at 19 different Constant Rate Factor (CRF) values (from 24 to 42, with lower CRF values implying better quality videos) to obtain a total of 570 encoded videos. CRF based encoding results in videos of different bitrates (file sizes) but of constant quality. CRF is not necessarily the preferred mode for encoding videos for streaming over the Internet, especially for streaming gaming videos[1]. However, the focus of this work is to study the difference, if any, between gaming and non-gaming videos and not necessarily consider the aspects of streaming. Hence, the videos are encoded using the CRF mode, which takes into account content complexity, to make sure that the videos are of almost constant quality. This, for our objective, will lead to a fair comparison and also compensate for the fact that the compared videos from the two categories do not necessarily have the same complexity. For encoding the videos, we used the *libx265* implementation of the HEVC standard in FFmpeg [85]. Table 4.1 summarizes the encoding parameters and settings used in this work.

4.3 Spatial and Temporal Complexity Comparison

We start the comparison study with the analysis of content complexity of the source videos in terms of SI and TI values. As per ITU-T Recommendation P.910 [22], spatial complexity is calculated in terms of SI for each frame (only luminance plane) at each time instant n by first filtering the frame using Sobel filter and then calculating the standard deviation over the pixels in each Sobel filtered frame in

[1] Gaming videos are usually encoded at CBR in real-time (either at the client or server).



(a) Elephants Dream

(b) Snow Mountain

(c) Big Buck Bunny

(d) Aspen

(a) Screenshots of some of the non-gaming videos used in this work.

(a) GTA V

(b) Need for Speed

(c) FIFA

(d) Dota 2

(b) Screenshots of some of the gaming videos used in this work.

Figure 4.1: Some of the sample non-gaming and gaming videos used in this work.

both vertical and horizontal direction [22]. The final score for the whole video sequence is the maximum value for the whole video sequence defined as:

$$SI = max_{time}(std_{space}[Sobel(F_n)]). \tag{4.1}$$

Table 4.1: Summary of video encoding parameters.

Parameter	**Value**
Duration	10 sec
Resolution	1080p
Frame Rate	24
Number of Reference Videos	30
Total Quality Levels	19
Number of Encoded Videos	570
Encoder	FFmpeg
Encoding Mode	CRF
Video Compression Standard	HEVC
Objective Quality Metrics	PSNR, SSIM [24], VIFP [25], VMAF [40]

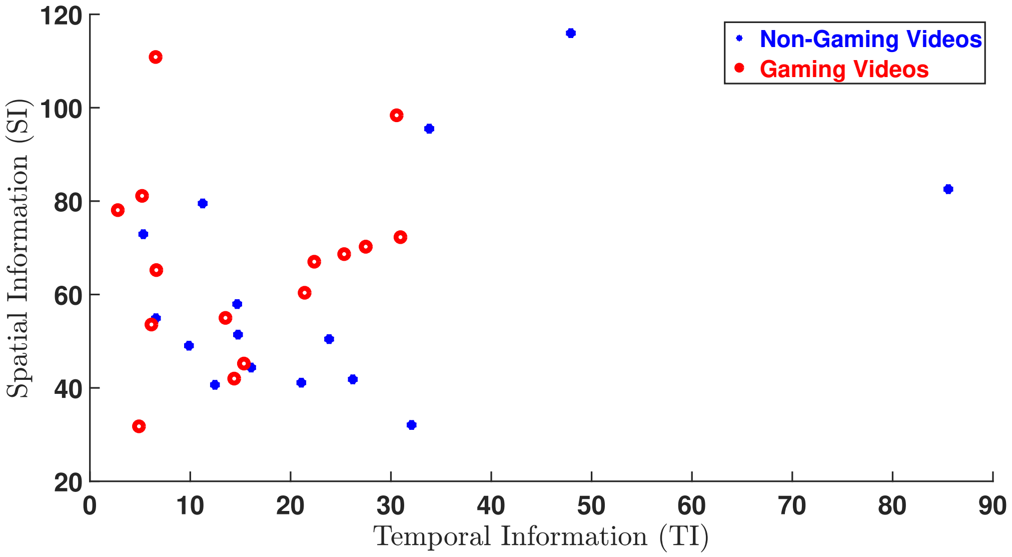


Figure 4.2: SI vs. TI plot of the videos used in this work.

Similarly, temporal complexity is quantified in terms of TI which is a measure of the amount of temporal changes in the video, with higher values indicating higher motion content. It is equal to the maximum over time of the standard deviation over space of the difference ($M_n(i, j)$) between corresponding pixel values in two adjacent frames (luminance component) defined as:

$$TI = max_{time}(std_{space}[M_n(i, j)]). \tag{4.2}$$

Figure 4.2 shows the plot of the spatial and temporal complexity of the videos used in terms of SI and TI respectively. Based on the figure, we can observe that the videos from both categories span over a wide range of different SI and TI values. Considering the maximum value as the representative of the content complexity is not always realistic, as also reported in other studies (such as [89]).

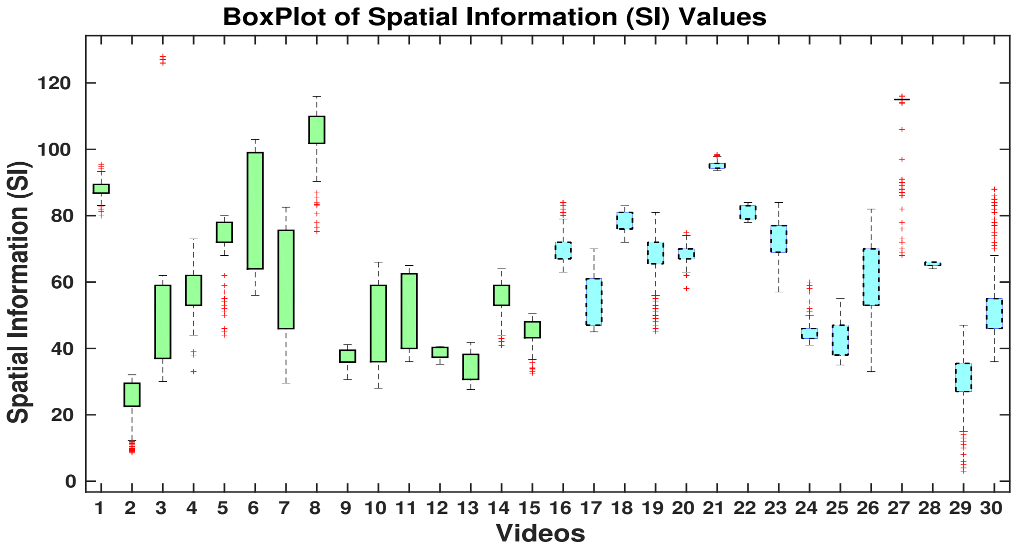


Figure 4.3: Box Plot for SI values of the 30 reference videos used in this work (Videos 1-15, Green: non-gaming videos; Videos 16-30, cyan, dashed boxes: gaming videos).

Hence, we performed detailed statistical analysis on the individual frame-level SI and TI values as discussed next.

4.3.1 Results for Spatial Complexity

Figure 4.3 shows the box plot considering all the individual frame-level SI values for all the 30 reference videos. The first 15 videos in green (number 1-15) correspond to the non-gaming videos while the last 15 videos in cyan, dotted boxes (number 16-30) correspond to gaming videos. It can be observed that, in general, SI for gaming videos has less variance when compared to non-gaming videos. One possible explanation could be the way games are designed. Since a game has almost the same level of abstraction, the spatial complexity of video games has less variation compared to non-gaming videos. Even complex games with a high level of abstraction such as CSGO (16), FIFA (21) and GTA (22) have much less variance[2].

4.3.2 Results for Temporal Complexity

Similarly to Figure 4.3, Figure 4.4 shows the box plot considering individual frame-level TI values. It can be observed that, unlike SI values, considering TI values, there does not exist any major difference between the non-gaming and gaming videos. In general, in contrast to SI, considering TI values, gaming videos show slightly higher variance when compared to non-gaming videos.

[2] The numbers in parenthesis indicates the corresponding game in the Figure

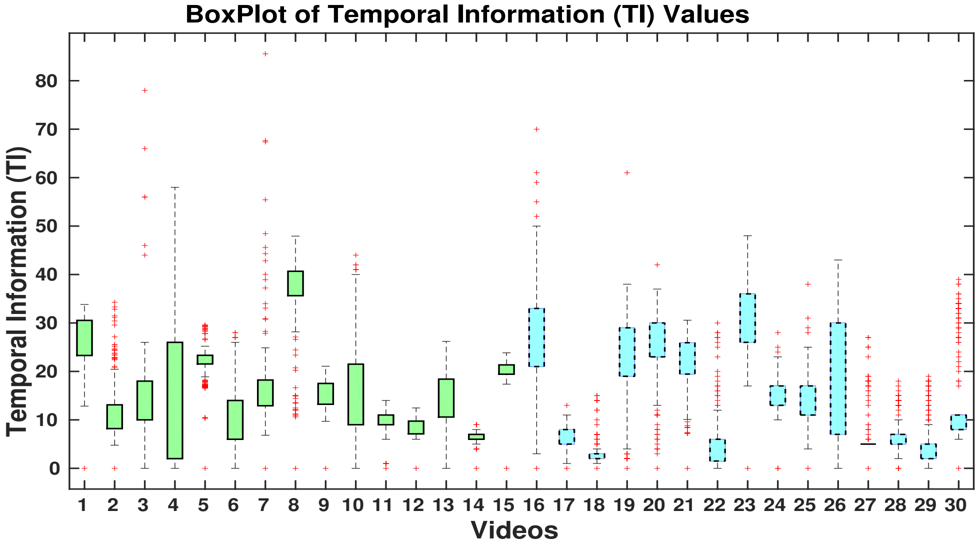


Figure 4.4: Box Plot for TI values of the 30 reference videos used in this work (Videos 1-15, Green: non-gaming videos; Videos 16-30, cyan, dashed boxes: gaming videos).

4.4 Objective Assessment

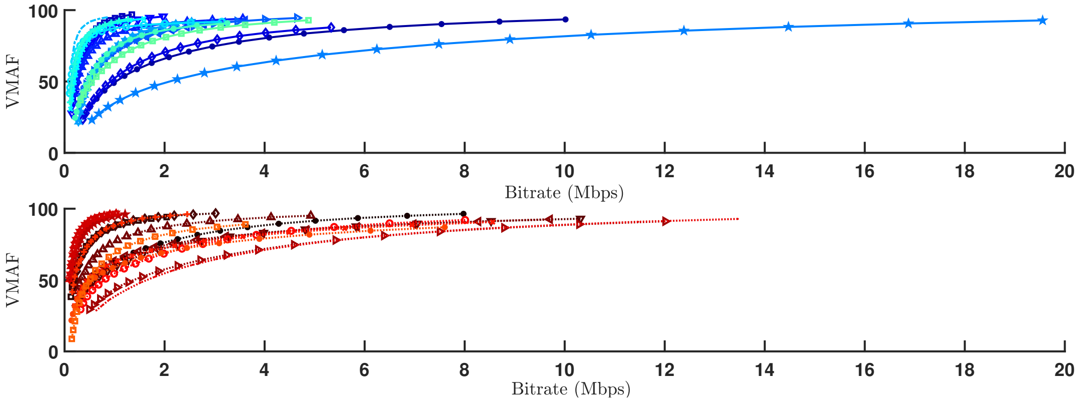


Figure 4.5: Quality (VMAF)-Bitrate curves for the 15 non-gaming videos (top) and the 15 gaming videos (bottom) encoded at 19 different CRF values.

The quality of the encoded videos is estimated using four objective metrics: PSNR, SSIM, VIFP and VMAF. Figure 4.5 shows the VMAF vs. bitrate curves of the non-gaming and gaming videos, encoded at 19 different quality levels. Similar curves are also obtained for the other quality metrics, PSNR, SSIM and VIFP but are not presented here due to lack of space. Since CRF encoding takes into account the content complexity, the bitrate values for the same CRF value may vary with the video. Based on the figure, it can be observed that there does not exist any major difference between non-gaming and gaming videos, with videos from both groups spanning the quality-bitrate curve. A statistical test based analysis for similarity/dissimilarity between the two groups cannot be used here as the data (videos) from both groups come from totally different distributions (the videos from different groups have totally different characteristics and hence cannot be compared with one another).

4.5 Subjective Assessment

To evaluate the quality of the videos as perceived by the end user, we performed subjective tests using a total of 12 videos (six each from gaming and non-gaming group from different genres), out of the total 30 reference videos used in this work. Out of the six videos from each category, two videos each correspond approximately to low, medium and high content complexity (in terms of SI and TI values). To keep the test duration short and avoid test subject exhaustion, we selected only five quality levels (CRF= 26, 30, 34, 38 and 42) out of the 19 quality levels used for objective assessment, each corresponding roughly to five quality levels on the ACR scale.

4.5.1 Test Setting and Environment

The subjective test was carried out in the Multidimensional Insight Lab, Yonsei University, South Korea, conforming to the controlled test environment settings as recommended in ITU-R Rec. BT.500 [23]. There were a total of 15 participants (14 male, one female) with an average age of approx. 28 years. All the test subjects were tested for proper eyesight and color blindness. There was a small training session using six gaming video sequences from games other than the ones used in the actual subjective test to help the test subjects get used to the test software interface and the objective of the subjective test. The test sequences were played on a FHD (1920x1080) 22″ LG Monitor using a different randomized playlist for each subject. The test evaluation methodology used was ACR on a scale of 1-5, five corresponding to the best quality.

4.5.2 Subjective Test Results

The results of the subjective tests in terms of MOS vs. bitrate did not show major differences between the non-gaming and gaming videos. We observed however that, on an average, gaming videos, at the highest CRF level (=26), achieved lower MOS values compared to non-gaming videos. A possible explanation for such behavior can be attributed to subject bias which is discussed later in Section 4.5.4.

Figure 4.6 shows MOS vs. objective metric scores for the four different objective metrics for the 12 video sequences. We can observe that, while the VMAF metric appears to estimate well the quality of both gaming and non-gaming videos, the SSIM metric provides a good estimation of the quality of non-gaming videos, but not of gaming videos.

4.5.3 Subjective and Objective Scores Correlation

To quantify the performance of the four objective metrics in relation to the subjective scores, we calculated the correlation among the four objective metrics and the MOS scores, separately for the gaming and non-gaming videos. As discussed in Chapter 2.5.2, in the field of image and video streaming, PLCC values can be used to estimate the prediction accuracy of a model between the predicted (objective metric values) and actual subjective scores (MOS scores). Similarly, SROCC values can be used to evaluate the prediction monotonicity of a model between the predicted (objective values) and actual (MOS) subjective scores. Table 4.2 presents the PLCC and SROCC values for all the four objective metrics. Based on the results, the following observations can be drawn:

- For both non-gaming and gaming videos, and in terms of both PLCC and SROCC, VMAF outperforms traditional objective metrics such as PSNR, SSIM and VIFP.
- For a given objective metric, the PLCC and SROCC values are less for the gaming videos when compared to that of non-gaming videos. One of the possible reasons for the same could be subject bias, as discussed in Section 4.5.4.
- In terms of PLCC values, out of all four objective metrics, SSIM performs the worst for gaming videos compared to non-gaming videos.
- In terms of SROCC values, while PSNR exhibits the worst performance for non-gaming videos, SSIM has the worst performance for gaming videos.

Based on the observations above, it can be concluded that unlike non-gaming videos, for gaming videos, PSNR performs better than SSIM. One of the possible reasons behind the low performance of the SSIM metric as compared to PSNR could be due to the contrast and luminance masking measurements used by SSIM for the quality estimation. Some games such as Dota 2, LoL, etc. might have many areas of the video which are dark due to the inherent nature of the game design.

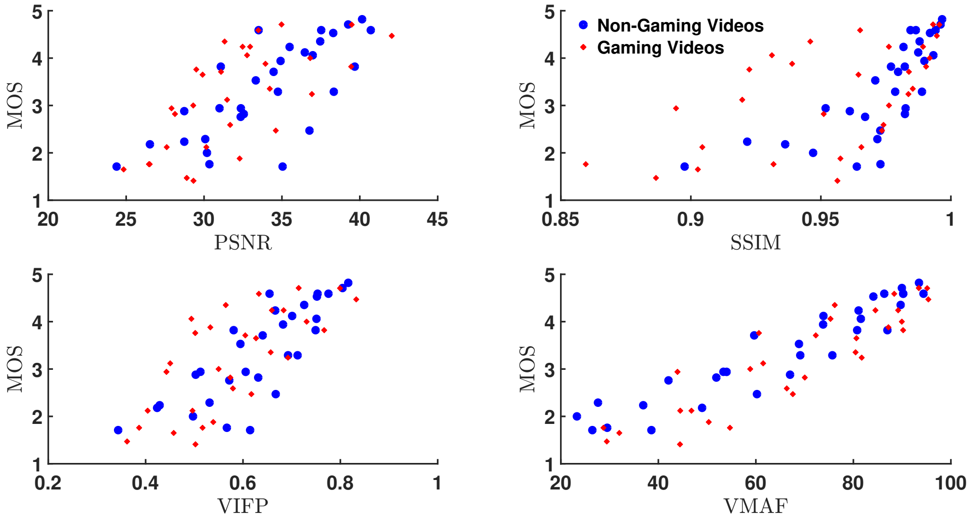


Figure 4.6: MOS vs. objective metrics for the twelve videos.

Table 4.2: PLCC and SROCC values of four objective metric scores with respect to MOS values

Objective Metrics	PLCC		SROCC	
	Non-gaming Videos	Gaming Videos	Non-gaming Videos	Gaming Videos
PSNR	0.7779	0.6980	0.7096	0.7153
SSIM	0.8189	0.5150	0.8936	0.5231
VIFP	0.8591	0.6693	0.8765	0.6567
VMAF	**0.9270**	**0.8810**	**0.9422**	**0.8928**

Hence, contrast and luminance masking might result in lower scores while the video itself might be of good quality. A more detailed investigation is required in this direction for more conclusive results and observations. It is observed that the newly proposed video quality metric VMAF performs much better than traditionally used image-quality based metrics such as PSNR and SSIM. This advocates for more work in the direction of fusion-based metrics which can lead to the design of much better objective metrics.

4.5.4 Discussion

Performance of the objective quality metrics

As discussed in the previous section, it is observed that two of the widely used image quality metrics for video quality assessment, PSNR and SSIM, do not perform well in terms of correlation with subjective scores for both gaming and non-gaming videos as compared to the newer quality metric VMAF. The performance is even worse for the gaming videos. To investigate the reason behind such behavior, we use Partial PSNR (PPSNR) to estimate the quality variation over the frames of the encoded video sequences. PPSNR is measured by calculating the PSNR at 8x8 block level as:

$$PPSNR_{i,j} = 10log_{10}(\frac{R^2}{MSE_{i,j}}) \tag{4.3}$$

where i and j denote the position of a certain block in the frame, R is the maximum pixel intensity value (=255 for 8 bpp frame) and MSE is the mean square error between the reference and distorted block pixels.

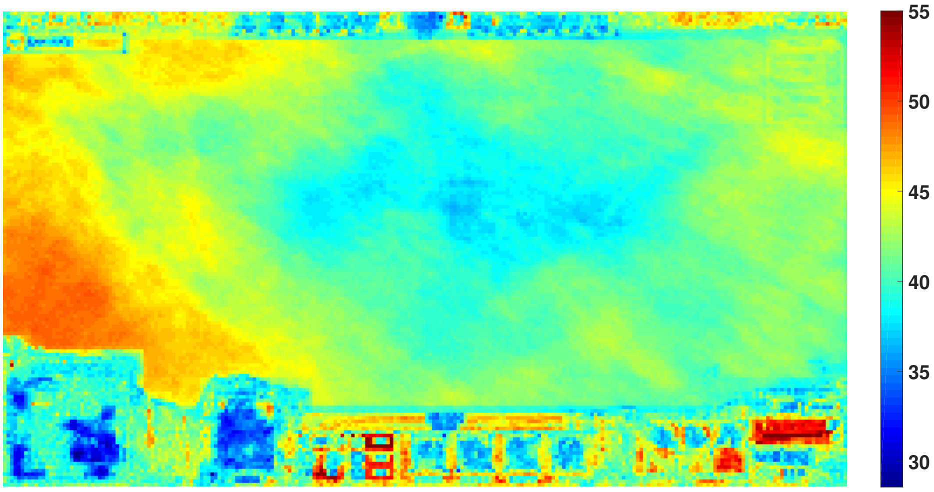


Figure 4.7: The heatmap of the PPSNR values (dB) for a sample gaming video sequence, Dota2.

Fig 4.7 shows the heatmap of the PPSNR averaged over all the frames of the sequence for the encoded video sequence "Dota 2". It can be observed that there are some regions (the bottom of the scene as well as the top middle side of the scene) in the game where no changes appear during the gameplay (commonly referred to as feedback elements). Feedback elements communicate the detail about the game's inner states to the player, such as the status of a particular resource. This information usually stays constant over several frames, hence the local quality in terms of PPSNR is higher since temporal prediction is accurate. On the other side, the areas with high motion (mainly at the center of each frame) have a lower local quality, while these are also the areas attracting higher subject's attention.

The first three objective quality metrics do not take such information (constant regions) into account, neither the fact that participants usually pay attention to the center part of a video game [9], and merely average the individual frame-level quality scores to obtain the final score, resulting in lower correlation with the subjective rating. However, VMAF takes into account motion and hence results in much better quality prediction. Similar results are obtained for other gaming videos but are not presented here due to lack of space. This indicates that ROI based encoding strategies can result in very high quality encoded videos for gaming services, as also discussed in [9] for gaming applications and in [69] for medical video streaming applications.

Subject Bias

While in the previous section we discussed the possible reason for the bad performance of the older metrics such as PSNR and SSIM when compared to newer metrics such as VMAF, here we investigate the possible reasons for the lower correlation scores for gaming videos compared to that of non-gaming videos. As briefly mentioned previously in Section 4.5.2 and Section 4.5.3, one of the possible reasons for lower ratings and hence lower values of correlation of subjective scores with objective scores for gaming videos compared to non-gaming videos could be subject bias. In general, since several users do not watch gaming videos on a regular basis, they may not be used to the graphical and unrealistic content of gaming videos. To investigate the possible subject bias, we asked the test participants to fill in a questionnaire. In addition to their age and other video watching habits, the subjects were asked primarily three questions regarding their habits: whether they watch non-gaming videos, gaming videos and if they play games on a regular basis. We present the data obtained, using set theory, in Figure 4.8. It can be observed that while almost all of the test participants are used to playing games and watch non-gaming videos streamed on the Internet, only a few of them (one-fifth), actually watch gaming videos streamed on the Internet. It is important to note here that though the majority of the test participants play games, they are mostly limited to a very specific genre and hence this does not necessarily indicate familiarity with all other game types considered in this study.

Since most of the subjects are not used to watching gaming videos, they may be reluctant to give such “artificial” (graphical and unrealistic) videos, higher ratings. Also, in the absence of prior experience of watching gaming videos, not all users have an idea about the source quality of such games. For example, games like Dota 2 or League of Legends may naturally look dull, and hence people who are not used to watching and/or playing such games, may rate them lower, even if the actual quality is high. Based on these results, we can conclude that for subjective

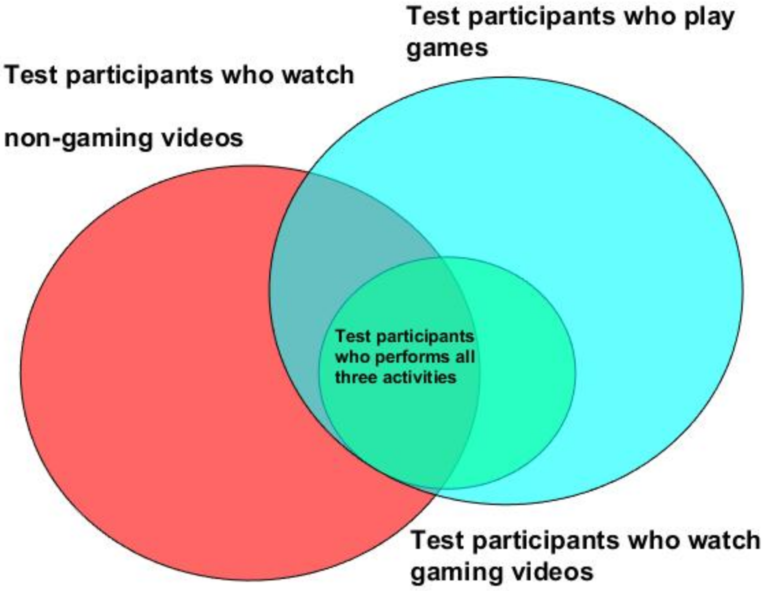

Test participants who play games
Test participants who watch non-gaming videos
Test participants who performs all three activities
Test participants who watch gaming videos

Figure 4.8: Venn Diagram illustrating the percentage of test participants interested in the three activities.

quality assessment of gaming videos, other test methodologies, such as with hidden reference (e.g., ACR with hidden reference) or paired comparison (e.g., DSCQS or DSIS), would be more appropriate. Well designed training sessions may also help in mitigating such subject bias.

4.6 Conclusion

In this work we performed a comparative study of non-gaming and gaming videos using objective and subjective measurements. It was found that gaming videos are in many aspects similar to non-gaming videos. When considering some characteristics, there exist certain differences between gaming and non-gaming videos. In terms of spatial complexity, it was found that gaming videos, in general, have less variance and lower average SI values compared to that of non-gaming videos. No particular difference was observed for TI values. Among the four quality metrics used in this work, VMAF results in best PLCC and SROCC values for both non-gaming and gaming videos. It is also interesting to note that, for gaming videos, SSIM performs worse than PSNR. In terms of the correlation between objective and subjective scores, it was found that, in general, non-gaming videos achieved a higher correlation score (both PLCC and SROCC) when compared to that of non-gaming videos. Our initial investigation in this direction using a questionnaire-based study suggests a possible subject bias. The hypothesis that lower subjective ratings for gaming videos vs. non-gaming videos can be due to the difference in the complexity of the videos can be ruled out due to the choice of CRF as the encoding setting (as CRF mode of encoding takes into account the video complexity so as to achieve a consistent quality). Hence, we propose possible usage of subjective test methods such as with hidden reference or paired comparison or proper training sessions to overcome the potential subject bias. A systematic and exhaustive study in this direction can lead to more conclusive evidence in this direction.

"Know how to solve every problem that has ever been solved."

— Richard Feynman

5
An Objective and Subjective Quality Assessment Study of Passive Gaming Video Streaming

Contents

The work presented in this chapter has been published as (with minor modifications): N. Barman and S. Zadtootaghaj and S Schmidt and M. G. Martini and S. Möller, "An Objective and Subjective Quality Assessment Study of Passive Gaming Video Streaming", *International Journal of Network Management.*

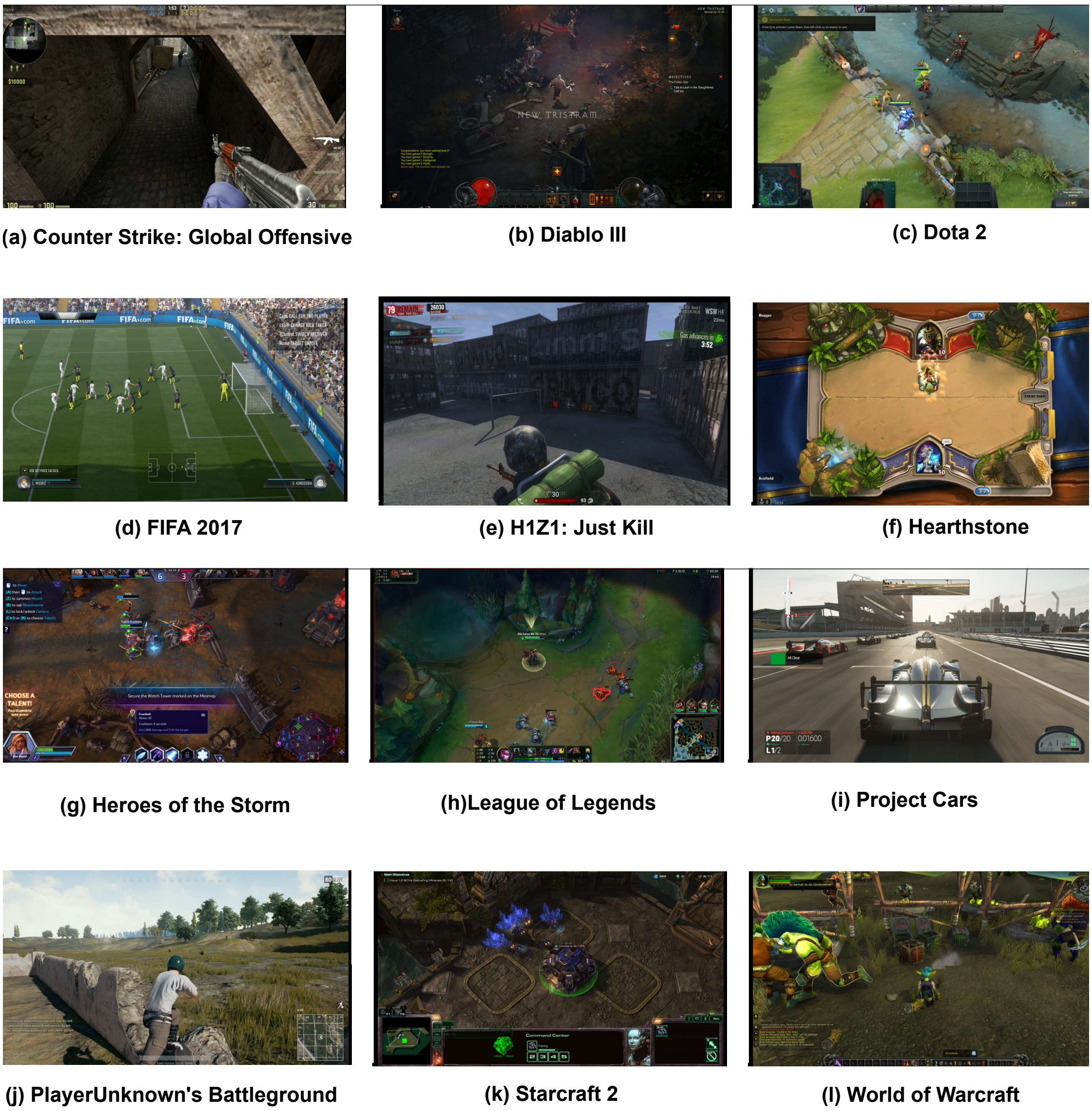
5.1 Introduction

Based on the results presented in the previous chapter (Chapter 4), we observe that the performance of the existing VQA metrics is different for gaming videos as compared to the non-gaming content. Recent years have seen tremendous advancement in the field of objective VQA metrics, with the development of models that can predict the quality of the videos streamed over the Internet. As discussed in Chapter 2.4.2, depending on the availability and the amount of source information, objective VQA metrics can be categorized into FR, RR, and NR. So far, these metrics have been developed and tested for non-gaming videos, usually considering VOD streaming applications. Also, some of the metrics such as NIQE and BRISQUE are based on qualities inherent to natural images (for details see Section 2.4.2). Gaming videos, on the other hand, are artificial and synthetic in nature and have different streaming requirements (1-pass, CBR). A study on the performance of objective VQA on gaming videos is still missing. Towards this end, we present in this chapter an objective and subjective quality assessment study on gaming videos considering passive streaming applications. In Section 5.2 we present a dataset and the evaluation methodology. The subjective quality assessment study resulting in subjective ratings for 90 stimuli generated by encoding six different video games in multiple resolution-bitrate pairs are presented in Section 5.3. Objective quality performance evaluation considering eight widely used VQA metrics is performed using the subjective test results and on a dataset of 24 reference videos and 576 compressed sequences obtained by encoding them in 24 resolution-bitrate pairs is presented in Section 5.4. The results of the metric performance for various temporal pooling strategies and complexity based game classification strategy are also presented and discussed. Section 5.5 finally concludes the chapter with a discussion of how the results and observations reported in this study can be used in future studies.

5.2 Dataset and Evaluation Methodology

5.2.1 Description of the Games

Gaming videos streamed over the Internet cover a wide range of games from different genres of varying degree of encoding complexity. For this study, we recorded 24 video sequences from a total of 12 games (each game two sequences) taking into account the genre, popularity (number of viewers on Twitch.tv) and video encoding complexity, as shown in Figure 5.1 and summarized in the Table 5.1.



(a) Counter Strike: Global Offensive

(b) Diablo III

(c) Dota 2

(d) FIFA 2017

(e) H1Z1: Just Kill

(f) Hearthstone

(g) Heroes of the Storm

(h)League of Legends

(i) Project Cars

(j) PlayerUnknown's Battleground

(k) Starcraft 2

(l) World of Warcraft

Figure 5.1: Screenshots of the gaming videos used in this work

5.2.2 Source Videos

The video sequences were captured losslessly in the RGB format at 30 frames per second (fps) using FRAPS[1] (version 3.5.99). For subjective video quality tests using non-gaming video content, typically a very short stimulus duration of only 10-15 seconds is used. However, as presented and discussed by the authors in various works [90–92], such a short duration may not be sufficient to cover a representative scene of a game, as the complexity over the length of the video sequence may vary drastically due to player behavior. Therefore, for this study, in line with the earlier

[1] http://www.fraps.com/

Table 5.1: Summary of selected games, respective genre and Twitch.tv ranking (RPG: Role Playing Game; MOBA: Multiplayer Online Battle Arena; MMORPG: Massively Multiplayer Online Role-playing Game)

Game	Genre	Twitch.tv Ranking
Counter Strike Global Offensive (CSGO)	FPS	3
Diablo III (Diablo)	RPG	31
Defense of the Ancients 2 (Dota2)	MOBA	4
FIFA 2017 (FIFA)	Sports	18
H1Z1: Just Survive (H1Z1)	Survival	53
Hearthstone (HS)	Collectible Card Game	8
Heroes of the Storm (HoTS)	MOBA	21
League of Legends (LoL)	MOBA	1
Project Cars (PC)	Racing Simulator	100+
PlayerUnknown's Battleground (PUBG)	Battle Royale	5
Startcraft 2 (SC)	Strategy	25
World of Warcraft (WoW)	MMORPG	9

recommendations, we selected a stimulus duration of 30 seconds. The scenarios for the captured sequences were chosen in such a way that they represent a common player behavior and are also representative of the actual game characteristics as usually streamed by the OTT services and watched by viewers. The captured sequences were then processed and converted into YUV format using FFmpeg[2].

As discussed earlier, for selection of game as well as for the selection of the respective video sequences, the video complexity of recorded sequences was taken into account. SI and TI values as defined in ITU-T Rec. P.910 can be used as an approximate measure of complexity, with high SI and high TI values representing a high level of complexity. From Figure 5.2, it can be observed that SI and TI values do not vary much for video sequences for the same game considering the fact that they represent different scenarios from the game (for some cases, even different levels).

In order to further analyze the temporal and spatial behavior of the recorded video sequences of our dataset, we calculate and plot the variation of SI and TI over the total duration of the video sequences (900 frames) instead of using single maximum value, as defined originally by ITU-T Rec. P.910 [22]. Figure 5.3 presents the Box Plot considering the individual frame level SI and TI values of the games. It can be observed from Figure 5.3 that there is a small variation of SI and TI over the duration of the video for games with omnipresent perspectives (players view and simultaneously influence the entire set of resources under their control [90]), such as Dota2, LoL and Starcraft 2 (SC). In addition, for games with the first-person perspective (the camera location is synonymous with the avatar's eyes and the

[2] http://www.ffmpeg.org/

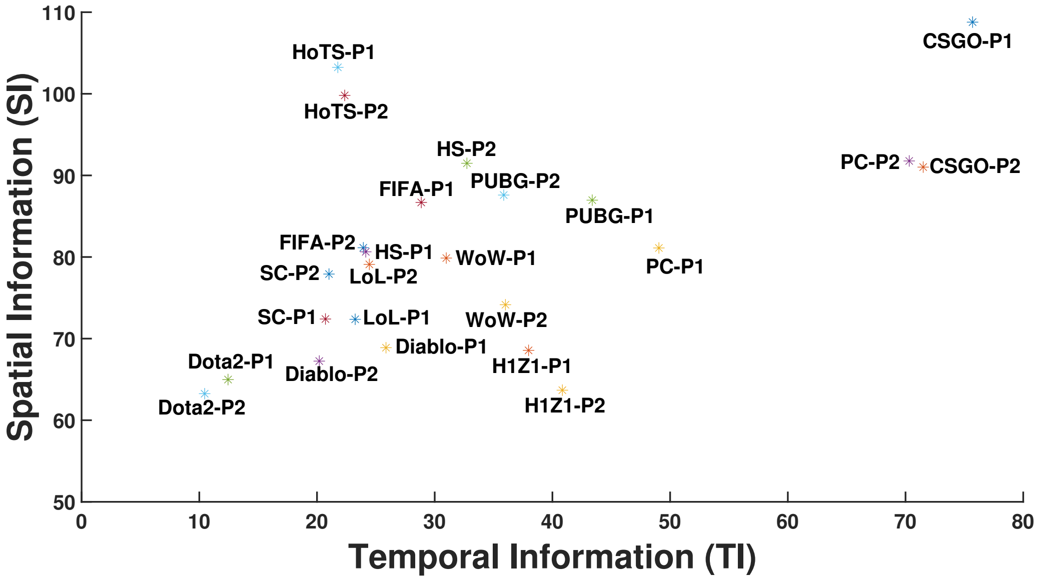


Figure 5.2: SI and TI plot for the 24 gaming video sequences.

game world objects appear smaller and closer together the farther they are from the camera location [90]) such as CSGO and H1Z1: Just Survive (H1Z1), high variation of SI and TI is observed. A comprehensive analysis of video game complexity has been carried out in [93] which was used as a reference in the selection of video sequences for the subjective quality assessment as discussed later in Section 5.3.1.

5.2.3 Encoding Settings

For streaming video games over the Internet, usually CBR is used as a rate control mode. CBR is usually selected due to the inherent feature of video games that can result in highly dynamic scenes followed by dull moments of gameplay. Using other rate control modes of encoding such as VBR, can result in stalling of the video playback at the end-user when high dynamic scenes appear after long dull moments of gameplay, leading to lower QoE[3]. Based on the discussion provided in Chapter 3, we select the encoding settings based on the recommendations by various OTT service providers. Our preliminary studies (not reported here) have shown no observable difference in the performance of VQA metrics when considering other encoding settings (VBR, 2-pass). However, in line with the industry wide used settings and recommendations, we choose CBR and 1-pass encoding settings as summarized in Table 5.2. Table 5.3 describes the resolution-bitrate pairs considered in this work.

[3] https://help.twitch.tv/customer/en/portal/articles/1253460-broadcast-requirements

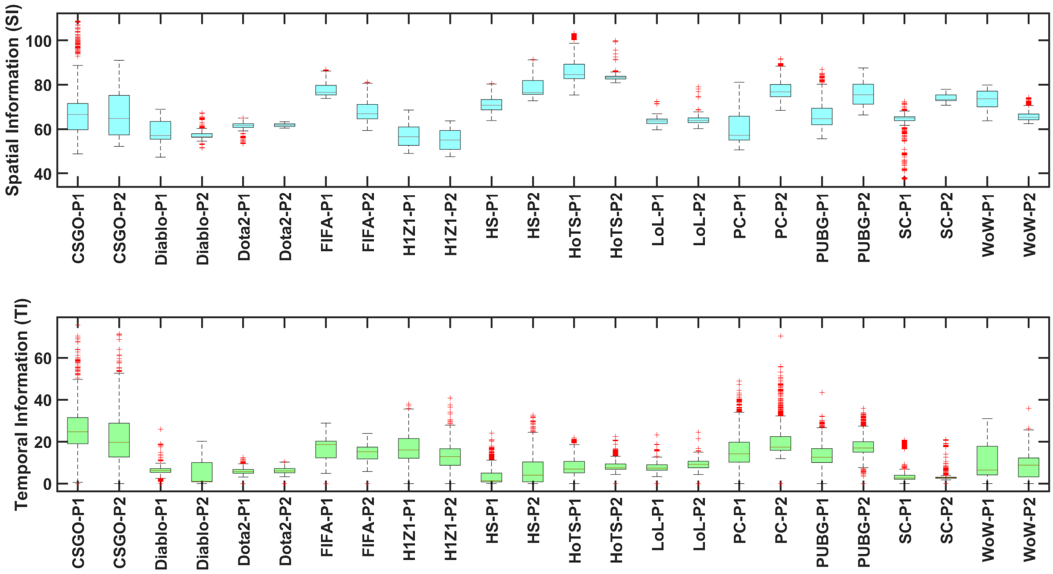


Figure 5.3: Box plot of SI (top) and TI (bottom) values for short duration gaming video sequences.

Table 5.2: Summary of video encoding parameters.

Parameter	Value
Duration	30 sec
Resolution	1080p, 720p, 480p
Frame Rate	30
Encoder	FFmpeg
Encoding Mode	CBR
Video Compression Standard	H.264, Main 4.0
Preset	veryfast

Table 5.3: Resolution-Bitrate pairs used to obtain distorted (compressed) video sequences. The bitrates in bold text refer to the bitrates used in the subjective quality assessment.

Resolution	Bitrate (kbps)
1080p	**600**, **750**, 1000, **1200**, 1500, **2000**, 3000, **4000**
720p	**500**, **600**, 750, 900, **1200**, 1600, **2000**, 2500, **4000**
480p	**300**, 400, **600**, 900, **1200**, **2000**, **4000**

5.3 Subjective Assessment

To assess the video quality of the encoded gaming video sequences, we carried out a subjective test in which ratings for different gaming videos encoded at different resolutions and bitrates are obtained from human observers. We discuss next the subjective test settings and methodology used in this work.

5.3.1 Test Environment and Set up

The subjective test was conducted at the Berlin Institute of Technology in standardized test room according to ITU-R Rec. BT.500 [23]. In order to avoid test participant fatigue, we limit the test duration by selecting six video sequences, three resolutions and five bitrates at each resolution. The resolution-bitrate pairs were chosen with the aim to cover a broad range of quality degradations without reaching saturation. The selection of video games was made after an in-depth analysis of video content (genre, popularity etc.) and considering game video complexity classification presented by Zadtootaghaj *et al.* [93]. The selected six games are as follows: CSGO and H1Z1 (high complexity), FIFA, LoL, and Project Cars (PC) (medium complexity) and HS (low complexity). Table 5.3 presents the selected resolution-bitrate pairs used in this study and with the resolution-bitrates pairs used for the subjective assessment highlighted in bold.

In order to avoid any unexpected artifacts of re-scaling of the videos by the video player, the downscaled, encoded MP4 video sequences at 720p and 480p resolutions were decoded and rescaled to 1080p, YUV videos using the *bilinear* scaling filter. The decoded raw YUV videos were then put in an *.avi* container for playback on a 24″ ViewSonic display monitor using the VLC player[4]. The order of resolution-bitrate pairs, as well as the order of the games, were randomized to avoid learning effects. A training session was conducted prior to the test in order to get users familiar with the test set up and the interface of the tool. For the training session, four video sequences from two games (Diablo and WoW, which are different from the ones considered for the subjective tests) were used. The visual acuity and color blindness of all participants were checked by using Snellen charts and Ishihara plates, respectively. The subjective ratings of the test participants who did not fulfill the visual capability requirements were removed from the dataset. In the end, the dataset consisted of the subjective scores of a total of 25 subjects with a median age of 29 years. For use cases such as codec evaluation, quality metric performance evaluation and comparison, cross-lab validation studies, etc., subjective video quality ratings are of very high interest to the research community. Hence, we make available the results as well as the reference and distorted videos as an open source dataset, GamingVideoSET[5].

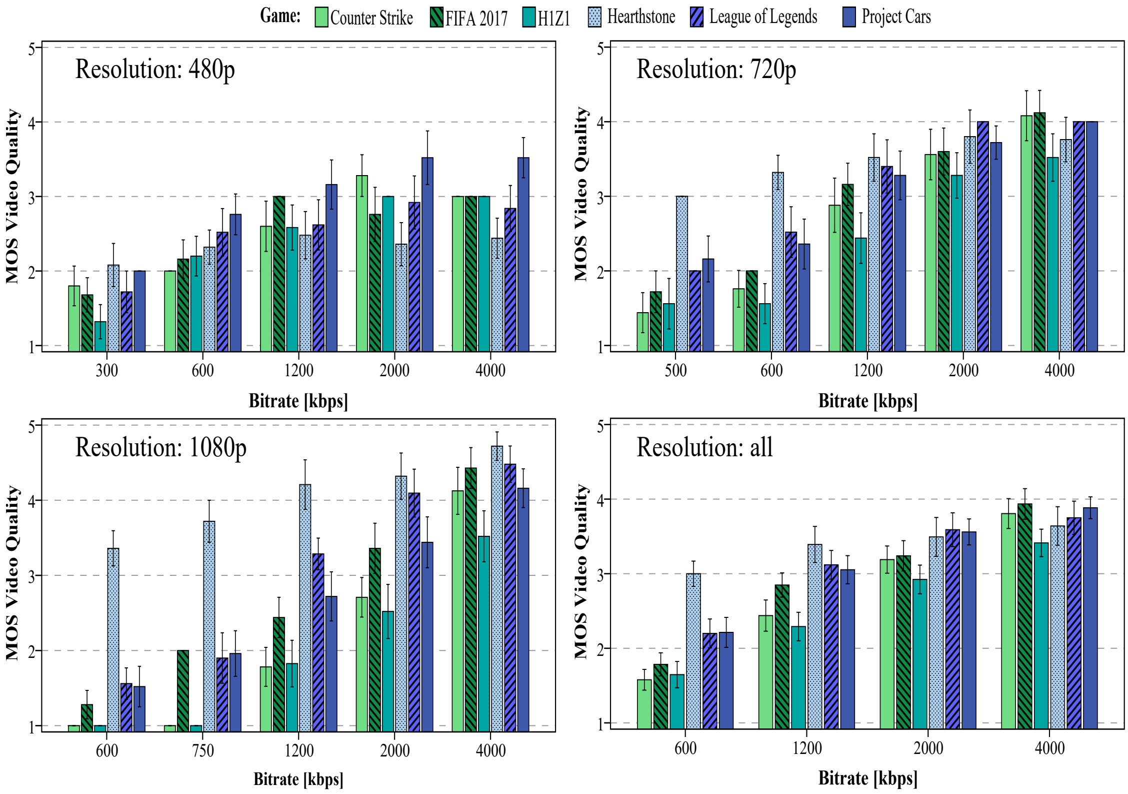


Figure 5.4: Barplots of video quality ratings for the six selected games for the used bitrates at different resolutions.

5.3.2 Subjective Scores

Figure 5.4 presents the bar plot of the video quality ratings in terms of MOS for the six gaming video sequences at 480p, 720p, and 1080p resolutions as well as for all resolutions combined. Based on the Figure 5.4, the impact of video complexity is apparent since for high complex games, H1Z1 and CSGO, at 1080p resolution and 600 kbps, no participant rated the video quality better than "bad" (1) while it does not hold true for other games. In addition, it can be observed that at 480p resolution, the video quality gets saturated for bitrates equal and higher than 2000 kbps. Furthermore, the video quality ratings for low complexity game HS are significantly higher than that for the other games at low bitrates for resolutions of 720p and 1080p. Lastly, the ratings of H1Z1 even for 4000 kbps at a resolution of 1080p, are lower than that of the other games.

Along with the subjective opinion ratings measured on a 5-point ACR scale, the acceptance of the respective stimulus on a binary scale (yes or no) was also

[4] https://www.videolan.org/vlc/

[5] https://kingston.box.com/v/GamingVideoSET

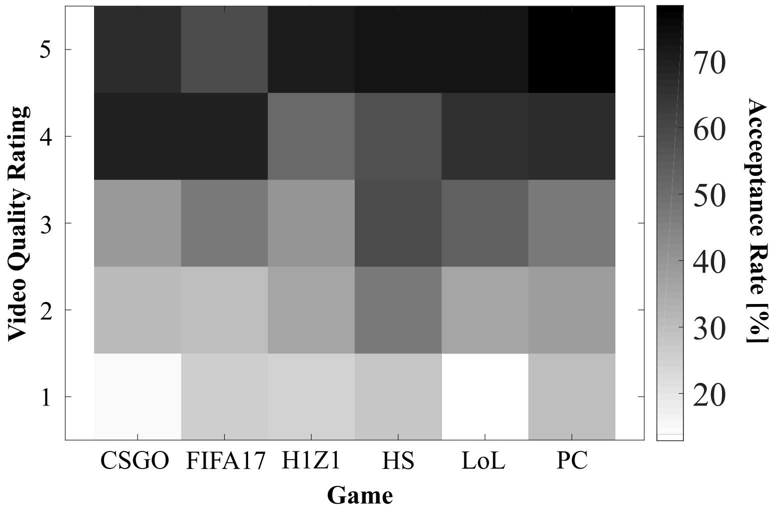


Figure 5.5: Heat map of acceptance rate at each video quality level (regardless of condition) for six gaming video sequences.

measured. A heat map of acceptance rate created based on the percentage of acceptance at each quality level (ranging from 1 to 5, regardless of condition) for six gaming video sequences used in the subjective test is presented in Figure 5.5. It can be observed that for low quality ratings, the acceptance rate for HS was higher than that for the other games and that the acceptance rate for PC for the highest score is higher than that for the other games. These differences might be caused by user factors (e.g., game preferences).

5.4 Objective Assessment

5.4.1 Objective Metric Performance Evaluation

In this work, we measure the performance of the objective metrics in two phases. In the first phase, we compare the performance of the VQA metrics with subjective scores considering the subjective dataset. In the second phase, for a comprehensive evaluation of the VQA metrics on the full dataset, we compare the VQA metric performance with a benchmark VQA metric. Since the encoded videos available are MP4, for FR and RR metric calculations, we instead use the decoded, raw YUV videos obtained from the encoded MP4 videos. The videos at 480p and 720p resolution were rescaled to 1080p YUV format using *bilinear* scaling filter as was done for subjective quality assessment. For NR metric calculations we use the encoded videos at their original resolution (without scaling 480p and 720p videos to 1080p). PSNR and SSIM calculation were done using the VQMT tool

Table 5.4: Comparison of the performance of the VQA metric scores with MOS ratings in terms of PLCC and SROCC values. *All Data* refers to the combined data of all three resolution-bitrate pairs. The best performing metric is shown in bold.

Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.67	0.64	0.80	0.78	0.86	0.87	0.74	0.74
	SSIM	0.58	0.44	0.81	0.78	0.86	0.90	0.80	0.80
	VMAF	**0.81**	0.74	**0.95**	**0.94**	**0.97**	**0.96**	**0.87**	**0.87**
RR Metrics	STRREDOpt	-0.67	-0.56	-0.86	-0.89	-0.83	-0.96	-0.75	-0.77
	SpEEDQA	-0.69	-0.55	-0.88	-0.90	-0.80	-0.96	-0.75	-0.77
NR Metrics	BRISQUE	-0.40	-0.37	-0.76	-0.80	-0.79	-0.76	-0.44	-0.46
	BIQI	-0.41	-0.35	-0.72	-0.70	-0.83	-0.82	-0.42	-0.45
	NIQE	-0.77	**-0.76**	-0.77	-0.73	-0.84	-0.84	-0.72	-0.71

available in [39] while for VMAF (version: VMAF_VF0.2.4b-0.6.1) we used the Linux based implementation in [40]. For ST-RREDOpt, SpEEDQA and BIQI we used the implementation made available by the authors using the default settings. NIQE[6] and BRISQUE[7] calculations were done using the inbuilt MATLAB function (version: R2017b).

Comparison of VQA metrics with MOS

Table 5.4 shows the correlation values of the eight VQA metrics with respect to MOS scores. The results are reported separately for each resolution as well as all three resolution-bitrate pairs combined (*all data*). It can be observed that VMAF results in the highest performance in terms of both PLCC and SROCC values across all three resolutions and *all data* except for SROCC value at 480p. The two RR metrics have a similar performance in terms of correlation values across all resolution-bitrate pairs considered separately as well as combined. Hence for applications where an increased speed of computation is of high importance, SpEEDQA can be selected as the preferred RR metric as it is almost seven times faster than ST-RREDOpt. Among the NR metrics, NIQE performs the best. For 1080p, BIQI and NIQE result in almost similar correlation values. For other resolutions and *all data*, BRISQUE and BIQI perform very similar.

Impact of resolution on VQA metrics

It can be observed that in general, the performance of the VQA metrics varies across different resolutions. For the FR and NR metrics, the performance of

[6] https://de.mathworks.com/help/images/ref/niqe.html

[7] https://de.mathworks.com/help/images/ref/brisque.html

Table 5.5: Comparison of the performance of the VQA metric scores with VMAF scores in terms of PLCC and SROCC values. *All Data* refers to the combined data of all three resolution-bitrate pairs. The best performing metric is shown in bold.

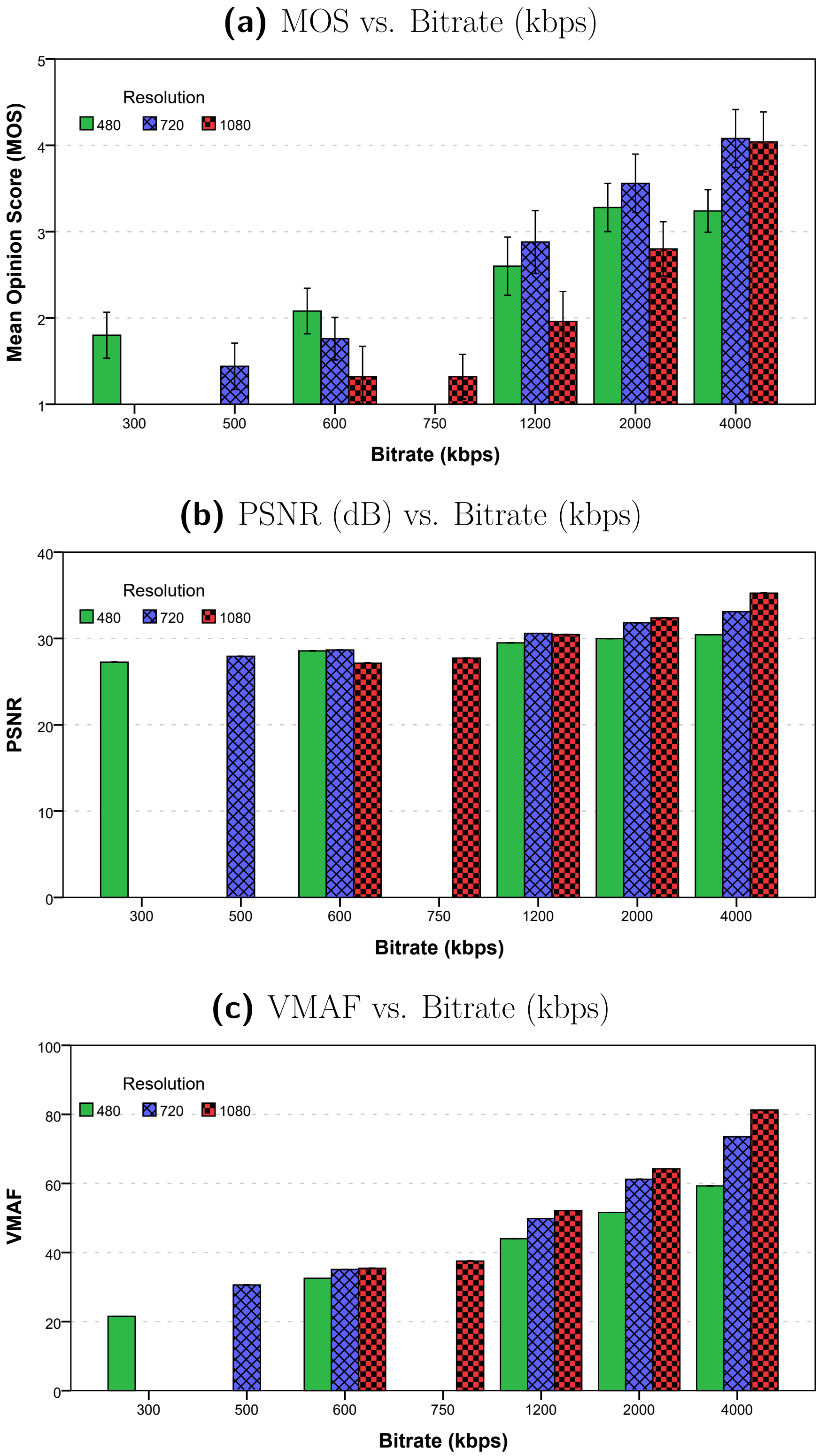
Metrics		480p		720p		1080p		All Data	
		PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
FR Metrics	PSNR	0.62	0.60	0.79	0.77	**0.91**	0.92	**0.87**	**0.87**
	SSIM	0.56	0.57	0.69	0.70	0.81	0.84	0.71	0.74
RR Metrics	STRREDOpt	-0.70	-0.88	-0.79	-0.92	-0.80	-0.93	-0.56	-0.63
	SpEEDQA	-0.71	**-0.90**	**-0.80**	**-0.95**	-0.79	**-0.95**	-0.58	-0.65
NR Metrics	BRISQUE	-0.62	-0.62	-0.75	-0.74	-0.74	-0.75	-0.11	-0.11
	BIQI	-0.55	-0.51	-0.70	-0.70	-0.67	-0.68	-0.04	-0.04
	NIQE	**-0.75**	-0.77	-0.79	-0.79	-0.77	-0.75	-0.41	-0.42

the metrics for higher resolution videos is significantly better than that for lower resolution videos. In contrast, both RR metrics resulted in higher correlation in terms of PLCC with MOS scores for 720p resolution videos, followed by 1080p and 480p resolution videos. Fisher's Z-test[8] to assess the significance of the difference between two correlation coefficients indicates that the difference between 720p and 1080p is not statistically significant, while the difference between 720p and 480p is significant, $Z = 2.954$, $p < 0.01$. For all eight VQA metrics, the performance for the 480p resolution (cf. Table 5.4) is considerably lower compared to the same VQA metric performance for the 720p and 1080p resolutions. Also, the decrease in performance for some metrics is higher than others. In figure 5.6b this observation is illustrated using PSNR and VMAF as an example. It can be seen that PSNR for different bitrates at 480p resolution is not able to capture the variation in MOS (cf. Figure 5.6a) as its values for the 480p resolution almost remain constant even at higher bitrates. VMAF, on the other hand, as evident from Figure 5.6c, captures this variation quite well and hence results in increased performance overall and also across each individual resolutions.

Comparison of VQA metrics with VMAF

In the previous section, we presented and evaluated the performance of the eight VQA metrics based on the subjective ratings using six reference gaming video sequences and 15 resolution-bitrate pairs. It was found that across all conditions, VMAF resulted in the highest performance among all eight VQA metrics in terms of both PLCC and SROCC values. Thus, in the following section, we will consider VMAF values as reference scores (ground truth) to estimate the quality of the

[8] http://psych.unl.edu/psycrs/statpage/biv_corr_comp_eg.pdf



(a) MOS vs. Bitrate (kbps)

(b) PSNR (dB) vs. Bitrate (kbps)

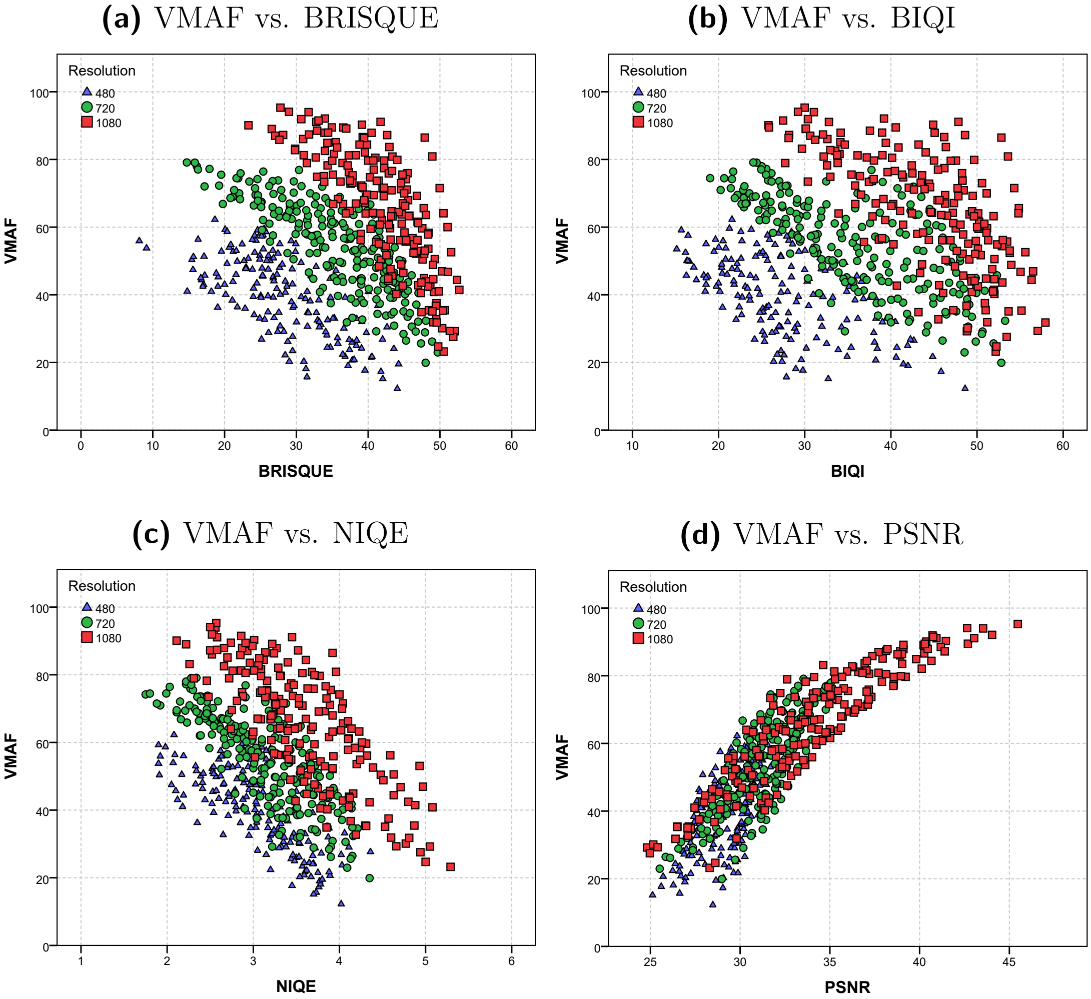
(c) VMAF vs. Bitrate (kbps)

Figure 5.6: MOS (with 95% confidence interval), PSNR and VMAF values for the CSGO video sequence at different resolution-bitrate pairs. A similar behavior is observed for other video sequences (relevant results not reported here due to lack of space).

remaining resolution-bitrate pairs which were not assessed during the subjective tests. We then evaluate the rest of the seven VQA metrics on the full dataset (24 reference video sequences and a total of 24 resolution-bitrate pairs, resulting in a total of 576 encoded video sequences). Table 5.5 shows the PLCC and SROCC correlation values for the seven VQA metrics with VMAF scores. It can be observed that PSNR results in the highest correlation followed by SSIM. Similar to the correlation values with MOS as reported in Table 5.4, both RR metrics result in similar performance. Also, it is observed that similar to the results reported in Table 5.4, for some metrics the correlation values vary significantly over different resolutions. At 1080p, PSNR results in the highest PLCC scores and SpEEDQA results in higher SROCC values. At 720p, SpEEDQA results in the highest PLCC and SROCC correlation values. At 480p, NIQE results in the highest PLCC scores and SpEEDQA results in the highest SROCC values. These results indicate towards the high potential for the use of RR and NR metrics for quality evaluations for applications limited to a single resolution and where full reference information is not available.

Comparative performance analysis of NR metrics

While the VQA metrics, in general, perform quite well, when considering multiple resolutions their performance decreases. Compared to FR and RR metrics, the performance degradation of NR metrics for *all data* was considerably high. We investigate the reason behind such performance degradation across multiple resolution-bitrate pairs using Figure 5.7 which shows the scatter plot of BRISQUE, BIQI and NIQE with VMAF scores considering all three resolutions. It can be observed from Figure 5.7 that, when considering individual resolutions, the variation of the NR metric values with respect to VMAF values are somewhat well correlated and increases linearly and hence results in reasonable PLCC scores. When considering all resolution-bitrate pairs, however, the spread of values is no longer linear, hence the lower correlation scores. PSNR on the other hand still has a linear correlation when considering all resolution-bitrate pairs and thus results in higher correlation scores. Among the three NR metrics, NIQE results in a much smaller spread for each resolution and when considering *all data* as compared to BIQI and BRISQUE. Hence, NIQE results in a higher overall prediction quality when using both MOS scores and VMAF scores as the benchmark. BRISQUE, on the other hand, results in almost similar performance as NIQE for 1080p and 720p resolutions but the correlation values decrease for 480p (a wider spread of the scores) and *all data*. BIQI performs the worst among all three.



(a) VMAF vs. BRISQUE

(b) VMAF vs. BIQI

(c) VMAF vs. NIQE

(d) VMAF vs. PSNR

Figure 5.7: Scatter plot showing the variation of the NR metrics and PSNR wrt. VMAF scores considering all three resolutions over the whole dataset.

The difference in values per resolution can be attributed to the fact that, while for FR and RR metric calculations we used the rescaled YUV videos, for 720p and 480p resolutions, for NR metric calculations we used the downscaled, compressed videos. This, along with lack of proper training with videos consisting of different resolutions, as well as the absence of parameters in the models which can capture the differences due to change in resolution results in lower correlation scores when considering all resolution-bitrate pairs.

5.4.2 Effect of Temporal Pooling on VQA Performance

The VQA metrics considered in this work provide a score for each frame of the video as most of them like SSIM, NIQE, BRISQUE etc. were initially designed for image quality assessment. For VQA, the usual practice is to consider the average of the individual per-frame level quality scores as the final score of the respective quality metric. Over the years, many different pooling methods such as *Minkowski Summation*, *Mean Last Frames* etc. have been proposed, which are shown to improve the correlation scores for various metrics when compared to simple averaging. An evaluation of six pooling methodologies was initially carried out by Drjle *et al.* [94]. It was found that a proper choice of pooling strategies can significantly improve the prediction quality of a VQA metric. Based on the performance results of various pooling strategies, the authors concluded that taking into account the recency effect and emphasizing the higher importance to low quality segments for the pooling strategies can increase the performance of the VQA metrics. However, this work was limited to six pooling strategies and considered videos of only 352x288 resolution of 12 seconds duration. Therefore, Seufert *et al.* [95] evaluated thirteen pooling methods on long duration videos (100 seconds) considering adaptive streaming application scenarios. Based on the results obtained, it was concluded that none of the complex pooling strategies performed significantly better than that of simple averaging.

Towards this end, we evaluated various temporal pooling strategies considering gaming video streaming application scenario in the present work, which are described in the following. In addition to averaging the individual per-frame level quality scores (*simple mean*), a total of seven different temporal pooling methods were evaluated. The *I-frame mean* pooling method considers the average of all I-frame scores only. In *Mean last frames*, the mean value of the N-last frames is considered (here $N = 10$). The *Minkowski Summation* pooling is defined as $\left[\frac{1}{T}\sum_{i=1}^{T} x^p(t)\right]^{\frac{1}{p}}$, where $x(t)$ is the frame level VQA metric value and p emphasizes the influence of the highest quality frames (here p=2). To take into account the recency effect, the Minkowski Exponential Summation is defined as $\left[\frac{1}{T}\sum_{i=1}^{T} \exp^{\frac{t-T}{\tau}} x^p(t)\right]^{\frac{1}{p}}$, where additionally the weighting factor related parameter τ (here, $\tau = 2$) is introduced. The pooling method *N Successive Frames* calculates the minimum value for mean values of N successive frames (here $N = 10$). Last but not least, the strategies *P Percentile Lowest_10* and *P Percentile Lowest_25* correspond to the mean value of the P percentile lowest quality frames with $P = 10\%$ and $P = 25\%$, respectively. A summary of the performance of the applied pooling methods is shown in Table 5.6.

Similar to the results reported by Seufert *et al.* [95], it can be observed that alternative pooling strategies can increase the performance of VQA metrics. However, we could not find a particular pooling method which improves the performance of all VQA metrics or a majority of them. An interesting observation is that, for RR and NR metrics, on an average, the *I-frame mean* pooling strategy

Table 5.6: Comparison of the performance of the VQA metric scores with MOS scores in terms of PLCC values for various pooling strategies. The pooling strategy with the highest score for a given VQA metric is shown in bold.

	Objective metrics															
Pooling Method	**PSNR**		**SSIM**		**VMAF**		**STRREDOpt**		**SpeedQA**		**BIQI**		**BRISQUE**		**NIQE**	
	PLCC	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**	**PLCC**	**SROCC**
Simple Mean	0.74	0.74	0.80	0.80	0.87	0.87	-0.75	-0.77	-0.74	-0.77	-0.42	-0.45	-0.44	-0.46	-0.72	-0.71
I Frame Mean	0.64	0.61	0.68	0.70	0.79	0.79	**-0.79**	**-0.80**	**-0.76**	**-0.80**	-0.46	**-0.48**	**-0.47**	**-0.50**	-0.69	-0.69
Mean Last Frames	0.52	0.54	0.76	0.78	0.78	0.77	-0.58	-0.66	-0.51	-0.66	-0.43	-0.47	-0.45	-0.52	-0.69	-0.70
Minkowski Summation	0.73	0.73	**0.80**	**0.80**	0.86	0.85	-0.74	-0.76	-0.73	-0.77	-0.41	-0.44	-0.44	-0.47	**-0.72**	**-0.71**
Minkowski Exponential Summation	0.53	0.53	0.75	0.77	0.79	0.79	-0.57	-0.66	-0.50	-0.65	-0.45	-0.47	-0.45	-0.51	-0.69	-0.69
P Percentile Lowest_25	**0.81**	**0.81**	0.78	0.79	**0.92**	**0.91**	-0.77	-0.76	-0.76	-0.76	-0.47	-0.45	-0.42	-0.45	-0.65	-0.66
P Percentile Lowest_10	0.79	0.78	0.77	0.78	0.91	0.90	-0.76	-0.75	-0.76	-0.74	**-0.47**	-0.47	-0.39	-0.43	-0.62	-0.63
N Successive Frames	0.77	0.75	0.74	0.78	0.87	0.87	-0.75	-0.73	-0.71	-0.71	-0.43	-0.42	-0.35	-0.38	-0.58	-0.61

results in a higher correlation with subjective ratings as compared to *simple mean* pooling strategy. Considering that in this work an I-frame occurs every two seconds, using the I-frame mean pooling can significantly reduce the computational complexity while resulting in similar performances compared to the other strategies, as the metric evaluation needs to be done only once per two seconds.

5.4.3 Effect of Content Complexity on VQA Performance

The complexity of a video sequence has a significant effect on the efficiency of video compression. Therefore, for subjective video quality assessment, ITU recommends selecting video sequences that cover a wide range of spatial and temporal complexity [22] as it was done in the present work (see Section 5.2.2). Consequently, the video complexity may affect the performance of video quality metrics. In this section, we analyze the impact of video complexity on the performance of the metrics that we used. As discussed in Section 5.3.1, for subjective assessment, we considered six video sequences from three complexity classes: high (CSGO and H1Z1), medium (FIFA, PC) and low (Dota2, HS). Based on this classification, we evaluated the performance of quality metrics for different content complexity classes and pooling methods, which are summarized in Table 5.7. In addition to the pooling strategies discussed in Section 5.4.2, we used here another temporal pooling strategy *TI pooling*, where the frame level VQA scores are averaged after weighing them using TI scores.

Based on the result in Table 5.7, for FR metrics we observe a higher performance for PSNR and VMAF when the video complexity is low, while SSIM performs better for the high video complexity class. One potential reason behind the poor

Table 5.7: Comparison of the performance of the VQA metric scores with MOS scores in terms of PLCC values for various pooling strategies considering different complexity classes. The pooling strategy with the highest score for a given VQA metric for a certain complexity class is shown in bold.

VQA Metrics	Game Complexity Class	Pooling Strategy								
		Simple Mean	I Frame Mean	Mean Last Frames	Minkowski Summation	Minkowski Exponential Summation	P Percentile Lowest_25	P Percentile Lowest_10	N Successive Frames	TI Pooling
PSNR	High	0.58	0.49	0.25	0.56	0.24	**0.73**	0.67	0.63	0.61
	Medium	0.82	0.62	0.87	0.81	0.87	0.93	**0.94**	0.93	0.86
	Low	0.85	0.78	0.88	0.85	0.88	0.89	**0.89**	0.89	0.85
SSIM	High	0.89	0.72	0.67	**0.89**	0.66	0.85	0.85	0.87	0.86
	Medium	0.77	0.62	0.93	0.77	**0.92**	0.79	0.78	0.73	0.79
	Low	0.79	0.76	0.79	0.79	0.79	**0.80**	0.79	0.79	0.79
VMAF	High	0.81	0.76	0.59	0.78	0.62	0.93	**0.93**	0.88	0.85
	Medium	0.87	0.74	0.89	0.86	0.88	0.92	**0.93**	0.93	0.88
	Low	0.92	0.88	0.91	0.92	**0.92**	0.91	0.90	0.89	0.92
STRREDOpt	High	-0.84	-0.86	-0.52	**-0.87**	-0.51	-0.78	-0.77	-0.86	-0.84
	Medium	-0.86	**-0.89**	-0.77	-0.83	-0.78	-0.88	-0.89	-0.88	-0.84
	Low	-0.64	**-0.65**	-0.60	-0.64	-0.60	-0.64	-0.64	-0.63	-0.64
SpEEDQA	High	-0.81	-0.82	-0.48	-0.83	-0.47	-0.76	-0.77	**-0.86**	-0.81
	Medium	**-0.88**	-0.81	-0.80	-0.86	-0.78	-0.87	**-0.88**	-0.87	-0.86
	Low	-0.63	**-0.66**	-0.60	-0.63	-0.60	-0.63	-0.63	-0.62	-0.63
BRISQUE	High	-0.67	-0.69	-0.63	-0.67	-0.65	**-0.71**	-0.71	-0.68	-0.68
	Medium	-0.45	-0.48	-0.54	-0.45	**-0.56**	-0.47	-0.47	-0.46	-0.45
	Low	-0.05	**-0.08**	-0.07	-0.05	-0.07	-0.04	-0.05	-0.04	-0.05
BIQI	High	-0.72	-0.78	-0.55	-0.71	-0.58	-0.82	**-0.85**	-0.81	-0.74
	Medium	-0.34	-0.42	-0.42	-0.33	-0.49	-0.44	-0.49	**-0.53**	-0.35
	Low	-0.05	-0.08	**-0.11**	-0.05	-0.09	-0.07	-0.07	-0.08	-0.06
NIQE	High	-0.83	-0.83	-0.77	-0.83	-0.78	-0.83	-0.83	**-0.86**	-0.83
	Meduim	-0.56	**-0.59**	-0.59	-0.56	-0.59	-0.51	-0.49	-0.48	-0.56
	Low	-0.76	-0.67	-0.74	-0.76	-0.73	-0.75	-0.75	-0.76	**-0.77**

performance of PSNR and VMAF for more complex videos could be the HVS characteristic called visual masking [96]. Visual masking plays a vital role on the perception of distortions in videos as explained by Choi *et al.* [97] who argue that *"more presence of spatial, temporal, or spatiotemporal distortions does not imply a corresponding degree of perceptual quality degradation, since the visibility of distortions can be strongly reduced or completely removed by visual masking"*. Choi *et al.* [97] analyzed the motion influences on the performance of VQA metrics by dividing LIVE VQA database [98] into two subsets of small motions and large motions. Their results revealed that some metrics such as PSNR perform poor in case of large motions. However, they did not evaluate VMAF and SSIM. For RR metrics, medium complexity classes result in highest correlation values followed by the high and then low complexity classes.

For NR metrics, we observe that the performance of metrics severely decreases when moving from highly complex content to low complex content, especially for BRISQUE and BIQI. The correlation of BIQI and BRISQUE with subjective video quality ratings for the low complexity class is very low (PLCC of 0.05 for the simple mean method). Although NIQE also performs worse for low and medium complex clusters, the decrease is not as significant as for BIQI or BRISQUE. A possible explanation for the weak performance of NR metrics in the low complex class could be the usage of NSS features in all these three NR metrics which needs to be investigated further. Finally, with respect to the effect of temporal pooling, similar to the results discussed in Section 5.4.2, there does not exist any particular temporal pooling method which results in the highest correlation value across all metrics even when considering different complexity classes. Also, for RR and NR metrics, similar to earlier, *I-frame mean* pooling seems to perform almost identical to that of *simple mean* pooling across different complexity classes.

5.5 Conclusion

We presented in this chapter an objective evaluation and analysis of the performance of eight different VQA metrics on gaming video considering a passive, live streaming scenario. Towards this end, GamingVideoSET dataset consisting of 24 reference gaming videos of 30 seconds duration and 576 distorted sequences was created. Subjective tests were carried out on a subset of the dataset consisting of 90 video sequences. The performance of the VQA metrics was evaluated on the subjective dataset in terms of PLCC and SROCC values. It was found that VMAF results in the highest correlation with subjective scores. Both RR metrics performance was found to be similar. Among the NR metrics considered in this work, NIQE performed the best. It was observed that many metrics failed to capture the MOS variation at lower resolutions, hence resulting in lower correlation values. Then we evaluated the performance of the rest of the VQA metrics against VMAF on the full test dataset. The performance of the NR metrics decreased when considering different resolution-bitrate pairs together.

Additionally, we analyzed the effect of various temporal pooling strategies on the performance of the VQA metrics. It was found that while for most metrics there exists a pooling strategy which results in a higher correlation score than that of commonly used simple averaging, we did not find any pooling strategy which performed significantly better than that of simple averaging across a wide range of metrics. It was found that for RR and NR metrics, I-frame mean pooling strategy resulted in a similar or even better prediction quality as compared to simple averaging, which when used can significantly reduce the complexity as such scores needs to be calculated only a few times over the length of the video, that too only at fixed intervals.

Lastly, we study the performance of the VQA metrics across different complexity classes. It was observed that for VMAF and PSNR, the prediction quality increases while moving across high to low complexity classes while a reverse trend is observed for the NR metrics. SSIM, on the other hand, has a different behavior that the highest correlation appears in the high complexity class followed by low and then medium complexity classes. For RR metrics, medium complexity class performed the best followed by high and low complexity classes. While possible reasons behind this have been discussed in this chapter, a further investigation is required.

"The trouble with weather forecasting is that it's right too often for us to ignore it and wrong too often for us to rely on it".

— Patrick Young

6

No-reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications

Contents

The work presented in this chapter has been published as (with minor modifications): N. Barman, E. Jammeh, S. A. Ghorashi and M. G. Martini, "No-reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications", *IEEE Access*, vol. 7, pp. 74511-74527, 2019.

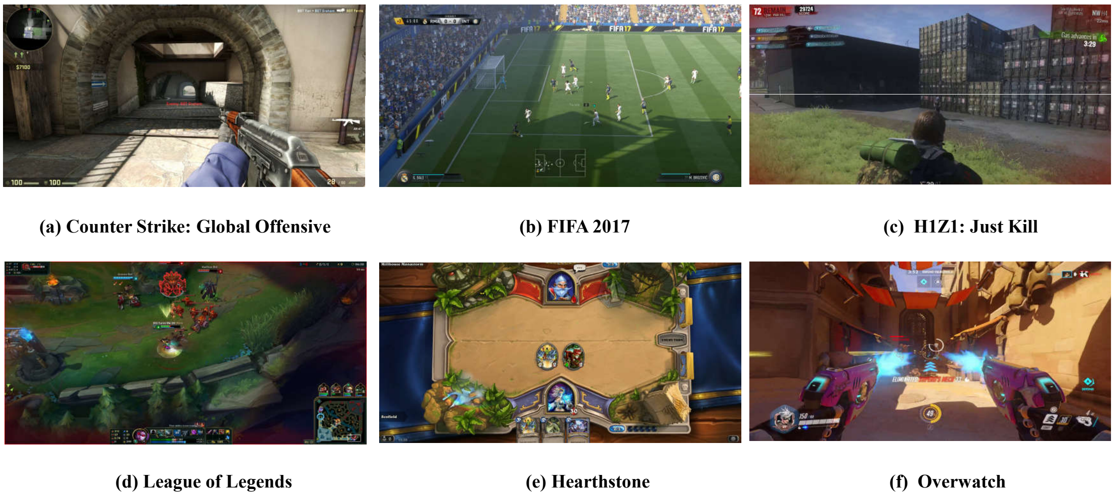
6.1 Introduction

Over the past two decades, researchers have investigated methods and techniques to estimate audio, image and video quality as perceived by the end users. In the previous chapters we discussed how due to the disadvantages of subjective quality assessment methodologies for practical applications such as real-time network monitoring and resource allocation, there has been a rising interest and work in the field of objective VQA metrics. Our performance evaluation of the eight most popular VQA metrics in the previous chapter indicated a not so satisfactory performance of the state-of-the-art NR VQA metrics. In the absence of any NR gaming video quality metric/model that meets existing requirements of high accuracy and low computational complexity to estimate accurately the quality of gaming videos in real time, we present in this chapter two machine learning based lightweight[1] gaming video quality estimation models. Both models due to their low complexity nature can, therefore, be used as the first stage of an optimized online gaming QoE management system, even on thin clients. For model design, we extract different features such as contrast, blur, blockiness, etc. from the distorted video sequences which are then used as an input to train two machine learning algorithms using subjective and objective ratings as the target output, respectively. Feature selection is performed to decrease the number of features (without much loss of prediction performance) using feature selection methods, hence further reducing the complexity of the model. The main contributions of this chapter are as below:

1. We propose a Neural Network (NN) based No-Reference Gaming Video Streaming Quality Index (NR-GVSQI). The model is designed using subjective ratings (MOS) from two open-source datasets. The proposed model is shown to outperform existing state-of-the-art NR metrics.

2. We also present a SVR based model, No-Reference Gaming Video Streaming Quality Estimator (NR-GVSQE), which is designed using FR VQA scores from GamingVideoSET. Due to the limited availability of subjective ratings, using FR VQA scores as the target output to train a ML model is used in this model design, which allows for usage of a much bigger training and test dataset (as VQA metric scores can be easily calculated for a much bigger dataset compared to obtaining subjective ratings). Our test of the proposed model, NR-GVSQE, on an unseen dataset shows that the proposed model, although

[1] By lightweight, we refer to the fact that all the features used in this work can be extracted in real-time without the need for high computational power and hence can be used for real-time quality monitoring.



(a) Counter Strike: Global Offensive

(b) FIFA 2017

(c) H1Z1: Just Kill

(d) League of Legends

(e) Hearthstone

(f) Overwatch

Figure 6.1: Some of the sample videos used in this work.

no-reference, results in almost the same performance as the state-of-the-art full-reference VQA metric, VMAF.

3. Additionally, this chapter presents an open source dataset, KUGVD, which consists of both subjective (MOS) ratings and objective analysis considering six gaming videos.

The rest of this chapter is organized as follows. Section 6.2 describes the previously proposed NR machine learning based QoE models. In Section 6.3 we briefly describe the open source dataset, GamingVideoSET presented in Chapter 5 and also introduce our newly designed dataset, KUGVD. Section 6.4 describes the extracted features and the feature selection methods along with the model development methodology. Section 6.5 describes the development, testing and validation of the NR-GVSQI model to predict the MOS scores obtained via subjective tests (MOS). Section 6.6 describes the NR-GVSQE model which is developed using an existing state-of-the-art FR VQA metric (VMAF) as the target output. Section 6.8 concludes the chapter with a summary of key findings and possible future work.

6.2 Related Work

With the advancements in the field of machine learning, the field of quality assessment in recent years has seen many proposed quality metrics/models based on machine learning algorithms, using different types of quality impacting factors, such as jitter, packet loss, compression artifacts (blockiness, blurriness, flickering, etc.) and rescaling. Since this study is focused solely on the design of NR metrics, we provide a brief review of recent works which have used ML algorithms to predict image/video quality without using any reference information.

In one of the earliest works in this direction, the authors in [99] used a Back-Propagation Artificial Neural Network (BP-ANN) to estimate the PSNR of H.264/AVC encoded video and obtained 97.8% correlation between the predicted and the actual PSNR. However, PSNR has been shown not to correlate well with QoE [27] [100]. Jiang *et al.* in [101] used a three-layer BP-ANN to predict the quality of high definition video, using features such as image blur, entropy, blocking artifacts, frequency energy, chroma information, and temporal information. Choe *et al.* in [102] used a three-layer BP-ANN to predict subjective quality scores based on features that were extracted from the H.264 bit-stream on a frame-by-frame basis. The proposed method used features based only on compression impairments and not on network QoS. The authors in [103] used an adaptive network-based fuzzy inference system based hybrid ANN to train a NN to estimate the quality of video transmitted over a wireless local area network and universal mobile telecommunication system. The prediction model used content type, frame rate and sender bitrate as application layer parameters and block error rate and link bandwidth as physical layer parameters. Shahid *et al.* in [104] used a 2-layer BP-ANN to predict the PSNR, Perceptual Evaluation of Video Quality (PEVQ) and SSIM based on features such as bits per frame, percentage of inter blocks, average motion vector length, and average QP. Although, PSNR and PEVQ were accurately predicted, the SSIM score is predicted with less accuracy. Wang *et al.* in [105] used features such as picture size, bitrate (BR), frame rate, GOP structure, picture type, macroblock type, QP, motion vectors, coded block pattern, and DCT coefficient as inputs to a 3-layer BP-ANN for quality assessment of MPEG-2 video streams. However, they did not compare their method with other regression methods and they did not consider feature selection or Principal Component Analysis (PCA). Cherif *et al.* in [106] used features such as QP, base-layer loss rate, enhancement layer 1 and layer 2 loss rate as inputs to a 3-layer BP-ANN to estimate the QoE of H264/SVC bit stream.

Khattabi *et al.* [107] used a BP-ANN with 3 hidden layers, and features such as the average of differences, the standard deviation of discrete Fourier transform differences, the average and standard deviation of DCT differences, the variance of color energy, luminance, and chrominance, to predict both MOS and PSNR. The complexity was high, due to the high number of features as well as the NN structure, and the authors did not reduce the dimensionality. Singh *et al.* [108] used a 3-layer ANN for NR QoE monitoring of H.264/AVC encoded videos streamed using HTTP/TCP in the context of IPTV. In [109], the authors used SVR for quality prediction, compared the performance with different visual quality predictors and reported improvement in prediction accuracy. Sunala *et al.* [110] used bitrate, SSIM, and interframe transformation fidelity as inputs to an ANN for video quality estimation. The authors in [111] used a radial basis function network for QoE estimation of video streamed over wireless networks, using cross-layer features such as bitrate, frame rate and Resolution (RES) at the application layer, PLR at network layer, video content features and the screen size of the terminal equipment. Xue *et al.* in [112] introduced a NR ANN-based video quality metric to predict the quality by considering the impact of frame freezing due to packet loss and/or late arrival. They used features based on freezing events such as the number of freezes, freeze duration statistics, inter-freeze distance statistics, frame difference before and after the freeze, normal frame difference, and the ratio of them.

In [113], the authors used a low complexity Multilayer Perceptron (MLP) NN for video quality assessment for mobile streaming services which can be used in smartphones in 4G-LTE. In [114], delay, jitter, PLR, and mean loss burst size are used as the inputs to a three-layer BP-ANN to assess the QoE of video services in LTE networks. In [115], 16 features including blackout, blockiness, block loss, blur, brightness, contrast, exposure, flickering, freezing, interlacing, letter-boxing, noise, pillar-boxing, slicing, spatial activity, and temporal activity are used as the inputs of a BP-ANN for high definition video quality assessment. In [116], PLR, the percentage of damaged frames and the percentage of different temporal classification frame which loses the packet, are used to train a feed-forward BP-ANN wireless video quality assessment model. In [117], first, a 2D convolutional NN is used to learn the spatial quality features at the frame level. Then, at the sequence level, the motion information is extracted as a temporal quality feature. A multi-regression model is then used for video quality measurement. In [118] the authors use restricted Boltzmann machine as an unsupervised deep learning method for video quality assessment. BR, number of frames, scene complexity, video motion, blur mean, blockiness, and motion intensity are used as features. They achieved an average of 78 to 91 percent correlation with well-known FR degradation assessment model VQM. In terms of scalability, they reported that only nine samples from the original video content types were sufficient to accurately assess the remaining of 864 videos of the dataset. More recently, the authors in [119] presented a FR and NR IQA metric that has a superior performance with respect to the state-of-the-art NR and FR IQA metrics when its performance was evaluated using three publicly available databases. The authors in [120] proposed a NR deep neural network IQA metric (MEON) consisting of two sub-networks each catering for two sub-tasks (distortion identification and quality prediction) for quality assessment with dependent loss functions. Their model is shown to achieve superior performance over the existing NR IQA metrics including the one proposed in [119] considering four different publicly available datasets. Inspired by the MEON model, a deep neural network based NR VQA model called V-MEON is proposed by the authors in [121] which provides an estimation of both quality scores as well as codec type. A comparison with existing NR metrics is shown to achieve high performance on two publicly available datasets.

Although there are many recent works in the field of quality assessment - as described above - most of these studies are limited in one or more of the following: different context (IPTV, etc.), very high complexity ([107]), older/different codecs (SVC, MPEG-2, etc.), evaluation methodology (few videos/single datasets, no subjective ratings, etc.), design for image quality assessment, rather than video quality assessment. Furthermore, all of these studies are limited to non-gaming content, whereas, as discussed earlier, gaming content has different streaming requirements and is inherently different from non-gaming content. Our work, on the other hand, focuses solely on gaming video content and uses two different datasets with stimuli representing compression artifacts as currently used by various OTT service providers.

6.3 Datasets

In this work we use two datasets, one of which is the open source gaming video dataset GamingVideoSET which we introduced in Chapter 5. For an easier understanding of the reader, we briefly summarize the GamingVideoSET here. GamingVideoSET consists of a total of 24 reference videos of 30 seconds duration, encoded in 24 different resolution-bitrate pairs to obtain 576 distorted (compressed) video sequences. In addition, MOS ratings for 90 stimuli conditions (six videos, 15 multiple resolution-bitrate pairs) are provided. MOS values are calculated as the average of the ratings provided by individual test participants during a subjective test for a particular video sequence.

Since only 90 subjective ratings are available in the GamingVideoSET dataset, which may lead to overfitting of the data when building the model using subjective ratings, we created another dataset, Kingston University Gaming Video Dataset (KUGVD). In order to not include any new type of impairment to the dataset other

Table 6.1: Overview of the two datasets used in this work.

Parameter	**GamingVideoSET**	**KUGVD**
Number of reference videos	24	6
Number of distorted videos	576	144
Games for Subjective Assessment	**CSGO**, FIFA, *H1Z1*, HS, LoL, PC	**CSGO**, FIFA, *H1Z1*, HS, LoL, OW
Number of Test Subjects	25	17
Subjective Test Environment Condition	ITU-R BT.500	ITU-R BT.500
Subjective Test Methodology	ACR	ACR
Number of Stimulus for Subjective Quality Assessment	90	90

The game in bold indicates the same gaming video across the datasets.
The underlined games indicates different games across the two datasets.

than what the model would be trained on, we used the same encoding settings as in GamingVideoSET. We selected six gaming videos and encoded them in the same 24 resolution bitrate pairs as was done with the GamingVideoSET resulting in 144 stimuli. For subjective assessment, we selected 90 stimuli with the same resolution-bitrate pairs as in GamingVideoSET. Table 6.1 summarizes the parameters of the two datasets. Since the subjective test was carried out at different places using a different set of participants, we decided to keep one game, CSGO, across both datasets, which then acts as anchor conditions (see Section 6.3.3). Four of the games selected were the same (FIFA17, H1Z1, HS, and LoL) but a different part (scenario) of the game was considered. Depending on the stage/scenario of the game, the game content complexity can vary a lot. Hence, considering the same game but a different scenario (scene) will allow us to investigate whether the model designed using one dataset (considering a particular scenario) is robust enough to predict the quality with reasonable accuracy when considering a different scenario from the game. Additionally, we selected the game Overwatch (OW), a first-person shooting genre game, as it is more popular on Twitch.tv and is of high complexity. This allows us to introduce a totally unknown game in either of the datasets: Project Car (PC), a car racing game being the other one, with complexity and characteristics which will not be present during the training phase. This allowed us to design a robust model which can lead to satisfactory performance even when evaluating the quality of an unknown game type.

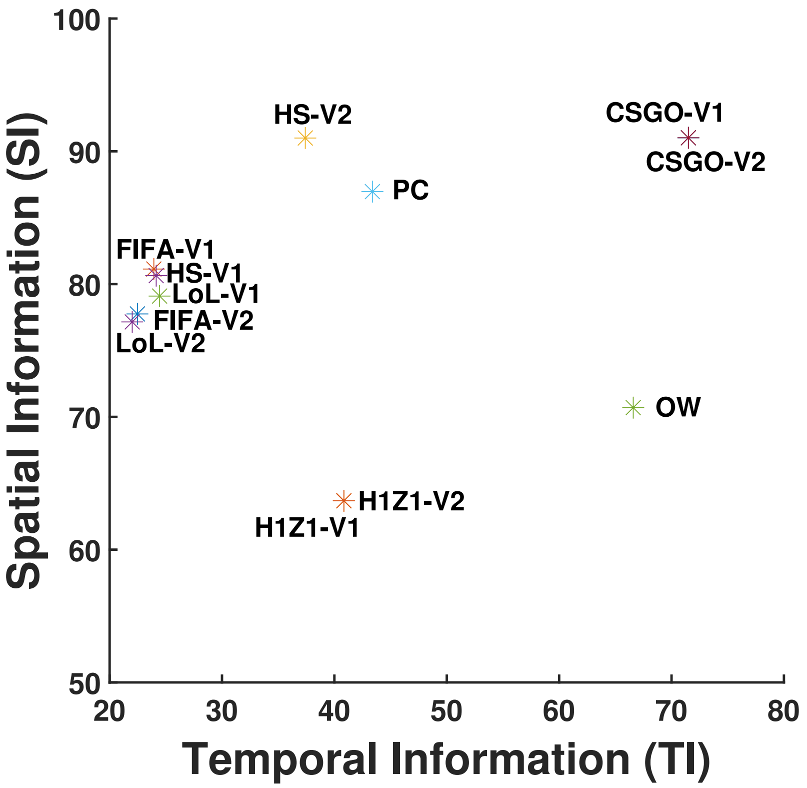


Figure 6.2: SI and TI plot for 12 gaming video sequences, six each from GamingVideoSET and KUGVD.

Table 6.2: Resolution-Bitrate pairs of compressed video sequences.

Resolution	**Bitrate (kbps)**
1080p	**600**, **750**, 1000, **1200**, 1500, **2000**, 3000, **4000**
720p	**500**, **600**, 750, 900, **1200**, 1600, **2000**, 2500, **4000**
480p	**300**, 400, **600**, 900, **1200**, **2000**, **4000**

6.3.1 Description of the Datasets

SI and TI as defined in ITU-T Rec. P.910 [22] are used as indicators of content complexity. Fig. 6.2 shows the SI vs. TI plots of the gaming videos considered for the subjective tests from both datasets. An interesting point to note is that the SI and TI for the video sequence from the game H1Z1 are the same even when the considered scenarios are different. LoL and FIFA are approximately of the same complexity while the HS sequence in KUGVD is of higher spatial and temporal complexity compared to the corresponding HS sequence in GamingVideoSET.

The videos were encoded at the same 24 resolution bitrate pairs (same as those used in GamingVideoSET, see Table 6.2) resulting in a total of 144 video sequences. Three resolutions and five bitrates from six videos from each dataset resulting in 90 stimuli were considered for subjective quality assessment which are shown in bold text in Table 6.2.

6.3.2 Test Environment and Set up

In line with the procedure followed for the creation of GamingVideoSET, we conducted a subjective quality assessment test at Kingston University, London, United Kingdom in a test lab adhering to ITU-R Rec. BT.500 standard [23]. The display monitor used was a 55″ Samsung 4K monitor. The 480p and 720p videos were upscaled and then, together with 1080p videos, decoded to raw YUV format. These were then put into an *.mp4* container for playback at 1080p resolution at the center of the display monitor with the rest of the pixels of the display fully black. The playlist was randomized in order to avoid learning effects. For training, we selected 4 videos from two games which were not part of the test, so as to make the test participants familiar with the test interface and the rating tool. The test participants were tested for visual acuity and color blindness using Snellen charts and Ishihara plates, respectively. After removing the ratings from test subjects who failed either of the visual tests, a total of 17 valid subjective test ratings were obtained.

6.3.3 Aligning Subjective Tests Scores

Since the subjective tests are conducted across different labs with different factors such as display, number and demographics of the test participants, the usual practice is to use anchor conditions (same test videos) across the different datasets and then use the MOS scores of these anchor conditions to determine a linear mapping function [122] which is then used to scale all MOS scores of the dataset(s). In our study, the gaming video sequence from the game CSGO is the same across both datasets (a total of 15 conditions taking into account three resolutions and 5 bitrates for each resolution). Considering the fact that GamingVideoSET contains MOS scores using more test participants as compared to KUGVD, we use the 15 MOS scores for CSGO from the GamingVideoSET as the reference scores for the anchor conditions and then use the linear mapping function $f(x) = mx + b$ as proposed in [122] to obtain the mapping between the anchor conditions. Using MOS scores of the anchor conditions, the coefficients m and b of the mapping function are obtained to be 0.9254 and -0.2613 respectively. Fig. 6.3 shows the scatter plot for the MOS scores and the linear fit. The goodness of fit scores obtained are as follows: SSE: 0.7114, R-square: 0.9443, Adjusted R-square: 0.94 and RMSE: 0.2339, indicating

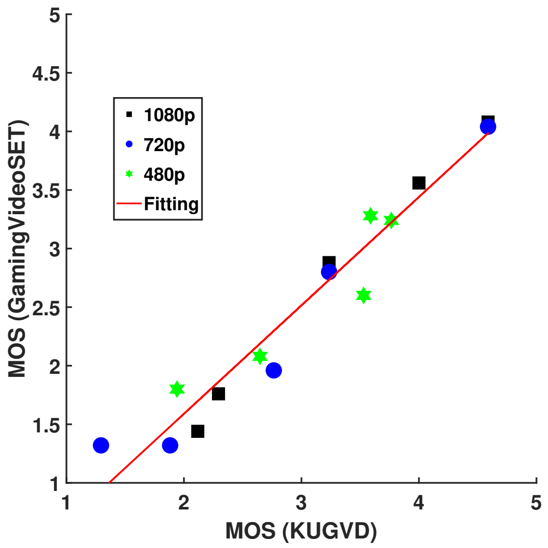


Figure 6.3: Scatter plot of MOS scores and the linear fit corresponding to the anchor conditions (15 conditions of CSGO sequence).

a good fit between the anchor MOS scores. The correlation between the anchor MOS conditions is obtained as 0.9875. As the fit is linear, there is no effect of the scaling of the MOS scores of KUGVD on the correlation scores with various metrics. Considering the fact that future work with third-party datasets may not have anchor conditions and that using linear scaling to adjust the MOS scores does not affect the performance of the VQA metrics in terms of their correlation with MOS scores, we finally decided to use the MOS scores from both datasets without any fitting.

Since an open dataset is of great use and interest to the research community, we have released the reference and distorted video sequences along with the scores for eight VQA metrics (3 FR, 2 RR and 3 NR) and subjective assessment scores (MOS ratings) as an open source dataset called KUGVD available at https://kingston.box.com/v/KUGVD. Henceforth, we will occasionally use Dataset 1 (D1) and Dataset 2 (D2) to refer to GamingVideoSET and KUGVD respectively.

6.4 Methodology

Fig. 6.4 shows the methodological framework that is used in this study to develop, test and validate the ML based gaming video quality estimation models. The key blocks of the methodology are feature extraction, feature selection, model development, and performance evaluation and validation. Datasets D1 and D2 were used in the development of NR-GVSQI and NR-GVSQE. For each model,

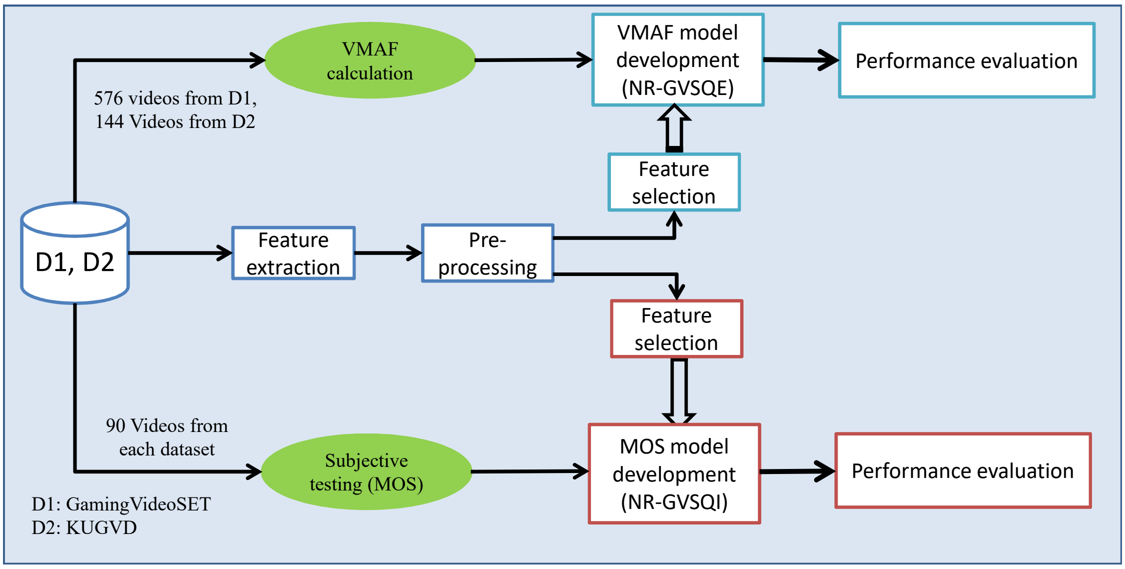


Figure 6.4: Methodological framework used in this work.

we extracted features and identified the best subset of features to use in model development. After training, validation was performed and each model was further tested on an external dataset which was not used in the model development process. A description of each individual step in the methodological framework is discussed next.

6.4.1 Feature Extraction

The performance of supervised ML-based predictive models is highly dependent on the features used in model development. Extracting relevant features for supervised learning is therefore critical. Previous statistical analysis has shown that video quality, as perceived by end users, is impacted by the combined influence of many factors such as the initial encoded video quality, content type, and the encoding parameters (e.g., frame rate, RES, BR and the QP) [123], [124]. Besides encoding, which determines the original encoded video quality, network QoS and client-side contexts, such as device type and resolution, further degrade the video quality. Since passive gaming video streaming applications such as Twitch.tv, YouTubeGaming, etc. use HAS technology, which is TCP based, they do not suffer from transmission-related impairments such as packet loss, bit error, etc. Hence, in this work we did not consider the impact of network and user context on the predicted quality. Since our goal is to build a NR model for quality estimation, we extracted features from only the distorted sequences, not relying on any reference information. We extracted 16 NR features for both datasets (GamingVideoSET and KUGVD) based

Table 6.3: Summary of the fifteen NR Features used in this work. The description of some of the features (those extracted using the tool [125]) is based on the description in [126] and [127].

NR Feature	Description
Blockiness	Resulting from the inherent nature of coding algoirthms which operate at the block level. It is one of the most common and visible artifact.
Blockloss	Loss of video data packets possibly during the transmission of the video.
Blur	Reduction of edge sharpness and spatial detail which results in a loss of high frequency information during coding.
Contrast	Difference in luminance and/or color that makes an object (or its representation in an image or display) distinguishable.
Exposure	Imbalance of brightness resulting from presence of frames that are too bright or too dark.
Flickering	A visible change in brightness which occurs between the screen refresh events.
Interlacing	Difference between consecutive pixels in columns, interlacing artefact is the result of special video compression where each frame is a connection between two frames in the original video.
Noise	Unwanted, uncontrolled or unprecedented pattern of intensity fluctuations.
Slicing	Artifact due to loss of packets, it occurs when a limited number of video lines is severely damaged.
Spatial Activity	Root mean square over space of the Sobel filtered values of a frame.
Temporal Activity	Root mean square over space of the difference in pixel values of the adjacent frames.
Spatial Information	Maximum over time of the standard deviation over space of the Sobel filtered values of a frame.
Temporal Information	Maximum over time of the standard deviation over space of the difference in pixel values of the adjacent frames.
RES	Resolution of the encoded video.
Bitrate	Number of bits per seconds of the video, reported in this chapter in kbps unless stated otherwise.

on the encoding process and on content. Hence, each sample of the dataset used in this study is described by 16 NR features based on content (spatial information (SI), temporal information (TI), spatial activity (SA), temporal activity (TA), exposure and contrast) and encoding process (RES, BR, blockiness, blockloss, blur, interlace, noise).

Additionally, we used the output scores of three NR metrics (NIQE, BRISQUE and BIQI) as input features for our model development (see Chapter 2.4.2 for more details on these metrics). Table 6.3 summarizes all the 15 NR features considered in this work. The first eleven NR features were computed using the tool provided by the authors in [125]. The tool provides per-frame scores for each video. We calculated the average of each of these 11 features, which are then used together with the other four features (SI, TI, RES and BR) and three NR metric outputs (which we consider as three features), to select a subset of features that was subsequently used for developing a model that maps these features onto an estimation of the video quality. Although some of these features are related and dependent (e.g., slicing and block loss; exposure and noise metrics; SI and SA; TI and TA), their combinations may result in improved prediction quality, as will be evident later. For a better understanding of these features, we guide the reader to the work in [126] and [127].

In addition, we also use the FR metric VMAF as an estimation of QoE because our earlier works presented in Chapter 4 and Chapter 5 in have shown that it estimates the subjective quality with high prediction accuracy for gaming videos. It should be noted that the three NR metrics were calculated on the downscaled encoded videos, whereas the rest of the features were calculated on upscaled, decoded raw videos.

Table 6.4: Correlation (PLCC) between the various features and MOS.

Feature	TI(1)	RES(2)	BR(3)	BIQI(4)	BRISQUE(5)	NIQE(6)	Blockiness (7)	SA(8)	Blockloss(9)	TA(10)	MOS(11)
TI (1)	1.00	-0.02	-0.03	0.32	0.32	0.31	0.22	0.14	0.16	0.85	0.09
RES (2)	-0.02	1.00	0.03	-0.62	-0.61	-0.30	-0.80	0.69	0.67	-0.06	0.04
BR (3)	-0.03	0.03	1.00	0.42	0.45	0.56	0.26	0.24	0.26	-0.12	0.71
BIQI (4)	0.32	-0.62	0.42	1.00	0.87	0.75	0.72	-0.09	-0.18	0.28	0.51
BRISQUE (5)	0.32	-0.61	0.45	0.87	1.00	0.82	0.77	-0.07	-0.06	0.36	0.54
NIQE (6)	0.31	-0.30	0.56	0.75	0.82	1.00	0.70	0.23	0.24	0.28	0.79
Blockiness (7)	0.22	-0.80	0.26	0.72	0.77	0.70	1.00	-0.31	-0.26	0.21	0.42
SA (8)	0.14	0.69	0.24	-0.09	-0.07	0.23	-0.31	1.00	0.79	0.28	0.51
Blockness (9)	0.16	0.67	0.26	-0.18	-0.06	0.24	-0.26	0.79	1.00	0.26	0.43
TA (10)	0.85	-0.06	-0.12	0.28	0.36	0.28	0.21	0.28	0.26	1.00	0.08
MOS (11)	0.09	0.04	0.71	0.51	0.54	0.79	0.42	0.51	0.43	0.08	1.00

6.4.2 Feature Selection

The nature of the data used to characterize the relationship between example data and the outcome measure may significantly affect the performance of predictive models. Noisy and unreliable data increase the difficulty of training machine-learning models. Removing redundant features to reduce the dimensionality of the data results in faster and more effective training and reduces the computational costs and the chance of overfitting [128]. The aim of feature selection is to select a subset X_S of the input features, $X = \{x_1, x_2, ..., x_N\}$, so that this subset can predict the outcome measure with a comparable performance with the case when the whole set of features X is used, but with less computational cost [129]. With N features, there are $2^N - 1$ possible subsets of features that should be tested if we want to use exhaustive search.

In order to reduce the complexity, different wrapper and filter feature selection methods [130] can be used to select a subset of the features discussed in Section 6.4.1. In the forward feature selection method, first the feature with the highest correlation with video quality is selected and progressively more features are added to create a larger subset of features with higher predictive power. Only features that increase the predictive power of the subset are retained. In the backward feature elimination method, all the features are selected as the starting subset and progressively the least promising feature that did not add any predictive power to the subset of features is eliminated [131]. The steps are repeated until a certain number of features remains or a certain performance level is reached. This reduces the complexity of the feature selection process by reducing the possible number of subsets from $2^N - 1$ to $N(N + 1)/2$.

6.4.3 Model Development

In this work, we propose two ML based models. The first one aims at estimating the MOS scores obtained via subjective testing, while the second aims at estimating the scores of a well-known objective quality metric (VMAF), that our previous studies identified as the best objective quality metric for gaming video streaming among the state-of-the-art ones analyzed. The first model, presented in Section 6.5, was designed using the MATLAB machine learning toolbox [132] while the second one, presented in Section 6.6, was designed using Waikato Environment for Knowledge Analysis (WEKA) [133]. As machine learning techniques, we used SVR, Gaussian Process (GP) regression, NN, and Random Forest (RF), which are representative machine learning algorithms that have been used in the domain of video quality prediction and modeling [134]. For an easier understanding of the presented models, we briefly describe the machine learning algorithms used in this work and for a more detailed discussion, we refer the reader to the individual references and to the work of Vega *et al.* [118] where a detailed description of the machine learning algorithms and their application in VQA is presented.

1. Neural Networks [135], commonly referred to as "artificial" neural networks is an information processing framework which is modeled after the biological nervous system such as the brain. They usually consist of a large number of interconnected elements (neurons) which work together to solve specific problems. In Section 6.5, we use a two-layer feed-forward network with sigmoid hidden neurons and linear output neurons that is able to fit multi-dimensional mapping problems. The Levenberg-Marquardt backpropagation algorithm is used to train the network. In Section 6.6, we use MLP which is a class of feed-forward artificial neural network consisting of at least three layers of nodes: an input layer, a hidden layer and an output layer which uses an iterative algorithm based on gradient descent as the backpropagation algorithm for supervised training. Using multiple layers and given the fact that all nodes except the input nodes is a neuron that uses a nonlinear activation function, it is different from a linear perceptron and hence able to distinguish data which is not linearly separable.

2. SVMs are supervised learning models which use learning algorithms for classification and regression analysis of the input data. Support-vector regression (SVR) is the version of SVM for regression which relies on a kernel function to fit the training data to a function with the error rate within a certain threshold.

3. A Gaussian Process [136] is a stochastic process where any finite subset of the range follows a multivariate Gaussian distribution. Gaussian process regression models are non-parametric kernel based probabilistic models.

4. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [137]. Random forests are based on the fact that while predictions made by a decision tree may not be accurate, using a combination of them will result in improved prediction accuracy.

The ML techniques used in this work have been selected due to their simplicity: one of the main goals of this work is to propose a light-weight NR metric, which is simple enough to be used in real-world applications. During the design of the metrics we have restricted ourselves to features which are of low complexity and can be extracted in real-time. The same low-complexity criterion is used during the selection of models so that the end metric can be used even on low computational power and energy constrained devices such as smartphones. The major objective of this work is to investigate different approaches for the selection of appropriate NR features and model design methodologies which can help build a NR quality model with high accuracy of subjective quality prediction specifically for gaming video streaming applications. Due to the inherent nature of the approach used in this work, where we already extract different features from the videos and only use them as input features to the machine learning algorithm, more complex ML algorithms such as Convolutional NN are not used in this approach. While different methods to train a machine learning algorithm can be used (for example providing as input the full video sequence to a CNN network for feature extraction and training by the ML algorithm), our initial work showed promising results with these simple models; hence such more complex solutions are not investigated. As will be evident later, our proposed models achieve better or similar performance as compared to deep-learning models proposed by other researchers.

6.5 NR Estimation of Subjective Scores (MOS)

6.5.1 Feature Selection and Model Development

For feature selection we used the backward elimination method explained in Section 6.4.2 for feature selection, and started with all of the 18 features (15 features mentioned in Table 6.3 and the three NR metrics) discussed in Section 6.4.1. At the first round of feature selection, to further reduce the complexity, the performance of all the combinations with 17 out of 18 features were examined by SVR, and it was observed that in eight cases the performance of MOS prediction is not significantly worse than the case with all of the 18 features. Then, these eight features were eliminated and the rest of the ten features (TI, RES, BR, BIQI, BRISQUE, NIQE, Blockiness, Spatial Activity (SA), Blockloss, and Temporal Activity (TA)) remained for further feature reduction.

We also evaluated feature derivation from existing features by PCA method and a mixed method (a combination of the selected features and derived features); however, feature selection using the backward elimination method showed better results. The subset that achieved the best prediction performance was selected for model development. The PLCC score was used to quantify the predictive power of subsets of features.

Table 6.4 summarizes the correlation scores (PLCC) between the remaining 10 features as well as with MOS scores (GamingVideoSET and KUGVD combined). For the feature selection part, both datasets (180 samples) were used, and then the model was trained and validated on GamingVideoSET, and its scalability (generalization) tested on KUGVD (and vice versa). The anchor conditions were used only during the training phase and were removed during the testing phase (since we had 15 anchor conditions in total, we had 90 samples for training and 75 samples for testing). Since TI and TA are calculated almost identically (see Table 6.3), it can be seen that they have a high correlation score of 0.85. As the videos are encoded at different resolutions, there is a high correlation between blockiness and RES, as expected. Also, since all three NR metrics are based on NSS, they have a high correlation among themselves. Also, Blockloss is found to have a high correlation with SA.

We evaluated the performance of both SVR and NN with all feature subset combinations for the two different training and test dataset combinations. Table 6.5 shows the performance results in terms of MSE, PLCC and SROCC scores of the best performing algorithm (NN) for various feature subsets. Here we have used the backward elimination method to reduce the number of features from 10 to 4 in 6 consecutive steps. It can be observed that for all feature subsets and training and test dataset combinations, NN performs better than SVR. For NN regression, the number of layers is 2 and the number of hidden neurons is 10. Levenberg-Marquardt is chosen as the training algorithm. During the training phase, 15% of the data is kept separately as the *validation* set. The algorithm training automatically stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples. The progress is monitored using the number of epochs (set to 1000) and the number of validation checks (set here to 6). The result of the trained network at the point of increasing error on validation is used for the test.

Table 6.5: Performance of various feature subsets for different training and test dataset combinations. The best performing cases are shown in bold.

Predictors (features)	**Training**			**Test**		
	MSE	**PLCC (%)**	**SROCC (%)**	**MSE**	**PLCC (%)**	**SROCC (%)**
1,2,3,4,5,6,7,8,9,10 (10F)						
Trained on D1, Tested on D2	0.10	0.93	0.94	0.58	0.84	0.85
Trained on D2, Tested on D1	0.05	0.98	0.98	0.22	0.87	0.86
1,2,3,4,5,7,8,9,10 (9F)						
Trained on D1, Tested on D2	0.11	0.93	0.93	0.42	0.85	0.86
Trained on D2, Tested on D1	0.07	0.97	0.97	0.22	0.86	0.86
1,2,3,4,5,7,8,10 (8F)						
Trained on D1, Tested on D2	0.12	0.92	0.93	0.42	0.85	0.85
Trained on D2, Tested on D1	0.07	0.97	0.97	0.22	0.87	0.86
1,2,3,4,5,8,10 (7F)						
Trained on D1, Tested on D2	0.14	0.91	0.91	**0.32**	**0.89**	**0.89**
Trained on D2, Tested on D1	0.08	0.96	0.96	**0.22**	**0.87**	**0.86**
1,2,4,5,8,10 (6F)						
Trained on D1, Tested on D2	0.22	0.86	0.85	0.83	0.74	0.75
Trained on D2, Tested on D1	0.13	0.94	0.94	0.99	0.73	0.73
1,2,4,8,10 (5F)						
Trained on D1, Tested on D2	0.20	0.87	0.86	0.94	0.70	0.72
Trained on D2, Tested on D1	0.21	0.90	0.91	0.41	0.73	0.72
1,4,8,10 (4F)						
Trained on D1, Tested on D2	0.22	0.85	0.85	0.80	0.71	0.73
Trained on D2, Tested on D1	0.26	0.88	0.88	0.42	0.72	0.72

Features selected are: TI(1), RES(2), BR(3), BIQI(4), BRISQUE(5), NIQE(6), Blockiness(7), SA(8), Blockloss(9), TA(10)
D1: GamingVideoSET; D2: KUGVD
Results are averaged over 100 iterations

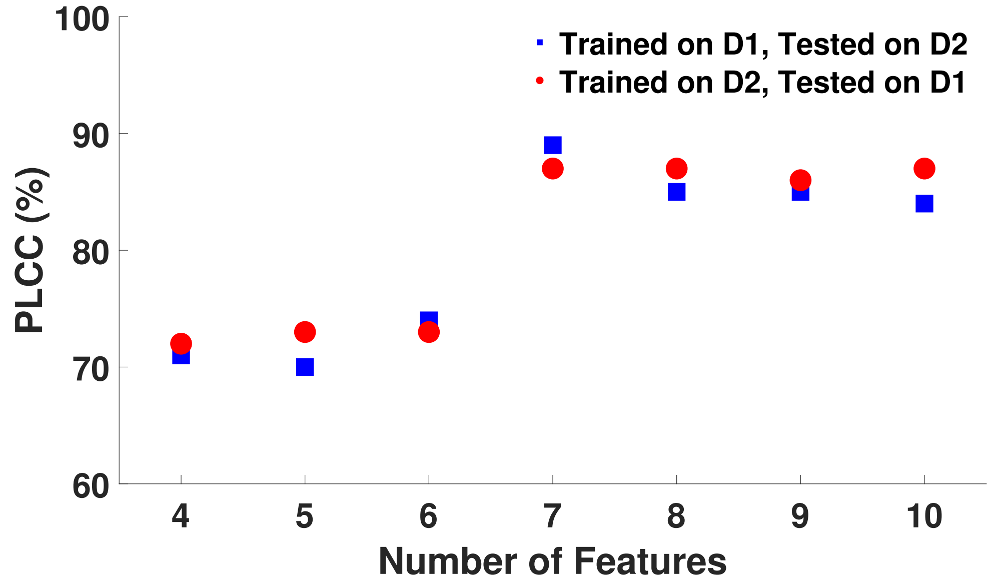


Figure 6.5: PLCC (%) variation with different number of features, for different training and test scenarios.

Fig. 6.5 shows the performance of the model in terms of PLCC scores with respect to the different number of features, considering two different training and test dataset combination scenarios as follows:

- *'Trained on D1, Tested on D2'*: Under test scenario, D1 and D2 were used as the "training" and "test" parts respectively consisting of 90 and 75 samples, respectively. This was repeated for 100 iterations for each feature subset combination.
- *'Trained on D2, Tested on D1'*: Same as *'Trained on D1, Tested on D2'* scenario but with D1 and D2 interchanged.

It can be observed that for both scenarios the prediction accuracy increases when the number of features is reduced from ten to seven. Further reduction in the number of features then reduces the prediction accuracy. The scalability (generalization) of the model is really tested when different datasets are used for training and testing. This is not considered in many research methodologies in QoE modeling and has resulted in optimistic performance results. Based on the results presented in Table 6.5 for the two different training and test scenarios, we consider that the NN model using 7 features results in the optimal performance which we refer to as NR-GVSQI.

Based on the results presented in Table 6.5 and Fig. 6.5, it can be observed that the performance results do not vary much when different datasets are used for training and testing (*'Trained on D1, Tested on D2'*, *'Trained on D2, Tested on D1'*) which leads us to the following conclusions:

- The proposed model, NR-GVSQI trained on a gaming video of a particular gameplay scene from a game is robust enough to predict the quality of another gaming video of another gameplay scene of the same game.
- When trained on a set of gaming videos from one dataset, NR-GVSQI is robust enough to predict the quality of a gaming video from a totally new game belonging to a different genre and complexity (both datasets consist of a game from a totally different genre; Project Cars in D1 and Overwatch in D2, see Fig. 6.2).

Table 6.6 shows the performance of nine state-of-the-art VQA metrics on the two datasets along with the performance of the proposed metric, NR-GVSQI. In addition to the three NR metrics considered during model development, we also compare the performance of our proposed metric with the recently proposed deep NN based metric MEON discussed in Section 6.2 which has been shown to outperform existing learning based as well as traditional NR metrics. Due to the use of proprietary SSIMplus in V-MEON, its the model implementation is no longer available publicly

Table 6.6: VQA metrics performance on the two datasets. The proposed metric's performance is shown in bold.

	GamingVideoSET	**KUGVD**
PSNR	0.74	0.80
SSIM	0.80	0.89
VMAF	0.87	0.92
STRREDopt	-0.71	-0.73
SPEEDQA	-0.71	-0.70
BRISQUE	-0.49	-0.62
BIQI	-0.43	-0.60
NIQE	-0.77	-0.85
NR-GVSQI	**0.87**	**0.89**

and hence could not be evaluated on our datasets. For the evaluation of MEON we have used the implementation and default settings as provided by the authors in the respective publication. The scores were computed on a per-frame basis and averaged over the whole video to obtain a final score. Due to non-availability of ground truth scores for our datasets, the model could not be re-trained and was evaluated using the trained weights provided by the authors.

It can be observed that the proposed metric NR-GVSQI results in the best performance on both datasets when compared to the four NR metrics. Considering GamingVideoSET, the proposed metric NR-GVSQI achieves a correlation of 0.87, which is almost the same as that achieved by the state-of-the-art FR metric, VMAF. For KUGVD, the metric achieves almost similar performance to SSIM, a widely used FR VQA metric. Comparing the performance across both datasets, it can be observed that while the performance of the state-of-art NR metric NIQE varies quite a lot, the performance achieved by NR-GVSQI is more stable across the two datasets. Among the four NR metrics, MEON results in the worst performance across both datasets (considering all 90 stimuli each) which is surprising given its high performance on different IQA datasets. This indicates that a machine learning based NR metric designed and tested on non-gaming videos does not necessarily perform well on gaming datasets and vice versa, hence establishing the need for customized NR metrics for gaming videos. Given the limited amount of training data we had, we expect that the proposed metric NR-GVSQI, when trained on a much bigger dataset, will result in an improved and more stable performance across different gaming videos and hence will be more generalizable to be used for quality estimation of gaming video streaming applications.

6.5.2 Discussion on Feature Selection

Based on the correlation values presented earlier in Table 6.4 it can be observed that the correlation between the features BIQI (4) and SA (8) with MOS is 0.51 and the correlation between the TI (1) and TA (10) with MOS is 0.09 and 0.08, respectively, which is very low. This implies that the selected features do not necessarily have a high correlation with the predicted entity. More interestingly, if we consider the correlation between the selected features, we can see for example that the correlation between TI(1) and TA(10) is 0.85, which as expected, is quite high. Hence, if feature selection would have been performed just based on the correlation values in Table 6.4, feature TI(1) and TA(10) both would not have been included in the same feature subset. Therefore, it can be concluded that the prediction performance of the features when considered individually and in a group can vary a lot. Feature selection using just domain knowledge or based on correlation with the prediction entity, as we saw here, is not enough and needs to be supplemented by feature selection methods such as backward elimination method as used in this work.

6.6 NR Estimation of FR Objective Metric (VMAF)

In the previous section, we proposed a machine learning based no reference model which was trained and evaluated in terms of its capability to estimate MOS. The limit of this approach is that the training and test data available consist of only 165 stimuli (90 stimuli from each dataset minus the 15 common conditions). Yet, conducting more subjective tests is time consuming, expensive and impractical, especially when creating a large dataset consisting of a large number of videos and encompassing various distortions. Hence, in this section we explore the possibility of developing a ML-based model to predict, rather than MOS, the best performing objective VQA score (still using NR features); a dataset covering a wide range of content and conditions is more practical and easier to create with VQA scores. We use the GamingVideoSET, which consists of a total of 576 distorted sequences, for model development and 120 sequences from KUGVD (excluding the 24 anchor conditions) for testing purposes. Since VMAF was found to have a very high correlation with MOS for both datasets we calculated for both datasets the VMAF scores, which are then used as the ground truth for the ML algorithms. Fig. 6.6 shows that the distribution of encoded video quality - in terms of VMAF - is well spread from low to high.

Table 6.7: Correlation (PLCC) between the various features considering all 576 sequences from GamingVideoSET.

Feature	RES	BR	SI	TI	BIQI	BRISQUE	NIQE	Blockiness	SA	TA	Blur	Noise	Exposure	Contrast	VMAF
RES	1.00	0.14	0.76	0.02	0.64	0.60	0.35	-0.82	0.77	0.06	-0.91	0.08	0.02	0.05	0.55
BR	0.14	1.00	0.23	0.04	-0.30	-0.36	-0.56	0.17	0.27	0.11	-0.21	0.46	0.03	0.02	0.68
SI	0.76	0.23	1.00	0.21	0.40	0.34	0.12	-0.46	0.92	0.27	-0.62	0.14	0.45	0.38	0.51
TI	0.02	0.04	0.21	1.00	0.32	0.27	0.24	-0.11	-0.06	0.86	0.10	-0.13	0.29	0.58	-0.20
BIQI	0.64	-0.30	0.40	0.32	1.00	0.89	0.77	-0.72	0.27	0.32	-0.55	-0.34	0.01	0.13	-0.04
BRISQUE	0.60	-0.36	0.34	0.27	0.89	1.00	0.82	-0.69	0.22	0.29	-0.50	-0.33	0.02	0.10	-0.11
NIQE	0.35	-0.56	0.12	0.24	0.77	0.82	1.00	-0.59	0.01	0.18	-0.23	-0.35	-0.04	0.02	-0.41
Blockiness	-0.82	0.17	-0.46	-0.11	-0.72	-0.69	-0.59	1.00	-0.43	-0.11	0.67	0.04	0.01	-0.07	-0.14
SA	0.77	0.27	0.92	-0.06	0.27	0.22	0.01	-0.43	1.00	-0.06	-0.69	0.22	0.32	0.22	0.65
TA	0.06	0.11	0.27	0.86	0.32	0.29	0.18	-0.11	-0.06	1.00	0.10	-0.08	0.46	0.61	-0.19
Blur	-0.91	-0.21	-0.62	0.10	-0.55	-0.50	-0.23	0.67	-0.69	0.10	1.00	-0.11	0.20	0.22	-0.62
Noise	0.08	0.46	0.14	-0.13	-0.34	-0.33	-0.35	0.04	0.22	-0.08	-0.11	1.00	0.05	-0.14	0.28
Exposure	0.02	0.03	0.45	0.29	0.01	0.02	-0.04	0.01	0.32	0.46	0.20	0.05	1.00	0.77	-0.05
Contrast	0.05	0.02	0.38	0.58	0.13	0.10	0.02	-0.07	0.22	0.61	0.22	-0.14	0.77	1.00	-0.04
VMAF	0.55	0.68	0.51	-0.20	-0.04	-0.11	-0.41	-0.14	0.65	-0.19	-0.62	0.28	-0.05	-0.04	1.00

6.6.1 Feature Reduction and Model Development

The Waikato Environment for Knowledge Analysis was adopted for feature reduction and model design for this metric. Based on our domain knowledge, as well as based on the results from our work in the previous section, we used 14 candidate predictors from the encoding process (BR, RES), content (SI, TI, SA, TA, Noise, Exposure and Contrast), compression artifacts (blur and blockiness) and no reference video quality metrics (BIQI, BRISQUE and NIQE) as the initial feature set. Four features considered for the earlier work (Blockloss, Interlacing, Flickering and Slicing) were not considered as they are not valid for the encoding conditions considered in the datasets used in this work.

Table 6.7 shows the correlation of features with VMAF and with different features in terms of PLCC scores. The features with the best correlation with VMAF are BR (0.68), SA (0.65), BLUR (-0.62), RES (0.55) and SI (0.51). No single feature is robust enough to predict the encoded video quality with acceptable accuracy. However, the correlation between BR with SA, BLUR, RES and SI is relatively low. A combination of BR and the three quality metrics may not yield high predictive power given that the correlation between BR and BIQI, BRISQUE and NIQE is relatively high. Combining BIQI, BRISQUE and NIQE to predict VMAF may not yield high prediction value due to the high inter-correlation between them, but as observed in results from model design in Section 6.5, this may not always be the case. Hence, as done previously in Section 6.5, it is necessary to conduct a feature selection process to determine a subset of features from the considered initial 14 features. Towards this end, we extracted the above mentioned initial 14 features for the full GamingVideoSET and KUGVD datasets using the same approach as was used previously. For model design and validation we used the extracted features from GamingVideoSET for feature selection purposes to determine subsets of features

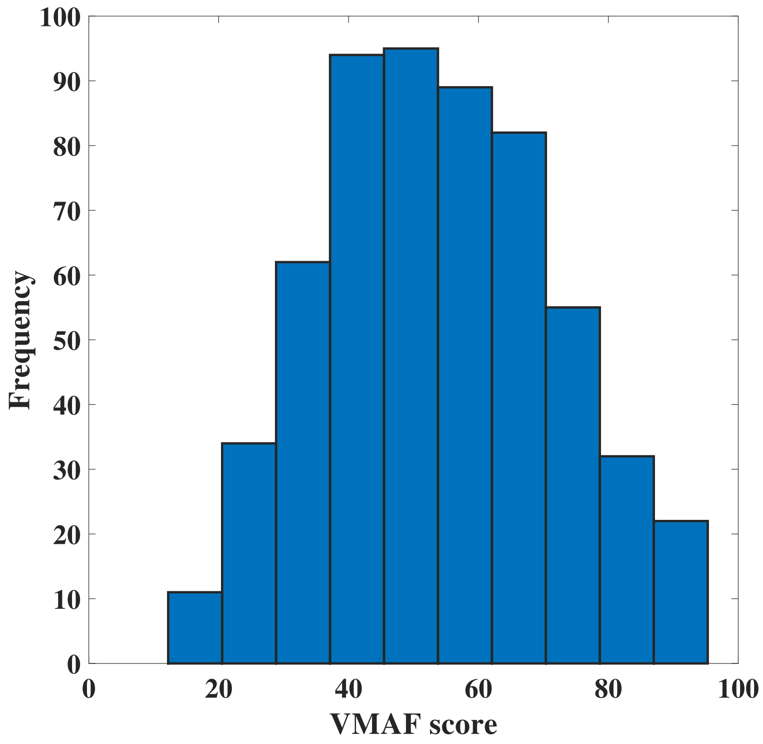


Figure 6.6: Histogram showing the distribution of the Quality (VMAF) of the encoded video sequences from GamingVideoSET used for training of the model.

Table 6.8: Results of model development for different sub-features using GP, MLP, SVR and RFML algorithms. The best performing result is shown in bold italics.

Features	GP			MLP			SVR			RF		
	PLCC	MAE	RMSE	PLCC	MAE	RMSE	PLCC	MAE	RMSE	PLCC	MAE	RMSE
1	0.77	9.04	11.17	0.64	11.22	14.01	0.67	10.58	13.16	0.79	8.74	10.90
2	0.87	7.18	8.64	0.76	9.60	11.82	0.87	6.97	8.76	0.85	7.38	9.39
3	0.89	6.51	8.07	0.80	9.01	11.49	0.88	6.43	8.41	0.87	6.72	8.65
4	0.93	4.97	6.40	0.88	7.12	9.09	0.95	4.18	5.77	0.93	4.79	6.29
5	0.94	4.77	6.19	0.91	6.15	7.82	0.96	3.45	4.91	0.95	4.52	5.82
6	0.96	4.16	5.24	0.92	5.83	7.28	0.99	2.06	2.92	0.96	4.05	5.10
7	***0.96***	***3.98***	***5.10***	***0.93***	***5.11***	***6.54***	***0.99***	***1.74***	***2.44***	***0.96***	***4.03***	***5.07***
8	0.96	3.83	4.91	0.93	5.13	6.46	0.99	1.35	2.02	0.97	3.74	4.71
9	0.97	3.73	4.78	0.96	4.21	5.25	0.99	1.36	2.00	0.97	3.77	4.71
10	0.97	3.74	4.80	0.95	4.74	5.99	0.99	1.24	1.81	0.97	3.77	4.73
11	0.97	3.34	4.24	0.98	3.16	3.97	1.00	1.11	1.68	0.97	3.38	4.34
12	0.98	3.23	4.11	0.98	2.72	3.40	1.00	1.05	1.60	0.98	3.24	4.15
13	0.98	3.18	4.08	0.98	2.53	3.10	1.00	1.07	1.61	0.97	3.44	4.34
14	0.98	3.19	4.09	0.98	2.75	3.48	1.00	1.05	1.60	0.97	3.36	4.26

for model development. We used the WEKA correlation based feature selection (CFS) function to select subsets of features [138]. This allowed us to reduce data dimensionality without having to manually evaluate all possible combinations of features. CFS evaluates the predictive worth of a subset of features by considering the predictive power of each feature together with the amount of redundancy between features. Preference is given to subsets that are highly correlated with VMAF whilst also having a low correlation between them. We selected subsets of features with the number of features in each subset increasing from 1 to 14 features. While this might not always be the optimal method for testing different feature set combinations, it works with reasonable accuracy considering our feature subset and model design as will be shown later. Each subset was then used as variables to develop four regression models using four different ML-based algorithms (GP, MLP, SVR and RF). These are some of the popular and frequently used ML algorithms in image/video quality prediction [113].

Using the extracted features from GamingVideoSET, we developed prediction models using the 10-fold cross validation methodology. The data were randomly divided into 10 sub-datasets. Nine sub-datasets were used for training a machine learning model and one was left out to test the model. This is repeated ten times until all sub-datasets have been used for training and testing. This methodology has the advantage that all data are used for training and testing. It is commonly used in machine learning to avoid overfitting [139] [140] [141], as was also used in the previous section. The performance of each model was averaged over the testing processes in order to determine the general performance.

Table 6.8 shows the performance of each subset of features in predicting video quality in terms of PLCC, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Increasing the number of features increases the prediction accuracy for all ML algorithms. However, there is an optimum number of features that provides an optimum balance between accuracy and complexity in terms of the number of features needed to achieve acceptable performance. Increasing the number of features beyond this number minimally improves the performance. However, this improvement is at the expense of increased complexity and increased computational requirement, which may limit usability in real time especially for thin clients.

For example, increasing the number of features used in the SVR model from 3 to 7 features increases PLCC by 12.5%. Yet doubling the number of features from 7 to 14 for the same algorithm increases PLCC by only 0.6%. Doubling the number of features does not improve the performance significantly, and only increases the model complexity (mainly due to increased feature extraction tasks as well as model computation time). Fig. 6.7 shows the relationship between the number of features and the performance for the four learning algorithms. The results clearly show saturation in performance with an increased number of features for all algorithms.

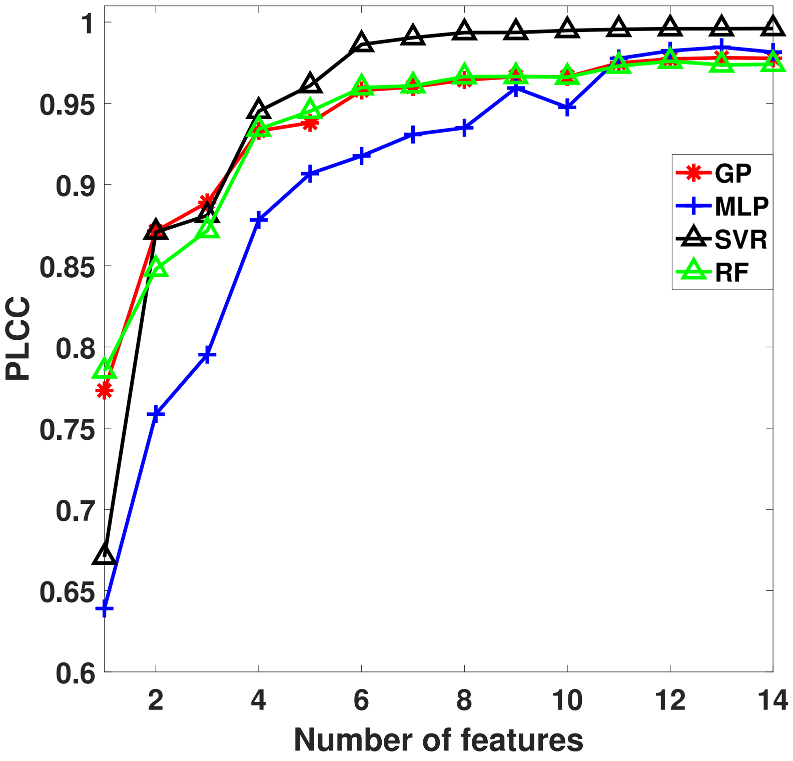


Figure 6.7: Impact of number of features on prediction accuracy for GP, MLP, SVR, and RF learning algorithms.

SVR obtained superior performance over all the other algorithms, with 7 features being optimal (*RES, BR, SI, TI, Contrast, Blur and Exposure*). We refer to this prediction model as NR-GVSQE. This is similar to our observation reported in Section 6.5 during our NR-GVSQI model design and evaluation using MOS ratings as the labels where also 7 features resulted in the optimum prediction model. The subset of features is different from those used in NR-GVSQI, which is not surprising as the relationship between the various features and MOS is not exactly the same as for VMAF. Henceforth, the VMAF scores as obtained from the distorted video sequences will be referred to as *(true) VMAF* while the VMAF scores as predicted using NR-GVSQE will be referred to as *predicted VMAF*.

Fig. 6.8 shows the performance of the models in terms of prediction accuracy during model development. The figure clearly shows that the quality prediction model based on SVR is of superior quality. We, therefore, selected this model for testing using the KUGVD dataset.

The selected SVR model was inherently validated during development due to the nature of 10-fold cross validation methodology. In practice, this is not usually

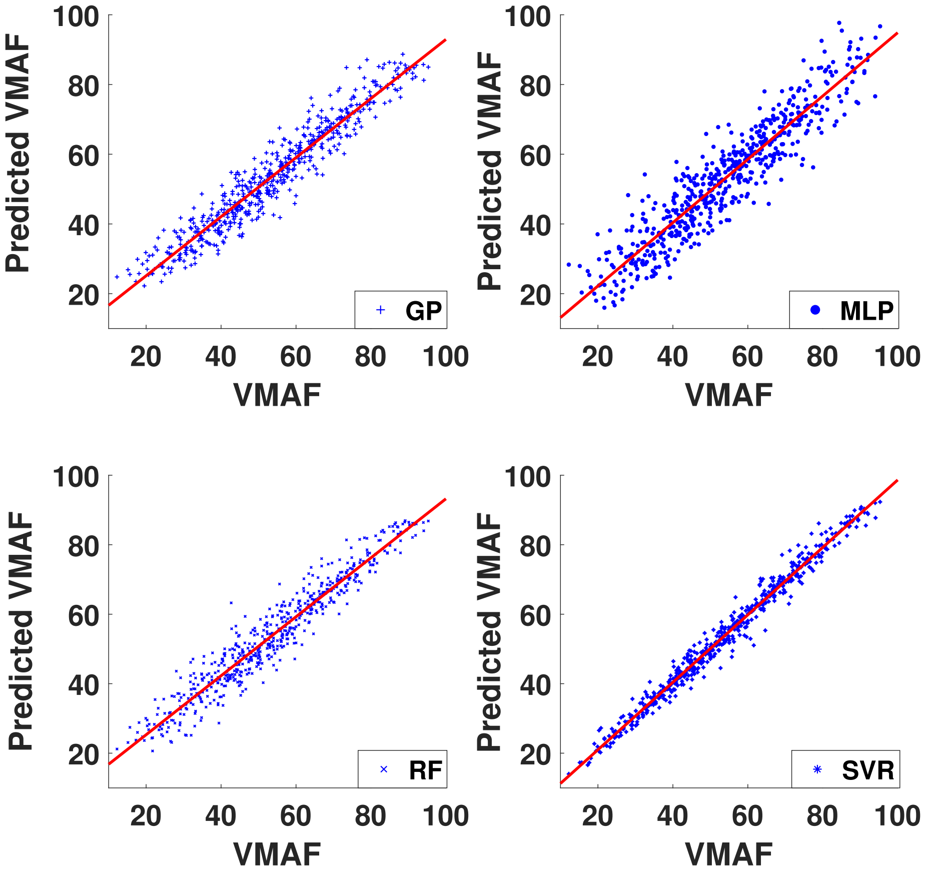


Figure 6.8: Prediction performance of GP, MLP, RF and SVR prediction models on the training dataset (GamingVideoSET).

enough and the final testing is usually conducted on an external dataset that was not used in model development. We externally validated and tested the model on KUGVD which is an unseen dataset and was not used in model development. The dataset has 120 samples (excluding the anchor video conditions) and has the same features as the training dataset. Fig. 6.9 shows the predicted VMAF by the model plotted against the (true) VMAF of KUGVD. The predicted VMAF is highly correlated with (true) VMAF, with a PLCC of 0.98.

Table 6.9 presents a comparative performance evaluation of our proposed metric versus popular FR (PSNR, SSIM), RR (STRREDopt, SpEED-QA) and NR (BIQI, BRISQUE, and NIQE) metrics to predict the VMAF of the KUGVD dataset. It can be observed that the proposed model outperforms these quality metrics by a huge margin in terms of correlation with VMAF. Since the ultimate goal of any IQA/VQA metric is to be able to predict the subjective quality as presented in

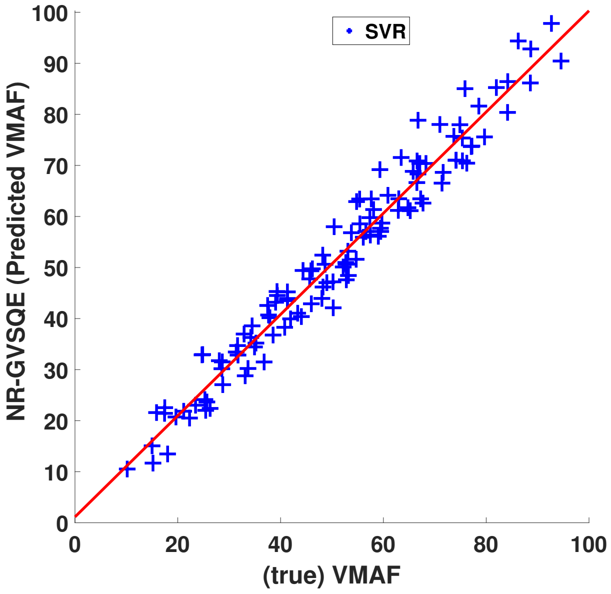


Figure 6.9: NR-GVSQE (Predicted VMAF) scores vs. (true) VMAF scores considering KUGVD dataset.

Table 6.9: Correlation (PLCC) of various VQA metrics and the proposed model, NR-GVSQE, w.r.t VMAF scores for KUGVD dataset. The best performing model is shown in bold.

Method	PLCC
PSNR	0.89
SSIM	0.84
STRREDopt	-0.66
SpeedQA	-0.64
BIQI	-0.35
BRISQUE	-0.4
NIQE	-0.74
NR-GVSQE	**0.97**

Section 6.5, we evaluate the performance of the model presented here - developed based on VMAF scores - in terms of correlation with respect to MOS scores from the subjective dataset.

Table 6.10 compares the performance of (true) VMAF and predicted VMAF scores vs. MOS on the KUGVD dataset, considering 75 stimuli (excluding anchor conditions). It can be observed that our proposed model, which was trained using VMAF scores from GamingVideoSET, when tested on an unknown dataset results in similar performance as (true) VMAF scores with respect to MOS ratings. It is

Table 6.10: Performance evaluation in terms of PLCC and SROCC of (true) VMAF and NR-GVSQE (predicted VMAF) scores w.r.t MOS scores from KUGVD dataset (excluding 15 anchor conditions).

Metric	PLCC	SROCC
(true) VMAF	0.930	0.934
NR-GVSQE	0.905	0.913

important to note here that our trained model utilizes NR features and hence is a NR metric, compared to VMAF which is a FR metric.

Compared to NR-GVSQI which was trained on MOS scores, the performance of NR-GVSQE is approximately 1.5% better on KUGVD in terms of PLCC scores. The improved performance of NR-GVSQE compared to NR-GVSQI can be attributed to the fact the the model design was performed using a much larger dataset due to the availability of objective VQA scores. The gain of 1.5% might not look a major improvement, considering the fact that both NR-GVSQI and NR-GVSQE use seven input features. It must, however, be noted that while NR-GVSQI uses the NR metrics BRISQUE and BIQI as input features, NR-GVSQE does not use any NR metric scores and uses only very basic NR features such as contrast, blur, and exposure along with basic features such as resolution, bitrate, SI and TI values. Hence, NR-GVSQE is of much lower complexity as compared to NR-GVSQI, with the added advantage that such model design can be performed on a huge dataset with multiple distortion artefacts without the need for any subjective (MOS) ratings.

6.7 Discussion

We will discuss in this Section the specificity of the model for gaming video and the comparison with other gaming video quality models.

The model design and performance on the gaming video datasets benefit from the inherent characteristics of the gaming videos (less variation in SI due to repetitive game elements [93], a difference in subjective opinions, etc. (see Chapter 4), which is not true for ordinary videos). For example, as discussed in [93], video games have special content characteristics in that they share the spatial and temporal features between different scenes of the same game. In fact, each game has a special motion pattern and a quite constant spatial complexity, as games are made of a pool of reused objects, which can be exploited by the machine learning algorithms, with possible increased performance for such gaming videos. In light of these factors, we argue that the while the proposed models are shown to work with high accuracy on the gaming video datasets considered in this work, it does not necessarily hold true for other non-gaming datasets (currently an ongoing work). As discussed before, gaming video streaming applications have so far not gained much attention from the research community. So far, in parallel to our work, there are two similar works carried out by the authors in [142] and [143] who also proposes machine learning based NR models: NR-GVQM and nofu, respectively.

NR-GVQM [142] is a SVR based model with Gaussian kernel which uses nine frame level input features and is trained and validated on GamingVideoSET. The model is trained using per-frame scores from 408 distorted video sequences (369000 frames) using VMAF scores as the target output, similar to the approach used in our proposed model NR-GVSQE (see Section 6.6). The model, when tested on the rest 144 distorted sequences, resulted in a correlation score of 0.98 with VMAF, while our proposed model NR-GVSQE achieves a correlation of 0.97 with VMAF. On the subjective dataset (90 video sequences), the model achieves a correlation of 0.89 with MOS. Compared to NR-GVQM, our model, NR-GVSQE achieves a higher correlation of 0.905 with MOS on an unknown dataset (KUGVD) using a lower number of features (seven compared to nine) and is of much lower complexity, as NR-GVSQE uses input features per video unlike NR-GVQM which uses per-frame scores for the final quality prediction.

Nofu [143] considers 12 different NR feature values per frame which are then divided into three equidistant groups, independent of the duration of the video. For each group, three values for each feature - the first value, the mean and standard deviation for each group is calculated, which results in a total of nine values per feature and a total of 108 pooled features values (considering the 12 selected features). The features are extracted from 360p center crop of the rescaled input video (irrespective of the native video resolution) after which the ExtraTreeRegressor method is used for feature selection using $0.5 \times mean$ as the threshold value and Random Forest as the choice of their regression algorithm. Similar to the aforementioned model, this model also uses VMAF scores (rescaled to 1-5) as target output. The model, when trained and tested on the GamingVideoSET via 10-fold cross validation, is shown to achieve a correlation of 0.96 as compared to 0.97 for our proposed model, NR-GVSQE. The proposed model, when tested on the subjective dataset part of GamingVideoSET (90 videos) via 10-fold cross validation, is shown to achieve a correlation of 0.91. In addition, the authors performed a source video based train and validation fold approach for subjective score prediction. For the 6 different video sources, they use 5 sources for training and 1 for validation, for which they achieved a correlation of 0.77. As discussed earlier, such an evaluation is hard because of the fact that each gaming video is from a different gaming genre and hence such an evaluation of a metric when tested on an unknown video(s) offers a more critical evaluation of the proposed model's performance for real-world applications. In contrast, our proposed model NR-GVSQE when trained on GamingVideoSET and tested on KUGVD containing different videos from the same as well as different games, achieves a correlation of 0.905.

Although both NR-GVQM and nofu appear to be promising models, due to lack of a second test dataset for the evaluation of the model performance, as discussed above, the actual performance of the models for real-world applications is not established. This also establishes the necessity of another open-source gaming dataset, such as KUGVD as presented in this chapter, which can be used for proper validation of proposed models for gaming streaming applications.

6.8 Conclusion

Subjective quality assessment of encoded gaming video is a necessity, yet it is time consuming, expensive, and not applicable in real time quality assessment scenarios. As a consequence, the development of objective quality assessment metrics is necessary. For some applications, such as passive gaming video streaming, FR and RR metrics are not suitable due to the unavailability of source information. On the other side, it has been shown that No-Reference (NR) quality metrics developed for natural video content are not suitable for compressed gaming video. Towards this end, we presented in this chapter two machine learning based NR metrics, NR-GVSQI and NR-GVSQE for gaming video quality prediction. Our proposed models, which are designed using supervised learning algorithms using MOS and FR Metric (VMAF) scores as the target output, are shown to perform better than the current state-of-the-art NR metrics, in the latter case achieving performance close to the state-of-the-art FR metric (VMAF). One of the major advantages of the proposed models is that they use a small number of features which can be extracted in real-time, hence the models can be used for real-time quality estimation of encoded gaming videos for live gaming video streaming applications.

Due to the inherent nature of the available datasets, the proposed models are limited to only compression and scaling artefacts. Also, currently both datasets are limited in scope considering the number of different games and the resolution-bitrate pairs considered. Since the datasets consist of videos compressed with the H.264 encoder, their performance on videos encoded with other newer encoders such as HEVC, VP9, or AV1 is an open question which we plan to explore in our future work, along with the creation of open-source datasets with an increased variety of games. While we have restricted ourselves here to known, pre-extracted features as an input to the machine learning algorithms, other approaches such as deep learning, 2D/3D convolutional NN approaches using transfer or self learning can also be investigated for possible increased prediction accuracy. Since such approaches require much bigger datasets than we have the MOS scores available for, the approach of using objective VQA metrics as used here in Section 6.6 can be investigated.

"There is no real ending. It's just the place where you stop the story."

— Frank Herbert (Frank Herbert interview, 3 February 1969, California State College, Fullerton)

7 Conclusions and Future Work

Contents

7.1 Overview

In this thesis, my research findings in the field of quality assessment for passive gaming video streaming applications using subjective and objective methods were presented. The literature review of the existing works on content and context aware coding and quality assessment approaches pointed out a huge gap in the field of quality assessment of passive gaming video streaming applications, which I tried to address based on various studies. The aims and objectives mentioned earlier in Chapter 1.2 were successfully achieved during the course of this research.

While on-demand OTT streaming applications such as Netflix, YouTube, Hulu etc. have gained much attention from the research community, live video streaming services and more specifically passive gaming video streaming applications such as Twitch.tv, YouTubeGaming etc. have not received much attention. For the continued success of such services, it is imperative to identify and study various quality assessment paradigms. The primary aim of this thesis was to address this gap towards which various studies were performed with a focus to study and evaluate the QoE of the end users. Our initial codec performance study pointed out a possible difference between gaming and non-gaming videos which was further verified by a comparative study of gaming and non-gaming videos. The results also indicated that the performance of existing VQA metrics on gaming videos might not be similar to their performance on non-gaming videos.

Towards this end, we evaluated the performance of eight most popular and widely used VQA metrics on a gaming video dataset, GamingVideoSET which was also created after realizing the lack of existing open-source gaming videos dataset. The GamingVideoSET consisting of 24 uncompressed reference videos, 576 distorted videos, and objective and subjective assessment results have now been made available as open source dataset for the research community. The evaluation of existing VQA metrics pointed out the poor performance of existing NR metrics. Since NR metrics are the most suitable for quality estimation of gaming videos streamed over the internet, we designed and proposed two novel machine learning based NR VQA metrics, NR-GVSQI and NR-GVSQE. To augment this study further, another open-source dataset, KUGVD was designed in line with the earlier designed gaming video dataset, GamingVideoSET. The performance evaluation of the proposed metric showed very high performance of the proposed metrics as compared to the existing NR metrics, with NR-GVSQE reaching almost similar performance as VMAF, a state-of-the-art FR metric.

7.2 Research Outcomes

In Chapter 2, QoE was introduced along with a discussion of influence factors and importance of QoE modeling from the perspective of different stakeholders and various subjective and objective quality evaluation methods. This was followed by a literature review of existing content and context aware video coding approaches which takes into account the end user QoE. The literature review highlighted promising works in this direction but almost all of them catered to typical VOD based streaming applications such as Netflix, OTT and/or broadcast video applications which streamed traditional video from different genres such as sports, action, drama, animation. Based on the review we identified a new but hugely popular and growing field of live passive gaming video streaming applications which so far had failed to catch the attention of the research community.

In Chapter 3, investigation regarding compression efficiency and encoding duration considering three most widely used codecs H.264, H.265 and VP9 was carried out using eight gaming videos of 10 seconds duration. This was done as a review of the existing works either did not consider real-time encoding settings as recommended for live gaming video streaming or they were carried out using non-gaming videos. Considering the fact that delay in encoding is critical for live streaming applications, encoding duration for the three codecs was also calculated and compared. Based on the findings it was observed that while the newer compression standard H.265/HEVC achieved a much higher compression gain compared to VP9 and H.264, the gain in performance was at the increased cost of encoding duration. The performance gain of VP9 and H.264 varied depending on the gaming video content which was not observed for non-gaming content. This indicated a possible difference between the gaming and non-gaming content which lead to our next work presented in Chapter 4.

In Chapter 4, a subjective and objective evaluation was carried out to study the possible difference between the gaming and non-gaming content. Using 15 video sequences each from gaming and non-gaming category and encoding them using HEVC at different quality levels, using spatial and temporal information, different objective quality metrics and subjective test, we studied how gaming and non-gaming videos are different. Our initial observation about the possible difference was that gaming videos, in general, had less variance in SI values as compared to the non-gaming videos which can be attributed to the fact how games are actually designed (same level of abstraction). It was also found that the objective VQA metrics resulted in lower correlation scores for gaming videos as compared to non-gaming videos. Also, image-based VQA metrics such as PSNR, SSIM resulted in lower performance than newer metrics such as VMAF as they failed to capture the motion content of videos which in case of gaming videos plays a critical role due to the inherent nature of the design of the games.

In Chapter 5, a significant contribution was made in terms of an open source gaming video dataset, GamingVideoSET to help address the problem of lack of existing open-source gaming videos dataset and encourage reproducible research. The dataset consisting of 24 reference videos, 576 distorted videos, and subjective and objective metric scores will serve as a valuable resource for interested researchers. A detailed performance evaluation of eight most popular and widely used VQA metrics was performed on a subset of the dataset considering multiple resolution-bitrate pairs. Our results showed that while the performance of FR VQA metric such as VMAF was good, the performance of NR metrics was not so satisfactory. A good performing NR metric is necessary for applications such as gaming video streaming as the reference information is usually not available in such services.

Chapter 6 presented the work on the design of NR machine learning based VQA metrics for passive live gaming video streaming applications. Two NR metrics are designed and evaluated in this work using different input features such as bitrate, blockiness, contrast etc. The first proposed VQA metric, NR-GVSQI is designed and tested using MOS scores as the target output. Three different training and testing methods and different machine learning algorithms are evaluated and the results for the best performing algorithm is presented for each of the train and test method. We observed that neural networks using seven features resulted in the best performance when considering results for the three train and test methods. Keeping in mind that the amount of data with subjective ratings is very limited, we investigated a novel approach of using state-of-the-art FR VQA scores as the target output. Similar to before, we tested the different combination of features and several machine learning based algorithms but on a much larger number of videos as calculating VQA scores on a higher number of videos is comparatively much easier than conducting and gathering opinion scores. Based on the results, it was observed that SVR resulted in the highest performance for seven number of features. The final model referred to as NR-GVSQE when trained on a dataset and tested on another dataset on MOS scores is shown to result in an almost similar performance as the state-of-the-art FR metric. The promising results show potential in machine learning based model design for quality estimation as well as using state-of-the-art VQA metric scores as the target output allowing for training on a much larger dataset.

7.3 Future Work

The work presented this thesis is a contribution towards the identified research gap in the existing literature on quality assessment of passive live gaming video streaming applications. Based on the research work presented in this thesis along with the newly designed and freely available open source gaming videos dataset, the below mentioned future work could be of interest to other researchers:

- One of the initial work was the codec compression performance using three codecs can be extended to newer codecs such as AV1 and also to hardware encoders. Also, live gaming video streaming simulations can be carried out to check the impact of high encoding delays with newer codecs such as HEVC on the end-user QoE.

- Our work comparing gaming and non-gaming videos indicated subject bias as a possible reason behind the lower correlation scores for gaming videos as compared to non-gaming videos. As mentioned before, a more systematic study could be conducted to investigate the magnitude of such effects. Also, based on the discussion about the possible lower performance of image based quality assessment metrics, different ROI based encoding strategies and their effect on the performance of the VQA metrics can be evaluated.

- The datasets presented in this work were limited to 1080p resolution. With the advent of 4K gaming and increasing cheaper HDR enabled consoles and games, an interesting future work could be to create a dataset considering higher resolutions (4K) and HDR games. Also, the current dataset includes 12 different games which may further be extended to include an increased number of diverse games, encoding conditions and other compression standards.

- The proposed machine learning based NR metrics were shown to perform quite well on different datasets and their combinations. The work can be extended further by considering datasets with more games and other resolution-bitrate pairs. The future works can consider other NR features to further improve the prediction accuracy.

- The proposed NR metrics and other newly proposed work can be evaluated using a live gaming video streaming testbed along with more realistic subjective test setups to evaluate the performance and suitability of the NR metrics proposed in this work and other newly designed metrics.

"Literature is the original Internet – every footnote, every citation, every allusion is essentially a hyperlink to another text, to another mind."

— Maria Popova, Writer, blogger, and critic,
BrainPickings.org

References

[1] A. S. Dias, S. Schwarz, M. Siekmann, S. Bosse, H. Schwarz, D. Marpe, J. Zubrzycki, and M. Mrak, "Perceptually Optimised Video Compression", in *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, Turin, Italy, Jun. 2015, pp. 1–4.

[2] Netflix, *Netflix Per-Title Encode Optimization*, `https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2`, [Online: Accessed 15-Jan-2019], Dec. 2015.

[3] Netflix, *Dynamic optimizer - A perceptual video encoding optimization framework*, `https://medium.com/netflix-techblog/dynamic-optimizer-a-perceptual-video-encoding-optimization-framework-e19f1e3a277f`, [Online: Accessed 15-Jan-2019], Mar. 2018.

[4] IETF RFC 4445, *A Proposed Media Delivery Index (MDI)*, Apr. 2006.

[5] ITU-T Rec. P.1201, *Parametric non-intrusive assessment of audiovisual media streaming quality*, ITU-T Recommendation, Oct. 2012.

[6] ITU-T Rec. P.1203, *Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport*, ITU-T Recommendation, Nov. 2016.

[7] GlobalWebIndex, *Gaming: An Examination of How the Gaming Landscape Is Changing, Q3 2017*, `www.globalwebindex.net`, [Online: Accessed 17-Dec-2018].

[8] Streamlabs, *Tipping up 33%, Twitch viewers up 21%, Fortnite dominates - Q1'18 Streamlabs Report*, `https://blog.streamlabs.com/tipping-up-33-twitch-viewers-up-21-fortnite-dominates-q118-streamlabs-report-52f60450af5a`, [Online: Accessed 07-Dec-2018], Apr. 2018.

[9] H. Ahmadi, S. Zadtootaghaj, M. R. Hashemi, and S. Shirmohammadi, "A Game Attention Model for Efficient Bit Rate Allocation in Cloud Gaming", *Multimedia Systems*, vol. 20, no. 5, pp. 485–501, Oct. 2014.

[10] Cisco, *Cisco Visual Networking Index: Forecast and Methodology, 2016–2021*, `https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf`, [Online: Accessed 10-Feb-2019], Jun. 2017.

[11] *Adobe HTTP Dynamic Streaming (HDS)*, `https://www.adobe.com/devnet/hds.html`, [Online: Accessed 17-Dec-2018], 2017.

[12] *Apple HTTP Live Streaming*, `https://developer.apple.com/streaming/`, [Online: Accessed 17-Dec-2018], 2017.

[13] *Microsoft Silverlight Smooth Streaming*, `https://www.microsoft.com/silverlight/smoothstreaming/`, [Online: Accessed 17-Dec-2018], 2017.

[14] *ISO/IEC 23009-1:2014 Preview Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats*, https://www.iso.org/standard/65274.html, [Online: Accessed 17-Jan-2019], 2017.

[15] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet", *IEEE MultiMedia*, vol. 18, no. 4, pp. 62–67, Apr. 2011.

[16] J. Kua, G. Armitage, and P. Branch, "A Survey of Rate Adaptation Techniques for Dynamic Adaptive Streaming Over HTTP", *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1842–1866, Thirdquarter 2017.

[17] N. Cranley, P. Perry, and L. Murphy, "User perception of adapting video quality", *International Journal of Human-Computer Studies*, vol. 64, no. 8, pp. 637–647, 2006.

[18] S. Egger, B. Gardlo, M. Seufert, and R. Schatz, "The Impact of Adaptation Strategies on Perceived Quality of HTTP Adaptive Streaming", in *Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming*, ser. VideoNext '14, Sydney, Australia: ACM, 2014, pp. 31–36.

[19] P. L. Callet, S. Möller, and A. Perkis, "Qualinet White Paper on Definitions of Quality of Experience (2012)", *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, vol. 1.2, Mar. 2012.

[20] ITU-T Rec. P.10/G.100, *Vocabulary for performance and quality of service. Amendment 5: New definitions for inclusion in Recommendation ITU-T P.10/G.100*, ITU-T Recommendation, Jul. 2016.

[21] L. Skorin-Kapov and M. Varela, "A Multi-dimensional View of QoE: the ARCU model", in *Proceedings of the 35th International Convention MIPRO*, Opatija, Croatia, May 2012, pp. 662–666.

[22] ITU-T Rec. P.910, *Subjective video quality assessment methods for multimedia applications*, ITU-T Recommendation, Apr. 2008.

[23] ITU-T Rec. BT.500-13, *Methodology for the subjective assessment of the quality of television pictures*, ITU-T Recommendation, Jan. 2012.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity", *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[25] H. R. Sheikh and A. C. Bovik, "Image information and visual quality", *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[26] Netflix, *Toward A Practical Perceptual Video Quality Metric*, https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652, [Online: Accessed 25-Jan-2019], Jun. 2016.

[27] B. Girod, "Digital images and human vision", in, A. B. Watson, Ed., Cambridge, MA, USA: MIT Press, 1993, ch. What's Wrong with Mean-squared Error?, pp. 207–220. [Online]. Available: http://dl.acm.org/citation.cfm?id=197765.197784.

[28] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments", *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[29] I. Katsavounidis and L. Guo, "Video codec comparison using the dynamic optimizer framework", *Proc. SPIE 10752, Applications of Digital Image Processing XLI*, 107520Q, Sep. 2018.

[30] R. Soundararajan and A. C. Bovik, "Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[31] A. C. Bovik, R. Soundararajan, and C. Bampis, *On the Robust Performance of the ST-RRED Video Quality Predictor*, http://live.ece.utexas.edu/research/Quality/ST-RRED/, [Online: Accessed 07-Dec-2018].

[32] C. G. Bampis, P. Gupta, R. Soudararajan, and A. Bovik, *Source code for optimized Spatio-Temporal Reduced Reference Entropy Differencing Video Quality Prediction Model*, http://live.ece.utexas.edu/research/Quality/STRRED_opt_demo.zip, [Online: Accessed 07-Dec-2018].

[33] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial Efficient Entropic Differencing for Image and Video Quality", *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[34] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain", *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[35] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer", *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[36] A. K. Moorthy and A. C. Bovik, "A Two-Step Framework for Constructing Blind Image Quality Indices", *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, May 2010.

[37] Matlab, *Train and Use a No-Reference Quality Assessment Model*, https://www.mathworks.com/help/images/train-and-use-a-no-reference-quality-assessment-model.html/, [Online: accessed 12-Aug-2019], 2018.

[38] ITU-T Rec. J.340, *Reference algorithm for computing peak signal to noise ratio of a processed video sequence with compensation for constant spatial shifts, constant temporal shift, and constant luminance gain and offset*, ITU-T Recommendation, Jun. 2010.

[39] Multimedia Signal Processing Group (MMSPG, EPFL), *VQMT: Video Quality Measurement Tool*, http://mmspg.epfl.ch/vqmt, [Online: Accessed 12-Feb-2019].

[40] Netflix, *VMAF - Video Multi-Method Assessment Fusion*, https://github.com/Netflix/vmaf, [Online: Accessed 12-Feb-2019].

[41] MathWorks, *Image Quality Metrics*, https://www.mathworks.com/help/images/image-quality-metrics.html, [Online: Accessed 17-Feb-2019], 2018.

[42] T. Hoßfeld, R. Schatz, M. Varela, and C. Timmerer, "Challenges of QoE management for Cloud Applications", *IEEE Communications Magazine*, vol. 50, no. 4, pp. 28–36, Apr. 2012.

[43] S. Baraković and L. Skorin-Kapov, "Survey and Challenges of QoE Management Issues in Wireless Networks", *Journal of Computer Networks and Communications*, vol. 2013, 165146:1–165146:28, Dec. 2013.

[44] W. Robitza, A. Ahmad, P. A. Kara, L. Atzori, M. G. Martini, A. Raake, and L. Sun, “Challenges of future multimedia QoE monitoring for internet service providers”, *Multimedia Tools and Applications*, pp. 1–24, Jun. 2017.

[45] A. Ahmad, L. Atzori, and M. G. Martini, “Qualia: A Multilayer Solution for QoE Passive Monitoring at the User Terminal”, in *IEEE International Conference on Communications (ICC)*, Paris, France, May 2017.

[46] *QUIC: A UDP-Based Secure and Reliable Transport for HTTP2*, https://tools.ietf.org/html/draft-tsvwg-quic-protocol-02, [Online: Accessed 07-Feb-2019], 2016.

[47] Y. Chen, K. Wu, and Q. Zhang, “From QoS to QoE: A Tutorial on Video Quality Assessment”, *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 1126–1165, Second quarter 2015.

[48] Video Quality Experts Group (VQEG), *VQEG FRTV Phase I Final Report*, https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx, 2000.

[49] Video Quality Experts Group (VQEG), *VQEG FRTV Phase II Final Report*, https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-ii/frtv-phase-ii.aspx, 2003.

[50] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[51] J. R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, “Comparison of the Coding Efficiency of Video Coding Standards-Including High Efficiency Video Coding (HEVC)”, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[52] F. Bossen, *Common Test Conditions and Software Reference Configurations for HM (JCTVC-L1100)*, ITU-T and ISO/IEC JTC 1 Joint Collaborative Team on Video Coding, Jan. 2013.

[53] L. Guo, J. D. Cock, and A. Aaron, “Compression Performance Comparison of x264, x265, libvpx and aomenc for On-Demand Adaptive Streaming Applications”, in *2018 Picture Coding Symposium (PCS)*, San Francisco, California, Jun. 2018, pp. 26–30.

[54] A. Zabrovskiy, C. Feldmann, and C. Timmerer, “A Practical Evaluation of Video Codecs for Large-Scale HTTP Adaptive Streaming Services”, in *2018 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, Oct. 2018, pp. 998–1002.

[55] J. S. Lee and T. Ebrahimi, “Perceptual video compression: A survey”, *IEEE Journal on Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 684–697, 2012.

[56] I. Himawan, W. Song, and D. Tjondronegoro, “Automatic region-of-interest detection and prioritisation for visually optimised coding of low bit rate videos”, in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, Tampa, FL, USA, Jan. 2013, pp. 76–82.

[57] S. Tavakoli, S. Egger, M. Seufert, R. Schatz, K. Brunnström, and N. García, “Perceptual Quality of HTTP Adaptive Streaming Strategies: Cross-Experimental Analysis of Multi-Laboratory and Crowdsourced Subjective Studies”, *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2141–2153, Aug. 2016.

[58] S. Hu, L. Sun, C. Gui, E. Jammeh, and I. Mkwawa, “Content-aware adaptation scheme for QoE optimized dash applications”, in *2014 IEEE Global Communications Conference*, Austin, TX, USA, Dec. 2014, pp. 1336–1341.

[59] H. T. Le, H. N. Nguyen, N. Pham Ngoc, A. T. Pham, H. Le Minh, and T. C. Thang, “Quality-driven bitrate adaptation method for HTTP live-streaming”, in *IEEE International Conference on Communication Workshop (ICCW)*, London, U.K., Jun. 2015, pp. 1771–1776.

[60] S. Wang, J. Fu, Y. Lu, S. Li, and W. Gao, “Content-aware layered compound video compression”, in *2012 IEEE International Symposium on Circuits and Systems*, Seoul, South Korea, May 2012, pp. 145–148.

[61] S. Hu, R. A. Cohen, A. Vetro, and C. C. J. Kuo, “Screen content coding for HEVC using edge modes”, in *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 1714–1718.

[62] M. G. Martini, “Wireless broadband multimedia health services: Current status and emerging concepts”, in *2008 IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, Cannes, France, Sep. 2008, pp. 1–6.

[63] J. Xue and C. W. Chen, “Mobile video perception: New insights and adaptation strategies”, *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 390–401, Jun. 2014.

[64] A. Aqil, A. O. F. Atya, S. V. Krishnamurthy, and G. Papageorgiou, “Streaming Lower Quality Video over LTE: How Much Energy Can You Save?”, in *IEEE 23rd International Conference on Network Protocols (ICNP)*, San Francisco, CA, USA, Nov. 2015, pp. 156–167.

[65] E. Liotou, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, “Enriching HTTP adaptive streaming with context awareness: A tunnel case study”, in *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[66] N. Barman, S. Valentin, and M. G. Martini, “Predicting link quality of wireless channel of vehicular users using street and coverage maps”, in *IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.

[67] P. C. Cosman, R. M. Gray, and R. A. Olshen, “Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy”, *Proceedings of the IEEE*, vol. 82, no. 6, pp. 919–932, Jun. 1994.

[68] M. Razaak, “Quality evaluation of medical ultrasound videos for e-health and telemedicine applications”, PhD thesis, 2016. [Online]. Available: http://eprints.kingston.ac.uk/id/eprint/35852.

[69] M. G. Martini and C. T. E. R. Hewage, “Flexible Macroblock Ordering for Context-aware Ultrasound Video Transmission over Mobile WiMAX”, *International Journal of Telemedicine and Applications*, vol. 2010, no. 127519, Jan. 2010.

[70] S.-C. Lin, Y.-C. Lin, W.-S. Feng, J.-M. Wu, and T.-J. Chen, “A Novel Medical Image Quality Index”, *Journal of Digital Imaging*, vol. 24, no. 5, pp. 874–882, Oct. 2011. [Online]. Available: https://doi.org/10.1007/s10278-010-9353-y.

[71] M. Razaak and M. G. Martini, “CUQI: cardiac ultrasound video quality index”, *Journal of Medical Imaging*, vol. 3, pp. 3–10, 2016. [Online]. Available: https://doi.org/10.1117/1.JMI.3.1.011011.

[72] The WebM Project, *VP9 Video Codec*, http://www.webmproject.org/vp9/, [Online: Accessed 20-Jan-2019].

[73] D. Fitzgerald and D. Wakabayashi, *Apple Quietly Builds New Networks*, https://www.wsj.com/articles/apple-quietly-builds-new-networks-1391474149, [Online: Accessed 07-January-2019], Feb. 2014.

[74] FFmpeg, *Encoding for streaming sites*, https://trac.ffmpeg.org/wiki/EncodingForStreamingSites, [Online: Accessed 10-Feb-2019].

[75] Twitch, *Twitch Streamers*, https://stream.twitch.tv/, [Online: Accessed 10-Feb-2019].

[76] J. De Cock, A. Mavlankar, A. Moorthy, and A. Aaron, “A large-scale video codec comparison of x264, x265 and libvpx for practical VOD applications”, *Proc. SPIE, Applications of Digital Image Processing XXXIX*, vol. 9971, no. 997116, Sep. 2016.

[77] M. Řeřábek and T. Ebrahimi, “Comparison of Compression Efficiency Between HEVC/H.265 and VP9 Based on Subjective Assessments”, *Proc. SPIE, Applications of Digital Image Processing XXXVII*, vol. 9217, 92170U, Sep. 2014.

[78] L. Mengzhe, J. Xiuhua, and L. Xiaohua, “Analysis of H.265/HEVC, H.264 and VP9 coding efficiency based on video content complexity”, in *2015 IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, China, Oct. 2015, pp. 420–424.

[79] J. Bienik, M. Uhrina, M. Kuba, and M. Vaculik, “Performance of H.264, H.265, VP8 and VP9 Compression Standards for High Resolutions”, in *2016 19th International Conference on Network-Based Information Systems (NBiS)*, Ostrava, Czech Republic, Sep. 2016, pp. 246–252.

[80] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, “Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders”, in *2013 Picture Coding Symposium (PCS)*, San Jose, California, USA, Dec. 2013, pp. 394–397.

[81] D. Grois, T. Nguyen, and D. Marpe, “Coding efficiency comparison of AV1/VP9, H.265/MPEG-HEVC, and H.264/MPEG-AVC encoders”, in *2016 Picture Coding Symposium (PCS)*, Nuremberg, Germany, Dec. 2016, pp. 1–5.

[82] D. Grois, D. Marpe, T. Nguyen, and O. Hadar, “Comparative assessment of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders for low-delay video applications”, *Proc. SPIE 9217, Applications of Digital Image Processing XXXVII*, 92170Q, Sep. 2014.

[83] M. Řeřábek, P. Hanhart, P. Korshunov, and T. Ebrahimi, “Quality evaluation of HEVC and VP9 video compression in real-time applications”, in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, Pylos-Nestoras, Greece, May 2015, pp. 1–6.

[84] M. Claypool, D. Farrington, and N. Muesch, “Measurement-based analysis of the video characteristics of Twitch.tv”, in *2015 IEEE Games Entertainment Media Conference (GEM)*, Toronto, Canada, Oct. 2015, pp. 1–4.

[85] *FFmpeg,* https://ffmpeg.org/, [Online: Accessed 11-Feb-2019].

[86] G. Bjontegaard, *Calculation of average PSNR differences between RD-curves (VCEG-M33)*, ITU-T Video Coding Experts Group (April 2001).

[87] X. Ma, C. Zhang, J. Liu, R. Shea, and D. Fu, “Live broadcast with community interactions: Bottlenecks and optimizations”, *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1184–1194, Jun. 2017.

[88] Xiph.org, *Xiph.org Video Test Media [derf's collection]*, https://media.xiph.org/video/derf/, [Online: Accessed 12-Feb-2019].

[89] A. Ostaszewska and R. Kłoda, “Quantifying the amount of spatial and temporal information in video test sequences”, in *Recent Advances in Mechatronics.* Springer Berlin Heidelberg, 2007, pp. 11–15.

[90] M. Claypool, “Motion and scene complexity for streaming video games”, in *Proceedings of the 4th International Conference on Foundations of Digital Games*, ACM, Orlando, Florida, USA, 2009, pp. 34–41.

[91] S. Schmidt, S. Möller, and S. Zadtootaghaj, “A comparison of interactive and passive quality assessment for gaming research”, in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Sardinia, Italy, May 2018, pp. 1–6.

[92] S. Möller, S. Schmidt, and S. Zadtootaghaj, “New ITU-T Standards for Gaming QoE Evaluation and Management”, in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Sardinia, Italy, May 2018, pp. 1–6.

[93] S. Zadtootaghaj, S. Schmidt, N. Barman, S. Möller, and M. G. Martini, “A Classification of Video Games based on Game Characteristics linked to Video Coding Complexity”, in *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, Amsterdam, Netherlands, Jun. 2018, pp. 1–6.

[94] S. Rimac-Drlje, M. Vranjes, and D. Zagar, “Influence of Temporal Pooling Method on the Objective Video Quality Evaluation”, in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Bilbao, Spain, May 2009, pp. 1–5.

[95] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, “To Pool or not to Pool: A Comparison of Temporal Pooling Methods for HTTP Adaptive Video Streaming”, in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt, Austria, Jul. 2013, pp. 52–57.

[96] G. Sperling, “Temporal and Spatial Visual Masking. I. Masking by Impulse Flashes”, *Journal of the Optical Society of America*, vol. 55, no. 5, pp. 541–559, May 1965.

[97] L. Choi and A. Bovik, “Video quality assessment accounting for temporal visual masking of local flicker”, *Signal Processing: Image Communication*, vol. 67, pp. 182–198, 2018.

[98] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video", *IEEE transactions on image processing*, vol. 19, no. 6, pp. 1427–1441, 2010.

[99] M. Slanina and V. Ricny, "Estimating PSNR without reference for real H.264/AVC sequence intra frames", in *2008 18th International Conference Radioelektronika*, Prague, Czech Republic, Apr. 2008, pp. 1–4.

[100] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment", *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[101] X. Jiang, F. Meng, J. Xu, and W. Zhou, "No-Reference Perceptual Video Quality Measurement for High Definition Videos Based on an Artificial Neural Network", in *2008 International Conference on Computer and Electrical Engineering*, Phuket, Thailand, Dec. 2008, pp. 424–427.

[102] J. Choe, K. Lee, and C. Lee, "No-reference video quality measurement using neural networks", in *2009 16th International Conference on Digital Signal Processing*, Santorini-Hellas, Greece, Jul. 2009, pp. 1–4.

[103] A. Khan, L. Sun, and E. Ifeachor, "Learning models for video quality prediction over wireless local area network and universal mobile telecommunication system networks", *IET Communications*, vol. 4, no. 12, pp. 1389–1403, Aug. 2010.

[104] M. Shahid, A. Rossholm, and B. Lövström, "A reduced complexity no-reference artificial neural network based video quality predictor", in *2011 4th International Congress on Image and Signal Processing*, vol. 1, Shanghai, China, Oct. 2011, pp. 517–521.

[105] C. Wang, X. Jiang, F. Meng, and Y. Wang, "Quality assessment for MPEG-2 video streams using a neural network model", in *IEEE 13th International Conference on Communication Technology*, Jinan, China, Sep. 2011, pp. 868–872.

[106] W. Cherif, A. Ksentini, and D. Négru, "No-reference Quality of Experience estimation of H264/SVC stream", in *IEEE Globecom Workshops*, Anaheim, CA, USA, Dec. 2012, pp. 1346–1351.

[107] H. E. Khattabi, D. Aboutajdine, and A. Tamtaoui, "Predict the MOS and the PSNR by the neural network", in *International Conference on Multimedia Computing and Systems*, Tangier, Morocco, May 2012, pp. 418–421.

[108] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of Experience Estimation for Adaptive HTTP/TCP Video Streaming using H.264/AVC", in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, NV, USA, Jan. 2012, pp. 127–131.

[109] M. Narwaria, W. Lin, and A. Liu, "Low-Complexity Video Quality Assessment Using Temporal Quality Variations", *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 525–535, Jun. 2012.

[110] S. L. Sunala and P. R. Anurenjan, "A novel video quality measurement using ANN", in *Annual International Conference on Emerging Research Areas and 2013 International Conference on Microelectronics, Communications and Renewable Energy*, Kanjirapally, India, Jun. 2013, pp. 1–4.

[111] Y. Kang, H. Chen, and L. Xie, “An artificial-neural-network-based QoE estimation model for Video streaming over wireless networks”, in *2013 IEEE/CIC International Conference on Communications in China (ICCC)*, Xi’an, China, Aug. 2013, pp. 264–269.

[112] Y. Xue, B. Erkin, and Y. Wang, “A Novel No-Reference Video Quality Metric for Evaluating Temporal Jerkiness due to Frame Freezing”, *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 134–139, Jan. 2015.

[113] M. T. Vega, E. Giordano, D. C. Mocanu, D. Tjondronegoro, and A. Liotta, “Cognitive no-reference video quality assessment for mobile streaming services”, in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, Pylos-Nestoras, Greece, May 2015, pp. 1–6.

[114] K. Zheng, X. Zhang, Q. Zheng, W. Xiang, and L. Hanzo, “Quality-of-experience assessment and its application to video services in lte networks”, *IEEE Wireless Communications*, vol. 22, no. 1, pp. 70–78, Feb. 2015.

[115] M. Juayek and R. Sotelo, “An Artificial Neural Network approach for No-Reference High Definition Video quality assessment”, in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Nara, Japan, Jun. 2016, pp. 1–3.

[116] J. Guo, K. Zheng, G. Hu, and L. Huang, “Packet layer model of HEVC wireless video quality assessment”, in *11th International Conference on Computer Science Education (ICCSE)*, Nagoya, Japan, Aug. 2016, pp. 712–717.

[117] C. Wang, L. Su, and Q. Huang, “CNN-MR for No Reference Video Quality Assessment”, in *4th International Conference on Information Science and Control Engineering (ICISCE)*, Changsha, China, Jul. 2017, pp. 224–228.

[118] M. T. Vega, D. C. Mocanu, J. Famaey, S. Stavrou, and A. Liotta, “Deep Learning for Quality Assessment in Live Video Streaming”, *IEEE Signal Processing Letters*, vol. 24, no. 6, pp. 736–740, Jun. 2017.

[119] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, “Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment”, *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[120] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, “End-to-end blind image quality assessment using deep neural networks”, *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[121] W. Liu, Z. Duanmu, and Z. Wang, “End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks”, in *Proceedings of the 26th ACM International Conference on Multimedia*, ser. MM ’18, Seoul, Republic of Korea: ACM, 2018, pp. 546–554. [Online]. Available: http://doi.acm.org/10.1145/3240508.3240643.

[122] Y. Pitrey, U. Engelke, M. Barkowsky, R. Pépion, and P. L. callet, “Aligning subjective tests using a low cost common set”, in *Euro ITV*, pp.irccyn contribution, Lisbonne, Portugal, Jun. 2011.

[123] L. Anegekuh, L. Sun, and E. Ifeachor, “End to end video quality prediction for HEVC video streaming over packet networks”, in *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, Sharjah, United Arab Emirates, Feb. 2013, pp. 1–6.

[124] G. Cermak, M. Pinson, and S. Wolf, "The Relationship Among Video Quality, Screen Resolution, and Bit Rate", *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 258–262, Jun. 2011.

[125] *AGH Video Quality Indicators*, http://vq.kt.agh.edu.pl/metrics.html, Last Accessed: 15 January 2019.

[126] M. Leszczuk, M. Hanusiak, I. Blanco, A. Dziech, J. Derkacz, E. Wyckens, and S. Borer, "Key indicators for monitoring of audiovisual quality", in *22nd Signal Processing and Communications Applications Conference (SIU)*, Trabzon, Turkey, Apr. 2014, pp. 2301–2305.

[127] M. Leszczuk, M. Hanusiak, M. C. Q. Farias, E. Wyckens, and G. Heston, "Recent developments in visual quality monitoring by key performance indicators", *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10 745–10 767, Sep. 2016. [Online]. Available: https://doi.org/10.1007/s11042-014-2229-2.

[128] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media, 2012, vol. 454.

[129] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011. [Online]. Available: https://books.google.co.uk/books?id=bDtLM8CODsQC.

[130] G. Chandrashekar and F. Sahin, "A survey on feature selection methods", *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[131] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of machine learning research*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968.

[132] R. Duin, "PRTools - Version 3.0 - A Matlab Toolbox for Pattern Recognition", version 3.0, Jan. 2000, Pattern Recognition Group, Delft University of Technology.

[133] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278.

[134] M. T. Vega, D. C. Mocanu, S. Stavrou, and A. Liotta, "Predictive no-reference assessment of video quality", *Signal Processing: Image Communication*, vol. 52, pp. 20–32, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092359651630176X.

[135] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[136] C. E. Rasmussen, "Gaussian Processes in Machine Learning", in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 63–71. [Online]. Available: https://doi.org/10.1007/978-3-540-28650-9_4.

[137] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324.

[138] M. A. Hall, “Correlation-based Feature Subset Selection for Machine Learning”, PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1998.

[139] V. Menkovski, A. Oredope, A. Liotta, and A. C. Sánchez, “Predicting quality of experience in multimedia streaming”, in *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*, ACM, Kuala Lumpur, Malaysia, 2009, pp. 52–59.

[140] A. Bouzerdoum, A. Havstad, and A. Beghdadi, “Image quality assessment using a neural network approach”, in *Proceedings of the Fourth IEEE International Symposium on Signal Processing and Information Technology, 2004.*, Rome, Italy, Dec. 2004, pp. 330–333.

[141] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, “Developing a predictive model of quality of experience for internet video”, in *ACM SIGCOMM Computer Communication Review*, ACM, vol. 43, 2013, pp. 339–350.

[142] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Möller, “NR-GVQM: A No Reference Gaming Video Quality Metric”, in *2018 IEEE International Symposium on Multimedia (ISM)*, Taichung, Taiwan, Dec. 2018, pp. 131–134.

[143] S. Göring, R. R. R. Rao, and A. Raake, “nofu — A Lightweight No-Reference Pixel Based Video Quality Model for Gaming Content”, in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2019, pp. 1–6.