

Temporal Video Quality Prediction Using Multi-Way Data Analysis

Clemens Horch



Diploma Thesis

Temporal Video Quality Prediction Using Multi-Way Data Analysis

Clemens Horch

15. March 2013



Institute for Data Processing
Technische Universität München



Clemens Horch. *Temporal Video Quality Prediction Using Multi-Way Data Analysis*.
Diploma Thesis, Technische Universität München, Munich, Germany, 2013.

Supervised by Prof. Dr.-Ing. Klaus Diepold and Dipl.-Ing. Christian Keimel; submitted
on 15. March 2013 to the Department of Electrical Engineering and Information Tech-
nology of the Technische Universität München.

© 2013 Clemens Horch

Institute for Data Processing, Technische Universität München, 80290 München, Ger-
many, <http://www.ldv.ei.tum.de>.

This work is licenced under the Creative Commons Attribution 3.0 Germany License. To
view a copy of this licence, visit <http://creativecommons.org/licenses/by/3.0/de/>

Abstract

Currently the most reliable method of determining the visual quality of video sequences is subjective testing with human observers. Since this is both time-consuming and expensive, there is ongoing research on objective video quality metrics that allow to measure the visual quality without carrying out subjective tests. A very promising approach in the field of no-reference video quality metrics are metrics based on data analysis rather than on modeling the human visual system. It has been shown that the results improve when taking the temporal structure of the video sequence into account using multi-way data analysis methods.

In this thesis I show a way of refining multi-way quality metrics based on the extraction of H.264/AVC bitstream features by considering the internal GOP-structure of the coded video. Normally, metrics based on multi-way data analysis require both the sequences in the training set and the unknown sequence whose quality is to be predicted to have exactly the same length. By splitting the video sequences into their GOPs, the metric can be made length-independent and therefore more suitable for real-life applications while at the same time maintaining the performance.

Furthermore, the GOP-based quality metric can be used to predict the temporal progression of the visual quality. In order to evaluate this quality estimation, I present a new subjective test method that allows the assessment of quality fluctuations in short video sequences since existing methods like SSCQE are not suitable for this task. Instead of directly asking the observers about the continuous quality, I ask them about their overall quality impression, their impression of the quality fluctuation strength, and give a choice of different patterns that represent possible shapes of quality curves. From the answers to these three questions, the curve of the temporal quality can be reconstructed. Finally, I compare the results of this subjective test to the quality estimations of the GOP-based metric. It turns out that the results of both quality measurements are highly correlated.

Zusammenfassung

Die derzeit zuverlässigste Methode, die visuelle Qualität von Videosequenzen zu bestimmen, ist die Durchführung subjektiver Tests mit Testpersonen. Da dies sowohl zeitaufwändig als auch teuer ist, gibt es laufend Forschung zu objektiven Videoqualitätsmetriken, die die Messung von Videoqualität ohne subjektive Tests ermöglichen. Ein vielversprechender Ansatz im Bereich der No-Reference Videoqualitätsmetriken sind Metriken, die auf Datenanalyse und nicht auf der Modellierung des menschlichen Sehsystems (Human Visual System, HVS) basieren. Es konnte gezeigt werden, dass sich die Ergebnisse verbessern, wenn man die zeitliche Struktur einer Videosequenz mittels multi-way Datenanalysemethoden mit in die Modellierung einbezieht.

In dieser Arbeit zeige ich, wie man multi-way Qualitätsmetriken, die auf der Extraktion von Merkmalen aus H.264/AVC-Bitstreams basieren, verbessern kann, indem man die interne GOP-Struktur des kodierten Videos berücksichtigt. Normalerweise setzen Metriken, die auf multi-way Datenanalysemethoden basieren, voraus, dass die Sequenzen im Training-Set und die Sequenz deren Qualität zu messen ist, die gleiche Länge besitzen. Teilt man die Videosequenzen in ihre GOPs auf, kann man die Metrik längenunabhängig machen und damit ihre Anwendbarkeit auf reale Probleme verbessern, ohne dabei ihre Leistung zu verschlechtern.

Darüber hinaus können GOP-basierte Qualitätsmetriken verwendet werden, um den zeitlichen Verlauf der visuellen Qualität zu bestimmen. Um diese Qualitätsschätzung zu untersuchen, stelle ich eine neue subjektive Testmethode vor, die es erlaubt, Qualitätsschwankungen in kurzen Videosequenzen zu messen, da bereits existierende Methoden wie SSCQE in diesem Fall nicht ausreichend gut funktionieren. Anstatt die Testpersonen direkt nach ihrer Einschätzung der zeitkontinuierlichen Qualität zu fragen, stelle ich drei Fragen: eine nach der Qualität des Videos insgesamt, eine nach ihrer Einschätzung der Schwankungsstärke der Qualität und eine nach dem groben Verlauf der Qualität anhand von verschiedenen vorgegebenen Optionen. Aus den Antworten auf diese Fragen lässt sich der zeitliche Verlauf der Qualität rekonstruieren. Abschließend vergleiche ich die Ergebnisse aus diesem Test mit der Qualitätsschätzung der GOP-basierten Metrik. Es zeigt sich, dass die Ergebnisse beider Verfahren hoch korreliert sind.

Contents

List of Abbreviations	9
List of Symbols	11
1. Introduction	13
1.1. Visual Video Quality	13
1.1.1. Subjective Video Quality Assessment	13
1.1.2. Continuous Quality Assessment	14
1.2. Objective Video Quality Metrics	15
1.2.1. Objective Video Quality Metrics	15
1.2.2. Metrics Using Multi-Way Data Analysis	16
2. Design of GOP-based Video Quality Metrics	19
2.1. H.264/AVC Bitstream Feature Extraction	19
2.2. Two-Way Regression Analysis	21
2.2.1. Data Pre-Processing	21
2.2.2. Multiple Linear Regression (MLR)	22
2.2.3. Principal Component Regression (PCR)	23
2.2.4. Bilinear Partial Least Squares Regression (PLS1)	24
2.3. Three-Way Regression Analysis	25
2.3.1. Two-Dimensional Principal Component Regression (2D-PCR)	25
2.3.2. Trilinear Partial Least Squares Regression (Tri-PLS1)	27
2.4. Video Quality Prediction	28
2.4.1. GOP-based Temporal Quality Prediction	28
2.4.2. Length-Independent Quality Prediction	30
2.5. Data Post-Processing: Sigmoid Correction	31
3. Evaluation of GOP-based Video Quality Metrics	33
3.1. Evaluation Methodology	33
3.1.1. Statistical Performance Measures	33
3.1.2. Cross Validation	34
3.1.3. Full-Reference Metrics for Comparison	34
3.2. Evaluation of Length-Independent Quality Prediction	36
3.2.1. Datasets Used for Evaluation	36
3.2.2. Performance Evaluation	40
3.3. Evaluation of Temporal Quality Prediction	43
3.3.1. Design of the Subjective Test	43

Contents

3.3.2. Reconstruction of Temporal Quality Progression	46
3.3.3. Choice of Video Sequences	47
3.3.4. Results of the Subjective Test	48
3.3.5. Results of the Temporal Quality Prediction	52
4. Summary	57
A. Additional Tables and Figures	59
A.1. Results of the Subjective Test	60
A.2. Additional Scatter Plots	63
A.2.1. Length-Independent Quality Prediction	63
A.2.2. Temporal Quality Prediction	66
A.3. Plots of the Temporal Quality Prediction	67
List of Figures	73
List of Tables	75
Bibliography	77

List of Abbreviations

2D-PCR	Two-Dimensional Principal Component Regression
AVC	Advanced Video Codec
CIF	Common Intermediate Format
DCR	Degradation Category Rating
DSCQS	Double Stimulus Continuous Quality Scale
DSCS	Double Stimulus Comparison Scale
DSUR	Double Stimulus Unknown Reference
DVB	Digital Video Broadcast
fps	frames per second
GOP	Group of Pictures
HC	High Complexity
HD	High Definition
HDTV	High Definition Television
HVS	Human Visual System
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
IT-IST	Instituto de Telecomunicações, Instituto Superior Técnico
ITU	International Telecommunication Union
JVT	Joint Video Team
LC	Low Complexity
LDV	Institute for Data Processing (Lehrstuhl für Datenverarbeitung)
MSE	Mean Square Error
MLR	Multiple Linear Regression
MOS	Mean Opinion Score
MPEG	Motion Pictures Experts Group
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS1	Bilinear Partial Least Squares Regression

Contents

PLSR	Partial Least Squares Regression
PSNR	Peak Signal-to-Noise Ratio
QP	Quantization Parameter
RMSE	Root Mean Square Error
SD	Standard Definition
SNR	Signal-to-Noise Ratio
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity
SSIS	Single Stimulus Impairment Scale
SSMM	Single Stimulus MultiMedia
SVD	Singular Value Decomposition
SVT	Sveriges Television (Sweden's Television)
Tri-PLS1	Trilinear Partial Least Squares Regression
TUM	Technische Universität München
UHD	Ultra High Definition
VCEG	Video Coding Experts Group
VCL-NALU	Video Coding Layer Network Abstraction Layer Unit
VQEG	Video Quality Experts Group

List of Symbols

a	a scalar
\mathbf{a}	a column vector
\mathbf{a}^\top	a row vector
\mathbf{A}	a matrix (or two-way array)
\mathbf{A}^\top	a transposed matrix
$\underline{\mathbf{A}}$	a three-way array (or multi-way array)
$\bar{\mathbf{a}}, \bar{\mathbf{A}}$	arithmetic mean of the elements in \mathbf{a} or \mathbf{A}
$\check{\mathbf{a}}, \check{\mathbf{A}}, \check{\underline{\mathbf{A}}}$	centered version of \mathbf{a} , \mathbf{A} , or $\underline{\mathbf{A}}$ (cf. 2.2.1 on page 21)
$\hat{\mathbf{A}}, \hat{\underline{\mathbf{A}}}$	autoscaled version of \mathbf{A} or $\underline{\mathbf{A}}$ (cf. 2.2.1 on page 21)
$\hat{a}, \hat{\mathbf{a}}, \hat{\mathbf{A}}, \hat{\underline{\mathbf{A}}}$	estimate of a , \mathbf{a} , \mathbf{A} , or $\underline{\mathbf{A}}$
N	number of videos sequences (1 st mode)
M	number of features (2 nd mode)
T	number of frames per sequence (3 rd mode)
G	GOPs per sequence
T_G	frames per GOP = $\frac{T}{G}$
R	number of components used in a regression model
S	number of participants in a subjective test
n, m, t, g, r, s	running indices for N, M, T, G, R, S
$\mathbf{X}, \underline{\mathbf{X}}$	feature matrix or multi-way array
\mathbf{y}	vector of subjective test results
$\mathbf{X}_u, \underline{\mathbf{X}}_u$	feature matrix of multi-way array of unknown sequences
$\mathbf{P}, \underline{\mathbf{P}}$	loadings matrix or multi-way array
$\mathbf{T}, \underline{\mathbf{T}}$	score matrix or multi-way array
\mathbf{X}_{Sct}	scatter or covariance matrix
$\mathbf{b}, \underline{\mathbf{B}}$	weight vector or multi-way array

1. Introduction

At the beginning of this thesis I want to outline the importance of video quality assessment in general and the shortcomings of popular quality metrics. This is at the same time the motivation for refining existing video quality metrics.

1.1. Visual Video Quality

Since the beginnings of processing moving pictures, it has always been of major importance to improve the visual quality of the image reproduction. Nowadays, most of the video recording, processing, transmission, and playback take place in the digital domain and the fields of application are almost uncountable – they reach from the more traditional forms like cinema or television to the emerging technologies of the last decades such as video streaming or mobile communications. Unlike in the times of analog processing, the transmission and storage of video themselves are no longer an issue for the visual quality of video; however, both storage capacities and transmission bandwidths now require the use of data compression techniques. Video compression can be seen as a key technology for many applications. Without nowadays' lossy video coding almost every application of digital video would still be unimaginable. Whereas the capacity of storage media and the bandwidth of internet connections are still growing fast, just like ten years ago, transmitting and storing uncompressed video data remains unfeasible. Of course, the fastest available consumer internet connections at the moment are in theory capable of transmitting uncompressed video data in real time – but only in Standard Definition (SD), which was state-of-the-art more than ten years ago. For example 576p (720×576 , 25 fps) video data with 4:2:0 chroma subsampling has a bit rate of 98,88 MBit/s. Currently, private internet connections with 100 MBit/s or even more are available in some urban areas. Nevertheless, today, television is already broadcast in HD resolution and the first 4K UHD displays with a resolution of 3840×2160 have already been announced. So as long as the image resolutions keep growing, there will always be a strong need for video compression, especially when thinking of wireless transmission for example via cellular networks.

1.1.1. Subjective Video Quality Assessment

On the one hand, lossy video compression enables many applications, on the other hand it naturally comes with quality degradations, which make it very important to think about visual video quality and its measurement in detail. From imperfect cameras with visible sensor noise, over a digital transmission that may result in the loss of data packets, to an uncalibrated monitor that the video is displayed on: there are many factors besides

1. Introduction

lossy coding that contribute to a possible loss in video quality. When investigating visual quality of digital video these factors are often eliminated by using high-end camera and display technologies and avoiding transmission errors.

The standard procedure for determining the visual quality of video material is a subjective test. Winkler [47, p. 51] states that “[s]ubjective experiments represent the benchmark for vision models in general and quality metrics in particular”. For the assessment of visual quality with subjective tests there are several standards that formalize the viewing conditions and test procedures. This helps to create reproducible and meaningful results. The oldest standard is the ITU-R Recommendation BT.500 “Methodology for the subjective assessment of the quality of television pictures” [12]. Its first version is from 1974, and currently ITU-R Rec. BT.500-13 from 2012 is in force. Whereas ITU-R Rec. BT.500 focuses on television, ITU-T Rec. P.910 “Subjective video quality assessment methods for multimedia applications” [15] from 1996 is meant for multimedia applications. Both standards define different testing procedures that are applied in different situations and there are even more standards for related fields like audiovisual quality assessment.

1.1.2. Continuous Quality Assessment

Most of the subjective test methods that are defined by the standards are so-called single-rating methods. That means that the test subject watches a video sequence, typically about 10 seconds long, and rates the visual quality on a rating scale. There are differences in the design of the rating scales and some methods include displaying an undistorted reference sequence, but essentially all methods ask for the overall quality impression of a short video sequence. While this is fine in cases where the visual quality is constant during the test sequences, problems can occur when longer sequences are to be evaluated. Aldridge et al. [1] presented 30 seconds of video in a subjective test and found out that the observers were strongly influenced by the quality of the last 10 seconds of each sequence. So it does not make sense to use sequences much longer than 10 seconds. Nevertheless, as Winkler [47, p. 54] states, the assessment of short video sequences is in many cases not sufficient:

All single-rating methods [...] share a common drawback [...]: changes in scene complexity, statistical multiplexing or transmission errors can produce substantial quality variations that are not evenly distributed over time; severe degradations may appear only once every few minutes.

To cover such cases, the ITU-R Recommendation BT.500 proposes the *Single Stimulus Continuous Quality Evaluation* (SSCQE) method. Instead of rating the quality after watching a sequence, the viewers use a slider to express their perception while watching the video sequence. Typically the displayed sequence is about 20–30 minutes long. On the one hand, this method allows to determine the temporal dimension of visual quality; on the other hand, the results are more complicated to analyze, for example because of different reaction times or context effects [47]. Pinson and Wolf [36] compared SSCQE

to the single-rating methods *Double Stimulus Continuous Quality Scale* (DSCQS) and *Double Stimulus Comparison Scale* (DSCS) from ITU-R Rec. BT.500. They found that SSCQE is capable of producing results similar to DSCQS and DSCS when the test is designed properly. Nevertheless, they also state that on average the observers need 6 seconds to adapt the slider position to a new quality level; so statements about quick quality changes seem to be difficult to obtain with SSCQE.

Gauss et al. [8] give an example for the application of SSCQE for determining temporal quality changes of video sequences. The distortions were caused by packet loss and the individual video sequences of 30 seconds each were combined to a 30 minute program. Although their experiment proved to be suitable for their purposes, they had to deal with several problems regarding the accuracy of the test results. Only after an advanced selection process in which the data from over half of all observers were discarded, the results fulfilled their expectations.

So apparently, there is no subjective method that allows to assess the temporal progression of visual quality at a high sampling rate. This motivates the development of a new subjective quality assessment method that enables the evaluation of quick changes in visual quality in short video sequences. This new method will be explained in section 3.3.1 on page 43.

1.2. Objective Video Quality Metrics

Although subjective testing is the most reliable method for determining video quality, there is a strong need for objective quality assessment. Subjective video tests are both time-consuming, and costly and it is nearly impossible to use subjective tests as quality assurance in a professional production workflow.

1.2.1. Objective Video Quality Metrics

Objective video quality metrics try to estimate the visual quality from the video data as a human observer would perceive it. These metrics can be classified into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR) metrics [45].

Full-reference metrics use both the distorted video data and the data of the undistorted reference as input. Obviously, this is the optimal case for a quality metric as it has the most data at its disposal. The most popular FR metric is probably the Peak Signal-to-Noise Ratio (PSNR), although its results are not satisfying at all. The reasons why it is still so wide-spread are mainly its simplicity and computational performance [45]. Another very popular FR metric is the Structural Similarity (SSIM) index [44], which outperforms the PSNR in most cases. There are many more full-reference metrics, some of them are also available as commercial products. The more advanced methods try to model the Human Visual System (HVS) in order to simulate the human perception. [45, 47] offer an overview of those techniques. The *Multimedia Phase I* and *HDTV Phase I* projects of the Video Quality Experts Group (VQEG) tried to evaluate objective video quality metrics and decided to standardize several FR metrics. The resulting recommendations are for example the ITU-T Recommendation J.247 [14] and the ITU-T

1. Introduction

Recommendation J.341 [16]. A big advantage of FR metrics is also their independence of technologies such as coding because they are based on the raw pixel data.

The class of reduced-reference quality metrics does not require the full pixel data of the undistorted reference. It rather relies on reference features that are extracted from the reference video. This is useful in case of transmission over a channel with limited bandwidth. The VQEG also developed standards for RR metrics; namely ITU-T recommendation J.246 [13] and ITU-T Recommendation J.342 [17]. There are publications that propose reference metrics that try to estimate continuous video quality as well. For example, Masry and Hemami [31, 32] use a wavelet transform-based approach to model the temporal properties of the HVS.

The third class of the no-reference metrics is very interesting because of its high flexibility. In many applications there is no undistorted reference available and therefore no-reference metrics are the only possible option. NR metrics often depend on a special coding technology. For example, metrics based on the measurement of blocking artifacts (as described by Wang et al. [43]) can only be used on video data that has been coded with a block-based coding technology. Even more specialized are metrics based on H.264/AVC, like for example [7, 40, 41] or the work by Keimel et al. [20, 21, 24], the latter of which I will improve in this thesis. In the field of NR metrics there are attempts to measure temporal quality as well: Kawayoke and Horita [19] presented one example of such a metric.

1.2.2. Metrics Using Multi-Way Data Analysis

The Data Analysis Approach

The traditional approach for designing objective video quality metrics tries to model the HVS in order to reproduce the perception of a human observer. Examples for such metrics are the SSIM [44], the quality metrics by Wolf and Pinson [48] or Watson [46]. A general problem of this approach is that it requires a sufficient understanding of the HVS, which we currently do not possess: “In general, the development of computational HVS-models itself is still in its infancy” [47, p. 151]. An alternative to understanding the HVS in even more detail is the so-called data-driven approach, which does not require this knowledge.

Data analysis methods are very common in the field of chemometrics, an introduction to data analysis as used in this science can be found in Martens and Martens [29]. Transferring this approach to the domain of video quality, the general idea is to extract as many feature data as possible from the video sequence for evaluation. This feature data is then used to train a regression model with the help of ground-truth data from subjective quality tests. After the training, the model is able to predict the quality of an unknown sequence. There have already been some contributions using a data analysis approach to build a no-reference video quality metric. Principal Component Regression (PCR) as a common regression method was first used in the design of a video quality metric by Miyahara [34].

H.264/AVC Bitstream Features

Some very promising metrics are based on the extraction of H.264/AVC bitstream features from the video data [20, 21, 24]. Features extracted directly from the H.264/AVC bitstream have been used by Eden [7] to estimate the PSNR of interlaced HDTV video or Slanina et al. [41], who estimate the PSNR of video sequences in CIF resolution. Rossholm and Lövsström [40] used bitstream features to estimate some other quality metrics additionally to the PSNR. These approaches allow to determine the full-reference value PSNR with a no-reference method. If one is interested in the visual quality, this is only helpful to a certain extent because – as discussed above – the PSNR does not correlate very well with the perceived quality of a human observer. Therefore, Keimel et al. [20, 21, 24] refined the usage of bitstream features in order to directly estimate the visual video quality.

H.264/AVC¹ [11, 18] is probably the most popular and most widely used video coding technology in the world at the moment. It is used for coding the video on BluRay disks, for the television broadcast via satellite, cable or terrestrial transmitters according to DVB and other standards, and is also the most common codec for internet streaming. For detailed information about the functionality of H.264/AVC refer to the books by Richardson [38, 39].

The usage of features extracted directly from the H.264/AVC bitstream is very convenient because no decoding or other format conversion steps are involved. While encoding video using H.264/AVC, the encoder performs many different calculations on the temporal and spatial structure of the video content in order to decide how to encode the data. Many of these decisions can be reconstructed from the resulting bitstream and since they affect the visual quality fundamentally, it seems to be reasonable to use this data as feature data for the prediction of visual quality.

Extension by Multi-Way Data Analysis

Recently, Keimel et al. [24] have shown that the inclusion of the temporal dimension by using multi-way data analysis further improves the prediction. Instead of temporally pooling the feature data of all frames of a video sequence, the complete set of feature data can be processed by using a multi-way regression method. One example for such a method is the two-way version of principal component regression 2D-PCR, which was proposed by Yang et al. [49] in order to improve face recognition algorithms. Examples for the use of 2D-PCR for video quality metrics can be found in Keimel et al. [24, 27]. Besides PCR, Partial Least Squares Regression (PLSR) can also be used to design quality metrics [22, 23] as well as its extension, the multilinear PLSR as introduced by Bro [5]. The latter is demonstrated in [20].

¹The H.264/AVC development was started by the working group Video Coding Experts Group (VCEG) of the International Telecommunication Union (ITU). In 2001 they joined forces with the ISO/IEC Motion Pictures Experts Group (MPEG) and formed the Joint Video Team (JVT). In 2003 both groups published the standard: ITU as Recommendation H.264 [18] and MPEG as ISO/IEC 14496-10 MPEG-4 Part 10, Advanced Video Codec (AVC) [11]. For this reason I call the codec H.264/AVC in this thesis.

1. Introduction

The price that has to be paid by including the temporal dimension into the quality prediction model is that this requires all training sequences to consist of the same number of frames, and also the sequence whose quality is to be predicted needs to match this length. The reason for this is mathematical and will be explained in section 2.4.2 on page 30. This problem impedes the application of the metric in most fields as it is not feasible to train different regression models for each occurring length of video sequences.

The main goal of my thesis is to extend these no-reference data analysis-based metrics using H.264/AVC bitstream features in order to support the prediction of temporal quality. As it turns out, this will also solve also the problem of the length-dependence of metrics using multi-way data analysis.

2. Design of GOP-based Video Quality Metrics

In this chapter I will start with a brief explanation of the H.264/AVC bitstream features that will act as input data for the video quality estimation. Then, I will explain one-way and two-way regression methods that will be the mathematical foundation of the proposed quality metrics. Finally, I will point out how to design video quality metrics using these methods and in what ways they are an improvement on existing metrics.

2.1. H.264/AVC Bitstream Feature Extraction

The H.264/AVC bitstream features are extracted from the video data by a modified version of the *H.264/AVC JM Reference Software* developed by Klompke et al. [28]. In what follows, the $M = 17$ used bitstream features are outlined. In general, the features are calculated on a per-slice basis. The H.264/AVC standard [11, 18] allows frames to be partitioned into several slices. Since this option is not very common and not used in the videos selected for this thesis, from here on the terms *slice* and *frame* are used synonymously. For further details regarding the bitstream feature extraction refer to [21, 28, 40].

Slice Features

<i>Slice Type</i>	Each slice is either an <i>I</i> -, <i>P</i> -, or <i>B-Slice</i> . This information is mapped to integer numbers as follows: $I \rightarrow 0$, $P \rightarrow 1$ and $B \rightarrow 2$.
<i>kBit</i>	This feature contains the size of the VCL-NALU in kilobits.
QP_{avg}	Each slice has an initial QP which may be altered on the macroblock level. QP_{avg} contains the average QP for the current slice.
ΔQP_{avg}	This is the average of the differences between each macroblock's QP and the initial QP of the slice.

2. Design of GOP-based Video Quality Metrics

Macroblock Features

<i>intra</i>	Percentage of intra-macroblocks in relation to the total number of macroblocks in this slice.
<i>inter</i>	Percentage of inter-macroblocks, both predicted (P) and bi-predicted (B), in this slice.
<i>skip</i>	Percentage of macroblocks that have the skip flag enabled and therefore do not contain any image data. The decoder will use the image data from the referenced frame here.
$I_{16 \times 16}$	Percentage of intra-macroblocks with a size of 16×16 pixels in relation to the total number of macroblocks in this slice.
$I_{8 \times 8}$	Percentage of intra-macroblocks with a size of 8×8 pixels in relation to the total number of macroblocks in this slice.
$I_{4 \times 4}$	Percentage of intra-macroblocks with a size of 4×4 pixels in relation to the total number of macroblocks in this slice.
$P_{16 \times 16}$	Percentage of inter-macroblocks subdivided into smaller blocks. Note that the P in $P_{16 \times 16}$ (or P_8 and P_4) does not only refer to predicted (P) macroblocks but rather to all inter-coded macroblocks, including bi-predicted (B) types.
P_8	Percentage of macroblocks divided into 16×8 , 8×16 or 8×8 partitions. I.e. P_8 is the total number of macroblocks minus $P_{16 \times 16}$.
P_4	Percentage of 8×8 macroblocks in the slice that were subdivided into sub-macroblocks of the sizes 8×4 , 4×8 or 4×4 .

Motion Vector Features

MVl_{max}	The length of the longest motion vector in this slice. The motion vector length MVl is calculated from the predicted coordinates MV_{pred} and the difference between the predicted and the actual coordinates ΔMV which can both be read from the bitstream:
-------------	---

$$MVl = \sqrt{(MV_{pred,x} + \Delta MV_x)^2 + (MV_{pred,y} + \Delta MV_y)^2}. \quad (2.1)$$

MVl_{avg}	The average length of all motion vectors (cf. equation 2.1) used in this slice.
ΔMV_{max}	The maximum difference between the predicted and the actual motion vectors.
ΔMV_{avg}	The mean difference between the predicted and the actual motion vectors.

2.2. Two-Way Regression Analysis

In general, regression analysis can be used to determine the relationship between dependent variables \mathbf{y} and independent variables \mathbf{X} . In the case of video quality prediction \mathbf{y} is the $N \times 1$ column vector of the subjective quality values (the results of a subjective test) for all N sequences. The feature extraction step discussed above results in a $1 \times M$ *feature vector* \mathbf{x} per video sequence. All feature vectors can be combined into the $N \times M$ *feature matrix* \mathbf{X} . The relationship between the features and the subjective quality as ground truth is represented by the $M \times 1$ column vector \mathbf{b} , also called *weight vector*:

$$\mathbf{y} = \mathbf{X}\mathbf{b}. \quad (2.2)$$

In a first step a model has to be trained or calibrated using a set of known video sequences and the corresponding subjective test results. Mathematically speaking, in this calibration step the aim is to find \mathbf{b} . In general, \mathbf{X} will neither be square nor have full rank so one can only try to find a good estimate $\hat{\mathbf{b}}$ for the weight vector. There are several different regression models available to achieve this – some of them will be discussed below. After the calibration the model can be used to predict the quality \mathbf{y}_u of unknown video sequences from their features \mathbf{X}_u :

$$\hat{\mathbf{y}}_u = \mathbf{X}_u \hat{\mathbf{b}}. \quad (2.3)$$

Models like this can be called two-way models. This refers to the number of *ways* or *modes* of the feature matrix \mathbf{X} . This number is not to be confused with the number of rows and columns, which is often called *dimensionality*. Thus, a normal matrix can also be called a two-way array in order to distinguish the multi-way arrays that

2.2.1. Data Pre-Processing

In order to improve the performance of the following regression models some data pre-processing is performed. Smilde et al. [42] describe the different possibilities of data pre-processing for multi-way analysis in detail. In this thesis I use the so-called *autoscaling*, which is the combination of *centering* across the first mode and *scaling* to unit standard deviation within the second mode.

Step 1: Centering

In the centering step constant offsets are removed from \mathbf{X} and \mathbf{y} by subtracting the column-means. With x_{nm} and y_n being the respective elements of \mathbf{X} and \mathbf{y} , this can be written as

$$\tilde{x}_{nm} = x_{nm} - \bar{x}_m = x_{nm} - \frac{1}{N} \sum_{n=0}^N x_{nm} \quad (2.4a)$$

2. Design of GOP-based Video Quality Metrics

and

$$\begin{aligned}\check{y}_n &= y_n - \bar{y} = y_n - \frac{1}{N} \sum_{n=0}^N y_n \\ &= y_n - y_0.\end{aligned}\tag{2.4b}$$

The results are the centered feature matrix $\check{\mathbf{X}}$ with its elements \check{x}_{mn} and the centered vector $\check{\mathbf{y}}$. After the prediction step the subtracted offset y_0 will be added to the predicted quality value \hat{y} again.

Step 2: Scaling

When scaling within the second mode, every column of the centered feature matrix $\check{\mathbf{X}}$ is divided by the standard deviation of this column. This results in the autoscaled matrix $\tilde{\mathbf{X}}$ with the elements

$$\tilde{x}_{nm} = \frac{\check{x}_{nm}}{\sqrt{\frac{1}{N-1} \sum_{n=1}^N \left(\check{x}_{nm} - \frac{1}{N} \sum_{n=0}^N \check{x}_{nm} \right)^2}}.\tag{2.5}$$

For the dependent variables $\check{\mathbf{y}}$ no scaling step is required since in a one-way vector there is no second mode one could scale within.

In what follows, for reasons of simplicity, I will always write \mathbf{X} instead of $\check{\mathbf{X}}$ and \mathbf{y} instead of $\check{\mathbf{y}}$. Nevertheless, I will always mean the autoscaled versions of these input variables.

2.2.2. Multiple Linear Regression (MLR)

The simplest way to find an estimate $\hat{\mathbf{b}}$ of the weight vector is to minimize the residual $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}}$ through least squares regression. The least squares problem can be formulated as following:

$$\min_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2.\tag{2.6}$$

The ordinary least squares estimator gives a solution for the minimization problem:

$$\hat{\mathbf{b}} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y}.\tag{2.7}$$

This regression vector can then be used for predicting the quality \hat{y}_u of an unknown video sequence by multiplying it with its feature vector \mathbf{x}_u . Furthermore, the offset y_0 that was subtracted in the centering step (equation (2.4b)) is added to the result again. Thus, the quality prediction for the unknown sequence is

$$\hat{y}_u = y_0 + \mathbf{x}_u \hat{\mathbf{b}}.\tag{2.8}$$

For further information about MLR, refer to Draper and Smith [6].

2.2.3. Principal Component Regression (PCR)

One problem of MLR is that all features in \mathbf{X} are of equal importance in the regression. By applying Principal Component Analysis (PCA) on the data matrix, redundancy can be removed. After performing PCA it is sufficient to use the first few Principal Components (PCs) instead of all features to express most of the variance in \mathbf{X} . Furthermore, the resulting regression model becomes more statistically stable when using PCA [29]. There are several different ways of calculating the PCs. In the following, I will explain the method used by Keimel et al. [24]. Although this may not be the most efficient method, it is quite easy to understand.

It is assumed that there are more video sequences than features ($N > M$) so there will be M principal components. The feature matrix $\mathbf{X} \in \mathbb{R}^{N \times M}$ can be decomposed using Singular Value Decomposition (SVD) as

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{P}^\top. \quad (2.9)$$

The resulting matrix $\mathbf{P} \in \mathbb{R}^{M \times M}$ is called *loadings matrix* and its column vectors are eigenvectors of $\mathbf{X}^\top\mathbf{X}$. In a second step the *score matrix* $\mathbf{T} \in \mathbb{R}^{N \times M}$ is defined as

$$\mathbf{T} = \mathbf{U}\mathbf{D} = \mathbf{X}\mathbf{P}. \quad (2.10)$$

From now on, $\mathbf{T}_R \in \mathbb{R}^{N \times R}$ denotes the matrix of the first R columns of \mathbf{T} and accordingly $\mathbf{P}_R \in \mathbb{R}^{M \times R}$, the matrix of the first R columns of \mathbf{P} , limiting these matrices to the R largest principal components. With these two matrices the estimation $\hat{\mathbf{X}}_R$ of \mathbf{X} based on these principal components can be written as

$$\hat{\mathbf{X}}_R = \mathbf{T}_R\mathbf{P}_R^\top. \quad (2.11)$$

To determine the influence of these principal components of \mathbf{X} on \mathbf{y} the new regression model is

$$\mathbf{y} = \mathbf{T}_R\mathbf{c}. \quad (2.12)$$

Hereby, \mathbf{c} is an unknown weight vector which can be estimated as $\hat{\mathbf{c}}$ by MLR analogously to equation (2.7):

$$\hat{\mathbf{c}} = \left(\mathbf{T}_R^\top\mathbf{T}_R\right)^{-1}\mathbf{T}_R^\top\mathbf{y}. \quad (2.13)$$

Finally the estimated weight vector $\hat{\mathbf{b}}$ can be calculated by the transformation of $\hat{\mathbf{c}}$ back into the feature space:

$$\hat{\mathbf{b}} = \mathbf{P}_R\hat{\mathbf{c}}. \quad (2.14)$$

Then, the quality prediction of an unknown sequence works analogously to MLR and equation (2.8).

2.2.4. Bilinear Partial Least Squares Regression (PLS1)

One problem of PCR is that the largest R principal components may describe the variance in \mathbf{X} very well but do not necessarily give the best description of the variance in \mathbf{y} . Therefore, Partial Least Squares Regression (PLSR), as an improvement on PCR, uses \mathbf{X} and \mathbf{y} simultaneously in modeling.

PLSR is the generic term for many types of regression methods – the most basic one is Bilinear Partial Least Squares Regression (PLS1). *Bilinear* refers to the dimensionality of the data matrix \mathbf{X} and the 1 in PLS1 denotes that \mathbf{y} is a column vector. For further details regarding PLSR, refer to [29, 30].

Algorithm 1 Bilinear Partial Least Squares Regression (PLS1)

```

center  $\mathbf{X}$  and  $\mathbf{y}$ 
 $\mathbf{X}_1 = \mathbf{X}, \mathbf{y}_1 = \mathbf{y}$ 
 $r = 1$ 
1: repeat
2:    $\mathbf{w}_r = \frac{\mathbf{X}_r^\top \mathbf{y}_r}{\|\mathbf{X}_r^\top \mathbf{y}_r\|}$ 
3:    $\mathbf{t}_r = \mathbf{X}_r \mathbf{w}_r$ 
4:    $\hat{\mathbf{c}}_r = \frac{\mathbf{t}_r^\top \mathbf{y}_r}{\mathbf{t}_r^\top \mathbf{t}_r}$  and  $\mathbf{p}_r = \frac{\mathbf{X}_r^\top \mathbf{t}_r}{\mathbf{t}_r^\top \mathbf{t}_r}$ 
5:    $\mathbf{X}_{r+1} = \mathbf{X}_r - \mathbf{t}_r \mathbf{p}_r^\top$  and  $\mathbf{y}_{r+1} = \mathbf{y}_r - \mathbf{t}_r \hat{\mathbf{c}}_r$ 
6:    $r = r + 1$ 
7: until  $r = R$ 

```

Listing 1 shows an iterative algorithm for PLS1; a more detailed explanation can be found at Martens and Næs [30]. The results of the algorithm are the $M \times R$ matrices \mathbf{W} and \mathbf{P} with their columns \mathbf{w}_r and \mathbf{p}_r , and the $R \times 1$ column vector $\hat{\mathbf{c}}$. The regression vector $\hat{\mathbf{b}}$ is then defined as

$$\hat{\mathbf{b}} = \mathbf{W} \left(\mathbf{P}^\top \mathbf{W} \right)^{-1} \hat{\mathbf{c}} \quad (2.15a)$$

and a corresponding constant offset as

$$\hat{b}_0 = \bar{\mathbf{y}} - \bar{\mathbf{X}}^\top \hat{\mathbf{b}}. \quad (2.15b)$$

The algorithm above is only one possibility to do the PLS1, but it is the simplest and because it is sufficient for a basic understanding of the method, it is the only one shown here. Another very common algorithm is the NIPALS-Algorithm (Nonlinear Iterative Partial Least Squares) [29].

The regression step that results in the quality prediction \hat{y}_u of an unknown sequence with the (centered) feature vector \mathbf{x}_u is similar to MLR and PCR; only the constant model offset \hat{b}_0 has to be added as well:

$$\hat{y}_u = y_0 + \hat{b}_0 + \mathbf{x}_u \hat{\mathbf{b}}. \quad (2.16)$$

2.3. Three-Way Regression Analysis

Digital video itself can be considered as (at least) three-dimensional data. Each pixel of a frame has two spatial coordinates u and v and a temporal coordinate t which denotes the number of the frame it belongs to. To estimate the video quality using a data analysis approach, some feature data needs to be extracted from the video data. When using pixel-based features like *blocking* [43] or *predictability* as described by Keimel et al. [22], the feature values can be calculated separately for each frame (or at most with looking at the preceding and/or the following frame). The H.264/AVC bitstream features have per se more temporal information since an H.264/AVC encoder uses several frames to predict the preceding or succeeding frames.

In both cases it is common to perform some kind of temporal pooling to combine the feature data of each frame resulting in a handy feature vector. In practice this means mostly averaging the features over the temporal dimension. In doing so, one naturally loses all the temporal information, which is the crucial thing that distinguishes video from still images. In fact, this temporal pooling has a negative effect on the quality estimation, which is shown in [23, 24]. To overcome the negative effect of temporal pooling, multi-way data analysis can be used. Both PCR and PLSR can be extended to process multi-way input data.

Before discussing these extensions in detail, I will introduce the mathematical notation for multi-way data structures as used in this thesis. As far as it is appropriate I will adopt to the *MATLAB*-inspired notation used by Keimel et al. [24]. The list of symbols on page 11 gives an overview of the variables and notations I chose. In the multidimensional case the $1 \times M$ feature vector \mathbf{x} becomes a $M \times T$ matrix, where T denotes the number of frames per video sequence. Accordingly, the $N \times M$ feature matrix becomes the $N \times M \times T$ feature array (or tensor) $\underline{\mathbf{X}}$. Figure 2.1 on the next page shows a graphical representation of this feature cube $\underline{\mathbf{X}}$ and its different slices. In general, $\underline{\mathbf{X}}_{N::}$ denotes the matrix of $\underline{\mathbf{X}}$ with the fixed dimension $n = N$, and $\underline{\mathbf{X}}_{NM}$: the vector with two fixed dimensions $n = N$ and $m = M$.

As in the one-way case I will use autoscaling as described in section 2.2.1 on page 21 for preprocessing the data. Centering and scaling work very similarly in the two-way case, so I will not repeat the description here. Again, I will write $\underline{\mathbf{X}}$ and \mathbf{y} instead of $\tilde{\underline{\mathbf{X}}}$ and $\tilde{\mathbf{y}}$, meaning the autoscaled versions of the data variables.

2.3.1. Two-Dimensional Principal Component Regression (2D-PCR)

One way to include the additional dimension into the regression model is the extension of PCR to two-way data proposed by Yang et al. [49]. This method will be referred to as *Two-Dimensional Principal Component Regression* (2D-PCR). The basic idea is to perform the PCA step on the mean of the covariance matrices of $\underline{\mathbf{X}}$, which describes the average covariance of the temporal dimension. The application of 2D-PCR on video quality metrics has first been described by Keimel et al. [24, 27].

The first step is to calculate this covariance or scatter matrix $\mathbf{X}_{Set} \in \mathbb{R}^{M \times M}$ from the

2. Design of GOP-based Video Quality Metrics

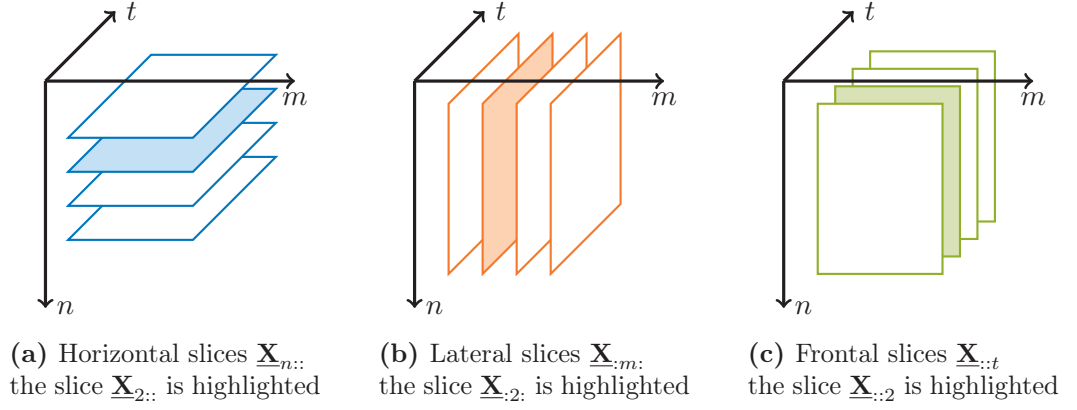


Figure 2.1.: The three-way feature cube $\underline{\mathbf{X}} \in \mathbb{R}^{N \times M \times T}$ and its different slices

feature array $\underline{\mathbf{X}}$ by averaging over the covariance matrices of each temporal slice of $\underline{\mathbf{X}}$:

$$\mathbf{X}_{Sct} = \frac{1}{T} \sum_{t=1}^T \underline{\mathbf{X}}_{::t} \underline{\mathbf{X}}_{::t}^\top \quad (2.17)$$

The scatter matrix is now used for PCA and therefore SVD will be performed on \mathbf{X}_{Sct} to get the loadings matrix $\mathbf{P} \in \mathbb{R}^{M \times M}$:

$$\mathbf{X}_{Sct} = \mathbf{U} \mathbf{D} \mathbf{P}^\top \quad (2.18)$$

In the two-way case the score matrix \mathbf{T} is calculated, but here the three-way scores array $\underline{\mathbf{T}} \in \mathbb{R}^{N \times M \times T}$ is defined per frontal slice as

$$\underline{\mathbf{T}}_{::t} = \underline{\mathbf{X}}_{::t} \mathbf{P} \quad \forall t = 0, \dots, T. \quad (2.19)$$

In this and in all following equations the multiplication of a multi-way array and a matrix (or two multi-way arrays) is given in a per-slice form. As with one dimensional PCR both the loadings matrix and the score array are reduced to the first R principal components, resulting in $\mathbf{P}_R \in \mathbb{R}^{M \times R}$ and $\underline{\mathbf{T}}_R \in \mathbb{R}^{N \times R \times T}$. Analogously to equation 2.13 now the estimation for the weighting factors is given by

$$\hat{\underline{\mathbf{C}}}_{::t} = \left(\underline{\mathbf{T}}_{R,::t}^\top \cdot \underline{\mathbf{T}}_{R,::t} \right)^+ \cdot \underline{\mathbf{T}}_{R,::t}^\top \cdot \mathbf{y} \quad \forall t = 0, \dots, T. \quad (2.20)$$

The transformation of $\hat{\underline{\mathbf{C}}}$ back into the original feature space results in the estimation for the weight array $\hat{\underline{\mathbf{B}}} \in \mathbb{R}^{M \times 1 \times T}$:

$$\hat{\underline{\mathbf{B}}}_{::t} = \mathbf{P}_R \hat{\underline{\mathbf{C}}}_{::t} \quad \forall t = 0, \dots, T. \quad (2.21)$$

To calculate the quality prediction \hat{y}_u for an unknown sequence, the corresponding $M \times T$ feature matrix \mathbf{X}_u with its column-vectors $\mathbf{x}_{u,t}$ has to be multiplied by the

corresponding slice of the weight array estimation $\hat{\mathbf{B}}$. After adding the centering offset, this results in the vector of quality estimations per video frame $\hat{\mathbf{y}}_u \in \mathbb{R}^T$ that consists of the elements

$$\hat{y}_{u,t} = y_0 + \mathbf{x}_{u,t} \hat{\mathbf{B}}_{:,t}. \quad (2.22)$$

To get an overall quality prediction $\hat{y}_u \in \mathbb{R}$, the average of these per-frame-values is calculated as

$$\hat{y}_u = \frac{1}{T} \sum_{t=0}^T \hat{y}_{u,t}. \quad (2.23)$$

2.3.2. Trilinear Partial Least Squares Regression (Tri-PLS1)

Trilinear Partial Least Squares Regression (Tri-PLS1) was introduced by Bro [5] as multi-dimensional extension of PLS1. In Tri-PLS1, the components are determined depending on weights gained along both the m and the t dimension, whereas in PLS1 the components are only dependent on the m dimension.

Listing 2 shows an iterative algorithm that describes the decomposition of \mathbf{X} into its components \mathbf{w}^M and \mathbf{w}^T along both feature dimensions. \mathbf{Z} in step 2 of the algorithm represents the matrix of all z_{mt} with

$$z_{mt} = \sum_{n=0}^N y_n x_{nmt}. \quad (2.24)$$

The scores t_n corresponding to each sample n can then be written with the components as

$$t_n = \sum_{m=0}^M \sum_{t=0}^T x_{nmt} w_m^M w_t^T. \quad (2.25)$$

Algorithm 2 Trilinear Partial Least Squares Regression (Tri-PLS1)

- center \mathbf{X} and \mathbf{y}
 - $\mathbf{X}_1 = \mathbf{X}, \mathbf{y}_1 = \mathbf{y}$
 - $r = 1$
 - 1: **repeat**
 - 2: calculate \mathbf{Z}
 - 3: determine $\mathbf{w}_r^m, \mathbf{w}_r^t$ by SVD of \mathbf{Z}
 - 4: calculate \mathbf{t}_r . $\mathbf{T} = [\mathbf{t}_1 \cdots \mathbf{t}_r]$
 - 5: $\mathbf{b}_r = (\mathbf{T}^\top \mathbf{T})^{-1} \mathbf{T} \mathbf{y}_r$
 - 6: $\mathbf{X}_{r+1} = \mathbf{X}_r - t_r \mathbf{w}_r^M (\mathbf{w}_r^T)^\top$ and $\mathbf{y}_{r+1} = \mathbf{y}_r - \mathbf{T} \mathbf{b}_r$
 - 7: $r = r + 1$
 - 8: **until** proper description of \mathbf{y}_r
-

From the extracted components and scores an estimate of the $T \times M$ weight matrix $\hat{\mathbf{B}}$ and the model offset \hat{b}_0 can be obtained for direct regression of an $1 \times M \times T$ slice of

2. Design of GOP-based Video Quality Metrics

the feature array \mathbf{X}_u representing the features of an unknown sequence. The elements of the vector of the quality estimations $\hat{\mathbf{y}}_u$ for this unknown sequence can be written as

$$\hat{y}_{u,t} = y_0 + \hat{b}_0 + \mathbf{x}_{u,t} \hat{\mathbf{b}}_t. \quad (2.26)$$

where $\mathbf{x}_{u,t}$ denotes the t -th column vector of \mathbf{X}_u and $\hat{\mathbf{b}}_t$ the t -th row vector of $\hat{\mathbf{B}}$. The overall quality prediction \hat{y}_u can again be obtained by averaging these per-frame-values:

$$\hat{y}_u = \frac{1}{T} \sum_{t=0}^T \hat{y}_{u,t}. \quad (2.27)$$

2.4. Video Quality Prediction

All the regression models described in the previous chapter can be used to predict the quality of unknown video sequences. As Keimel et al. [20, 21, 23, 24, 27] have described, the metrics built on these regression methods lead to comparatively good results. In the following, I will show how these metrics can be improved further.

2.4.1. GOP-based Temporal Quality Prediction

Since $\hat{\mathbf{y}}_u$ in the equations (2.22) or (2.26) already contains quality predictions for each frame of the unknown video sequence, what comes to mind is the idea of using these values to make a statement about the quality progression during the video sequence. To explain why this first idea cannot lead to practical results, some deeper knowledge about the nature of the used feature data is required. As discussed in section 2.1 on page 19, most of the features describe either the percentages of different macroblock-types or the statistics of the motion vectors in each frame. Both types of features do not work with I-Frames because there are no macroblocks other than $I_{16 \times 16}$ and $I_{4 \times 4}$ and therefore also no motion vectors. In fact, only 5 of the $M = 17$ features can take values different from 0 in I-Frames.

To perform an appropriate temporal quality prediction, I propose to split up each video sequence into individual Groups of Pictures (GOPs). A GOP is a repeating sequence of slice types in an H.264/AVC video stream. The exact structure of a GOP and its parameters such as length or the number of reference frames can be configured during the encoding process. For the application of quality prediction, it is only important that the video is encoded using fixed GOP-lengths and that all videos in the training set share the same GOP-length T_G . Figure 2.2 on the facing page shows an example of a possible GOP structure as it is used in the IT-IST dataset, which will be used in chapter 3 to evaluate the quality prediction performance. One can clearly see the inter-frame dependencies within the shown GOP.

Subsequently, each GOP is considered as a separate sequence of the length T_G when training the regression model. As a consequence, the video sequences can differ in their length. For the sake of simplicity, all videos in this thesis were of equal length and consisted of $G = \frac{T}{T_G}$ GOPs. This means that instead of an $N \times M$ feature matrix I now

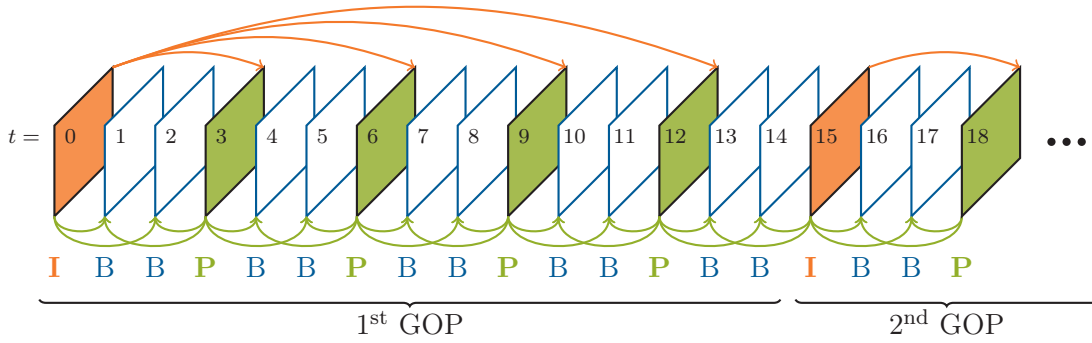


Figure 2.2.: Example for a Group of Pictures (GOP) with length $T_G = 15$, two reference frames and the structure $IBBPBBPBBPBBPBB$. These settings are also used in the videos of the IT-IST dataset as described in section 3.2.1 on page 36.

use an $NG \times M$ matrix as input data for the regression model. In the case of three-way models, this results in an $NG \times M \times T_G$ feature array instead of one with the dimensions $N \times M \times T$. Figure 2.3 shows the difference between the traditional and the GOP-based approach by illustrating the changes of dimensionality of the feature cube.

The regression itself is not different from the way it was described above. The prediction of the quality of unknown sequences now results in a quality value per GOP and these values actually reflect the perceived progression of visual quality quite well, as shown below. In case of three-way methods the quality prediction $\hat{y}_{u,g}$ of unknown

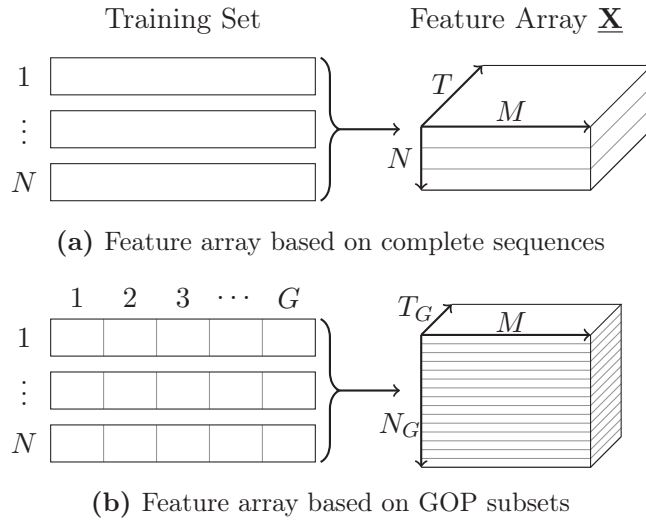


Figure 2.3.: Changes in the dimensionality of the feature array $\underline{\mathbf{X}}$ when moving from the traditional feature cube based on the features of the complete video sequences to a GOP-based approach.

2. Design of GOP-based Video Quality Metrics

GOPs is calculated analogously to equations (2.23) and (2.27):

$$\hat{y}_{u,g} = \frac{1}{T_G} \sum_{t=0}^{T_G} \hat{y}_{u,t}. \quad (2.28)$$

2.4.2. Length-Independent Quality Prediction

Besides the possibility to predict the quality progression of video, a major advantage of the extended GOP prediction model is the possibility to do quality predictions for video sequences of different lengths with the same training set. With the conventional method, matching dimensionality is required to evaluate the equations (2.22) or (2.26). Provided that the GOP-length of the unknown video is equal to the T_G used in the training set, a prediction becomes possible. In practice, it is much more feasible to train prediction models for all possible GOP lengths than for all occurring video lengths.

Figure 2.4 shows an example of a temporal quality prediction of a much longer sequence than the ones used as training data. Although this example has not been evaluated by subjective testing, it illustrates quite well that even the prediction of three minutes of video with a model trained by sequences of 10 seconds each apparently results in plausible curves.

If one is only interested in an overall quality prediction \hat{y}_u , this value can easily be derived from the predictions on GOP level by calculating the arithmetic mean:

$$\hat{y}_u = \frac{1}{G} \sum_{g=0}^G \hat{y}_{u,g}. \quad (2.29)$$

As will be shown later, this prediction is equally significant as an estimation made by the conventional method predicting the quality of the whole video sequence at once.

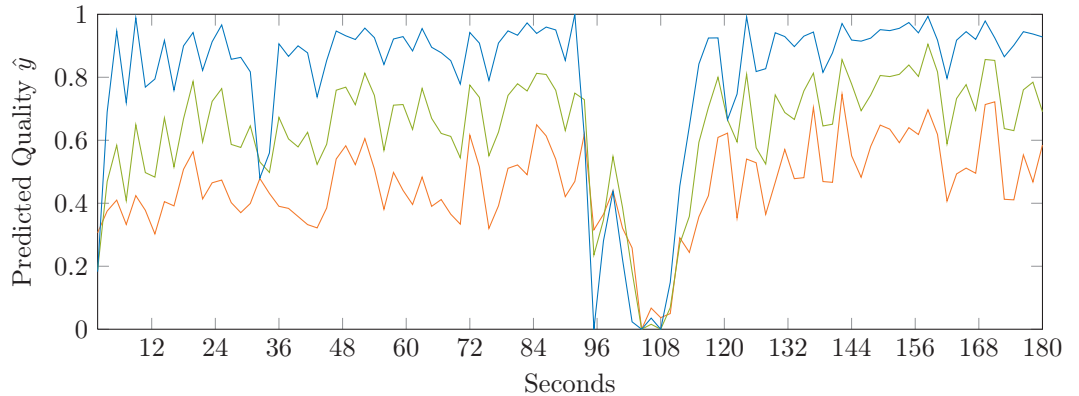


Figure 2.4.: Temporal quality prediction of the first three minutes taken from an episode of a popular TV-series encoded with three different bit rates using H.264/AVC. The 2D-PCR model was trained with the IT-IST dataset. Dimensions in the training step were $N = 48$, $M = 17$, $G = 17$, $T_G = 15$ and dimensions in the prediction step: $N = 3$, $M = 17$, $G = 500$, $T_G = 15$.

The break-in around second 100 can be assumed to be due to the opening credits at this point.

2.5. Data Post-Processing: Sigmoid Correction

Regardless of which of the quality metrics described above is used, after the prediction of a quality value the last step is to apply the so-called *Sigmoid Correction* as described by Keimel et al. [22]. Hereby, the nonlinear correction function

$$\hat{y}_{sc} = \frac{a}{1 + e^{-\frac{\hat{y}-b}{c}}} \quad (2.30)$$

is applied to the prediction values with the parameters set to $a = 1.0, b = 0.5, c = 0.2$. Figure 2.5 shows a plot of this function. The purpose of this post-processing is to adapt the prediction values to the nature of subjective MOS data. The correction function is nearly linear over a wide range, but cuts the values near 0 and 1 since these extrema are not likely to occur in a real test.

If not stated otherwise, all prediction values used in this thesis – especially all values used for evaluation in chapter 3 – have been corrected with the Sigmoid Correction.

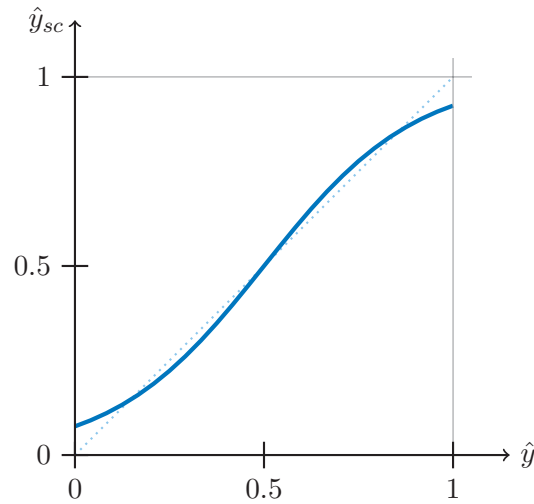


Figure 2.5.: Sigmoid Correction Function

3. Evaluation of GOP-based Video Quality Metrics

In order to verify the improvements on the video quality metrics described in the previous chapter, their prediction performance needs to be evaluated. In this chapter I will first outline the mathematical performance measures and methodology used for the evaluation. Then, I will compare the prediction performance of the length-independent metric to a corresponding length-dependent metric. Finally, I will discuss the evaluation of the temporal quality prediction. This requires a new subjective quality assessment method, which I will describe before I analyze the evaluation results.

All quality metrics and all evaluation functions were implemented in *MATLAB* using the *N-Way Toolbox* by Andersson and Bro [2] for all PLSR-related code.

3.1. Evaluation Methodology

3.1.1. Statistical Performance Measures

The prediction performance of video quality metrics is measured by statistical functions, namely Pearson's correlation coefficient, the Spearman rank correlation coefficient, and the Root Mean Square Error (RMSE), which will be explained in the following.

Pearson's correlation coefficient

Pearson's correlation coefficient measures the linear dependence between two variables. The coefficient r between n samples of the variables a and b is defined as

$$r = \frac{\sigma_{ab}}{\sigma_a \sigma_b} = \frac{\sum_{i=0}^n (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=0}^n (a_i - \bar{a})^2} \sqrt{\sum_{i=0}^n (b_i - \bar{b})^2}} \quad (3.1)$$

where σ_a and σ_b denote the sample standard deviation of a and b , and σ_{ab} the covariance of a and b . The correlation coefficient can take values between -1 and $+1$. The correlation between subjective test results and the prediction of a quality metric is an important benchmark: The nearer r is to $+1$, the better the prediction performance is.

Spearman's Rank Correlation Coefficient

Spearman's correlation coefficient ρ is a measure of how well a monotonic function describes the relationship between two variables. If used on video quality metrics, a correlation of $\rho = 1$ means that the evaluated metric can order all sequences correctly by their

3. Evaluation of GOP-based Video Quality Metrics

subjective quality. This metric can give perfect answers to the question which video of two has better quality but may not predict perfect absolute quality ratings.

Mathematically, Spearman’s rank correlation is the Pearson correlation coefficient between the ranks of the samples of two variables a and b . Simplified, the rank is the position of the i -th sample in a sorted table of all values.

Root Mean Square Error (RMSE)

The RMSE of n predictions \hat{a} of the variable a is defined as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^n (a_i - \hat{a}_i)^2} \quad (3.2)$$

and can be used as a measure for the prediction accuracy. Clearly, in this case, the best value is 0.

In contrast to the correlation measures described above, the RMSE is scale-dependent. In this thesis all quality values are normalized to 1 which has to be considered when comparing the RMSE-values to those from other publications.

Scatter Plots

Since the three numerical measures described above cannot explain all dependencies between two variables, a so-called *scatter plot* can be helpful. Especially nonlinear relationships that are not taken into account by correlation measures become visible in a scatter plot. Therefore, I will present scatter plots with all correlation values.

3.1.2. Cross Validation

In the evaluation of data analysis methods it is important not to use the same data for the training and the validation. Otherwise, the results would be over-optimistic and misleading. Ideally, one would use two different datasets for the training and the validation. Since the size and number of suitable datasets are limited and it is very time-consuming to create new datasets, it is not affordable to use only one half of a dataset for the training and the other for validation.

In order to avoid this problem, I performed cross validation. This means the model is trained with the data of the complete dataset except one sequence and all that share the same content. Afterwards, the quality of the sequences that had been left out is predicted using the resulting model. In doing so, I can obtain quality predictions for all sequences without using the same data for training and validation and making use of the whole dataset at the same time. For a more theoretical view on cross validation refer to Martens and Martens [29].

3.1.3. Full-Reference Metrics for Comparison

Along with the results of the evaluated quality metrics, I provide the prediction results of two well-known full-reference metrics.

Peak Signal-to-Noise Ratio (PSNR)

The Signal-to-Noise Ratio (SNR) is a very common measure – especially in telecommunications engineering – for the quality of a noisy signal. Since an image can be interpreted as a signal, the SNR can be applied to image data as well. The PSNR is the adoption of this concept to image processing.

In order to obtain the PSNR between an image I and its coded version I_c with dimensions U, V , first the Mean Square Error (MSE) of both images is calculated as

$$MSE = \frac{1}{UV} \sum_{u=0}^U \sum_{v=0}^V [I(u, v) - I_c(u, v)]^2. \quad (3.3)$$

The PSNR is then the logarithmic ratio of the RMSE and the maximum pixel value MAX_I of the image I :

$$PSNR = 10 \cdot \log_{10} \frac{MAX_I}{\sqrt{MSE}} \text{ dB}. \quad (3.4)$$

With 8-bit video data the maximum pixel value is $2^8 - 1 = 255$. The formulas above only apply to grayscale images. All color data used in this thesis was converted to $YCrCb$ color space and the PSNR is only applied to the luma channel. The PSNR is calculated separately for each frame of the video and the average is taken afterwards.

The correlation between PSNR and the perceived quality is not very high, which makes PSNR not the best choice as a video quality metric. There are examples of images with very different PSNR without any noticeable quality difference and the other way round [45, 47]. But as the PSNR is rather widespread and often used in the development of codecs, I will use it for comparison.

Structural Similarity (SSIM)

To overcome the shortcomings of PSNR, Wang et al. [44] proposed the so-called Structural Similarity Index (SSIM) as a new metric for image quality. The underlying assumption is that the HVS is highly adapted to extracting structural information from a scene.

The SSIM is calculated on small windows of an image. The index of two windows x and y is defined as

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3.5)$$

where μ_x and μ_y are the averages, σ_x^2 and σ_y^2 the variances, and σ_{xy} the covariance of x and y . The constants C_1 and C_2 are set to $C_1 = (0.01 \cdot 255)^2$ and $C_2 = (0.03 \cdot 255)^2$ by default for 8-bit data. For further details, refer to [44]. As with the PSNR, the SSIM index is calculated only on the luma channel and separately for each video frame.

In general, the SSIM correlates much better to subjective quality than PSNR. Therefore, I consider it the reference metric whose results have to be exceeded with the proposed no-reference metrics.

3.2. Evaluation of Length-Independent Quality Prediction

In order to verify the validity of the length-independent quality prediction as described in section 2.4.2, I compare its results to the prediction results of the conventional metric that predicts the quality of a complete sequence at once and some other metrics.

3.2.1. Datasets Used for Evaluation

Since the length-independent quality prediction can easily be validated on existing data without conducting subjective tests, I will apply the metric on three different datasets which are described in what follows.

IT-IST CIF Dataset

The first dataset is provided by IT-IST and has been used first by Brandão and Queluz [4]. The videos in this dataset were encoded with H.264/AVC and have CIF resolution (352×288 pixels). The frame rate is 30 fps except for the sequence *Australia*, which has 25 fps. From the the available sequences I chose 4 different rate points in the range from 32 kbit/s to 2048 kbit/s for each of the 12 different video sequences in the dataset. Table 3.1 on the facing page shows the complete list of all $N = 48$ video sequences. The sequences had been encoded with a fixed GOP-length of $T_G = 15$ frames. I removed the first GOP, as it consists of only 13 frames, and also the last few frames because the last GOP is incomplete. In total there are $T = 240$ frames per sequence and $G = 16$ GOPs per sequence which gives $N_G = 768$ subsets of video to train the model with.

IT-IST provides the results of subjective quality assessment for all video sequences in their dataset. The test was conducted with 42 participants using Degradation Category Rating (DCR) as described in ITU-T Recommendation P.910 [15]. It is assumed that the Mean Opinion Score (MOS) values are equally valid for the slightly shortened sequences I use here.

The biggest advantage of this dataset is the fact that it consists of many different sequences. When using cross validation, the training dataset is still quite large and still covers very different content types. The dataset also covers a broad range of visual quality from high data rates where almost no distortions are noticeable to low quality where it is hard to recognize any of the video content. Both factors make the dataset well suited for quality metrics based on data analysis.




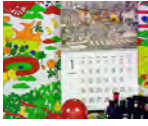








TUM 1080p25 High Definition Dataset

The TUM 1080p25 HD dataset consists of the 4 different video sequences *CrowdRun*, *ParkJoy*, *IntoTree* and *OldTownCross* from the SVT multi format test set [9]. As described by Keimel et al. [26], the 25 fps version has been generated from the original 50 fps material by dropping every second frame. All sequences are in 1080p HD resolution (1920×1080 pixels).

The dataset was originally designed to compare different coding technologies. Therefore, all sequences were encoded with H.264/AVC using two different sets of encoder

3.2. Evaluation of Length-Independent Quality Prediction

Table 3.1.: Used subset of the IT-IST test set with MOS values from [4]

Sequence	kBit/s	MOS	First Frame	Sequence	kBit/s	MOS	First Frame
Australia	32	0.32		Foreman ^a	64	0.01	
	64	0.61			128	0.49	
	128	0.85			256	0.77	
	256	0.99			512	0.98	
City ^a	128	0.56		Mobile ^b	64	0.08	
	200	0.65			128	0.12	
	256	0.80			256	0.73	
	512	0.98			512	0.92	
Coastguard	64	0.22		Silent ^b	64	0.14	
	128	0.55			200	0.69	
	256	0.85			400	0.89	
	512	0.92			1024	1.00	
Container ^b	64	0.67		Stephan ^b	128	0.01	
	128	0.70			256	0.39	
	256	0.95			512	0.84	
	512	0.99			1024	0.98	
Crew ^b	128	0.06		Table ^a	64	0.05	
	200	0.26			128	0.47	
	400	0.69			256	0.89	
	1024	0.99			512	0.98	
Football ^a	256	0.19		Tempete ^a	128	0.39	
	512	0.58			200	0.63	
	1024	0.74			400	0.86	
	2048	0.99			750	0.99	

^a Used also in the test phase of the evaluation of temporal quality prediction (cf. chapter 3.3)

^b Used also in the training phase of the evaluation of temporal quality prediction

parameters at four different rate points each. There are some more sequences that were encoded using the wavelet-based Dirac codec which are not suitable for H.264/AVC feature extraction and therefore not used in this thesis. Hence, the used dataset consisted of $N = 32$ sequences in total.

One problem of this dataset is that the two encoder settings HC (high complexity) and LC (low complexity) result in different GOP lengths. The HC sequences have 13 frames per GOP, the LC sequences only 12. As a workaround I discarded the feature data of the last frame (a B-frame) in each GOP of the HC sequences. The result are feature cubes with $T_G = 12$ frontal slices per GOP. This is obviously not the optimal solution, but can also be seen as an example of how to master this or similar challenges in real-life applications. The good results discussed later will justify this workaround.




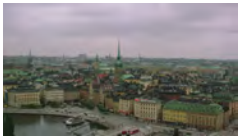
The original sequences consist of 250 frames. After removing the first GOP of the LC sequences (which is only 10 frames long), the last frame of each GOP in the HC sequences

3. Evaluation of GOP-based Video Quality Metrics

and the last incomplete GOP in all sequences, $T = 228$ frames or $G = 19$ GOPs per sequence remain. In total there are $N_G = 608$ GOPs available for training the regression model.

The subjective data of the dataset were gained in a subjective test using the Double Stimulus Unknown Reference (DSUR) method. This method is a variation of the standard Double Stimulus Continuous Quality Scale (DSCQS) test method [12] and has been proposed by Baroncini [3]. Table 3.2 gives an overview of the used video sequences and the corresponding MOS values. There another disadvantage of this dataset becomes visible: it does not cover low visual quality well. All HC sequences have an MOS above 0.5 and even 75% of the LC sequences have an MOS in the upper half of the scale. The sequence *OldTownCross* is the most drastic example – the worst subjective quality that has been measured is 0.69. Therefore, one cannot expect the same prediction performance as with the IT-IST dataset.

Table 3.2.: TUM 1080p25 Dataset

Sequence	MBit/s	MOS HC	MOS LC	First Frame
CrowdRun	8.4	0.51	0.26	
	12.7	0.69	0.56	
	19.2	0.83	0.62	
	28.5	0.93	0.78	
ParkJoy	9.0	0.68	0.19	
	12.6	0.75	0.21	
	20.1	0.89	0.54	
	30.9	0.96	0.84	
IntoTree	5.7	0.73	0.43	
	10.4	0.86	0.62	
	13.1	0.92	0.63	
	17.1	0.93	0.62	
OldTownCross	5.4	0.89	0.69	
	9.6	0.90	0.78	
	13.7	0.94	0.79	
	19.0	0.96	0.82	

TUM 1080p50 High Definition Dataset



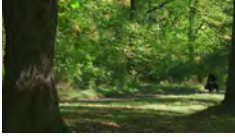


The 5 sequences of the third dataset are also taken from the SVT multi format test set [9]. Both datasets share the sequence *CrowdRun*, but the 1080p50 dataset additionally contains the sequences *TreeTilt*, *PrincessRun*, *DanceKiss* and *Flag/Shoot*. The dataset with 1080p HD resolution (1920×1080 pixels) has been created by Redl et al. [37] for determining the influence of different viewing devices on the perceived visual quality and is also described in [26]. All sequences have the full 50 fps of the original footage.

3.2. Evaluation of Length-Independent Quality Prediction

The sequences were encoded with H.264/AVC at four different rate points. Table 3.3 shows the resulting sequences together with their bit rates and the MOS value. For the subjective evaluation the Single Stimulus MultiMedia (SSMM) method, a variation of the standard SSIS test method proposed by Oelbaum et al. [35], had been used. The encoding of the sequences in the dataset resulted in GOPs of the fixed length $T_G = 24$. As in the other datasets, I had to remove the first GOP, because it contained only 19 frames. The last and incomplete GOP was also removed so that $G = 19$ GOPs remained. In total, every sequence still had $T = 456$ frames.

In contrast to the TUM 1080p25 dataset, this dataset covers the low end of the quality range much better. Nevertheless, the main drawback of the TUM 1080p50 dataset is its small total number of $N = 20$ sequences or $N_G = 380$ GOPs: With cross validation only 16 sequences are used in the training of the regression model.

Table 3.3.: TUM 1080p50 Dataset

Sequence	MBit/s	MOS	First Frame
CrowdRun	8	0.19	
	20	0.59	
	30	0.80	
	40	0.81	
TreeTilt	2	0.30	
	3	0.56	
	6	0.89	
	10	0.89	
PrincessRun	8	0.19	
	20	0.56	
	30	0.74	
	40	0.70	
DanceKiss	2	0.29	
	3	0.54	
	6	0.82	
	10	0.81	
Flag/Shoot	2	0.25	
	3	0.55	
	6	0.80	
	10	0.77	

Summary

After discussing the three datasets, it is expected that the IT-IST dataset will show the best performance. But it is still worthwhile looking at the results for the other data sets

3. Evaluation of GOP-based Video Quality Metrics

as they will show the limits of the data-analysis-based quality metrics and cover the more relevant HD resolution. Table 3.4 summarizes the most important parameters of the tree datasets.

Table 3.4.: Parameters of the used datasets

Dataset	fps	Resolution	N	T_G	G	T	N_G
IT-IST	30/25	352×288	48	15	16	240	768
TUM 1080p25	25	1920×1080	32	12	19	228	608
TUM 1080p50	50	1920×1080	20	24	19	456	380

3.2.2. Performance Evaluation

Compared Quality Metrics

In order to validate the GOP-based approach for length-independent quality prediction, I compare the results of the proposed metric to the corresponding method based on the training with the complete sequences. For the datasets IT-IST and TUM 1080p25 I used metrics based on Tri-PLS1, since this regression method provides the best results in these cases. For the TUM 1080p50 set, a 2D-PCR model shows the best results. For the sake of convenience, I call the GOP-based metrics Tri-PLS1-GOP and 2D-PCR-GOP.

The number of PLSR factors or PCR components R is given in the tables with the results. In general, I chose the R that lead to the best prediction performance for the complete-sequence-metric and also used the same number for the GOP metric.

Keimel et al. [24, p. 48] claim that “more dimensions are really better”. To support this statement, I also applied the corresponding two-way metrics after temporally pooling the feature data. For the datasets IT-IST and TUM 1080p25 the regression method is PLS1, for the TUM 1080p50 dataset simple PCR.

For further comparison, the results for the full-reference metrics PSNR and SSIM as discussed in section 3.1.3 on page 34 are provided along with the metrics described above. The correlation and RMSE values for all datasets and all metrics are shown in Table 3.5 on the next page. All corresponding scatter plots can be found in appendix A.2.1 on page 63.

Discussion of the Results

First of all, the correlation values increase with all datasets when using the three-way model instead of the two-way model. The component number R decreases or at least remains the same, which indicates that the three-way models are able to describe the variance of quality more accurately. So yes, “more dimensions are really better” [24, p. 48].

When looking at the GOP-based metrics, one notices that both the Pearson and the Spearman coefficients suggest a decrease in prediction performance compared to the models trained with the complete sequences. The same statement can be made when

3.2. Evaluation of Length-Independent Quality Prediction

Table 3.5.: Performance measures of the different quality metrics applied on the three datasets

(a) IT-IST				
	R	Pearson	Spearman	RMSE
PSNR		0.723	0.777	
SSIM		0.850	0.871	
PLS1	3	0.935	0.919	0.117
Tri-PLS1	2	0.951	0.962	0.108
Tri-PLS1-GOP	2	0.947	0.955	0.125
(b) TUM 1080p25				
	R	Pearson	Spearman	RMSE
PSNR		0.738	0.717	
SSIM		0.859	0.805	
PLS1	5	0.849	0.816	0.113
Tri-PLS1	5	0.910	0.866	0.092
Tri-PLS1-GOP	5	0.900	0.849	0.092
(c) TUM 1080p50				
	R	Pearson	Spearman	RMSE
PSNR		0.468	0.391	
SSIM		0.848	0.908	
PCR	7	0.692	0.556	0.200
2D-PCR	3	0.888	0.782	0.137
2D-PCR-GOP	3	0.844	0.746	0.149

looking at the RMSE. With the IT-IST and TUM 1080p25 datasets, the difference of the correlations is small and generally the correlation is on a very high level with both methods. This is also illustrated by the two scatter plots in Figure 3.1 on the next page. At least at first glance, there is no visible qualitative difference between the two metrics. Thus, in the cases of these two datasets, one can hardly speak of a real disadvantage of the GOP-method – especially when considering the advantage of length-independence.

In the TUM 1080p50 data, the drop of the correlation values is clearer. However, as mentioned above, of the three this dataset is probably the most challenging for data analysis metrics. Nonetheless, the performance is still much better than with the two-way metric PCR, so that one cannot take this result as a counterargument against GOP-based metrics in general.

In the IT-IST and TUM 1080p25 datasets, both three-way metrics clearly outperform

3. Evaluation of GOP-based Video Quality Metrics

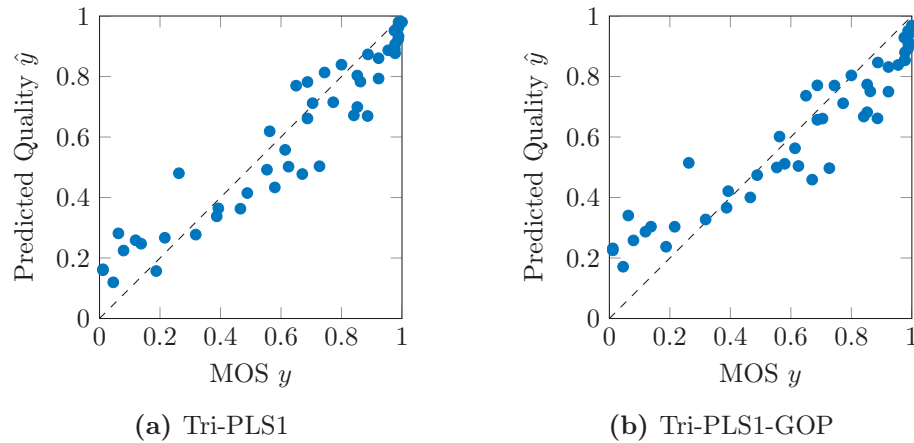


Figure 3.1.: Scatter plots of the quality prediction of the metrics Tri-PLS1 and Tri-PLS1-GOP for the IT-IST dataset

the full-reference metrics PSNR and SSIM, and the two-way metrics are at least on the same level as the SSIM. For the TUM 1080p50 dataset, the PCR metric as well as the 2D-PCR-GOP metric are outperformed by the SSIM index. Although the 2D-PCR has a higher Pearson correlation and a better RMSE, the rank correlation of the SSIM is much higher. This can again be explained with the mentioned low number of sequences in the dataset. The SSIM as a full-reference metric does not depend on the size of a training set and although its results are on about the same level as with the other datasets, it can outperform the data analysis metrics that suffer from the little amount of training data here.

A further conclusion of the results above is that generally PLSR-based metrics perform better than the ones based on simple PCR. As mentioned above, I chose the best regression method for each dataset and with two of the three datasets this were the PLSR metrics.

Conclusion

To summarize the discussion above, I can say that the proposed GOP-based quality metrics lead to very good results compared to other data analysis methods. Especially Tri-PLS1 seems to be an appropriate regression method to achieve the best performance. But as with all data analysis approaches it is important to have a big set of training data that covers a broad range of image contents as well as quality levels.

Figure 3.2 on the facing page displays an example for the quality prediction of a video sequence from the IT-IST dataset. It shows how close the quality predictions of the two three-way metrics are to the true MOS.

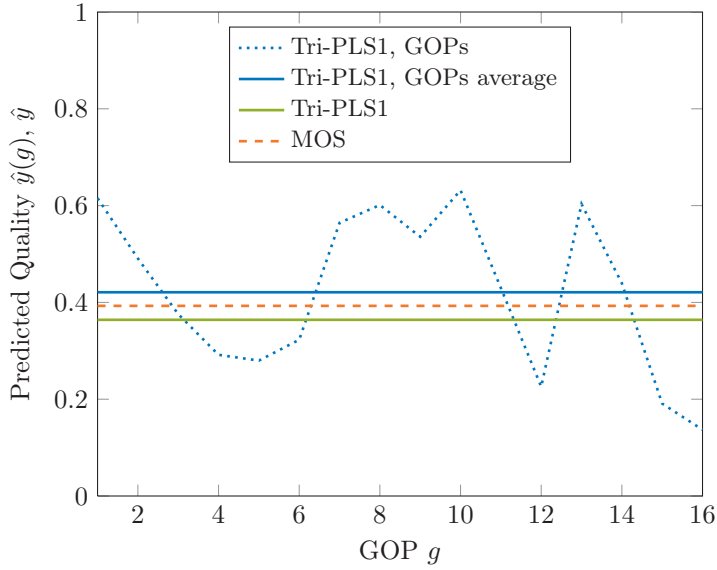


Figure 3.2.: Examples for predictions of the different metrics on the video sequence *Stephan* from the IT-IST dataset at 265 kbit/s

3.3. Evaluation of Temporal Quality Prediction

As with the length-independent metrics, subjective test results are required in order to examine and evaluate the temporal prediction performance of the metrics as discussed above. Since there is currently no data available that would provide a suitable subjective and time-continuous quality measurement, I had to carry out my own test. This required the design of a new subjective test method for determining the temporal progression of subjective video quality of short video sequences, which will be explained in the following.

3.3.1. Design of the Subjective Test

Quality Assessment Method

Since there are no established methods for the subjective assessment of temporal quality progression in short video sequences (cf. section 1.1.2 on page 14), I had to create a new assessment method. The SSCQE method as described in ITU-R Recommendation BT.500 [12] is only suitable for much longer sequences and cannot be adapted to sequences this short: The average observer would not be able to follow the quick changes in visual quality with his or her finger.

Therefore, I divided the task into three separate questions. The first one asked the participant about his or her overall quality impression and is answered on a continuous scale using an on-screen slider. The scale is similar to the ITU-R Rec. BT.500 SSCQE-scale. The second question asked the observer to rate the strength of the quality fluctuation during the video sequence. This question is answered on a continuous scale as well. Un-

3. Evaluation of GOP-based Video Quality Metrics

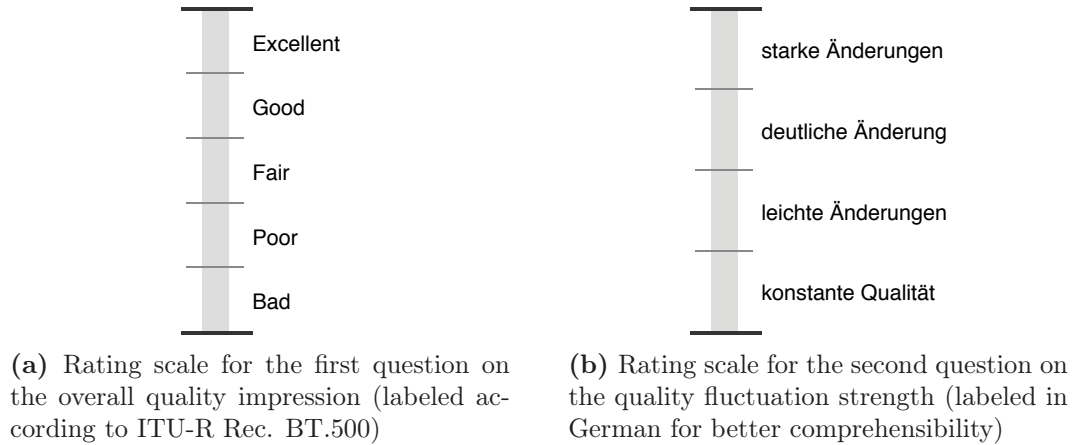


Figure 3.3.: Rating scales for temporal quality assessment

like the first standardized scale this was labeled in German because all test subjects were German native speakers. Both scales can be seen in Figure 3.3.

In a third question the test subjects were asked to categorize the quality progression during the sequence. Each of the available six categories was represented by the small symbolic plot as shown in the right column of Table 3.6 on page 48. For the design of the first five categories, polynomial functions with increasing degrees from 0 to 2 were chosen. The sixth category is meant for all sequences with fast but noticeable quality fluctuations that do not fit into the first five categories. The answers to the three questions can then be used to reconstruct the perceived quality progression as it will be described in section 3.3.2 on page 46.

The advantage of this assessment method is that the observer can make a statement about the temporal quality progression even for short video sequences. But this comes with the compromise that not all possible quality patterns can be represented due to the limitation to the six available categories. The longer the sequences get, the more complex the quality progression pattern can become and the more crucial this limitation becomes. In general, more patterns would allow a better representation of the observer's perception, but would also be much more demanding and time-consuming for the test subject. The six patterns appeared to be a reasonable compromise and led to acceptable results for the video sequences of the dataset.

Since it is difficult to focus on all three questions at the same time while watching the video sequence, the observer is allowed to watch the video more than once. Instead of displaying each sequence three times or even more often to cover all questions, a replay-button was made available and the participants were told that they could use this button as often as required. To avoid any further complexity, a single stimulus method was chosen.

Three questions instead of one per sequence make the training phase before the actual test even more important than with traditional assessment methods. Especially the second and the third question need to be explained thoroughly, in order to make sure that

3.3. Evaluation of Temporal Quality Prediction

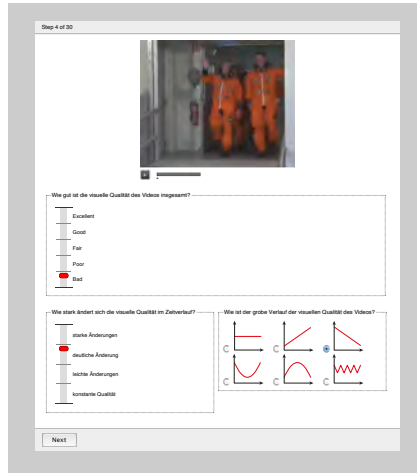
all participants have an equal understanding of the rating scales. Although I performed the training carefully, the variance of the answers of the second question was quite high. This will be discussed in more detail later.

Test Conditions

The test was conducted in the video laboratory of the Institute for Data Processing (LDV) at the Technische Universität München (TUM), which is compliant to ITU-R Recommendation BT.500 [12]. The videos were displayed using a *Sony BVM-L230* reference display (23") at a visible height of 8cm and with a viewing distance of about 60cm. $S = 21$ persons with ages in the range from 14 to 28 participated in the test, most of them were university students and non-experts in video processing.



(a) Video Laboratory at LDV



(b) Screenshot QualityCrowd 2

Figure 3.4.: Photography of the test setting in the LDV video lab and a screenshot of the QualityCrowd 2 software from the conducted test

The software framework *QualityCrowd 2* was used to provide an interactive user interface for the test. *QualityCrowd 2* is a web-based tool which allows the test operator to compile a test batch. This batch defines all displayed texts and the order of the videos. The test runs in a web-browser and therefore the videos are played back by *Adobe Flash Player* from lossless compressed files. Figure 3.4 shows a typical screen from the conducted test. Since the software did not support the assessment method described above, I modified it in order to provide the required functionality. For further information regarding *QualityCrowd* see Horch et al. [10] and Keimel et al. [25].

Description of the Test

The *QualityCrowd 2* software organizes a test in separate steps. The conducted test consisted of 30 steps which are described in the following.

3. Evaluation of GOP-based Video Quality Metrics

Step 1, Welcome Screen This screen was shown during the arrival and welcoming of the test person.

Steps 2–3, Training Phase I In this phase two videos were presented, one with very bad and one with excellent quality, so the test person could learn how to use rating slider and video player and got to know the expected quality range.

Steps 4–6, Training Phase II In these steps the two additional questions were introduced and explained. The three videos demonstrated three different quality patterns and different grades of fluctuation strength. The operator told the participants about the possibility to watch the video more than once and that they could and should use the complete rating scale.

Step 7, Pause Time for open questions before the actual test started and the operator left the test room.

Steps 8–9, Stabilization Phase To further improve the participant’s interpretation of the rating scales, two videos were displayed that were to be repeated in the test phase. The ratings of this phase are discarded and not included in the evaluation.

Steps 10–29, Test Phase Each video sequence of the test set was displayed once and rated by the test person. The sequences were shown in such an order that two successive videos never shared the same content or bit rate.

Step 30, Final Screen The test operator thanked the participant and said goodbye.

On average the test took 15.9 minutes – 4.1 minutes for the training phase, 1.2 minutes for the stabilization, and 10.7 minutes for the test phase.

3.3.2. Reconstruction of Temporal Quality Progression

Reconstruction per Test Subject

To evaluate the temporal quality progression, each of the six different patterns the test subjects had to choose from (third question), is modeled using a simple function. Table 3.6 on page 48 shows the patterns and the assigned reconstruction functions. The quality value of the g -th GOP according to the rating of the s -th test subject q_{gs} is calculated in dependency of the overall quality rating \bar{q}_s (answer to the first question) and the degree of quality fluctuation f_s (answer to the second question). The functions have been designed with respect to two conditions:

- The distance between the highest and lowest value q_{gs} equals the fluctuation strength f_s :

$$\max(q_{gs}) - \min(q_{gs}) = f_s. \quad (3.6)$$

- The mean of all q_{gs} per sequence equals the overall quality rating \bar{q}_s :

$$\frac{1}{G} \sum_{g=0}^G q_{gs} = \bar{q}_s. \quad (3.7)$$

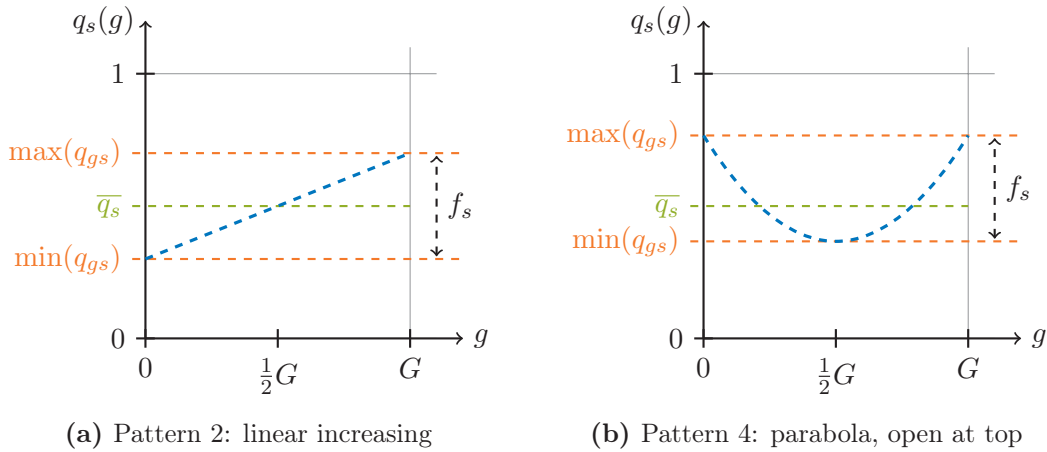


Figure 3.5.: Examples of the reconstruction functions of pattern 2 and pattern 4 as shown in Table 3.6 on the next page

These design criteria are visualized exemplarily in Figure 3.5 for two patterns. G denotes the number of GOPs of the current video sequence and S the number of test subjects.

Reconstruction per Video Sequence

After that, the reconstructed curves for each observer are combined by calculating the arithmetic mean q_g of all q_{gs} :

$$q_g = \frac{1}{S} \sum_{s=0}^S q_{gs}. \quad (3.8)$$

The quality curves q_g are then the values that are later compared to their corresponding predictions \hat{y}_g . Figure 3.6 on page 49 shows an example of how the reconstructions per subject form the main reconstruction.

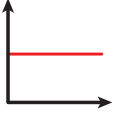
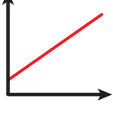
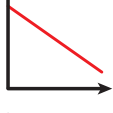

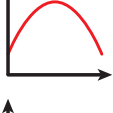
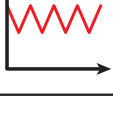
3.3.3. Choice of Video Sequences

The most important requirement for suitable video sequences were clearly visible quality changes over time. I ran the temporal quality prediction on all three datasets described in 3.2.1 on page 36 and compared the variance in the quality progression per sequence. Since the IT-IST dataset clearly showed the most fluctuation over time, this dataset was chosen for the evaluation of the temporal quality prediction.

One has to admit that the CIF-resolution of the IT-IST dataset is far from state-of-the-art by now, but the two High Definition (HD) datasets TUM 1080p25 and TUM 1080p50 did not fulfill the requirement of visible quality fluctuation. Presumably the main reason for this is the comparatively high quality of all sequences in these datasets. It would have been possible to create a new dataset with sequences that show some visible quality

3. Evaluation of GOP-based Video Quality Metrics

Table 3.6.: Quality Patterns and Reconstruction Functions

Pattern	Reconstruction Function	Icon
1 constant	$q_{gs} = \bar{q}_s$	
2 linear increasing	$q_{gs} = \bar{q}_s - \frac{f_s}{2} + \frac{f_s}{G} \cdot g$	
3 linear decreasing	$q_{gs} = \bar{q}_s + \frac{f_s}{2} - \frac{f_s}{G} \cdot g$	
4 parabola, open at top	$q_{gs} = \frac{4f_s}{G^2} \cdot g^2 - \frac{4f_s}{G} \cdot g + \bar{q}_s + \frac{2f_s}{3}$	
5 parabola, open at bottom	$q_{gs} = -\frac{4f_s}{G^2} \cdot g^2 + \frac{4f_s}{G} \cdot g + \bar{q}_s - \frac{2f_s}{3}$	
6 oscillating	$q_{gs} = \frac{f_s}{2} \cos(g) + \bar{q}_s$	

changes, but due to the expected high expenditure of this option I decided to work with the tried and tested IT-IST dataset. It should still be possible to prove the basic concept of GOP-based video quality prediction.

Since the full amount of 48 sequences would have been too much for a single subjective test, I selected a subset of the full dataset consisting of the five sequences *City*, *Football*, *Foreman*, *Table* and *Tempete*. In addition, a broad range of both MOS values and expected quality fluctuation strengths was covered. This led to the number of $N = 20$ video sequences with in total $N_G = 320$ GOPs. To avoid the negative effects of a decreased number of training sequences, the complete set of 48 sequences was used to train the regression models.

3.3.4. Results of the Subjective Test

In order to discuss the results of the subjective test, I will deal with the three questions of the test and the corresponding answers separately. Section 3.3.5 on page 52 will then discuss the results of the curve reconstruction.

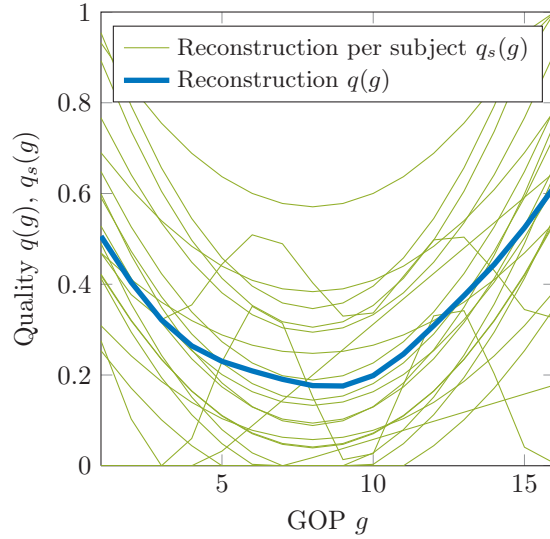


Figure 3.6.: Example for the reconstruction of temporal quality curves from the test answers. The green lines are the results of the reconstruction per test subject (cf. section 3.3.2), the blue line is the average of the green curves. The video sequence used in this figure is *Football* from the IT-IST dataset at the rate point 256 kBit/s

First Question: Overall Quality Impression

The validity of the subjective test can easily be checked by calculating the so-called inter-lab correlation. Therefore, I compared my test results to those from the test by Brandão and Queluz [4]. Both subjective tests had similar test conditions, but with DCR [15] they used a double stimulus method. Nevertheless, both the correlation coefficients and the RMSE as shown in Table 3.7 indicate a highly significant similarity between the results of both tests. The corresponding scatter plot is shown in Figure 3.7 on the following page. The complete list of the raw subjective data is shown in the appendix in Table A.1 on page 60.

This high inter-lab correlation is important, as I will compare the prediction results

Table 3.7.: Inter-lab correlation of the MOS calculated from the answers to the first question of my subjective test and the corresponding MOS values from IT-IST by Brandão and Queluz [4].

Sequence	Pearson	Spearman	RMSE
City	0.970	1.000	0.112
Football	0.996	1.000	0.091
Foreman	0.997	1.000	0.098
Table	0.997	1.000	0.144
Tempete	0.998	1.000	0.072
all	0.976	0.959	0.106

3. Evaluation of GOP-based Video Quality Metrics

of a regression model that has been trained with the subjective values from Brandão and Queluz [4] to the reconstructed curves using my subjective results. Furthermore, it shows that at least the first question of my test provides the high-quality results that are expected from a subjective laboratory test. This indicates that the the two additional questions and the self-operation of the test do not distract the observers from the judgement of the overall video quality.

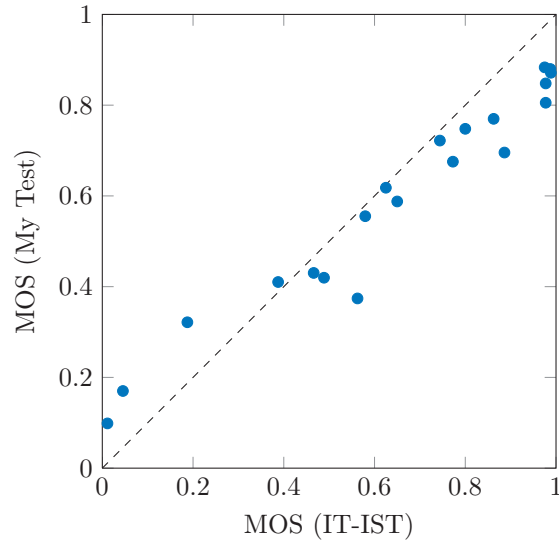


Figure 3.7.: Scatter plot corresponding to the values shown in Table 3.7

Second Question: Quality Fluctuation Strength

The answers to the second question cannot be validated easily. This is because there is no similar subjective data I could compare my results with. Consequently, I can only consider how plausible the ratings as shown in Table A.2 on page 61 are.

The bars in Figure 3.8 on the next page show the averages of the fluctuation ratings per sequence. Several things attract attention: First of all, the fluctuation strength seems to be strongly correlated with the subjective quality (Pearson correlation coefficient of -0.79). This is no surprise, as the better quality levels naturally show less quality fluctuations. Furthermore, the highest bit rates are always considered to have very little fluctuations. These things make the answer data to this question very plausible.

On the other hand, the plot shows that the statistical dispersion of the answers is fairly high. Presumably this is due to the complexity of the question. Despite the fact that the rating scale was explained in detail to each participant, it is still much less trivial than just asking for an overall quality impression. While watching the test subjects, I noticed that most of them started a test step by answering the first and third question. Then they often watched the video again and tried to rate the fluctuation strength (second question). Most observers made a less confident impression when moving the slider for this question than when rating the overall quality. I think at this point it becomes clear

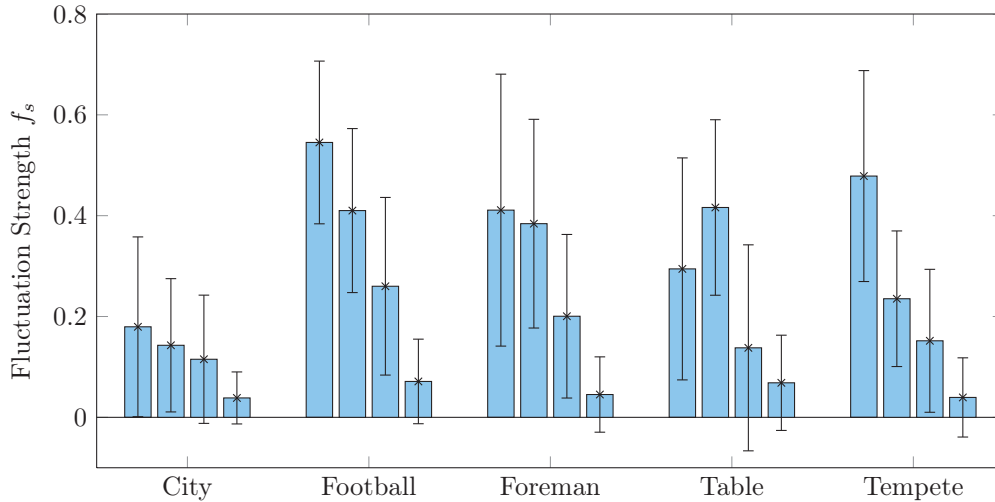


Figure 3.8.: Mean and standard deviation of the answers to the second question about the strength of the quality fluctuation for the 5 sequences in the IT-IST subset. Each bar corresponds to one of 4 different rate points from the lowest to the highest bit rates.

that the test method I chose is probably the maximum complexity one can ask of a non-expert test participant. Since the capacity of the human working-memory is limited (cf. Miller [33]), it is not possible to concentrate on several different things at once. Nevertheless, since the average results are plausible, the results seem to be useful for the intended purpose of reconstructing the quality progression.

Third Question: Quality Progression Pattern







The third question poses the same problem as the second one. There is no data available for comparison, so once again I will focus on the plausibility of the results. The complete set of answer data can again be found in the appendix in Table A.2 on page 61.

Since this question is not based on a rating scale that results in continuous values, statistical methods are not helpful when evaluating the data. The question can be considered rather a choice than a rating on a scale and therefore I calculated the percentages of the options that were chosen for each sequence. Table 3.8 on the next page shows the pattern selected most often per sequence and the corresponding percentage. One can see in this table that in 17 out of 20 cases more than half of all observers agreed on one pattern – 15 times it were even more than two thirds. This indicates that the six options that were provided were a reasonable choice, as apparently the test subjects were able to express their perception quite well with the help of the six patterns.

At first, it seems to be a good idea to discard all votes that do not agree with the majority, but actually I rather think it is important to take them into account as well. Although a particular observer might not agree with the most common voting, his or her perception might still be valuable for the curve reconstruction. If, for example, the real quality progression is similar to the addition of two patterns, some observers

3. Evaluation of GOP-based Video Quality Metrics

Table 3.8.: Percentages of the answer data of the third question about the quality pattern. Rate point (RP) 1 corresponds to the lowest bit rate, RP 4 to the highest.

Sequence	RP							Majority
City	1	71.4	4.8				23.8	1
	2	66.7	14.3	4.8			14.3	1
	3	66.7	9.5	4.8			19.0	1
	4	100.0						1
Football	1		9.5		81.0		9.5	4
	2	4.8	9.5		66.7	9.5	9.5	4
	3	23.8	9.5		66.7			4
	4	76.2		4.8	19.0			1
Foreman	1	19.0		76.2	4.8			3
	2	14.3	9.5	14.3	42.9	4.8	14.3	4
	3	47.6	4.8	14.3	28.6		4.8	1
	4	90.5		4.8	4.8			1
Table	1	33.3		33.3	4.8		28.6	1/3
	2	4.8	57.1		23.8		14.3	2
	3	71.4	9.5		9.5		9.5	1
	4	81.0	9.5		9.5			1
Tempete	1	4.8		4.8			90.5	6
	2	33.3	4.8	4.8			57.1	6
	3	66.7	4.8	14.3			14.3	1
	4	90.5			4.8		4.8	1
all sequences		48.3	7.9	9.0	18.3	0.7	15.7	

might vote for one of the two patterns and the remaining observers for the other. In the reconstruction step, the two options are then combined and a much more accurate reconstruction will be the result.

3.3.5. Results of the Temporal Quality Prediction

As discussed in the section about length-independent quality prediction, I chose the regression method and the number of used factors R that results in the best prediction performance. For the evaluation of the temporal quality prediction Tri-PLS1-based prediction with $R = 3$ factors was used.

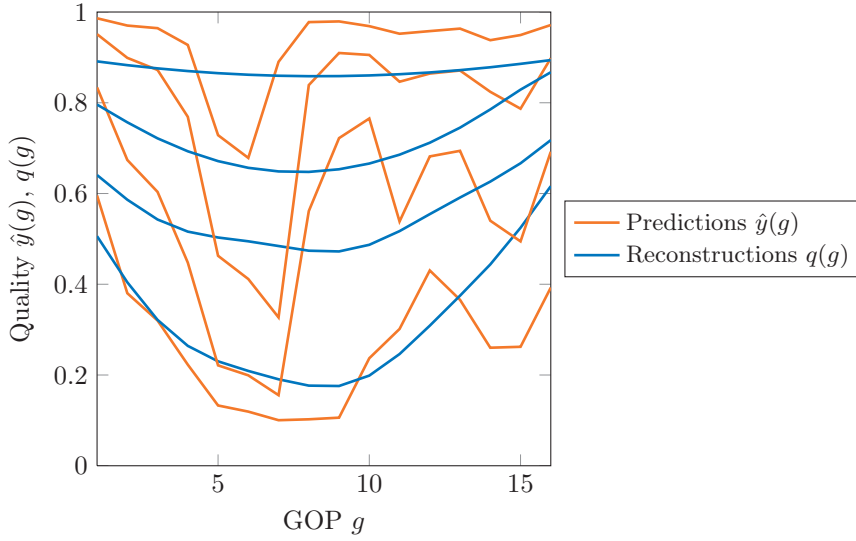


Figure 3.9.: Quality curve reconstructions and predictions of all rate points of the sequence *Football*

Quality Curve Reconstruction

The reconstruction of the quality curves from the test results was done as described in section 3.3.2 on page 46. The plots of the resulting curves and the corresponding predictions are shown in appendix A.3 on page 67.

Figure 3.9 shows the quality progression curves for the sequences with the content *Football* as an example. This sequence is the most interesting as it contains the strongest quality fluctuation strength of the whole dataset. One can see that the reconstructed and the predicted curves do look quite similar but the prediction contains more details. Obviously, the reconstructed curves are not able to represent the exact shape of the quality progression as it is predicted; more complex curve shapes are not possible because of the limitation to the six patterns. In addition, it is arguable whether a human observer could resolve such high frequency changes of visual quality.

Statistical Evaluation

In order to also numerically compare the temporal quality prediction to the reconstructed quality curves, the correlation coefficients and the RMSE of the quality values are calculated per GOP. That means for all $N_G = 320$ GOPs a pair of \hat{y}_i and q_i is built and used for statistics. Table 3.9 on the next page shows the results for each video sequence and Figure 3.10 on the following page, the corresponding scatter plot. The individual scatter plots for each sequence can be found in appendix A.2.2 on page 66.

First of all, the correlation coefficients are on the very high level of about 0.9 and the RMSE is reasonably low, which are a very good values for a no-reference video quality metric. The individual sequences show even higher correlation – except for the sequence

3. Evaluation of GOP-based Video Quality Metrics

Table 3.9.: Correlation coefficients and RMSE between reconstructed and predicted quality progression curves

Sequence	Pearson	Spearman	RMSE
City	0.989	0.963	0.142
Football	0.871	0.877	0.149
Foreman	0.942	0.944	0.126
Table	0.963	0.963	0.085
Tempete	0.947	0.965	0.103
all	0.893	0.898	0.124

Football. As stated above, this sequence has the strongest quality changes over time, and especially the heavy drop in quality around GOP 7 cannot be represented by the subjective test method. At this point of time the camera pans very quickly and at the lower bit rates the visual quality becomes really bad.

Conclusion

In general, the results of the temporal quality prediction look very promising. Of course, it is somehow problematic to evaluate an objective quality metric with an untested subjective test methodology and vice versa; in this case it was the only option. Particularly with video sequences as short as the ones used here the proposed test method seems to work really well. The fact that the quality prediction per GOP provides good results has

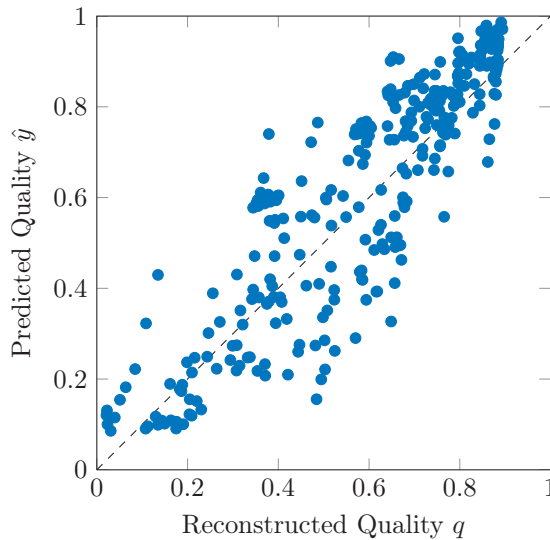


Figure 3.10.: Scatter plot of the reconstructed and predicted GOP quality values for the temporal quality metric

3.3. Evaluation of Temporal Quality Prediction

already been stated in the section about length-independent quality prediction, so it is very plausible that the quality prediction for the individual GOPs provides reasonable results as well.

A general problem of the temporal quality prediction as described here is that there is no temporal subjective data available to train the model with. It can be assumed that the performance would further improve if real temporal data was used instead of the overall MOS, that can be measured by the known subjective methods.

4. Summary

In this thesis I was able to achieve three things. At first, I showed how to improve the design of video quality metrics using multi-way data analysis by making the training and the quality prediction independent of the length of the video sequences. The proposed no-reference metric is based on features extracted from H.264/AVC bitstreams and makes use of the GOP-structure of H.264/AVC encoded video. It turns out that averaging the per-GOP estimated quality values of a video sequence results in a quality estimation that correlates very well with the perceived quality as measured in subjective tests. The statistical results show that this metric performs equally well as a corresponding length-dependent metric and outperforms common full-reference metrics. Apart from that, the main advantage of the presented metric is its improved universality when it comes to real world application. What remains is the drawback that all video sequences need to be encoded with the same GOP-length, but this is less inconvenient than demanding equal lengths for the complete sequences. What is more, this is very common in broadcasting applications.

Then I developed a new method for subjective tests of visual video quality that allows the testing of temporal quality progression. The method is well-suited for the short video sequences (about 10 seconds) that most of the popular datasets consist of. At least the approximate shape of the temporal quality changes in a video sequences can be represented by three comparatively easy questions. A subjective test using this method needs to be software-based as the test subjects should be able to control the test autonomously. Therefore, I modified the existing software QualityCrowd 2 in order to support the test method.

Finally, I showed that the temporal prediction per GOP is highly correlated to the subjective results obtained by the presented method. This creates the possibility to use data analysis-based video quality metrics in order to predict the visual quality with a high sampling rate. As with the length-independent quality prediction this does not require the introduction of new mathematical concepts; it is only necessary to split the video sequence into small subsets. In order to validate the temporal quality prediction, I carried out a subjective test according to the proposed test method with 21 participants and reconstructed the temporal quality from their answers.

Additionally, my results confirm the validity of the multi-way PLSR-based metric as presented by Keimel et al. [20] and the 2D-PCR based metric as discussed in [27]. By applying the metrics to different datasets than in the original publications, I can confirm that by taking into account the temporal dimensions metrics based on data analysis improve.

Despite the fact that overall the results presented in this thesis are very positive, there are some open questions left. Especially the temporal quality prediction and the

4. Summary

new subjective test method require some more investigation. Although I did argue that SSCQE is not a suitable method for the short video sequences used here, one should apply the temporal metric presented here on longer sequences and compare the results to SSCQE data. My subjective method needs to be validated in more tests; the one I carried out is of course not sufficient to finally prove the validity of the method. The length-independent quality prediction seems to work very well but, as I have discussed, there can be some problems when using datasets, that are too small. In general, I think there is considerable need for a more detailed evaluation of the influence of the size and nature of training sets on data analysis-based quality metrics.

A. Additional Tables and Figures

A.1. Results of the Subjective Test

Table A.1.: Test results of the first question on the overall quality rating \bar{q}_s

Sequence	RP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
City	1	0.38	0.29	0.22	0.43	0.23	0.58	0.25	0.49	0.60	0.46	0.44	0.37	0.52	0.34	0.16	0.47	0.32	0.39	0.23	0.36	0.32
	2	0.69	0.48	0.78	0.71	0.62	0.59	0.65	0.48	0.58	0.70	0.64	0.41	0.68	0.51	0.34	0.65	0.51	0.50	0.78	0.59	0.44
	3	0.83	0.92	0.87	0.77	0.82	0.68	0.81	0.88	0.68	0.82	0.74	0.60	1.00	0.54	0.43	0.64	0.81	0.71	0.83	0.60	0.72
	4	0.99	1.00	0.84	0.99	0.81	0.82	1.00	0.81	0.78	0.82	0.67	1.00	1.00	0.82	0.84	1.00	0.91	1.00	0.86	0.82	0.77
Football	1	0.73	0.42	0.16	0.28	0.21	0.50	0.21	0.23	0.39	0.56	0.52	0.21	0.03	0.29	0.17	0.08	0.34	0.34	0.31	0.18	0.61
	2	0.85	0.27	0.50	0.59	0.62	0.58	0.61	0.71	0.49	0.70	0.62	0.37	0.38	0.53	0.62	0.56	0.60	0.63	0.44	0.60	0.39
	3	0.94	0.59	0.85	0.68	0.98	0.86	0.61	0.81	0.69	0.92	0.88	0.73	0.39	0.88	0.67	0.27	0.79	0.71	0.56	0.64	0.71
	4	1.00	0.79	0.94	0.95	0.83	0.97	1.00	0.89	0.86	1.00	0.90	0.78	0.73	0.84	0.71	0.78	1.00	0.93	0.86	0.77	0.76
Foreman	1	0.05	0.09	0.00	0.07	0.08	0.32	0.00	0.18	0.06	0.03	0.19	0.00	0.00	0.19	0.06	0.00	0.11	0.15	0.13	0.20	0.15
	2	0.47	0.33	0.26	0.47	0.36	0.73	0.37	0.51	0.41	0.46	0.46	0.38	0.39	0.34	0.21	0.47	0.39	0.51	0.37	0.51	0.43
	3	1.00	0.51	0.48	0.73	0.77	0.87	0.72	0.91	0.54	0.80	0.52	0.71	0.66	0.73	0.53	0.62	0.61	0.61	0.61	0.66	0.60
	4	1.00	0.89	0.74	0.90	0.91	0.81	1.00	1.00	0.78	0.96	0.71	0.87	1.00	0.81	0.70	0.62	1.00	1.00	0.86	0.57	0.68
Table	1	0.17	0.09	0.00	0.25	0.18	0.21	0.11	0.29	0.28	0.26	0.24	0.13	0.05	0.12	0.16	0.07	0.13	0.00	0.27	0.33	0.22
	2	0.56	0.29	0.33	0.38	0.31	0.64	0.41	0.47	0.51	0.39	0.45	0.48	0.50	0.38	0.17	0.62	0.42	0.42	0.43	0.51	0.38
	3	0.71	0.37	0.64	0.65	0.65	0.80	0.81	1.00	0.74	0.81	0.66	0.61	0.64	0.69	0.61	0.98	0.63	1.00	0.50	0.56	0.56
	4	1.00	0.70	0.86	0.71	0.83	0.93	1.00	0.92	0.77	0.78	0.71	0.86	0.73	0.78	0.51	0.99	1.00	0.79	0.73	0.62	0.70
Tempete	1	0.58	0.52	0.11	0.39	0.28	0.64	0.49	0.39	0.40	0.35	0.44	0.51	0.34	0.34	0.09	0.49	0.39	0.51	0.39	0.53	0.42
	2	0.87	0.49	0.66	0.63	0.58	0.71	0.92	0.59	0.69	0.64	0.57	0.35	0.48	0.43	0.42	1.00	0.71	0.62	0.66	0.47	0.49
	3	0.83	0.61	0.96	0.85	0.64	0.85	1.00	0.72	0.68	0.63	0.64	0.96	0.73	0.76	0.86	1.00	0.78	0.68	0.77	0.64	0.58
	4	1.00	0.87	0.84	0.97	0.83	0.88	1.00	0.81	0.76	0.92	0.63	1.00	0.72	0.87	0.81	0.92	1.00	1.00	0.99	0.93	0.74

Table A.2.: Test results of the second question on the fluctuation strength rating f_s

Sequeunce RP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
City	1	0.00	0.11	0.12	0.29	0.65	0.44	0.00	0.00	0.28	0.31	0.00	0.08	0.00	0.03	0.04	0.28	0.40	0.04	0.24	0.23	0.22
	2	0.52	0.10	0.12	0.18	0.21	0.36	0.11	0.00	0.15	0.02	0.00	0.02	0.26	0.02	0.26	0.22	0.00	0.18	0.04	0.12	0.12
	3	0.27	0.00	0.02	0.10	0.27	0.39	0.16	0.00	0.07	0.07	0.00	0.00	0.00	0.04	0.18	0.07	0.00	0.11	0.11	0.16	0.42
	4	0.00	0.00	0.04	0.03	0.12	0.06	0.00	0.00	0.00	0.08	0.00	0.00	0.01	0.04	0.02	0.07	0.00	0.01	0.04	0.12	0.19
Football	1	0.47	0.38	0.58	0.57	0.66	0.62	0.50	0.84	0.60	0.77	0.40	0.49	0.30	0.48	0.33	0.41	0.58	0.29	0.67	0.72	0.79
	2	0.38	0.38	0.77	0.57	0.23	0.36	0.26	0.38	0.35	0.30	0.00	0.51	0.51	0.26	0.52	0.37	0.34	0.58	0.61	0.48	0.46
	3	0.21	0.38	0.36	0.53	0.01	0.31	0.00	0.44	0.10	0.25	0.00	0.19	0.30	0.00	0.50	0.58	0.13	0.34	0.29	0.26	0.27
	4	0.00	0.09	0.03	0.11	0.03	0.06	0.00	0.00	0.03	0.07	0.00	0.14	0.26	0.02	0.12	0.00	0.00	0.01	0.23	0.07	0.23
Foreman	1	0.46	0.21	0.01	0.89	0.76	0.84	0.26	0.51	0.02	0.36	0.71	0.78	0.00	0.48	0.50	0.26	0.47	0.35	0.15	0.30	0.34
	2	0.44	0.41	0.27	0.52	0.78	0.45	0.76	0.00	0.33	0.64	0.00	0.23	0.41	0.16	0.46	0.16	0.43	0.34	0.34	0.43	0.51
	3	0.00	0.38	0.10	0.24	0.24	0.27	0.42	0.00	0.13	0.58	0.00	0.19	0.37	0.02	0.13	0.11	0.00	0.12	0.31	0.28	0.32
	4	0.00	0.14	0.08	0.00	0.04	0.12	0.00	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.04	0.02	0.00	0.00	0.00	0.06	0.31
Table	1	0.34	0.39	0.26	0.33	0.58	0.38	0.00	0.00	0.26	0.29	0.00	0.00	0.46	0.03	0.59	0.63	0.62	0.00	0.46	0.28	0.28
	2	0.64	0.51	0.27	0.56	0.76	0.49	0.50	0.00	0.33	0.64	0.28	0.28	0.37	0.32	0.42	0.18	0.33	0.35	0.53	0.56	0.44
	3	0.34	0.56	0.27	0.08	0.01	0.12	0.00	0.00	0.01	0.11	0.00	0.00	0.00	0.00	0.27	0.00	0.00	0.00	0.02	0.48	0.63
	4	0.00	0.19	0.04	0.27	0.06	0.05	0.00	0.00	0.06	0.03	0.00	0.00	0.02	0.00	0.07	0.01	0.00	0.04	0.06	0.23	0.30
Tempete	1	0.48	0.36	0.75	0.36	0.94	0.51	0.26	0.77	0.40	0.63	0.21	0.61	0.52	0.32	0.04	0.52	0.44	0.65	0.27	0.41	0.60
	2	0.23	0.28	0.37	0.15	0.22	0.33	0.12	0.00	0.22	0.27	0.21	0.23	0.43	0.02	0.50	0.00	0.30	0.27	0.13	0.28	0.39
	3	0.31	0.20	0.03	0.12	0.32	0.12	0.00	0.43	0.26	0.14	0.00	0.00	0.29	0.04	0.27	0.00	0.00	0.05	0.04	0.17	0.39
	4	0.00	0.12	0.08	0.00	0.00	0.12	0.00	0.00	0.03	0.04	0.00	0.00	0.01	0.00	0.06	0.03	0.00	0.00	0.01	0.00	0.34

Table A.3.: Test results of the third question on the quality patterns p_s

Sequence	RP	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
City	1	1	1	1	6	6	6	1	1	1	6	1	1	1	1	2	6	1	1	1	1	1	
	2	2	1	1	1	1	6	6	1	2	1	1	1	6	1	3	2	1	1	1	1	1	
	3	6	1	1	1	6	6	6	1	1	1	1	1	1	1	3	2	1	1	1	1	2	
	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Football	1	4	6	4	4	4	4	4	4	4	4	4	4	2	4	4	4	4	4	4	4	6	4
	2	4	6	4	4	4	4	4	4	4	4	1	2	6	4	4	5	4	4	4	4	4	4
	3	4	4	4	4	4	1	4	1	2	1	4	4	2	1	4	4	4	4	4	4	4	4
	4	1	4	4	1	1	1	1	1	1	1	1	1	4	1	3	1	1	1	4	4	1	4
Foreman	1	3	4	1	3	3	3	3	3	1	3	3	3	1	3	3	3	3	3	1	3	3	3
	2	4	4	2	4	6	3	2	1	5	4	1	3	4	4	6	1	4	3	6	4	4	4
	3	1	4	1	1	1	3	2	1	4	4	1	4	1	3	3	1	1	1	6	4	4	4
	4	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3	1
Table	1	4	6	3	3	6	6	1	1	3	3	1	1	6	1	3	6	6	1	3	1	3	3
	2	2	6	2	4	2	2	2	1	6	2	6	2	2	4	2	2	2	2	4	4	4	4
	3	6	4	2	1	1	1	1	1	1	1	1	1	1	1	6	1	1	1	1	1	4	4
	4	1	4	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	2
Tempete	1	6	6	6	6	6	6	6	6	6	6	6	6	6	6	1	6	6	6	6	6	3	6
	2	6	6	6	1	1	6	3	1	6	6	6	6	2	1	6	1	6	6	1	1	1	6
	3	6	3	1	1	1	1	1	6	2	1	1	1	3	1	3	1	1	1	1	1	1	6
	4	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	6

A.2. Additional Scatter Plots

A.2.1. Length-Independent Quality Prediction

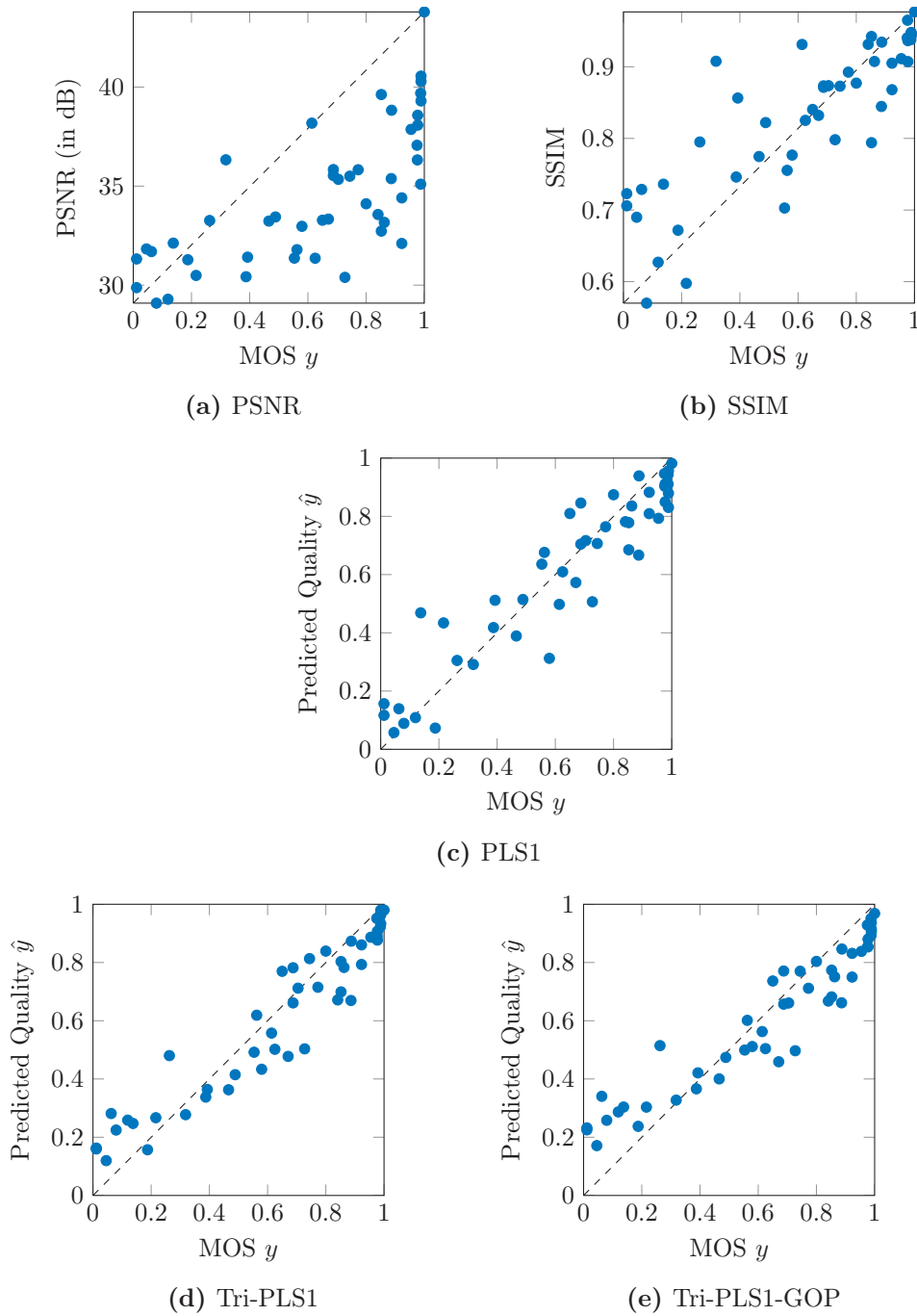


Figure A.1.: Scatter plots of the compared metrics on the IT-IST dataset

A. Additional Tables and Figures

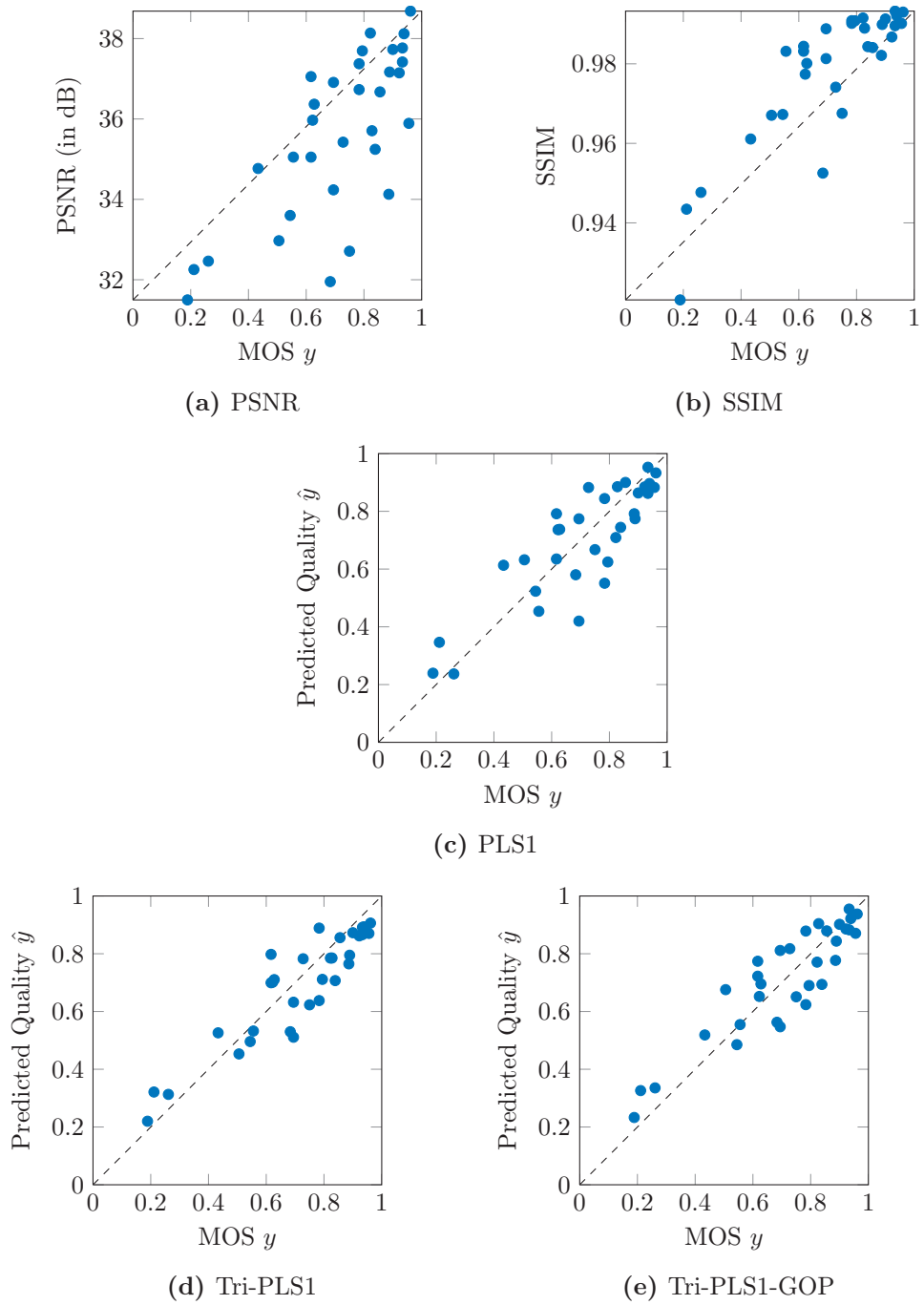


Figure A.2.: Scatter plots of the compared metrics on the TUM 1080p25 dataset

A.2. Additional Scatter Plots

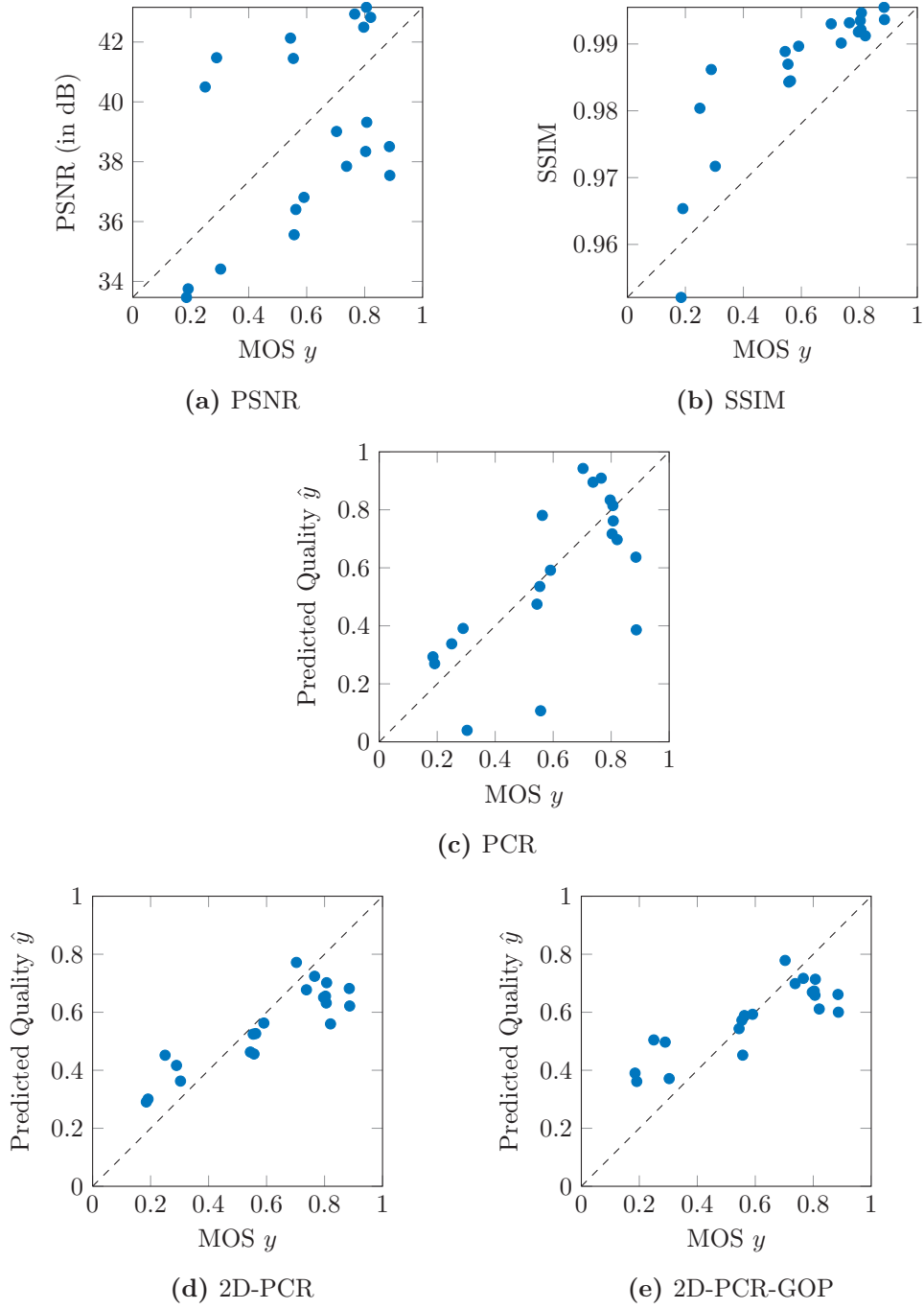


Figure A.3.: Scatter plots of the compared metrics on the TUM 1080p50 dataset

A.2.2. Temporal Quality Prediction

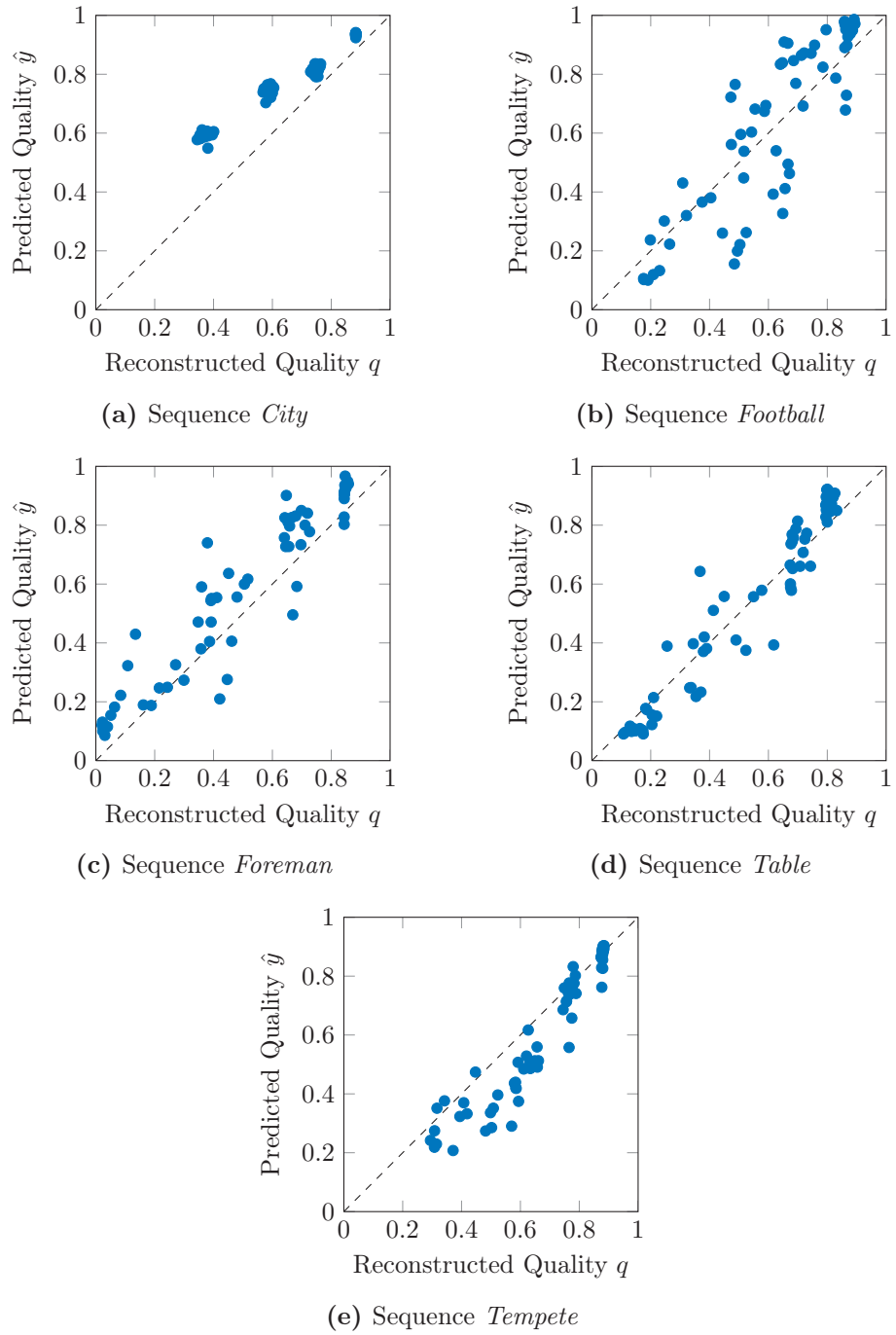


Figure A.4.: Scatter plots of the quality predictions and reconstructions of each GOP from the subset of the IT-IST dataset

A.3. Plots of the Temporal Quality Prediction

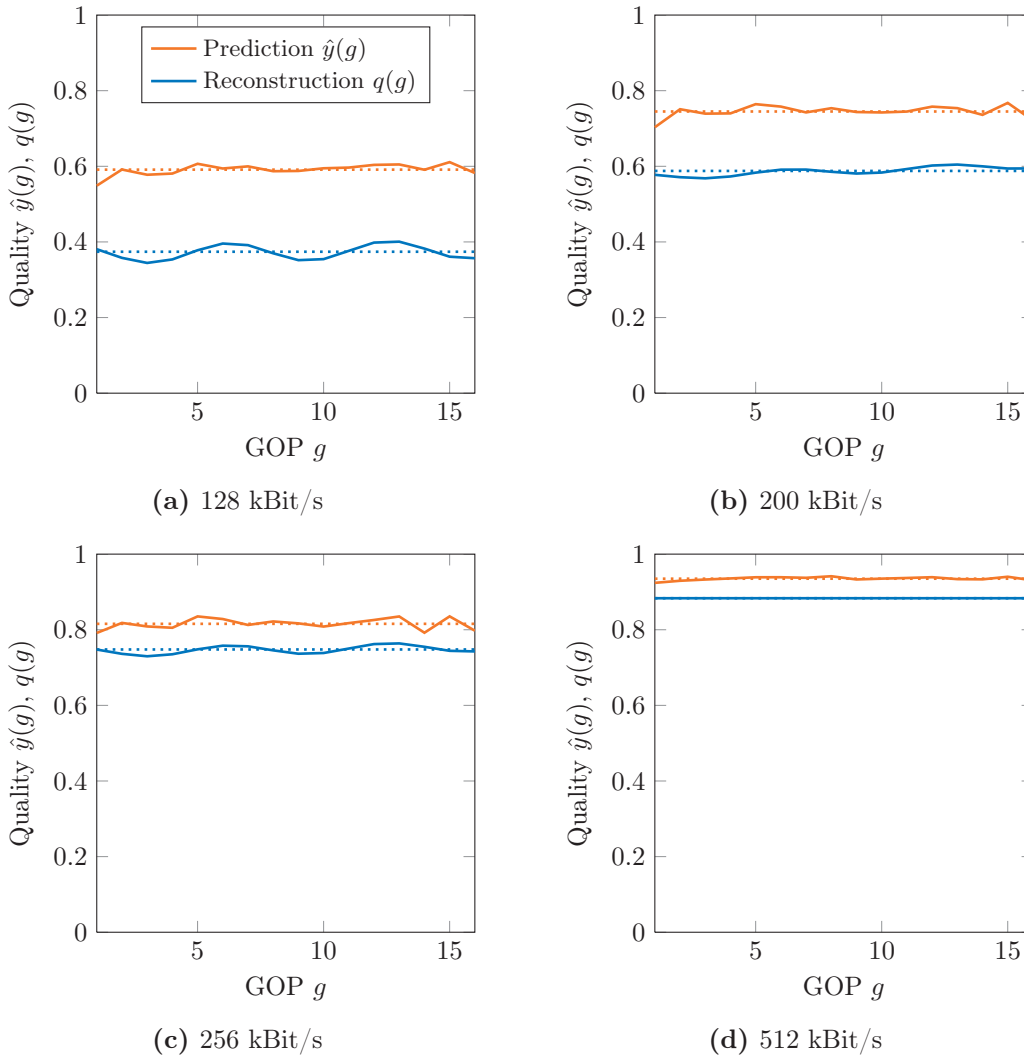


Figure A.5.: Quality curve predictions and reconstructions for the sequence *City* from the IT-IST dataset. The dotted lines represent the corresponding means.

A. Additional Tables and Figures

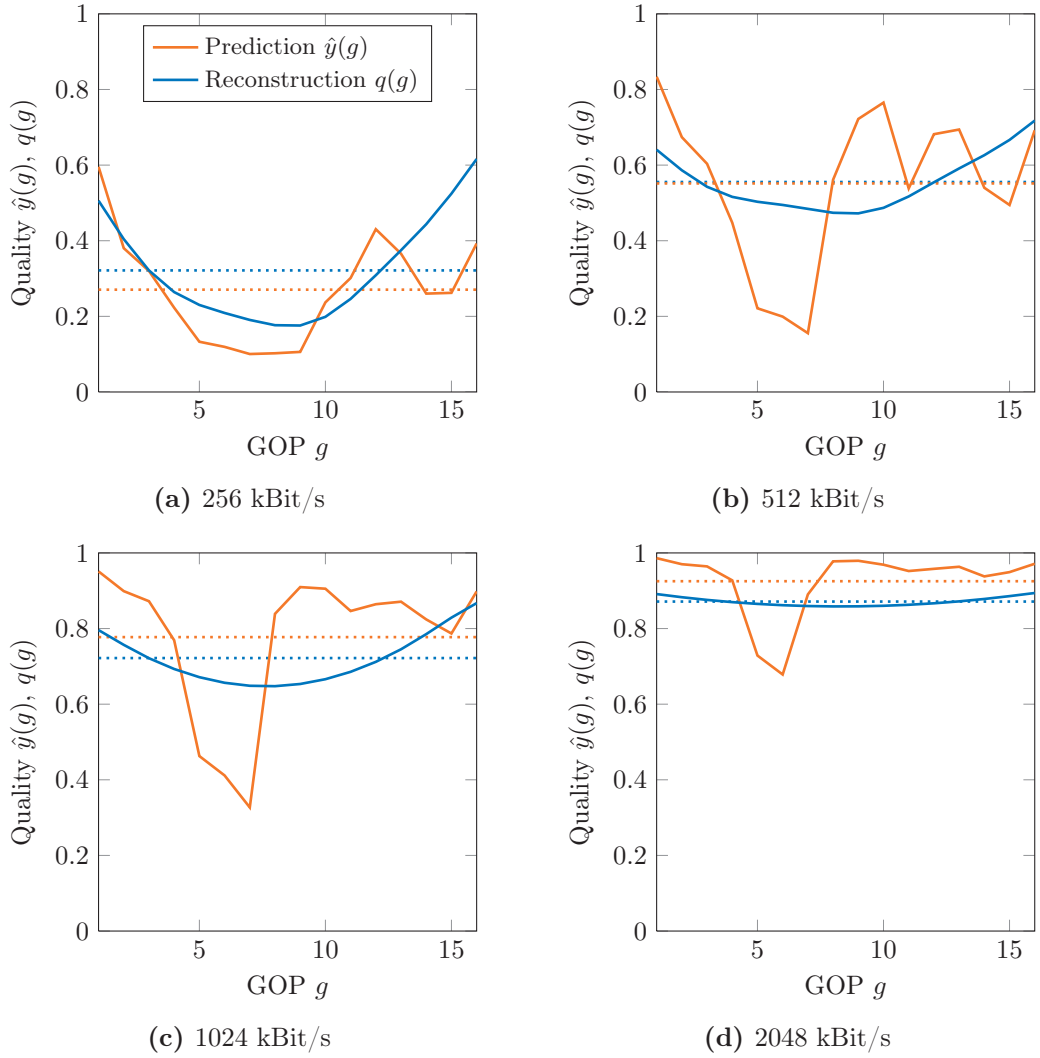
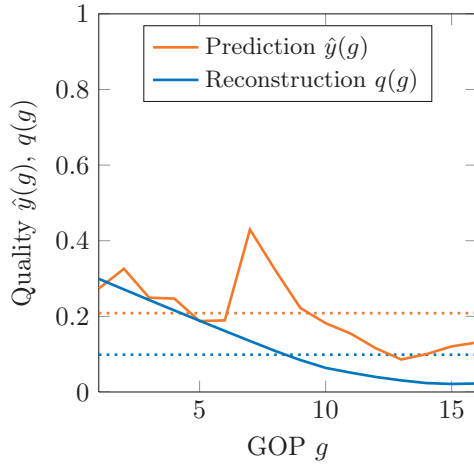
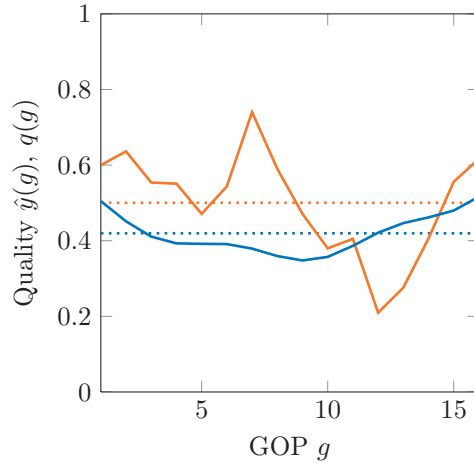


Figure A.6.: Quality curve predictions and reconstructions for the sequence *Football* from the IT-IST dataset. The dotted lines represent the corresponding means.

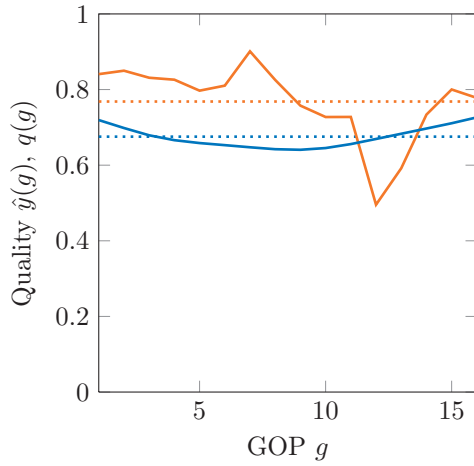
A.3. Plots of the Temporal Quality Prediction



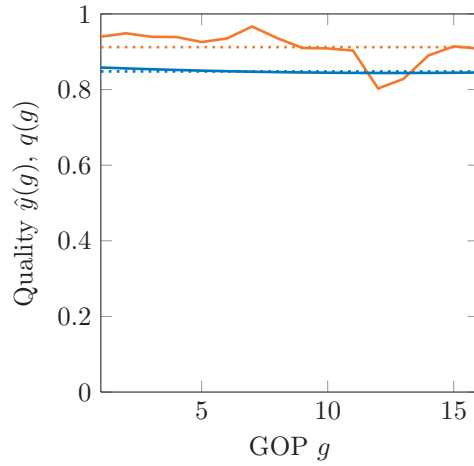
(a) 64 kBit/s



(b) 128 kBit/s



(c) 256 kBit/s



(d) 512 kBit/s

Figure A.7.: Quality curve predictions and reconstructions for the sequence *Foreman* from the IT-IST dataset. The dotted lines represent the corresponding means.

A. Additional Tables and Figures

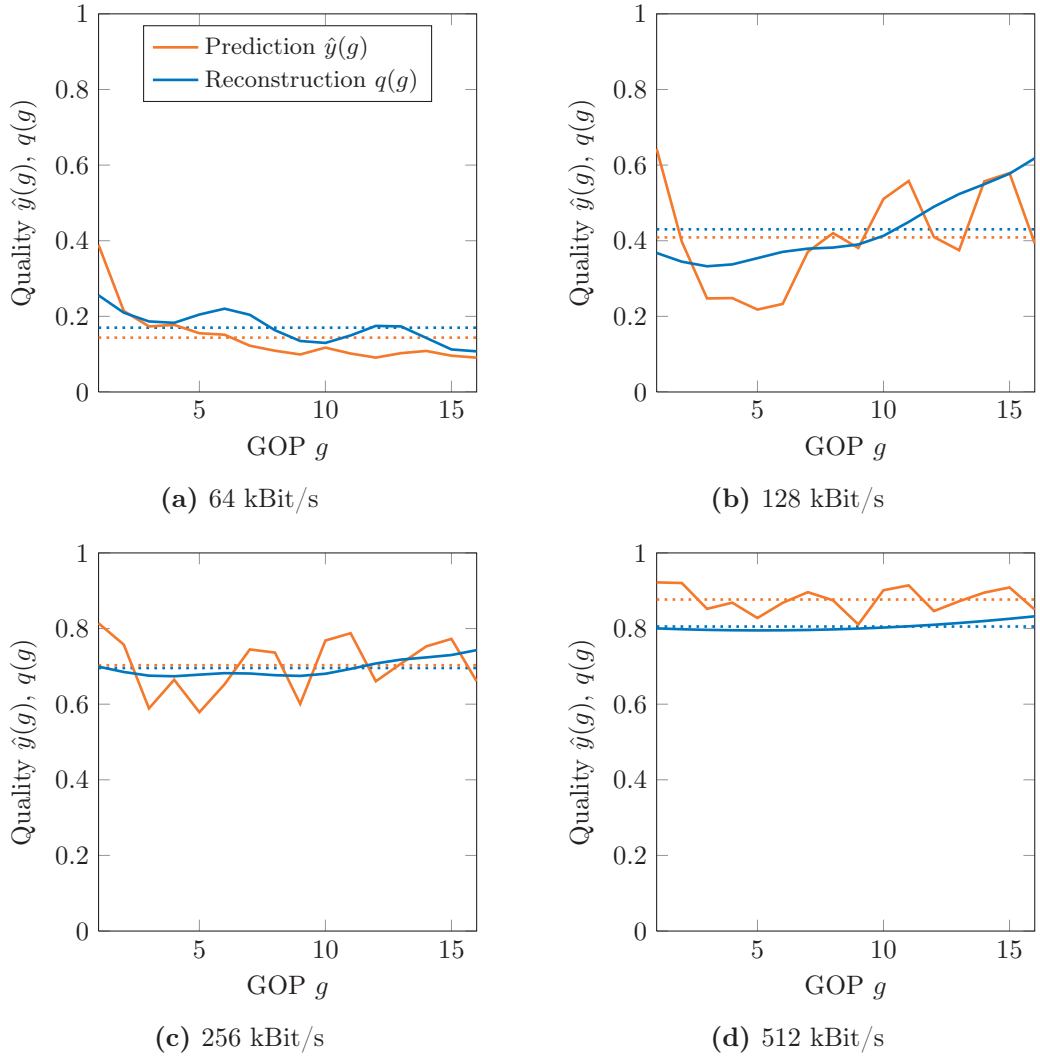


Figure A.8.: Quality curve predictions and reconstructions for the sequence *Table* from the IT-IST dataset. The dotted lines represent the corresponding means.

A.3. Plots of the Temporal Quality Prediction

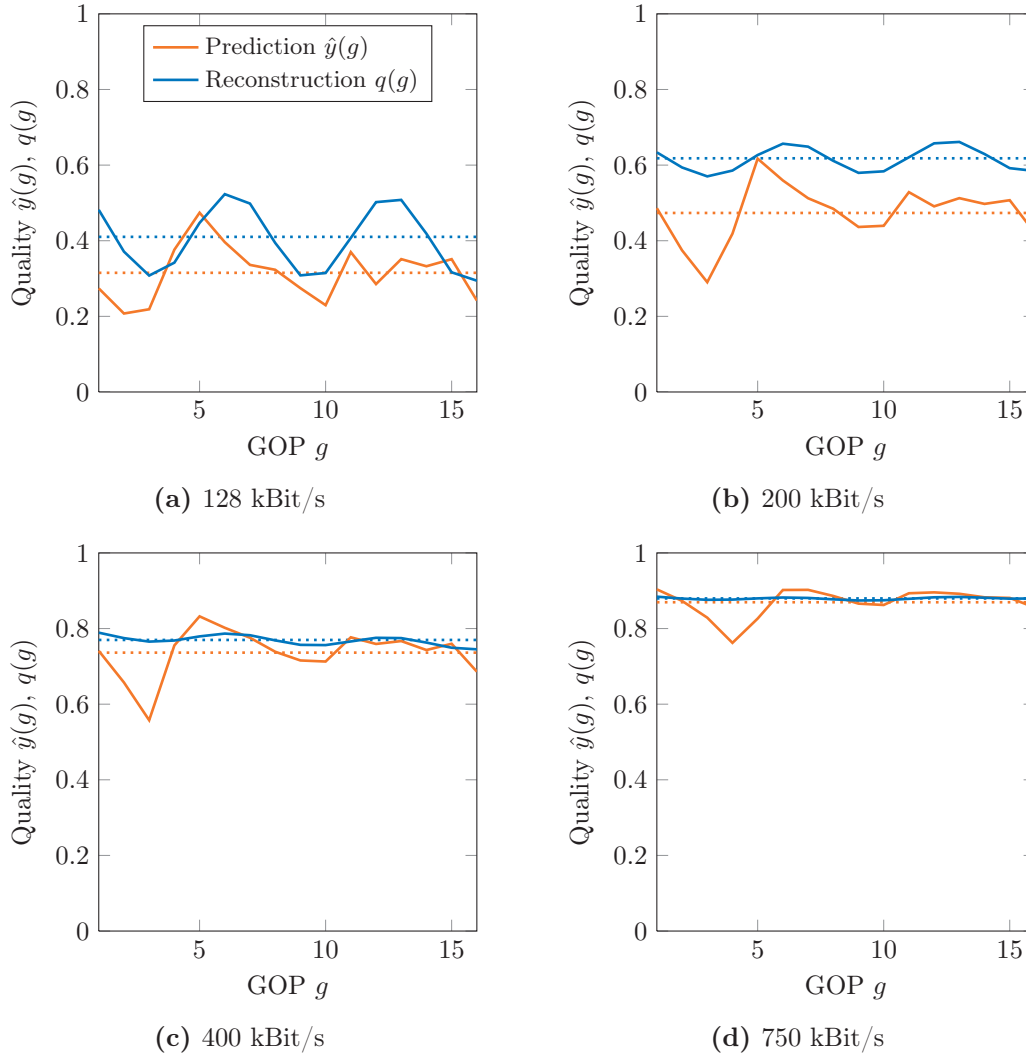


Figure A.9.: Quality curve predictions and reconstructions for the sequence *Tempete* from the IT-IST dataset. The dotted lines represent the corresponding means.

List of Figures

2.1.	The three-way feature cube and its different slices	26
2.2.	Example for a Group of Pictures (GOP)	29
2.3.	Changes in the dimensionality of the feature array when moving from a model based on complete video sequences to a GOP-based approach . . .	29
2.4.	Temporal quality prediction of the first three minutes taken from an episode of a popular TV-series encoded with three different bit rates . . .	30
2.5.	Sigmoid Correction Function	31
3.1.	Scatter plots of the quality prediction of the metrics Tri-PLS1 and Tri-PLS1-GOP for the IT-IST dataset	42
3.2.	Examples for predictions of the different metrics on the video sequence <i>Stephan</i> from the IT-IST dataset at 265 kbit/s	43
3.3.	Rating scales for temporal quality assessment	44
3.4.	Photography of the LDV video lab and a screenshot of the QualityCrowd 2 software	45
3.5.	Examples of the reconstruction functions of pattern 2 and pattern 4 . . .	47
3.6.	Example for the reconstruction of temporal quality curves from the test answers	49
3.7.	Scatter plot corresponding to the values shown in Table 3.7	50
3.8.	Mean and standard deviation of the answers to the second question about the strength of the quality fluctuation	51
3.9.	Quality curve reconstructions and predictions of all rate points of the sequence <i>Football</i>	53
3.10.	Scatter plot of the reconstructed and predicted GOP quality values for the temporal quality metric	54
A.1.	Scatter plots of the compared metrics on the IT-IST dataset	63
A.2.	Scatter plots of the compared metrics on the TUM 1080p25 dataset . . .	64
A.3.	Scatter plots of the compared metrics on the TUM 1080p50 dataset . . .	65
A.4.	Scatter plots of the quality predictions and reconstructions of each GOP from the subset of the IT-IST dataset	66
A.5.	Quality curve predictions and reconstructions for the sequence <i>City</i> from the IT-IST dataset	67
A.6.	Quality curve predictions and reconstructions for the sequence <i>Football</i> from the IT-IST dataset	68
A.7.	Quality curve predictions and reconstructions for the sequence <i>Football</i> from the IT-IST dataset	69

List of Figures

A.8. Quality curve predictions and reconstructions for the sequence <i>Football</i> from the IT-IST dataset	70
A.9. Quality curve predictions and reconstructions for the sequence <i>Tempete</i> from the IT-IST dataset	71

List of Tables

3.1. Used subset of the IT-IST test set with MOS values from [4]	37
3.2. TUM 1080p25 Dataset	38
3.3. TUM 1080p50 Dataset	39
3.4. Parameters of the used datasets	40
3.5. Performance measures of the different quality metrics applied on the three datasets	41
3.6. Quality Patterns and Reconstruction Functions	48
3.7. Inter-lab correlation of the MOS calculated from the answers to the first question of my subjective test and the corresponding MOS values from IT-IST	49
3.8. Percentages of the answer data of the third question about the quality pattern	52
3.9. Correlation coefficients and RMSE between reconstructed and predicted quality progression curves	54
A.1. Test results of the first question on the overall quality rating \bar{q}_s	60
A.2. Test results of the second question on the fluctuation strength rating f_s	61
A.3. Test results of the third question on the quality patterns p_s	62

Bibliography

1. R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson. Regency effect in the subjective assessment of digitally-coded television pictures. In *Proceedings of the 1995 IEEE International Conference on Image Processing and its Applications*, pp. 336–339. July 1995. doi:10.1049/cp:19950676.
2. C.A. Andersson and R. Bro. The N-way toolbox for MATLAB. In *Chemometrics and Intelligent Laboratory Systems*, 52(1), pp. 1–4, August 2000. doi:10.1016/S0169-7439(00)00071-X.
3. V. Baroncini. New tendencies in subjective video quality evaluation. In *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E89-A(11), pp. 2933–2937, November 2006. doi:10.1093/ietfec/e89-a.11.2933.
4. T. Brandão and M. Queluz. No-reference quality assessment of H.264/AVC encoded video. In *IEEE Transactions on Circuits and Systems for Video Technology*, 20(11), pp. 1437–1447, September 2010. doi:10.1109/TCSVT.2010.2077474.
5. R. Bro. Multiway calibration. multilinear PLS. In *Journal of Chemometrics*, 10(1), pp. 47–61, January 1996. doi:10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C.
6. N.R. Draper and H. Smith. *Applied Regression Analysis*. 3rd edition. Wiley, 1998. ISBN 978-0-471-17082-2.
7. A. Eden. No-reference estimation of the coding PSNR for H.264-coded sequences. In *IEEE Transactions on Consumer Electronics*, 53(2), pp. 667–674, May 2007. doi:10.1109/TCE.2007.381744.
8. S. Gauss, T. Muller, J. Wuenschmann, and A. Rothermel. Continuous subjective quality evaluation of terrestrial broadcast video. In *Proceedings of the 2011 IEEE International Conference on Consumer Electronics (ICCE 2011)*, pp. 356–360. September 2011. doi:10.1109/ICCE-Berlin.2011.6031838.
9. L. Haglund. SVT multi format test set version 1.0. February 2006. URL ftp://vqeg.its.blrdoc.gov/HDTV/SVT_MultiFormat/SVT_MultiFormat_v10.pdf.
10. C. Horch, C. Keimel, and K. Diepold. *QualityCrowd - Crowdsourcing for Subjective Video Quality Tests*. Technical Report, Technische Universität München, Institute for Data Processing, April 2011.

Bibliography

11. ISO/IEC. *14496-10 Information technology – Coding of audio-visual objects – Part 10: Advanced Video Coding*. International Organization for Standardization, January 2012.
12. ITU-R. *Recommendation BT.500: Methodology for the subjective assessment of the quality of television pictures*. International Telecommunications Union, Radiocommunication Sector, January 2012.
13. ITU-T. *Recommendation J.246: Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference*. International Telecommunication Union, Standardization Sector, August 2008.
14. ITU-T. *Recommendation J.247: Objective perceptual multimedia video quality measurement in the presence of a full reference*. International Telecommunication Union, Standardization Sector, August 2008.
15. ITU-T. *Recommendation P.910: Subjective video quality assessment methods for multimedia applications*. International Telecommunications Union, Standardization Sector, April 2008.
16. ITU-T. *Recommendation J.341: Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference*. International Telecommunication Union, Standardization Sector, January 2011.
17. ITU-T. *Recommendation J.342 : Objective multimedia video quality measurement of HDTV for digital cable television in the presence of a reduced reference signal*. International Telecommunication Union, Standardization Sector, April 2011.
18. ITU-T. *Recommendation H.264: Advanced video coding for generic audiovisual services*. International Telecommunications Union, Standardization Sector, January 2012.
19. Y. Kawayoke and Y. Horita. NR objective continuous video quality assessment model based on frame quality measure. In *Proceedings of the 2008 IEEE International Conference on Image Processing (ICIP 2008)*, pp. 385–388. October 2008. doi:10.1109/ICIP.2008.4711772.
20. C. Keimel, J. Habigt, M. Klimpke, and K. Diepold. Design of no-reference video quality metrics with multiway partial least squares regression. In *Proceedings of the Third International Workshop on Quality of Multimedia Experience (QoMEX 2011)*, pp. 49–54. September 2011. doi:10.1109/QoMEX.2011.6065711.
21. C. Keimel, M. Klimpke, J. Habigt, and K. Diepold. No-reference video quality metric for HDTV based on H.264/AVC bitstream features. In *Proceedings of the 2011 IEEE International Conference on Image Processing (ICIP 2011)*, pp. 3325–3328. September 2011. doi:10.1109/ICIP.2011.6116383.

22. C. Keimel, T. Oelbaum, and K. Diepold. No-reference video quality evaluation for high-definition video. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pp. 1145–1148. April 2009. doi:10.1109/ICASSP.2009.4959791.
23. C. Keimel, T. Oelbaum, and K. Diepold. Improving the prediction accuracy of video quality metrics. In *Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*, pp. 2442–2445. March 2010. doi:10.1109/ICASSP.2010.5496299.
24. C. Keimel, M. Rothbucher, H. Shen, and K. Diepold. Video is a cube. In *IEEE Signal Processing Magazine*, 28(6), pp. 41–49, September 2011. doi:10.1109/MSP.2011.942468.
25. C. Keimel, J. Habigt, C. Horch, and K. Diepold. Qualitycrowd – a framework for crowd-based quality evaluation. In *Proceedings of the 2012 Picture Coding Symposium (PCS 2012)*, pp. 245–248. May 2012. doi:10.1109/PCS.2012.6213338.
26. C. Keimel, A. Redl, and K. Diepold. The TUM high definition video data sets. In *Proceedings of the Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, pp. 97–102. July 2012. doi:10.1109/QoMEX.2012.6263865.
27. C. Keimel, M. Rothbucher, and K. Diepold. Extending video quality metrics to the temporal dimension with 2D-PCR. In S.P. Farnand and F. Gaykema (eds.), *Proceedings of the SPIE Conference on Image Quality and System Performance VIII*, volume 7867, pp. 786713:1–786713:10. SPIE, January 2011. doi:10.1117/12.872406.
28. M. Klimpke, C. Keimel, and K. Diepold. *Visuelle Qualitätsmetrik basierend auf der multivariaten Datenanalyse von H.264/AVC Bitstream-Features*. Technical Report, Technische Universität München, Institute for Data Processing, November 2010.
29. H. Martens and M. Martens. *Multivariate Analysis of Quality: An Introduction*. Wiley, 2001. ISBN 978-0-471-97428-4.
30. H. Martens and T. Næs. *Multivariate Calibration*. Wiley, 1989. ISBN 978-0-471-93047-1.
31. M.A. Masry and S.S. Hemami. CVQE: A metric for continuous video quality evaluation at low bit rates. In B.E. Rogowitz and T.N. Pappas (eds.), *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging VIII*, volume 5007, pp. 116–127. June 2003. doi:10.1117/12.477774.
32. M.A. Masry and S.S. Hemami. A metric for continuous quality evaluation of compressed video with severe distortions. In *Signal Processing: Image Communication*, 19(2), pp. 133–146, February 2004. doi:10.1016/j.image.2003.08.001.
33. G.A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. In *Psychological Review*, 63(2), pp. 81–97, March 1956. doi:10.1037/h0043158.

Bibliography

34. M. Miyahara. Quality assessments for visual service. In *IEEE Communications Magazine*, 26(10), pp. 51–60, October 1988. doi:10.1109/35.7667.
35. T. Oelbaum, H. Schwarz, M. Wien, and T. Wiegand. Subjective performance evaluation of the SVC extension of H.264/AVC. In *Proceedings of the 2008 IEEE International Conference on Image Processing (ICIP 2008)*, pp. 2772–2775. October 2008. doi:10.1109/ICIP.2008.4712369.
36. M.H. Pinson and S. Wolf. Comparing subjective video quality testing methodologies. In T.E.T. Sikora (ed.), *Proceedings of the SPIE Conference on Video Communications and Image Processing*, volume 5150, pp. 573–582. June 2003. doi:10.1117/12.509908.
37. A. Redl, C. Keimel, and K. Diepold. Influence of viewing device and soundtrack in HDTV on subjective video quality. In F. Gaykema and P.D. Burns (eds.), *Proceedings of the SPIE Conference on Image Quality and System Performance IX*, volume 8293, pp. 829312:1–829312:9. December 2012. doi:10.1117/12.907015.
38. I.E.G. Richardson. *H.264 and MPEG-4 video compression*. Wiley, 2003. ISBN 978-0-470-84837-1.
39. I.E.G. Richardson. *The H.264 Advanced Video Compression Standard*. 2nd edition. Wiley, 2010. ISBN 978-0-470-51692-8.
40. A. Rossholm and B. Lövström. A new video quality predictor based on decoder parameter extraction. In *SIGMAP 2008 - Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pp. 285–290. July 2008.
41. M. Slanina, V. Rícný, and R. Forchheimer. A novel metric for H.264/AVC no-reference quality assessment. In *Proceedings of the EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, pp. 114–117. June 2007. doi:10.1109/IWSSIP.2007.4381166.
42. A. Smilde, R. Bro, and P. Geladi. *Multi-way Analysis with Applications in the Chemical Sciences*. Wiley, 2004. ISBN 978-0-471-98691-1.
43. Z. Wang, A. Bovik, and B. Evan. Blind measurement of blocking artifacts in images. In *Proceedings of the 2000 IEEE International Conference on Image Processing (ICIP 2000)*, volume 3, pp. 981–984. September 2000. doi:10.1109/ICIP.2000.899622.
44. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. In *IEEE Transactions on Image Processing*, 13(4), pp. 600–612, April 2004. doi:10.1109/TIP.2003.819861.
45. Z. Wang, H.R. Sheikh, and A.C. Bovik. Objective video quality assessment. In B. Furht and O. Marques (eds.), *Handbook of video databases: design and applications*, pp. 1041–1078. CRC Press, September 2003. ISBN 978-0-849-37006-9.

46. A.B. Watson. Toward a perceptual video quality metric. In B.E. Rogowitz and T.N. Pappas (eds.), *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging III*, volume 3299, pp. 139–147. July 1998. doi:10.1117/12.320105.
47. S. Winkler. *Digital Video Quality – Vision Models and Metrics*. Wiley, 2005. ISBN 978-0-470-02404-1.
48. S. Wolf and M.H. Pinson. Spatial-temporal distortion metric for in-service quality monitoring of any digital video system. In A.G. Tescher, B. Vasudev, V.M. Bove Jr., and B. Derryberry (eds.), *Proceedings of the SPIE Conference on Multimedia Systems and Applications II*, volume 3845, pp. 266–277. November 1999. doi:10.1117/12.371210.
49. J. Yang, D. Zhang, A. Frangi, and J. yu Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1), pp. 131–137, January 2004. doi: 10.1109/TPAMI.2004.1261097.