METHODS FOR OBJECTIVE AND SUBJECTIVE VIDEO QUALITY ASSESSMENT AND FOR SPEECH ENHANCEMENT

Muhammad Shahid

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2014:15
Department of Applied Signal Processing
Methods for Objective and Subjective Video Quality Assessment and for Speech Enhancement

Muhammad Shahid
Methods for Objective and Subjective Video Quality Assessment and for Speech Enhancement

Muhammad Shahid

Doctoral Dissertation in Applied Signal Processing

Department of Applied Signal Processing
Blekinge Institute of Technology
SWEDEN
Abstract

The overwhelming trend of the usage of multimedia services has raised the consumers’ awareness about quality. Both service providers and consumers are interested in the delivered level of perceptual quality. The perceptual quality of an original video signal can get degraded due to compression and due to its transmission over a lossy network. Video quality assessment (VQA) has to be performed in order to gauge the level of video quality. Generally, it can be performed by following subjective methods, where a panel of humans judges the quality of video, or by using objective methods, where a computational model yields an estimate of the quality. Objective methods and specifically No-Reference (NR) or Reduced-Reference (RR) methods are preferable because they are practical for implementation in real-time scenarios.

This doctoral thesis begins with a review of existing approaches proposed in the area of NR image and video quality assessment. In the review, recently proposed methods of visual quality assessment are classified into three categories. This is followed by the chapters related to the description of studies on the development of NR and RR methods as well as on conducting subjective experiments of VQA. In the case of NR methods, the required features are extracted from the coded bitstream of a video, and in the case of RR methods additional pixel-based information is used. Specifically, NR methods are developed with the help of suitable techniques of regression using artificial neural networks and least-squares support vector machines. Subsequently, in a later study, linear regression techniques are used to elaborate the interpretability of NR and RR models with respect to the selection of perceptually significant features. The presented studies on subjective experiments are performed using laboratory based and crowdsourcing platforms. In the laboratory based experiments, the focus has been on using standardized methods in order to generate datasets that can be used to validate objective methods of VQA. The subjective experiments performed through crowdsourcing relate to the investigation of non-standard methods in order to determine perceptual preference of various adaptation scenarios in the context of adaptive streaming of high-definition videos.

Lastly, the use of adaptive gain equalizer in the modulation
frequency domain for speech enhancement has been examined. To this end, two methods of demodulating speech signals namely spectral center of gravity carrier estimation and convex optimization have been studied.

*Keywords*: Video Quality Assessment, No-Reference Methods, Reduced-Reference Methods, Subjective Experiments, Speech Enhancement, Adaptive Gain Equalizer
To my parents and siblings!
To Mahwish, the love of my life!
To Ayesha, the blessing of my life!
Acknowledgments

The years during the course of my PhD studies have been significantly important for shaping professional growth as well as a reformation of my personality. Undoubtedly, thinking like a researcher has given me a pragmatic sight to envisage the world differently. This dissertation, the knowledge that I gained, and the candid consuetude that I learnt has been made possible by an enormous support of many. I would like to pay my heartiest gratitude to all of them.

By all means, I consider that Dr. Benny Lövström played the most influential role in this awakening. Since the time I began my masters degree thesis back in 2010 and later on continued with my PhD under his supervision, his mentorship has always been there in the most of my research activities. Not only that, he has been the source of counseling for me in many of my daily life personal and social pursuits as well. I believe that I can not thank him enough in response to what he has been doing for me during these years. It has been a matter of great pleasure, learning, and prosperity that I have been endowed by his guidance. Many thanks, Benny!

Prof. Ingvar Claesson, my supervisor and examiner, has also been a great source of guidance for me. I still remember the time when I met him while he was heading the panel for my interview of admission in PhD studies. Since then, I have found him compassionate, supportive, and propitious in all matters, be it related to research work or the availability of resources for a comfortable work environment.

Besides guidance from my official supervisors, I had the great opportunity of learning from Prof. Hans-Jürgen Zepernick while writing articles with him. Many thanks to him for saying, "This sentence is correct but it hardly possesses any academic value!"

I would really like to mention Dr. Andreas Rossholm, a very good friend of mine, as one of those great people who played very important roles in my PhD studies. He has been a kind support during my masters degree thesis
and a considerate collaborator during the research work of my PhD studies. I should also acknowledge the kindness of all people from VQEG especially Kjell Brunnström. Many thanks to my research collaborators outside BTH including Katerina Pandremmenou, Rizwan Ishaq, Jacob Søgaard, Jeevan Pokhrel, and Samira Tavakoli.

Many thanks to Rector Anders Hederstierna and Vice-rector Eva Pettersson for their support regarding many official matters including IEEE BTH SB. Since the time of inception of IEEE BTH SB, I never fell short of a speaker on any topic as I always had the support of Prof. Wlodek Kulesza. Not only this, I would like to thank a lot for his comments and discussions on improving the presentation of research aims of this thesis. I really appreciate the support of Dr. Niklas Lavesson and Lena Vogelius regarding IEEE BTH SB matters.

Since the time of my masters degree at BTH, Prof. Lars Håkansson has been a really kind and supportive teacher to me. Many thanks to Dr. Siamak Khatibi for all the thoughtful discussions. Similarly, it has been really nice to have discussions with Prof. Mats Petterson. Dr. Sven Johansson has been a great support related to department and educational affairs. A very special thanks to my previous office mates Josef Ström Bartunek and Irina Gertsovich for being so cooperative. Many thanks to my previous and current colleagues/teachers including Dr. Nedelko Grbic, Dr. Jörgen Nordberg, Dr. Benny Sällberg, Dr. Bo Schankman, Dr. Thomas Sjögren, Dr. Jiandan Chen, Dr. Muhammad Gufran Khan, Mr. Imran Khan, Dr. Viet Thuy Vu, Dr. Hang Zettervall, and Dr. Åse Nygren. It has been really nice talking to you always: Mikael Åsman, Claes Jogreus, and Jan-Olof Gustavsson. Many thanks to Madeleine Jarlten, Mansour Mojtaba, Svetlana Zivanovic, Marie Ahlgren, and Lena Brandt Gustafsson for such a nice support in administrative matters.

My deepest gratitude to the local Pakistani community of students and families for providing very nice opportunities of social gatherings.

Muhammad Shahid
Karlskrona, Nov 10, 2014
Preface

This doctoral thesis presents my research work in the field of video quality assessment and speech enhancement, which I performed at the Department of Applied Signal Processing at Blekinge Institute of Technology, Sweden. The thesis is composed of four parts that present the following research contributions:


During my PhD studies, I have also been involved in production of the following research work that is not an integral part of this thesis.


**Paper VII** Tahir Minhas, Muhammad Shahid, Andreas Rossholm, Benny Lövström, Hans-Jurgen Zepernick, and Markus Fiedler, "QoE Rating Performance Evaluation of ITU-T Recommended Video Quality Metrics in the


I  On Review of No-Reference Image and Video Quality Assessment  35

2  No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches  37
   2.1  Review ............................................................. 39
       2.1.1  Introduction .................................................. 39
       2.1.2  Applications of no-reference image and video quality assessment ........................................ 41
       2.1.3  Single artifact NR-P-based methods ....................... 47
       2.1.4  Multiple artifacts NR-P-based methods .................. 61
       2.1.5  Features measures-based methods ........................ 66
       2.1.6  Bitstream-based methods .................................. 76
       2.1.7  Hybrid of NR-P and NR-B methods ....................... 83
   2.2  Conclusions ....................................................... 88
   2.3  Bibliography ....................................................... 92

II  On Objective Methods of No-Reference and Reduced-Reference Video Quality Assessment  115

3  A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor  117
   3.1  Introduction ....................................................... 119
   3.2  The inputs and targets for the model .......................... 121
       3.2.1  Selection of the parameters ................................. 121
       3.2.2  The quality metrics ......................................... 122
   3.3  The Proposed Quality predictor Model ......................... 124
       3.3.1  The architecture of the model ............................. 125
       3.3.2  The video sequences data set ............................. 126
   3.4  Experiments and Results ....................................... 126
   3.5  Conclusion ......................................................... 129
   3.6  Bibliography ....................................................... 130

4  A No-Reference Machine Learning Based Video Quality Predictor  133
   4.1  Introduction ....................................................... 135
   4.2  Related work ....................................................... 137
   4.3  Subjective Video Quality Assessment .......................... 138
       4.3.1  Test Video Sequences ...................................... 138
       4.3.2  Subjective Quality Assessment Setup ..................... 140
CONTENTS

4.4 The Proposed Method ........................................... 140
   4.4.1 Feature Extraction of H.264/AVC Bitstream Data .... 141
   4.4.2 Proposed Prediction Model ............................... 142
   4.4.3 Benchmark Measurements ............................... 144
   4.4.4 Simulation Results ..................................... 144
4.5 Comparison with Contemporary Techniques ....................... 144
4.6 Conclusion .................................................. 146
4.7 Bibliography ............................................... 147

5 Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models 149
   5.1 Introduction ............................................... 151
   5.2 Related Work ............................................. 154
   5.3 Features Related to Perceptual Video Quality .......... 157
      5.3.1 Examined Features ................................. 158
      5.3.2 Suitability of the Features ....................... 160
      5.3.3 Pre-processing of the Features .................. 163
   5.4 Video Quality Estimation Using Linear Regression Models 163
      5.4.1 Least Absolute Shrinkage and Selection Operator 164
      5.4.2 Model Learning .................................... 165
   5.5 Experimental Results .................................... 167
      5.5.1 Results ............................................. 167
      5.5.2 Feature Selection in LASSO ....................... 170
      5.5.3 Comparison with Related Work .................... 173
   5.6 Conclusions .............................................. 175
   5.7 Bibliography ............................................. 176

III On Subjective Methods of Video Quality Assessment 185

6 Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos 187
   6.1 Introduction .............................................. 189
   6.2 Subjective Assessment of the Test Stimuli ............ 192
      6.2.1 Generation of Test Stimuli ....................... 192
      6.2.2 Subjective Assessment ............................ 195
   6.3 Conclusion ............................................... 201
   6.4 Bibliography ............................................. 201
7 Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality 203
  7.1 Introduction .............................................. 205
  7.2 Test Stimuli and Subjective Video Quality Assessment ........... 210
    7.2.1 Encoding configuration ................................ 211
    7.2.2 Pre-processing the test sequences ...................... 212
    7.2.3 Subjective Video Quality Assessment Setup ............. 213
  7.3 Results ................................................... 214
    7.3.1 Observations from the MOS results ..................... 215
    7.3.2 Analysis of Variance Based Comparison ................... 215
  7.4 Conclusion ................................................. 216
  7.5 Bibliography ............................................... 216

8 Crowdsourcing Based subjective quality assessment of adaptive video streaming 219
  8.1 Introduction ............................................... 221
  8.2 Test Background .......................................... 222
  8.3 Results .................................................... 223
  8.4 Conclusion ................................................. 225
  8.5 Bibliography ............................................... 225

IV On Speech Enhancement in Modulation Frequency Domain 229

9 Modulation Domain Adaptive Gain Equalizer for Speech Enhancement 231
  9.1 Introduction ............................................... 233
    9.1.1 Literature Survey ...................................... 234
    9.1.2 Main Contributions ..................................... 235
  9.2 Modulation System ........................................ 235
    9.2.1 Spectral Center of Gravity Carrier Estimation ......... 237
  9.3 Adaptive Gain Equalizer System ............................ 238
    9.3.1 Gain Function .......................................... 238
    9.3.2 Modulation Domain AGE .................................. 239
  9.4 Evaluation of The Proposed System .......................... 240
    9.4.1 Spectrogram Analysis ................................... 243
  9.5 Conclusion ................................................ 244
10 Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization

10.1 Introduction ......................................................... 249
10.2 Background ......................................................... 251
10.3 Modulation Domain and AGE ................................. 251
   10.3.1 Spectral Center of Gravity Carrier Estimation ...... 253
   10.3.2 Adaptive Gain Equalizer System ......................... 253
10.4 Convex Optimization and the Proposed Model ............. 255
10.5 Comparative Performance Analysis ......................... 257
   10.5.1 Mean Opinion Score (MOS) ............................. 257
   10.5.2 Spectral Distortion ....................................... 258
   10.5.3 Signal to Noise Ratio Improvement (SNRI) ......... 258
   10.5.4 Spectrogram Analysis .................................. 259
10.6 Conclusion ......................................................... 260
10.7 Bibliography ....................................................... 261
Multimedia quality is important for modern communication systems in many aspects. Awareness of the consumers of multimedia services regarding perceptual quality has increased more than ever before due to the availability of a variety of modern devices capable of playing audio and video content. The consumption of multimedia content is expected to increase due to convenient access to the Internet and other networks of video and audio streaming. Moreover, both consumers and service providers want to make efficient use of these technologies. Thus, perceptual quality of multimedia services is a matter of concern for many stakeholders.

This thesis presents contributions in two key areas of multimedia services. Firstly, contributions on the development of objective methods of video quality assessment and on conducting subjective experiments of video quality assessment are presented. Afterwards, contributions on speech enhancement using adaptive gain equalizer are presented. Major part of the thesis is comprised of the contributions related to the former area.

This chapter is organized as the following. An introduction of the aforementioned two key areas is presented in Section 1.1 and Section 1.2 respectively. The thesis research objectives and scope are described in Section 1.3. The thesis outline and a summary of the presented research contributions are provided in Section 1.4. Lastly, conclusive remarks on the contribution of this thesis and an outlook of future research is given in Section 1.5.

In what follows this chapter, the full contents of the research contributions are presented in four parts. These parts are comprised of the chapters that are based on individual research articles.
1. Introduction

1.1 Introduction to Video Quality Assessment

Modern services of data transmission have made the delivery of videos much easier than ever before. Accordingly, there has been an abundant rise in the consumption of videos for requirements related to entertainment as well as for business. Multimedia services that have gained wide interest include digital television broadcasts, video streaming applications, and real-time audio and video communication over the Internet. The video portion of the mobile data traffic was 53% in 2013, and is expected to exceed 67% by 2018, as indicated in a recent study [1]. With this huge increase in exposure of videos to the human eye, an increase in the user awareness of video characteristics including its perceptual quality is obvious. This fact leads to a raise in the bar of quality with the availability of contemporary video services including storage discs such as DVD and Blu-ray, and display devices such as high definition TV. Not so commonly used storage cassettes like Video Head System (VHS) and display devices like cathode-ray TV are considered inferior to their modern counterparts. Evidently, these considerations give rise to the need for service providers to obtain knowledge of users’ Quality of Experience (QoE) and for users to obtain knowledge of the received quality.

Video quality can get degraded at any of the stages including capturing, compression, transmission, reproduction, and displaying. The underlying reasons can include: equipment malfunctioning, data-loss due to compression, fluctuations in transmission networks, and limitations of display terminals. In order to make an estimate of the degradations in the quality for practical scenarios, each of these reasons should be scrutinized and streamlined for their impact. For evaluation scenarios, generally, carefully designed experiments involving simulated or emulated behaviors of some of the aforementioned reasons are tested for determining their impact on the quality.

The legitimate judges of video quality are humans as end users, the opinions of whom can be obtained by subjective experiments of Video Quality Assessment (VQA). Subjective experiments involve a panel of participants which are usually non-experts, also referred to as test subjects, to assess the perceptual quality of given test material such as a set of images or videos. Subjective experiments of VQA are typically conducted in a controlled laboratory environment. Careful planning and several factors including assessment method, selection of test material, viewing conditions, grading scale, and timing of presentation have to be considered prior to a subjective experiment. The involvement of these factors turns the subjective experiments time consuming, tedious, and expensive.
1.1. Introduction to Video Quality Assessment

In order to alleviate some of the aforementioned issues related to laboratory based subjective experiments, crowdsourcing based subjective experiments have been introduced. This method involves collecting subjective assessment of quality through ubiquitous streaming of videos via the Internet. This enables the investigator to receive opinion from a vast variety of subjects; in a time-flexible, test-data size scalable, and swift manner. However, it is not possible to employ this methodology in applications that require real-time and online monitoring of the video quality.

Objective methods of VQA, which are usually computational models of quality estimation and can be automated, are considered time-efficient and more suitable for real-time applications. For an objective VQA method, the basic aim is to estimate the video quality as close to the corresponding human judgement as possible.

Objective methods can be categorized into different types following different considerations. For example, a rather crude estimation of video quality can be made by taking arithmetic difference between the original video and its processed version. Such type of fidelity measure methods are termed as data metrics, and Mean Square Error (MSE) is one of them [2]. MSE and its normalized logarithmic variant Peak-Signal-to-Noise-Ratio (PSNR) gained vast popularity due to their computational simplicity. However, as this type of methods remain agnostic to the perceptual contents of a video; they can produce highly unreliable results in terms of poor correlation with subjective assessment. For example, in Fig. 1.1 where the original image (b) has been processed with blurring and with salt and pepper noise; it is noted that image with much lower MSE given in (c) looks worse than the image with higher MSE as shown in (a). More of such examples are illustrated in [3]. These shortcomings of the data metrics motivate the requirement of objective methods that can take perceptual information of videos into account and such methods are termed as picture metrics [2]. The picture metrics entail fairly involved computations and complexity and examples include Structural SIMilarity (SSIM) [4] that is mainly targeted for images. One type of objective methods that can estimate perceptual video quality with reasonable accuracy and offer low computational complexity is known as bitstream-based methods [5]. Some of the contributions related to objective VQA that are presented in this thesis are related to bitstream-based methods.

This thesis presents an insight into studies that have been made to develop objective methods of video quality estimation and subjective experiments of VQA. The test videos used in these studies have been encoded using H.264/AVC standard [6]. This standard is built upon the coding tool known
1. Introduction

Figure 1.1: Original image (b) processed with blurring in (a) and with salt & pepper noise in (c)

as the hybrid Differential Pulse Code Modulation/Discrete Cosine Transform (DPCM/DCT) codec model, that incorporates motion estimation and compensation (DPCM), followed by transform (DCT) and entropy coding processes. In these stages, various decisions related to video frame (sub)partitioning, use of a suitable reference frame, and level of quantization are taken. These decisions are taken in the form of specific values of coding parameters that are utilized for (de)coding the video signal. It has been observed that certain features based on such coding parameters are instrumental in the estimation of the perceptual quality of the encoded video. Some of the contributions included in this dissertation contain details on this kind of quality estimation. Section 1.1.1 includes a brief introduction of the concepts related to H.264/AVC based video coding and the relationship of coding parameters with the video quality, followed by Section 1.1.2 that deals with an overview of video quality assessment. Afterwards, a brief survey of the publications related to reduced-reference (RR) and no-reference (NR) VQA methods is presented in Section 1.1.3.

1.1.1 H.264/AVC Based Compression and Coding Parameters

Video compression implicates a process of reducing the amount of data required for representing a video signal to a specific extent for efficient utilization of storage space and transmission channels. The fundamental problem of compression can be ascertained as either achieving the best possible signal fidelity level within the constraints of given bitrate or achieving the lowest pos-
1.1. Introduction to Video Quality Assessment

![Diagram](image)

Figure 1.2: Video coding and decoding chain

sible bitrate while a given signal fidelity is preserved [7]. Compression can be lossless or lossy which refer to the cases when the uncompressed (reconstructed after compression) signal is exactly the same as the original or an estimated version of the original. Reduction of various kinds of inherent redundancies such as spatial, temporal, and psychovisual can be performed to achieve compression. A system capable of such processing is called codec, consisting of a function coder and its complementary function decoder. Historically, video codec design problem has been realized through following an evolving series of international compression standards, where the standardization is primarily aimed at defining a specific syntax of the bitstream and decoding process, as shown in Figure 1.2. The mainstream standardization activities have been performed by Telecommunication standardization sector of International Telecommunication Union (ITU-T) together with joint teams of International Standard Organization (ISO) and International Electrotechnical Commission (IEC). Advanced video coding standard, commonly known as H.264/AVC, has been the most popular coding tool prior to the state-of-the-art standard known as High Efficiency Video Coding (HEVC) that was finalized in 2013.

Encoding procedure compliant to H.264/AVC standard proceeds as follows. Each frame of a video sequence is divided into blocks known as a macroblock (MB), i.e., a rectangular region of 16 × 16 pixels, followed by a prediction of each of the MBs as a whole or its smaller sub-divisions of size down to 4 × 4 pixels. Prediction can be performed within the frame, called intra prediction, or between frames in the video, called inter prediction and performed by doing motion estimation. Accordingly, motion compensation is performed and a residual signal is generated for transforming and entropy coding [6]. Overall, the operation of H.264/AVC can be described under two layers namely Video Coding Layer (VCL) and Network Abstraction Layer (NAL) as presented in the following [8].
1. **Introduction**

H.264/AVC VCL

VCL aims at efficient representation of the video content mainly by performing the block-based hybrid DPCM/DCT video coding scheme that includes the following concepts:

1. Macroblocks and Slices: The basic unit of coding process is macroblock that consists of luma and chroma components of the video picture. A region of macroblocks that can be decoded independently can be grouped together to form a slice. Each frame of a video can be split into one or more than one slices. Mechanisms for error resilience, segmentation of the data to fit in the transfer units of network, and parallel processing can take advantage from slicing a video frame. Based on the type of prediction, the slices can be coded as I, P, or B that refer to *intra prediction*, *forward-prediction*, and *bi-directional prediction* respectively.

2. Intra-picture Prediction: Irrespective of the slice coding type, intra prediction can be done for a complete macroblock, *Intra_16 × 16* or a block of 4 × 4 pixels, *Intra_4 × 4*, where the smaller block size is suitable for highly textured parts of a picture and smooth areas are more efficiently coded as macroblock size. There are nine possible modes in *Intra_4 × 4* based coding that differ in the direction used for selecting the neighbor pixels for prediction. In *Intra_16 × 16* based prediction, four coding modes are available.

3. Inter-picture Prediction: Inter prediction refers to the process of predicting a picture block by using a block from another picture, called as a reference picture. For either case of P or B based inter prediction, macroblocks can be partitioned as regions of block sizes 16 × 16, 16 × 8, 8 × 16, and 8 × 8. Furthermore, 8 × 8 block can be coded directly or sub-divided into 4 × 8, 8 × 4, or 4 × 4 sized blocks. Besides the block size information, Motion Vector (MV) is required for decoding. An inter macroblock can also be coded as *P_Skip* or *B_Skip* and the already decoded data is used to reconstruct the related macroblock.

4. Transform and Quantization: A residual signal is obtained from the aforementioned prediction procedures. A two-dimensional spatial transform is applied to each 4 × 4 regions of the residual picture. Specifically, an approximation of the Discrete Cosine Transform (DCT), known as integer DCT, is used. The obtained transform coefficients are quantized and this process is controlled through Quantization Parameter (QP) that can be chosen while encoding.
5. Entropy Coding: Two kinds of entropy coding is supported in H.264/AVC, Context-Adaptive Variable-Length Coding (CAVLC) and Context-Adaptive Binary Arithmetic Coding (CABAC). The former codes quantized transform coefficients using VLC tables and the latter adopts context-conditional probability estimates. Generally, CABAC is found to be more efficient in bitrate reduction than CAVLC at similar visual quality.

6. Profiles and Levels: The coding tools that are part of H.264/AVC are grouped into different profiles which specify the requirements on algorithmic capabilities of the decoder that can be used for a particular coded bitstream. The main advantage of the specification of profiles is the benefit of encoding a video that can be more suitable for a given application. For example, baseline profile is considered to be more applicable in the applications that require low-complexity and low-delay coding solutions. A particular level of the parameters such as frame resolution, frame-rate, and bitrate that a decoder is expected to support is specified by level. This way, a level defines a limit of the decoder capabilities.

**H.264/AVC NAL**

NAL offers a format and package to the VCL data for a variety of systems such as storage or transportation over a network. NAL is composed of the following three main building blocks:

1. NAL Units: Essentially, NAL units are integer sized packets of bytes that contain the coded video data (VCL) besides a header byte that indicates the data type. Other related information such as parameter sets and supplemental enhanced information is stored in non-VCL NAL units.

2. Parameter Sets: Header information applicable to many VCL NAL units is stored in parameter sets. Sequence parameter sets address a sequence of coded video frames and picture parameter set addresses one or more individual video frames. As the parameter sets contain information that is crucial for decoding of the video, they can be transmitted either within the channel of VCL NAL or by using a more reliable transport mechanism.

3. Access Units: Access unit refers to VCL and non-VCL NAL units data that belongs to a single decoded frame.
1. Introduction

Coding Parameters

At the stage of encoding a video by following the H.264/AVC standard, there are a variety of parameters whose values can be controlled based on the requirements of the target application. As the standard itself is limited for recommendations of the processing of the decoder and the syntax of the relevant bitstream format, an encoder has the possibility to freely use any tools that would result in a standard conforming coding. For example, Joint Model (JM) reference software [9] for H.264/AVC based encoding offers a huge set of parameter values that can be assigned in a configuration file that is used as input to the encoder. As a result, the encoder optimizes the coding performance to achieve a certain visual quality under the constraints of a certain bitrate. While performing this operation, various options of coding parameters are selected that are eventually used as the coded representation of the video data. At the decoding stage, the decoder has to extract all these coding parameters from the coded bitstream for reconstructing the original video. Some examples of such parameters are coding mode types, such as $P_{16 \times 16}$, motion vectors, and QP values. In various studies [5], it has been investigated that an ensemble of coding parameters can be used to estimate the visual quality. Briefly, Figure 1.3 presents an example of how various coding parameters are correlated with certain quality metrics. In literature, sometimes these parameters are also referred to as features. This concept is covered in more detail in Part I and II of this thesis.

1.1.2 Video Quality Assessment: An Overview

Video quality assessment typically refers to a process of estimation of perceptual quality of a video; mostly through an evaluation of the existence or absence of some artifacts that might distort the quality of an original video. The perceptual quality of a video can get degraded at any of the stages including acquisition, compression, transmission, reconstruction, and displaying due to different underlying reasons. This thesis is mainly concerned with artifacts introduced due to compression and transmission only. In state-of-the-art codecs such as H.264/AVC, the compression includes a process of quantization that is performed on the frequency domain components, typically DCT coefficients, of the video signal. This process of quantization is irreversible and can cause visible distortions in a video, depending on how severely it has been applied [10]. Consequently, a variety of artifacts can be introduced in the video including blocking due to block-based processing, blurring due to loss of spatial details and reduction of edge sharpness, and ringing due to rough quantization of
1.1. Introduction to Video Quality Assessment

Figure 1.3: Correlation of coding features and visual quality

high-frequency transform coefficients. When the compressed signal is trans-
mittted through a network (wired or wireless) in the form of data-packets, the
signal might need to pass through a variety of network components that may
differ in their throughput and availability. Depending upon the nature of the
issues that may arise in the transmission medium, a variety of artifacts can be
introduced in the video including loss of video data due to packet-loss, frame-
freeze due to non-availability of video frames, and jerkiness due to temporal
down-sampling of video in response to a fall in the capacity of the network [11].

Video quality assessment can be performed just after compression quan-
tifying the quality of the video varied or unvaried only due to compression
phenomenon. This situation applies to cases where the video data is com-
pressed for local usage, typically for saving it in some storage device. The
quality assessment can also be made after transmission to jointly quantify the
quality variations due to compression and transmission.

The most valid measure of video quality is obtained through collection of
ratings from a human panel by following certain recommendations, usually
referred to as *subjective experiments* and the involved humans are termed as
1. Introduction

The traditional method of subjective experiments, which is still the mainstream approach, is to conduct them in a controlled laboratory based environment following the standard recommendations. This involves taking into consideration of several factors about careful planning, assessment method, and selection of suitable test material. Opinion scores from each of the individual subject are used to compute a representative value known as Mean Opinion Score (MOS) and it serves as the ground-truth of quality for particular test-stimuli.

However, there are many challenges that researchers and practitioners might face while conducting the subjective experiments of VQA as they can be extremely laborious, time-consuming, and expensive. In an effort to circumvent related issues, the concept of crowdsourcing based VQA [13] has gained some attention recently. This method involves the collection of subjective VQA through ubiquitous delivery of the test-stimuli at the subjects’ premises via the Internet. Even though it can avoid some of the issues of laboratory based experiments, some new problems and issues are emerged that can be significantly difficult to resolve, such as, the problems inherent to the use of the Internet leading to limitations due to capacity and reliability of intermediate network. Other issues are the concerns regarding the performance of a remote subject in terms of correctly following the recommended guidelines.

In order to alleviate most of the issues that make the subjective VQA less applicable, computational models of VQA have been designed and are usually termed as objective methods (alternatively, metrics or models). The main goal of the objective methods is to estimate the visual quality, with a high correlation with the subjective assessment. With regards to the amount of the usage of information of the original video, objective methods are categorized as full-reference (FR), reduced-reference (RR), and no-reference (NR) methods as described in the following [3].

- **FR methods**: In this approach, the entire original video is available as a reference. Accordingly, FR methods are based on comparing processed video with the original video.

- **RR methods**: In this case, it is not required to give access to the original video but only provide representative features of the original video. Moreover, any other suitable information useful for quality estimation can be sent from the server side. The comparison of the reduced information from the original video with the corresponding information from
1.1. Introduction to Video Quality Assessment

Figure 1.4: Various types of video quality assessment methods

the processed video provides the input for RR methods.

- NR methods: This class of objective quality methods does not require access to the original video but searches for artifacts with respect to the pixel domain of a video, utilizes information embedded in the bitstream of the related video format, or performs quality assessment as a hybrid of pixel based and bitstream based approaches.

FR methods are more suitable in offline processing of videos such as local testing of coding tools or software for performance testing and comparisons. Also, FR methods are considered to be computationally complex as a processing on pixel-level for both of the videos is required to be performed in most cases. Applications that require real-time evaluation of the video quality do not have the possibility to access the original video and hence FR methods cannot be employed. In such cases, RR methods and especially NR methods are found to be the most applicable solution. In order to visualize the aforementioned three types of VQA methods in a typical encoder-transmitter-decoder scenario, Figure 1.4 depicts the involved processing in each type. The figure shows the suitability of FR methods as the viable approach for assessment of quality before the video is transmitted, in order to gauge the impacts of compression. Subjective assessment is shown to be conducted after the video signal has been received through a network to jointly gauge the impact of compression and
transmission artifacts. The input to a typical RR method is the reconstructed signal received after transmission and a set of suitable features that are preferably transmitted over a reliable channel. NR methods, however, need only the signal received after transmission and they may use the entire reconstructed video signal in the form of pixels (NR-P) or they may use only certain information of the received video from its bitstream (NR-B). Depending upon the requirement, the methods shown to perform after transmission can be used prior to transmission and vice versa. Moreover, pixel-based and bitstream-based information can be combined as well.

In the following, a description of standardization activities conducted in ITU for subjective and objective methods are presented briefly.

**Standardization of Subjective Methods**

In order to conduct reproducible and reliable subjective experiments of quality assessment, certain considerations are required to be taken into account. To this end, two divisions in ITU have put forward their series of recommendations. These divisions are ITU-R (Radiocommunication Sector of ITU) and ITU-T (Telecommunication standardization sector of ITU). Recommendations on various types of subjective experiments have been published in ITU-R BT.500 series that include details on single stimulus and double stimulus experiments. The former type refers to the case when only the processed version of stimuli is shown to the subjects and reference stimuli is also shown in the latter type. Moreover, it also includes recommendations on the laboratory setup, viewing conditions, and rating scales for the subjective experiments. The latest document of this series is BT.500-13 [14] namely *Methodology for the subjective assessment of the quality of television pictures*. Similarly, ITU-R BT.1788 recommends a methodology on subjective quality assessment of videos aimed for multimedia applications and ITU-R BT.2021 recommends subjective methods for stereoscopic 3DTV systems. ITU-T P.900 [15] series covers the domain of audiovisual quality in multimedia services. Recommendations on Absolute Category Rating (ACR), Degradation Category Rating (DCR), and Pair Comparison method (PC) are provided in [15]. It also includes details on characteristics of test sequences to be used such as the duration, the content, and the sample size. In view of the huge popularity of video services through online streaming, ITU-T P.913 [16] recommends methodologies for subjective assessment of video quality, audio quality, and audiovisual quality of Internet based videos.

Video Quality Experts Group (VQEG) [17] is a well-known global platform
1.1. Introduction to Video Quality Assessment

for its contributions in the field of video quality assessment in general and it has provided different datasets of video test-stimuli along with the corresponding subjective assessment results. The group is composed of experts from industry, academia, government organizations, and ITU. Two of its datasets are FR TV Phase 1 (2000) and HDTV Phase 1 (2010) for which the test plan, test-stimuli, and subjective assessment rating are freely available on its website. Based on the reports of the work performed under VQEG, several ITU recommendations have been published, such as ITU-T Rec. J.341 [18] is based on the work reported as HDTV Phase 1.

Standardization of Objective Methods

In order to standardize the development of objective methods of VQA, various activities have been made resulting in useful recommendations. ITU-T is the key organization that conducts the relevant studies to perform standardization which is crucial for, besides other benefits, interoperability of different information and communication technologies around the world. Depending upon the application area, different Study Groups (SG) of experts are established that conduct the required research to prepare the documents of recommendations. For instance, ITU-T SG12 [19] deals with standardization of performance, quality of service (QoS), and quality of experience (QoE) of multimedia services. Moreover, Video Quality Experts Group (VQEG) is a prominent platform where input from academia and industry is gathered to design planning and validation of objective methods. Based on the VQEG findings, ITU-T has published a variety of standards in the form of recommendations [20]. Both television and multimedia applications are covered by these reports and the related ITU standards.

As discussed in [21], objective methods of VQA can be classified into five types namely: Media-layer, Parametric packet-layer, Parametric planning, Bitstream, and Hybrid. An overview of this classification is provided in Table 1.1 where information about typical input, ITU-T standard and reference-based categorization of each type is briefly described. Media-layer methods process the video signal itself for an estimation of the quality and hence are more suitable in the applications such as offline testing, quality benchmarking, and codec comparisons. Examples include ITU-T J.247, which is related to multimedia VQA in the presence of a full reference, and ITU-T J.342, which applies to VQA of high-definition (HD) cable television in the presence of a reduced reference signal. ITU-T J.247 includes four methods of objective VQA for videos with
1. **Introduction**

Table 1.1: Overview of Standardization of Objective Methods of VQA

<table>
<thead>
<tr>
<th>Input</th>
<th>Media-layer</th>
<th>Parametric packet-layer</th>
<th>Parametric planning</th>
<th>Bitstream-layer</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITU-T</td>
<td>Video signal</td>
<td>Packet header</td>
<td>Design parameters</td>
<td>Packet header</td>
<td>Ensemble of others</td>
</tr>
<tr>
<td>Ref.</td>
<td></td>
<td></td>
<td></td>
<td>P.1202 [12]</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
</tbody>
</table>

QCIF (176 × 144), CIF (352 × 288), and VGA (640 × 480) resolution. ITU-T J.342 includes an edge-PSNR based method of VQA that relies on the edge-pixel information of the source video which is transmitted on an ancillary channel. Parametric packet-header methods get access to the header of the transported video packets for extracting information such as bitrate, frame-rate, and packet loss rate. Examples include ITU-T recommendation P.1201 that is related to NR quality assessment of audiovisual media streaming. This is actually an umbrella recommendation for two models, ITU-T P.1201.1 that is aimed at low resolution displays and ITU-T P.1201.2 that is aimed at applications involving HD resolution displays. Parametric planning methods, such as ITU-T G1070, have no access to the video bitstream and they rely only on the design parameters planned for the communication networks. Bitstream layer methods have full access to the bitstream for extracting a variety of video features that can be used for making an estimate of video quality and the related recommendation is ITU-T P.1202. Lastly, hybrid methods can be a composite of any of the earlier mentioned types and the related recommendation ITU-T J.bitvqm entitled Hybrid perceptual bitstream video quality assessment is yet under development phase.

1.1.3 **Overview of the Related Work**

In this thesis, the contributions related to objective VQA are focused on RR and especially on NR methods. In the following, some of existing techniques in RR and NR based methods of VQA are reviewed.
1.1. Introduction to Video Quality Assessment

Reduced-Reference Based Methods

As the RR methods require an ancillary channel that can be used from the server to the client for reliable transmission of the suitable features from server side, the available bandwidth of maintaining the additional channel can put constraints on the applicability of RR methods. In [29], an analysis has been performed on how various RR features can be reduced in size so that the required capacity of ancillary channel bandwidth can be lowered, based on authors’ earlier contribution [30]. Specifically, local harmonic strength related features, termed as *harmonic gain* and *harmonic loss* that correspond to blockiness and blurriness respectively, are computed from VQEG Test Phase-I database. In order to reduce the bitrate required for transmitting these features; quantization, temporal sampling, and feature matrix sub-sampling have been applied to study the tradeoff between quality prediction accuracy versus the bandwidth efficiency. The resulting setting point is reported to offer a huge compression factor in reducing the required bandwidth. However, the spatial *energy variation descriptor* and temporal *generalized Gaussian density* based method proposed in [31] reports to outperform [29] both in terms of higher accuracy and lesser requirement of ancillary channel bandwidth, while tested on H.264/AVC encoded videos. It is reported that the RR features, which are derived from spatial and temporal domains of the video, can be embedded in the encoded video or can be sent on a side channel.

The VQA method proposed by Wolf et al. [32] has been standardized and known as Video Quality Model (VQM) and has later been improved in [33]. These methods are based on a set of spatial and temporal features that represent the impact of compression artifacts. Impact of blur, blocking, and temporal artifacts on visual quality has been captured through representative features in reference [34] and the proposed method is tested for its performance on MPEG-2 encoded videos. Similarly, the impact of H.264/AVC compression artifacts, such as blur and blocking, has been captured to build a RR model in [35]. Distributed source coding theory has been applied in [36] to estimate the impact of network-induced artifacts on visual quality. In [37] the authors designed a RR measure, targeted for wireless applications. It is built based on the principle that humans tend to have different impairment perceptibility based on the spatial and temporal affected regions of a video sequence. The work in [38] presented a family of RR VQA models that differ in the amount of reference information required for video quality measurement, while [39] proposed a wavelet-based video distortion metric that can operate in FR or RR mode as required. RR based methods extract selected features from the origi-
nal video sequence and transmit them to the receiver either using an ancillary channel [40] or by watermarking [41]. Thus, they can be an alternative for FR metrics when the original video is not accessible. However, in some cases, the cost of maintaining an ancillary channel may be high for RR methods, or such methods may not meet the requirements of quality estimation in the event of a failure in features delivery to the receiver’s end. In such cases, NR methods are more applicable.

No-reference Based Methods

NR methods can be further classified into three types depending on the level of information about the processed video being used for making an estimate of the quality. These are pixel based, NR-P, bitstream based, NR-B, and a hybrid of both NR-P and NR-B methods. Bitstream based methods use readily available data from encoded bitstream of a processed video and do not require the full decoding of the video for making a quality estimate. The fundamental advantage of the NR-B methods is the requirement of lesser amount of computations for their mode of operation. A review of NR methods of VQA has been published in an article [11] and is presented as Chapter 2 in this thesis. Hence, it is referred to Chapter 2 for an overview of related work in NR based methods, in order to avoid repetition.

1.2 Introduction to Speech Enhancement

Speech is an essential component of most of the daily life communication services. Dedicated speech communication services include telephone (landline or mobile) and voice over IP services (Skype and Gtalk etc.). Speech can also be an integral part of multimedia services such as TV broadcasts. In all these cases, microphones are used to capture the speech signal for the purpose of storage or transmission. Unfortunately, depending upon the sensitivity of the microphone, it can also capture other acoustic sounds present in the vicinity of the source of speech signal. These additional unwanted signals or noise can degrade the original speech signal, even rendering it unintelligible. In many cases, it can be impossible to avoid the interference of noise in the speech signal. For example, a person working in a factory may always be surrounded by different noises that can disturb any phone calls. In order to circumvent such problems, signal processing methods have been devised that can boost the level of speech, thereby lowering the impact of the surroundings noise. Such methods are usually termed as speech enhancement methods [42].
1.2. Introduction to Speech Enhancement

There are a variety of methods by which speech enhancement can be performed. For example, spectral subtraction (frequently used for noise reduction) [43] and optimum Wiener filtering [44]. In spectral subtraction, the noise is assumed to be stationary or slowly varying and additive to the speech signal. The operation of Wiener filtering typically involves linear time-invariant filtering of the observed noisy process that is assumed to be stationary and additive.

Adaptive gain equalizer (AGE) is a time domain speech enhancement algorithm in which the speech signal is amplified based on signal-to-noise (SNR) estimates in subbands. A signal is divided into subbands for calculation of a gain which is independent for each band. The algorithm has shown advantages over contemporary techniques because of its low complexity implementation, no requirement of voice activity detector (VAD), and has no presence of musical noise as a result of controlled gains [45]. Additionally, hardware implementations of AGE [46] indicate its importance in speech processing applications. Besides the traditional time and frequency domains for signal processing, modulation domain was introduced as a two dimensional bi-frequency system, where time variation of the ordinary frequency is the second dimension [47]. There has been reasonably large interest in this domain for various tasks related to speech processing. Atlas et al. [48] used the concept of coherent modulation for the target talker enhancement in reference. They proved that processing in modulation domain can increase the speech intelligibility. Coherent modulation using the frequency reassignment has been used for speech enhancement and for demodulation of a signal into modulator and carrier [49].
The major part of this doctoral thesis is aimed at contributing to development in the area of VQA. In addition, one part aims at contributions in the area of speech enhancement. In the context of the aforementioned related work with regards to video quality assessment and introduction of speech enhancement, the following research objectives concisely describe the scope of the thesis.

Research Objective I: To classify and review recent approaches employed in the development of NR visual quality assessment methods.

Given the amount of published research in this area, it can be challenging to formulate a simple yet comprehensive approach of classification in a manner that most of the related publications can be reviewed. To this end, beginning with the general classification based on the domains of operation (pixel, bitstream, or a hybrid of these), a sub-classification with respect to the nature of data or features used as input for quality estimation enabled to present a review of an extensive number of contributions as presented in Chapter 2.

Research Objective II: To design, implement, and evaluate perceptually accurate bitstream-based methods of no-reference (NR) and reduced-reference (RR) video quality assessment.

For the design of a bitstream-based NR VQA method, the fundamental steps include the computation of perceptually relevant features that are extracted from the bitstream of a video and the presence of system or function that pools these features into an estimate of visual quality. Furthermore, the obtained estimate is used directly or it is mapped to a value of perceived quality. RR methods utilize additional inputs compared to NR methods in the form of suitable features obtained from server side. In order to evaluate a proposed method, standardized measures of performance as recommended by VQEG have been used. Chapter 3, 4, and 5 are dedicated to address this research objective.

Research Objective III: To apply a relevant subjective video quality assessment (VQA) method on a targeted set of video test-stimuli.

Various types of subjective VQA exist that may differ in how the content is presented to the subjects. Subjective VQA is performed for a number of reasons including the requirement to generate a ground-truth data that can be used for the evaluation and validation of objective methods; to determine perceptual preferences for various options of quality variations; and for investigation of a
certain method of subjective testing. Chapter 6 and 7 present subjective VQA performed by following ITU standards through laboratory based experiments and a non-standard method, namely crowdsourcing based subjective VQA, is investigated in Chapter 8.

Research Objective IV: To investigate the performance of the adaptive gain equalizer in the modulation frequency domain for speech enhancement.

The adaptive gain equalizer has been found to be a robust yet simple method of speech enhancement while implemented in time and frequency domains. Recently, there has been a trend to evaluate existing methods in the modulation frequency domain. This research objective entails a similar approach and has been addressed in Chapter 9 and 10. Essentially, the speech signal is divided into its carrier and modulator and the intended operation of speech enhancement is performed only on the modulators for improved results.
1.4 Thesis Outline: Summary of Research Contributions

In the context of the earlier mentioned research objectives, relevant contributions made in the form of research publications are presented in four parts of the thesis. A summary of the chapter(s) included in each of these parts is given in the following.

1.4.1 Part I: On Review of No-Reference Image and Video Quality Assessment

Chapter 2: No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

In view of the growing trends of research in perceptual quality assessment, it is imperative to obtain a perspective of the state-of-the-art methods. Especially, the NR based methods are more applicable in many real-time scenarios than RR and FR methods. This chapter presents a classification and review of recent approaches in the area of NR visual quality assessment. The NR methods of visual quality assessment considered for the review are classified into categories and subcategories based on the types of methods used for the underlying processing. Overall, the classification has been done into three categories, namely, pixel-based methods, bitstream-based methods, and hybrid methods of the aforementioned two categories. In order to review most of the existing recent approaches, a sub-classification up to feature or parameter level has been proposed. The chapter concludes with a description of future trends of research in this field, and research areas where there is a lack of studies are also identified.
1.4. Thesis Outline: Summary of Research Contributions

1.4.2 Part II: On Objective Methods of No-Reference and Reduced-Reference Video Quality Assessment

Chapter 3: A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

This chapter presents a study on the usage of bitstream-based features of an encoded video to estimate its perceptual quality. These features are related to H.264/AVC coding modes that essentially represent values assigned to various parameters while coding different parts of a video. These features have been selected from a previous study [50] where a correlational analysis of features was performed in order to avoid selecting any features that can potentially provide redundant information. Moreover, a further insight into the usability of these features is presented that is based on the contribution of an individual feature towards quality estimation. An Artificial Neural Network (ANN) based regression model has been trained on these features to predict video quality metrics such as PSNR, SSIM, and Perceptual Evaluation of Video Quality (PEVQ). By using a number of statistical measures, it has been shown that the proposed model can predict video quality better than a linear regression based model [50].

Chapter 4: A No-Reference Machine Learning Based Video Quality Predictor

It was noted that an ANN model has the capability to perform better than a linear regression model for quality assessment, as reported in Chapter 3. However, certain findings motivated to analyze the performance of regression models that can avoid inherent drawbacks of ANN such as overfitting. A rather popular kernel-based method of machine learning, known as Support Vector Machine (SVM), has been used to build a model of quality prediction that has been found to be more accurate than the ANN based quality prediction. Specifically, least-squares support vector machine (LS-SVM) that is a modified version of SVM, has been used. It has the possibility to avoid some of the computational requirements of the basic SVM technique. In addition to the set of features used in the previous study based on ANN, an enriched set of features has been used. The performance of the proposed model has been compared with similar techniques using standard methods of performance assessment as proposed by VQEG namely, Pearson correlation coefficient, Spearman rank order correlation coefficient, and outlier ratio.

Moreover, besides accounting for coding artifacts, a quality predictor also needs to consider the impact of the artifacts introduced through lossy trans-
mission. As it is hard to find an interpretable machine-learning based solution with respect to the relative use of the features in a model, the use of linear-regression techniques can be beneficial. To this end, Chapter 5 presents a study performed in order to investigate the use of features related to coding and network distortion for quality prediction based on regression techniques.

Chapter 5: Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

This chapter presents a study on the use of Reduced-Reference (RR) and No-Reference (NR) models for quality estimation of H.264/AVC videos impaired with packet losses. A variety of video features are examined for their impact on video quality by employing them in techniques of linear regression, namely Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) regression. In contrast to Ridge, the LASSO regression based models perform feature selection based on the impact of features on perceptual video quality, promoting not only the model’s sparsity but also accuracy. An insight into the feature selection done using LASSO with respect to different contents is also provided. Furthermore, the superiority of the proposed models has been highlighted by comparing them with related FR, RR, and NR based approaches.

1.4.3 Part III: On Subjective Methods of Video Quality Assessment

Chapter 6: Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

In order to validate the performance of an objective metric of quality assessment, it is essential to have access to a test-stimuli annotated with subjective assessment of quality and that can represent a variety of content characteristics. This chapter presents a study of generation of a set of test-stimuli and its subjective quality assessment performed following ITU standards. Particularly, six video sequences of different spatio-temporal characteristics with two frame-resolutions and two frame-rates were encoded at five different bitrates. After refining the results obtained in the subjective experiments performed in a standard compliant set-up, Mean Opinion Scores (MOS) were computed. The study verified the earlier known trends of perceptual quality variations in response to variations in bitrates, frame-rate, and frame-resolution. The encoded
bitstreams of the test-stimuli as well as the corresponding score were published online for potential use for the validation of objective models.

The dataset obtained in this study has been used in the validation of an objective VQA method, as given in Chapter 4. However, in view of the growing trend of use of adaptive streaming, it was realized that a more targeted test-set of video sequences is required. Thus, Chapter 7 presents a study where an investigation of the trade-off among quantization, frame-rate (temporal), and frame-resolution (spatial) variations with respect to perceptual preference has been performed.

Chapter 7: Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

Depending upon the context, simply increasing the compression-rate by increasing the level of quantization may not be perceptually preferable in response to decrease in the capacity of the transmission network. Instead, scaling down the video content spatially or temporally before the compression can cause lesser drop in perceptual quality. Such trade-offs of choosing an appropriate option for compression was not studied rigorously, especially in terms of comparison of a variety of frame-resolution and subjective assessment. This chapter presents a study where a trade-off among temporal, spatial, and quantization variations has been investigated with respect to subjective assessment of quality. The test-stimuli consists of five original videos encoded at four frame-resolutions, five bitrates, and three different frame-rates.

It was observed from this study that running subjective assessment experiments in a laboratory could be difficult even for a set of 190 videos, as it was required to do it in two sessions for each subject to avoid exceeding the recommended length of test-session. Moreover, the management of subjective experiments in a laboratory can be expensive, and laborious. In Chapter 8, a study on an alternative method of subjective assessment is presented that can avoid some of the challenges of the laboratory based experiments.

Chapter 8: Crowdsourcing based subjective quality assessment of adaptive video streaming

This chapter presents a study on a crowdsourcing based experiment of video quality assessment. This methodology mainly involves collecting subjective assessment of quality through ubiquitous streaming of test videos via the Internet. In contrast to the traditional laboratory based testing, crowdsourcing
1. Introduction

based subjective quality assessment offers more diversity in the participants, ratings based on a variety of the display devices, and it can be managed more economically. In this experiment, the test-stimuli consisted of HD videos encoded at four different bitrates. Additionally, these videos were processed to represent various adaptation scenarios of video streaming in response to fluctuating network conditions. The tested conditions include gradual or rapid decrease in the quality and buffering events with no or less decrease in the quality. This resulted in a set of 63 test videos that represent 9 different streaming scenarios. It turned out that crowdsourcing based quality assessment bears promising values of correlation with laboratory based testing and further improvements can be made in the crowdsourcing based experiments to employ them as a viable alternative to laboratory based experiments. The web interface used in this work has been published for public use at: https://github.com/J-Soegaard/PC-Video-Test-Interface

1.4.4 Part IV: On Speech Enhancement in Modulation Frequency Domain

Chapter 9: Modulation Domain Adaptive Gain Equalizer for Speech Enhancement

This chapter presents a study on the usage of the adaptive gain equalizer (AGE) technique of speech enhancement in modulation frequency domain. AGE implemented in time or frequency domain is a commonly used single-channel speech enhancement algorithm. It has shown its advantages already in digital, analog and hybrid domains due to its simplicity and low complexity. Modulation frequency domain deals with the analysis of a speech signal by splitting it into its modulator and carrier components through demodulation process. The implementation of the proposed system has been validated with various performance measurements, i.e., Signal to Noise Ratio Improvement (SNRI), predicted Mean Opinion Score (MOS), and Spectral Distortion (SD).

Chapter 10: Modulation frequency domain adaptive gain equalizer using convex optimization

In the process of demodulation, the possible modulator and carrier pairs are not unique and hence it could be problematic to find the most suitable pair using normal methods of demodulation. In order to resolve this matter, this chapter presents a study that evaluates the working of AGE in modulation fre-
Conclusions and Future Research Directions

After providing an introduction of two key areas of multimedia services, namely video quality assessment and speech enhancement, this doctoral thesis presents a number of contributions through research papers. These contributions aim at either extending existing work or proposing particular improvements of published studies. Briefly, an outlook of the conclusions of these contributions is presented in the following.

- A detailed review of the publications on no-reference based methods of image and video quality assessment is presented in Part I, in connection with Research Objective 1. The review is structured by following a classification that was introduced by extrapolating existing schemes of classification. A number of findings obtained from the review are reported, including the current research trends and possible future directions. This review can be instrumental for furthering research in a number of directions. Especially, it may equip young researchers of this field with a handbook of the related work as well as a quick means of gaining an overview of a particular category of methods.

- In response to Research Objective 2, three techniques of regression were investigated as presented in Part II. The results obtained through the use of non-linear techniques of regression, using artificial neural networks and support vector machines, supplement to the related work where other techniques of regression have been used. The use of a number of features in the regression based quality estimation models, as presented in Chapter 5, has been found promising where some of the features have been used for the first time. Furthermore, it has been found that some linear regression methods have competitive performance as compared to non-linear methods of VQA. Using linear regression based models, it is easier to interpret the usage of various features and the computational requirements are expected to be low as well.
1. **Introduction**

The promising results obtained through combining bitstream and pixel-based features for modeling an RR video quality estimator may instigate the development of more robust hybrid models built by following a similar approach.

- **Research Objective 3**, as addressed in Part III of the thesis, applies subjective assessment of video quality methods for different applications. Besides providing a ground-truth for evaluation and validation of objective models of quality assessment, the presented subjective VQA studies not only confirm the results of related prior studies but also present some intriguing findings. For example, there is a novel insight into perceptual preference of spatial resolution in the adaptation scenarios of tradeoff among quantization, temporal, and spatial levels. Crowdsourcing for subjective VQA has been performed, which is an emerging topic that has gained interest of a few research groups recently. Though the methodology itself remains to be standardized, the presented results in the thesis substantiate the potential of crowdsourcing as an alternative to the laboratory based experiments.

- **Finally, Research Objective 4**, as addressed in Part IV, aims at the investigation of the performance of adaptive gain equalizer in the modulation frequency domain as an alternative to the traditional time or frequency domain. The obtained results highlight the advantages of performing speech processing in this domain.

An outlook of the future research directions in the context of the presented contributions is given in the following.

- **Although Chapter 2 provides review of a large number of methods, further research can be made to integrate more approaches into the proposed classification. Moreover, best performing methods of each category can be found by doing performance comparisons that can ultimately give more insights for further improvements and inventions.**

- **In Part II, the proposed NR and RR methods based on various regression techniques have been evaluated on low or medium spatial resolutions. In view of the upcoming surge of the demand for HD and ultra HD, further studies can be made to generalize these methods for higher resolutions.**

- **As for the consumption of video services in a multitude of environments and display devices, the newly published ITU-T Recommendation**
1.5. Conclusions and Future Research Directions

P. 913 [51] addresses the subjective assessment of video quality in any environment. Accordingly, presentation of crowdsourcing based subjective VQA results is one of the auspicious aspects of Part III in this thesis and an extension of the related paper is already underway in order to maximally share the findings of this methodology. It would be advantageous to perform more investigations into the effectiveness of such alternative methods of subjective VQA.

- The processing of speech enhancement is performed only on the modulators of speech signal, using the AGE in modulation frequency domain as presented in Part IV. It can be useful to investigate the amount of reduction in computational complexity compared to the case of processing performed on the complete speech signal.

The remainder of this thesis presents the complete content of the aforementioned contributions. It is to be noted that minor formatting adjustments have been made in this thesis compared to the text of actual research papers.
1.6 Bibliography


[22] ITU-T Recommendation J.144: Objective perceptual video quality measurement techniques for digital cable television in the presence of a full
1. INTRODUCTION


1.6. Bibliography


1. Introduction


Part I

On Review of No-Reference Image and Video Quality Assessment
Two

No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

This chapter has been published as:

No-reference image and video quality assessment: a classification and review of recent approaches

Muhammad Shahid, Andreas Rossholm, Benny Lövström, and Hans-Jürgen Zepernick

Abstract

The field of perceptual quality assessment has gone through a wide range of developments and it is still growing. In particular, the area of no-reference (NR) image and video quality assessment has progressed rapidly during the last decade. In this article, we present a classification and review of latest published research work in the area of NR image and video quality assessment. The NR methods of visual quality assessment considered for review are structured into categories and subcategories based on the types of methodologies used for the underlying processing employed for quality estimation. Overall, the classification has been done into three categories, namely, pixel-based methods, bitstream-based methods, and hybrid methods of the aforementioned two categories. We believe that the review presented in this article will be helpful for practitioners as well as for researchers to keep abreast of the recent developments in the area of NR image and video quality assessment. This article can be used for various purposes such as gaining a structured overview of the field and to carry out performance comparisons for the state-of-the-art methods.

2.1 Review

2.1.1 Introduction

There has been a tremendous progress recently in the usage of digital images and videos for an increasing number of applications. Multimedia services that have gained wide interest include digital television broadcasts, video streaming applications, and real-time audio and video services over the Internet. The
global mobile data traffic grew by 81% in 2013, and during 2014, the number of mobile-connected devices will exceed the number of people on earth, according to predictions made by Cisco. The video portion of the mobile data traffic was 53% in 2013 and is expected to exceed 67% by 2018 [1]. With this huge increase in the exposure of image and video to the human eye, the interest in delivering quality of experience (QoE) may increase naturally. The quality of visual media can get degraded during capturing, compression, transmission, reproduction, and displaying due to the distortion that might occur at any of these stages.

The legitimate judges of visual quality are humans as end users, the opinions of whom can be obtained by subjective experiments. Subjective experiments involve a panel of participants which are usually non-experts, also referred to as test subjects, to assess the perceptual quality of given test material such as a sequence of images or videos. Subjective experiments are typically conducted in a controlled laboratory environment. Careful planning and several factors including assessment method, selection of test material, viewing conditions, grading scale, and timing of presentation have to be considered prior to a subjective experiment. For example, Recommendation (ITU-R) BT.500 [2] provides detailed guidelines for conducting subjective experiments for the assessment of quality of television pictures. The outcomes of a subjective experiment are the individual scores given by the test subjects, which are used to compute mean opinion score (MOS) and other statistics. The obtained MOS, in particular, represents a ground truth for the development of objective quality metrics. In ITU-R BT.500 and related recommendations, various types of subjective methods have been described. These types include either single stimulus or double stimulus-based methods. In single stimulus methods, the subjects are shown variants of the test videos and no reference for comparison is provided. In some situations, a hidden reference can be included but the assessment is based only on a no-reference scoring of the subjects.

Due to the time-consuming nature of executing subjective experiments, large efforts have been made to develop objective quality metrics, alternatively called as objective quality methods. The purpose of such objective quality methods is to automatically predict MOS with high accuracy. Objective quality methods may be classified into psychophysical and engineering approaches [3]. Psychophysical metrics aim at modeling the human visual system (HVS) using aspects such as contrast and orientation sensitivity, frequency selectivity, spatial and temporal pattern, masking, and color perception. These metrics can be used for a wide variety of video degradations but the computation is generally demanding. The engineering approach usually uses simplified metrics based
on the extraction and analysis of certain features or artifacts in a video but do not necessarily disregard the attributes of the HVS as they often consider psychophysical effects as well. However, the conceptual basis for their design is to do analysis of video content and distortion rather than fundamental vision modeling.

A set of features or quality-related parameters of an image or video are pooled together to establish an objective quality method which can be mapped to predict MOS. Depending on the degree of information that is available from the original video as a reference in the quality assessment, the objective methods are further divided into full reference (FR), reduced reference (RR), and no-reference (NR) as follows:

- **FR methods**: With this approach, the entire original image/video is available as a reference. Accordingly, FR methods are based on comparing distorted image/video with the original image/video.

- **RR methods**: In this case, it is not required to give access to the original image/video but only to provide representative features about texture or other suitable characteristics of the original image/video. The comparison of the reduced information from the original image/video with the corresponding information from the distorted image/video provides the input for RR methods.

- **NR methods**: This class of objective quality methods does not require access to the original image/video but searches for artifacts with respect to the pixel domain of an image/video, utilizes information embedded in the bitstream of the related image/video format, or performs quality assessment as a hybrid of pixel-based and bitstream-based approaches.

### 2.1.2 Applications of no-reference image and video quality assessment

In recent years, there has been increasing interest in the development of NR methods due to the widespread use of multimedia services in the context of wireless communications and telecommunication systems. Applications of NR methods include the following areas:

- Network operators and content providers have a strong interest to objectively quantify the level of service quality delivered to the end user and
inside the network nodes. NR methods will provide the data needed to adopt network settings such that customer satisfaction is secured and hence churn can be avoided.

- The involvement of multiple parties between content providers and the end users gives rise to establish service-level agreements (SLA) under which an agreed level of quality has to be guaranteed. In this respect, NR methods are a suitable choice for in-service quality monitoring in live systems.

- In general, NR methods are well suited to perform real-time objective quality assessment where resources are limited such as frequency spectrum in wireless communications. In such cases, RR methods have limited application as an ancillary channel is required to transmit the required features of the original video.

- Real-time communication and streaming services require quality adaptations using NR methods for collecting statistics of the delivered quality.

Related work: published reviews of objective visual quality methods

According to the framework introduced in [4] for NR visual quality estimation, three stages are present in an NR quality estimation approach. These stages are measurement of a physical quantity relevant for visual quality, also called as feature, pooling the measured data over space and/or time, and mapping the pooled data to an estimate of perceived quality. A survey of the measurement stage, which is essentially the main focus in much of the work done in NR quality estimation, has been provided in the same contribution. The survey in [4] divides the literature review into two main categories. In the first category, the methods estimating mean square error (MSE) caused by block-based compression, MSE caused by packet loss errors, and noise estimation methods to compute MSE have been discussed. The second category encompasses the approaches that are termed as feature-based. The feature-based methods are based on either a model developed for particular artifacts related to a visible degradation, or a model developed to quantify the impact of degradations on a specific set of attributes of the original uncorrupted image or video. A brief survey of NR methods of image quality assessment (IQA) based on the notion of quantifying the impact of distortions on natural scene statistics (NSS) is provided in [5]. Some NR methods of visual quality are discussed in [6] also under the categorization of features and artifacts detection. Similarly, a
review of the objective methods of video quality assessment (VQA) is provided in [7] including a classification of objective methods in general without specifying it for no-reference methods. In [7], the objective methods are classified as data metrics, pictures metrics, and packet or bitstream-based metrics. The review and performance comparison of video quality assessment methods in [8] present a classification of FR and RR methods only. A survey on visual quality assessment methods that are based on information theory is given in [9]. It was observed that information theory-based research for the development of NR methods is rather limited. The type of NR methods surveyed in [9] relies on an approach that employs Rényi entropy for determining the amount of randomness in the orientation of local structures in an image. NR methods have been reviewed in [10] by classifying them following three approaches. Firstly, a review of NR methods has been performed by classifying them based on the type of distortion that is estimated to formulate a quality value. The second approach used for the classification is based on methods that are designed for quantifying the artifacts produced by a specific compression standard. Lastly, a review of methods that are not designed specifically for a particular distortion has been performed. A broad survey of image and video quality methods, as well as a classification of the methods, was published during 2007 in [11]. This includes both NR and RR methods, and our article focuses on a classification and review of NR methods of IQA and VQA published after [11].

Our proposed classification

The current literature in the area of methods of NR image/video quality assessment is quite diverse. Hence, it is a challenging task to classify these methods into a well-structured and meaningful categorization. A good categorization of such methods should be concise enough to be properly understandable and also comprehensive enough to present most of the relevant methodologies. The aforementioned types of classifications cover a range of NR methods, but there is a need to broaden the categorization approaches in order to review currently existing methods in this area. Reibman et al. [12] classify NR methods as either stemming from statistics derived from pixel-based features and call them NR pixel (NR-P) type or computed directly from the coded bitstream and call them NR bitstream (NR-B) type. We believe that this is a useful classification which can serve as an effective basis for constructing a broader classification.

In the case of NR-P-based methods, one relevant method to classify available approaches is to investigate these in terms of the employment of certain
artifacts that are related to a specific kind of degradation of the visual quality. Quantification of such artifacts has been used as a measure for the quality assessment. The quality values may depend only on a single artifact or it may depend upon a combination of many artifacts. It is common that single artifact measure-based methods are developed by considering a given model of degradation, often simulated artifacts, and sometimes their performance remains unknown for realistic or more general scenarios. For example, most of the available blur methods are based on Gaussian or linear blur models, which may not adequately measure the blur produced by a complex relative motion between image capturing device and the object. Moreover, single-artifact-based quality methods may not have satisfactory performance in the assessment of the overall quality, in the presence of other artifacts. Therefore, methods have been introduced where estimation of a combination of artifacts is fused to generate a single quality score. Also, in the domain of NR-P-based methods, there are many methods which work beyond simple artifacts computation and the quality assessment is derived from the impact of distortions upon NSS (referring to statistical characteristics commonly found in natural images). Moreover, some quality-relevant features can be computed from the image/video pixels to formulate an estimation of the perceptual quality.

The NR-B-based methods are relatively simpler to compute than NR-P-based methods, and the quality values can often be computed in the absence of a full decoder. However, such methods can have limited scope of application as they are usually designed for a particular coding technique and bitstream format, e.g., H.264/AVC standard. Such methods are based on either the encoding information derived from the bitstream or the packet header information or a combination of both. These methods are quite suitable for network video applications such as IPTV and video conferencing.

Quality assessment performance can be compromised in NR-B-based methods to gain reduction in the computational complexity as compared to the NR-P-based methods. The performance of NR-B-based methods of quality assessment can be improved by adopting an approach of adding some input from NR-P-based quality assessment. Such composites of NR-P- and NR-B-based methods are called hybrid methods. These methods inherit the computational simplicity of NR-B-based methods and depend on NR-P-related data to gain further robustness.

In light of the discussion given above, our approach of a meaningful classification of NR objective visual quality methods is outlined in Figure 2.1. This
2.1. Review

No-Reference Image and Video Quality Assessment Methods

No-Reference Pixel (NR-P) Based Methods

- Artifacts Measure Based (Section 2.1.3 and 2.1.4)
- Natural Scene Statistics and other Spatial Features
- Blur, Blocking, Ringing, Noise, Temporal Impairments

No-Reference Bitstream (NR-B) Based Methods (Section 2.1.6)

- Features Measures Based (Section 2.1.5)
- Bitrate and Packet Loss Rate

Parametric Planning Model

- Bitrate, Frame-rate, Motion Vector, QP, Packet Loss Rate, etc.

Bitstream Layer Model

- Spatial Features and Bitstream Parameters
- DCT and Wavelet Coefficients Based Statistics

Hybrid of NR-P and NR-B Methods (Section 2.1.7)

- Pixel-based and Bitstream-based Features or Artifacts

Figure 2.1: An overview of NR image and video quality assessment methods. The second row of boxes gives a division into three main categories, further divided into subcategories in the next row. The bottom row gives examples of extracted features or information used for processing in each subcategory.

classification is formulated by considering the type and granularity of usage of the image or video data for the design of an objective method of quality. Thus, it offers the opportunity to present a discussion of most of recently published techniques of the NR visual quality assessment. It is to be noted that the main focus of this article is to review, in a systematic and structured manner, recent advancements in this area. Hence, a performance comparison of the reviewed methods on a comprehensive test database is out of the scope of this paper.

The preliminaries and organization of this paper

Most of the existing NR quality methods fall into NR-P or NR-B type methods or a hybrid of these two approaches. As shown in Figure 2.1, the following sections present an overview of the different classes of NR methods of IQA and VQA. In each section, we have presented a general idea used in computation of various types of methods of quality estimation using block diagrams.
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

Summaries of most of the discussed methods are shown in tables throughout the paper and in dedicated discussion sections. Mostly, the performance of an objective quality prediction model is reported by using measure of prediction accuracy, i.e., Pearson’s linear correlation coefficient, and measure of monotonicity, i.e., Spearman’s rank order correlation coefficient, as recommended by Video Quality Expert Group (VQEG) [13]. These measures have been used to report the performance of the reviewed methods in the tables. In these tables, some cells have been marked with a hyphen (-) in cases where the corresponding value has not been reported in the reference or some uncommon measure of performance has been used. Other than the explicit numerical values of the number of pixels used for stating the resolution of the test data, the following short forms are used:

- QCIF, Quarter Common Intermediate Format (176 × 144)
- CIF, Common Intermediate Format (352 × 288)
- SIF, Standard Interchange Format (320 × 240)
- SD, Standard Definition (720 × 480 or 720 × 576)
- HD, High Definition (1920 × 1080 or 1280 × 720)

For validation of the proposed method, some publicly available databases of images and videos have been used in most of the reference papers. In this paper, the reference to a public database of test media indicates that either a subset or the complete set of the available media has been used. These sources of the test media include the following:

- Laboratory for Image and Video Engineering (LIVE): LIVE offers databases of compressed images and videos with the corresponding data of the subjective assessment. The images have been encoded using Joint Photographic Experts Group (JPEG) and JPEG2000 standards. Moreover, some images have been generated using simulated conditions of certain artifacts such as Gaussian blur and white noise. The video database contains sets of videos encoded using Moving Picture Experts Group (MPEG)-2 and H.264/AVC. While we refer to the usage of test data from LIVE in the tables, the standard used for encoding shown in the column Processing indicates whether the used data is an image or a video. References to the publications based on the use of these databases have been provided at the source website [14].
2.1. Review

- Video Quality Experts Group (VQEG): VQEG has released its test data for public use which is available on their website [15]. The data contains standard definition television videos and the corresponding values of the subjective assessment.

- Tampere Image Database 2008 (TID2008): This database contains test data produced from 17 different types of distortion introduced in the given 25 reference images. The test images have been provided with the corresponding subjective assessment scores and values of many objective methods of quality estimation. More information on it is found in [16].

- Images and Video Communications (IVC): The IVC database contains a set of ten original images distorted by four types of processing and is supported by the corresponding quality scores as available in [17].

- Toyoma: This database consists of subjective assessment data and test stimuli generated through processing of 14 reference images using JPEG and JPEG2000 [18].

This article is organized as follows. For the pixel-based approaches, the methods that apply direct estimation of single and multiple artifacts are reviewed in Sections 2.1.3 and 2.1.4, respectively. The methods based on computation of various features and an evaluation of impacts of pertinent artifacts upon NSS are discussed in Section 2.1.5. Bitstream-based NR methods are reviewed in Section 2.1.6. The methods constructed as hybrids of pixel and bitstream-based approaches are discussed in Section 2.1.7. Finally, some conclusive remarks and a brief outlook of possible future works in this area are presented in Section 2.2.

2.1.3 Single artifact NR-P-based methods

Blurring, blocking, and ringing are considered to be the most commonly found spatial domain artifacts in images/videos compressed by lossy encoders [19]. Moreover, noise is also a common source of annoyance in images and videos. Transmission of videos over lossy networks gives rise to temporal artifacts such as frame freeze. In the following, we examine the recent methods which adopt the approach of quantifying a single artifact for perceptual quality estimation. The section is divided into subsections for each of these artifacts, and an overall discussion is provided at the end.
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

**Blurring**

Winkler defines blur as an artifact which appears as a loss of spatial detail and a reduction of edge sharpness [20]. The reasons for the occurrence of blur can be many, originating in the acquisition, processing, or compression [21]. The primary source of blur in compression techniques is the truncation of high-frequency components in the transform domain of an image. Other possible reasons of the blurring of an image or video can be out-of-focus capturing, relative motion between the camera and the object being captured, or limitations in the optical system. Traditional no-reference blur methods usually focus on a particular coding artifact for quality prediction and hence their performance is compromised in circumstances of more general blur. Moreover, there has been little work carried out to build methods which have the capability of assessing blur in natural scenarios, rather, most of the work is focused on the simulated blur. A basic schematic of NR blur assessment is shown by the flowchart given in Figure 2.2. In many NR methods of estimating the impact of blur on visual quality, the computations begin with measuring the spread of pixels present on the edges in an image. Usually, it involves the application of commonly used edge detectors such as Sobel and/or Canny for finding the edges in the image. The next step is typically the computation of the edge distortion value that can be used towards finding an estimate of the blur. Some methods, however, make use of HVS adaptation to the value of edge distortion to classify it as perceivable or not perceivable by a human subject.

A paradigm for blur evaluation has been presented in [22] that is mainly composed of four methods of blur quantification, given in [23, 24, 25] and [26], which have been integrated by an artificial neural network (ANN) powered multifeature classifier. In the method given in [23], an image quality measurement method in terms of global blur has been proposed. The method relies on histograms of discrete cosine transform (DCT) coefficients present in MPEG and JPEG encoded data to qualitatively encompass the distribution of null coefficients, given the fact that blurred images usually end up having a lot of high-frequency coefficients set to zero. This algorithm provides results which align with subjective assessment but it focuses only on out-of-focus blur and it does not perform well when there is a uniform background present or when an image is over-illuminated. The blur assessment algorithm proposed in [24] exploits the ability of the Haar wavelet transform (HWT) to distinguish edge types, and the method works both for out-of-focus and linear-motion blur. This method is however not tested for realistic blur. The method proposed in [25] presents a framework where global blur is measured in terms of averaged edge...
2.1. Review

Figure 2.2: A basic scheme for NR-P-based assessment of blur.

lengths. The authors considered only a small set of Gaussian blurred images for its evaluation. Nonetheless, the method has good correlation with subjective scores. An improved version of [25] is found in [26] where HVS properties have been added to get weighted edge lengths. It is to be noted that none of these four reference methods quantify realistic blur situations, but Ciancio et al. [22] have shown their method to be useable for measuring naturally occurring blur. Overall, [22] uses local phase coherence, mean brightness level, and variance of the HVS frequency response and contrast as additional inputs, together with the earlier mentioned four methods, to various ANN models designed for quality estimation. For input calibration, a five-parameter nonlinear mapping function was used for the types of blur including simulated Gaussian, simulated linear motion, a combination of both, and real blur. The proposed method outperforms the given four reference methods when tested on a fairly large database of 6,000 images corrupted by blur. Although the proposed method does not correlate so well with subjective scores in realistic blur scenarios, with a Pearson’s correlation coefficient of approximately 0.56, it performs better than the reference methods with respect to subjective rating. In an earlier paper, the same authors have used the idea of estimating image blur using local phase coherence [27] and a similar method proposed by Hassen et al. is found in [28].
It has been argued in [29] that blur below a certain threshold value of blur remains unperceived by the HVS and such a threshold value is termed as just noticeable blur (JNB). By incorporating the response of the HVS to sharpness at various contrast levels, the authors have proposed a measure of image sharpness. It is suggested that most of the existing no-reference blur assessment methods do not perform well for a variety of images and are rather limited to assess varying blur in a certain image. They have validated this argument by testing a set of 13 contemporary reference methods, which are based on different techniques of blur assessment used for quality assessment such as pixel-based techniques, statistical properties, edge-detection-based, and derivative-based techniques. The proposed method has higher correlation with subjective MOS than the given 13 objective methods of quality assessment when it has been tested on a public database of test images. In [29], the block size used for finding edge pixels is $64 \times 64$, and a similar contribution based on JNB from the same authors is reported in [30] where a block size of $8 \times 8$ has been used for finding the edge pixels. The method proposed in [30] has been improved in [31] by adding the impact of saliency-weighting in foveated regions of an image. Specifically, more weighting is given to the local blur estimates that belong to salient regions of an image, while spatial blur values are pooled together to compute an overall value of blur for the whole image.

A similar method found in [21,32] improves [29] by addition of the concept of commutative probability of blur detection (CPBD) so that the method should estimate the quality by including the impact of HVS sensitivity towards blur perception at different contrast levels. Testing the proposed method upon three public image databases having different blur types reveals that the method performance is considerably better than some of the contemporary sharpness/blur methods. However, this method gives a quality index in a continuous range of 0 to 1 and the authors have modified it in [33] where it gives a quality value on a discrete scale of 1 to 5, the usual five quality classes which are described from Bad to Excellent. Given that blur estimation methods most often work on the idea of measurement of edge-spread and blur manifests itself in smooth or diminished edges, some edges may remain undetected. Varadarajan et al. [34] improved the method proposed in [29] by incorporating an edge refinement method to enhance the edge detection and hence outperformed the blur assessment. The authors achieved as much as 9% increase in Pearson’s correlation coefficient.

In contrast to usual schemes of blur detection at the edges, the method proposed in [35] does an estimation of blur at the macroblock (MB) boundaries.
The overall blur of an image can be calculated by averaging the block level measure for the whole image. The authors have also used a content-sensitive masking approach to compensate the impact of image texture. As the method was designed for videos encoded following the H.264/AVC standard, it mainly quantifies the blurring effects from quantization and de-blocking filter. This method is essentially based on a method proposed for images [36] where an estimation of the blur in a video is made by taking an average measure of blur values for each frame.

A wavelet-based noise-resilient color image sharpness method is presented in [37]. The procedure is to compute a multiscale wavelet-based structure tensor which represents the multiscale gradient information of local areas in a color image (image gradient is defined as the directional change in the intensity or color in an image). The proposed tensor structure preserves edges even in the presence of noise. Thus, the sharpness method is defined by calculating the eigenvalues of the multiscale tensor once edges have been identified. A competitive correlation with subjective MOS is achieved when the proposed method is tested on LIVE image database [14], in comparison to a similar sharpness method.

Out-of-focus blur estimation without using any reference information has been given in [38] using the point spread function (PSF) which is derived from edge information. As the proposed algorithm works in the spatial domain, avoiding any iterations or involvement of complex frequencies, it is expected to operate fast and possible to be deployed in real-time perceptual applications. Based on the similar approach in [39], the method has been made workable to assess blurriness of conditions like added blur, realistic blur, and noise contamination.

Chen et al. [40] have claimed that their method works for any kind of blurriness, without being sensitive to the source of the blur. A gradient image is calculated from the given image pixel array. A Markov model is used and a transition probability matrix is computed. Finally, a pooling strategy is applied to the probabilistic values to obtain the blurriness measure.

Some of the other recently introduced no-reference blur assessment methods include the following: In [41] a method based on multiscale gradients and wavelet decomposition of images is given, an image sharpness based on Riemannian tensor mapping into a non-Euclidean space has been found in [42], radial analysis of blurred images in frequency domain is done in [43] to set an
image quality index for blur estimation, and reference [44] presents a perceptual blur method to assess quality of Gaussian blurred images. A method based on blur measure in salient regions has been presented in [45]. The perceptually relevant areas in an image are identified through elements of visual attention, namely, color contrast, object size, orientation, and eccentricity. Quality values in correlation with subjective scores are produced by localizing the degradation measure in these elements.

### Blocking

Blocking is an artifact which manifests itself as a discontinuity between adjacent blocks in images and video frames [3]. It is a predominant degradation that occurs after employment of block-based processing and compression techniques at high compression ratio conditions. In such techniques, transform is usually followed by quantization of each block individually leading to incoherent block boundaries in the reconstructed images or frames. Blockiness can be estimated in a region of an image, in general, by computing the difference between neighboring blocks and the amount of brightness around those blocks as shown in Figure 2.3. After the value of blockiness is determined in a certain region, it is important to estimate whether it would be significant for human perception or not by taking into account the impact from masking effects. This way, certain features that represent the input from HVS can be calculated. In general, blocking perception is affected by various factors including the blockiness strength (i.e., the difference between adjacent blocks), the local brightness around the blocks, and the local texture present in an image.

A frequency domain pixel-based bi-directional (horizontal and vertical) measure used to gauge blocking in images is presented in [46]. The authors claim that the proposed method can be used for any image or video format. Unlike the traditional no-reference blocking measures, this method does not require any *a priori* information about block origin, block offset or block-edge detection. The method has been evaluated on a large set of LIVE image and video database available as JPEG encoded images and MPEG-2 encoded videos. It outperforms a set of 13 contemporary blockiness methods in terms of prediction accuracy and monotonicity.

Liu et al. [47] presented an HVS-based blocking method to assess image quality using a grid detector to locate blocking. A local pixel-based blockiness measure which is calculated on the detected degraded regions is averaged to provide a blockiness value for the whole image. The main strength of this
2.1. Review

Figure 2.3: A basic scheme for NR-P-based assessment of blocking.

method in terms of computational efficiency and relevance to HVS response lies in the application of visual masking which makes the calculations perform only in the areas of blockiness visible to human perception. The authors took up the same method for further extensive evaluation in [48] under various conditions of comparison of performance where, for example, HVS models and grid detector are omitted or included. The results show that the proposed method performs better than some contemporary methods and can be a good candidate for realtime applications due to its simplified HVS model.

In [49], a blockiness assessment method is presented for block-based discrete cosine transform (BDCT) coded images. It is based on the estimation of noticeable blockiness. The so-called noticeable blockiness map is derived from luminance adaptation and texture masking in line with HVS response combined with a discontinuity map to quantify the visual quality. Along with its validated usability for deblocking of JPEG images, it has the potential of optimizing the codec parameters and similar other post-processing techniques.

Babu et al. presented their HVS related features-based blocking method in [50]. Blockiness as perceived by humans in JPEG encoded images is affected by a number of features such as edge amplitude around the borders of DCT blocks and edge length; the value of these increase in amount as compression rate is increased. It is also affected by the amount of background activity and background luminance as these have masking impact on possible blocking artifacts. The authors have used a sequential learning algorithm in a growing and
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

pruning radial basis function (GAP-RBF) network to estimate the relationship between the mentioned features and the corresponding quality measure. Babu et al. also proposed a method of determining block-edge impairment [51] using the idea that edge gradients of blocks in the regions of low spatial details would contribute towards the overall blocking in an image. The level of spatial details is estimated through edge activity that is computed through standard deviation measurement of each edge.

Other methods in this area include the blind measurement of blocking in low bit rate H.264/AVC encoded videos based on temporal blocking artifact measure between successive frames of a video presented in [52]. A weighted Sobel operator-based blocking method is presented in [53], in which the computation involves luminance gradient matrices of DCT-coded images. A method where a rather simple approach of taking abrupt change in pixel values as a signal of blocking has been proposed in [54] and it can be implemented both in pixel and DCT domain, and a method of blockiness estimation in natural scene JPEG compressed images has been presented in [55] which was influenced by the impact of multineural channels pattern of HVS for vision sensing.

Ringing

The ringing artifact is associated with Gibbs phenomenon and is observed along edges in otherwise smooth texture areas [20]. It has yet been relatively less investigated for NR perceptual quality measurements. This kind of degradation is caused by rough quantization of the high-frequency transform coefficients and is observed in the form of ripples around high contrast edges. A schematic block diagram of commonly used approaches for the estimation of perceptual ringing is shown in Figure 2.4. Certain features can be extracted from the edge maps to classify the image areas in terms of relevance towards ringing artifact. Masking effects of textured regions can be examined to check if the ringing would be visible to HVS perception. From the obtained data, a ringing map is generated for various regions and an overall value of perceptual ringing is obtained for the whole image. We have not found any publication on the NR estimation of ringing in videos.

Liu et al. have put forward HVS-based quality assessment methods which quantify ringing in compressed images in [56] and [57]. The work in [56] does not incorporate the masking effects of HVS properties. However, in [57], Liu et al. have improved the already existing method in multiple aspects. Edge
detection is crucial for locating the possible ringing artifact and is used along with consideration of HVS masking in designing of such a method. The HVS masking is integrated by adding human visibility index of ringing nuisance estimate inside the already detected distorted regions. This method has a performance level comparable to a full reference method and it outperforms the two given no-reference methods of ringing assessment while tested on JPEG compressed images. As the method does not use coding parameters like DCT coefficients, the authors argue that a slightly tuned version of the same method should perform similarly well when employed on other types of compressed images, e.g., JPEG2000.

Ringing may occur also as a result of an image restoration process, unlike the other artifacts which usually occur during compression. The ringing that occurs due to image restoration has different characteristics as compared to the one that occurs due to compression. Iterations of blind deconvolution in the image restoration process are likely to result in the generation of ringing [58]. A quality method to assess perceived ringing as a result of application of blind deconvolution methods for image restoration is proposed in [58] and in [59]. The authors claim that these methods evaluate ringing with no sensitivity to the image content type and any specific ringing process. In the method proposed in [58], a 2D Gabor wavelet filter and a line detector were used to quantify
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

ringing in restored images. A similar approach with enhancement is found in [59] where the authors have proposed to assess the degradation on image boundaries and image edges separately and then fuse the weighted results of the two values to have the overall ringing value. A 2D Gabor filter response image is used to calculate the perceived ringing at boundaries, and a Canny edge detector is used for locating ringing around edges in the image. The proposed method was tested on gray scale images restored from simulated blur. It has been found that the reported results are in line with subjective scores of quality assessment.

Noise

Besides the aforementioned unwanted components of an image or video that affect the perceptual quality, there can be other types of spatial noise as well. The mostly occurring types of spatial noise include salt and pepper noise, quantization noise, Gaussian noise, and speckle in coherent light situations. Mostly, the noise is considered to be an additive component, e.g., Gaussian noise, but in some situations the noise component is multiplicative, e.g., speckle noise [60]. Noise can be introduced during the image/video acquisition, recording, processing, and transmission [61]. Estimation of noise is required due to numerous reasons and applications in image processing such as denoising, image filtering, image segmentation, and feature extraction. For the estimation of noise signal, in most cases, it is assumed to be independent, identically distributed additive and stationary zero-mean signal, i.e., white noise [62]. Image noise estimation methods can be categorized into either smoothing-based approaches, where noise is computed using the difference between the input image and a smoothed version of it, or block-based approaches, where block variances of the most homogenous block in a set of image blocks is taken as noise variance [63]. Similar to the approaches used for estimation of other artifacts, computation of noise characteristics depends on the extraction of some features that are affected by noise. Figure 2.5 shows the basic scheme of a block-based approach of noise estimation where an image is divided into smooth areas. A variance higher than a certain threshold in those areas gives an estimate of the noise.

A block-based approach proposed in [64] uses statistical analysis of a histogram of local signal variances to compute an estimation of image noise variance. However, this method is challenged by high computational requirements due to its iterative processing, and [65] simplifies this technique by taking image structure into consideration. It uses high-pass directional operators to
determine the homogeneity of blocks besides using average noise variances. The performance of the improved method has been verified using highly noisy as well as good quality images. This method requires a full search of an image to determine the homogeneous areas in it. At the expense of decreased accuracy, spatial separations between blocks can be used to reduce the computational complexity. This approach has been adopted in [66] where particle filtering techniques have been used in the process of localization of the homogeneous regions. It has been shown that the proposed method reduces the number of required computations for homogeneity measurements while it outperforms [65] in accuracy. More examples of block-based approaches are found in [67,68,69] where noise level is computed by performing principal component analysis (PCA) of the image blocks.

**Temporal impairments**

Temporal impairments can be divided into two main categories: impairments caused by the encoding process and impairments caused by network perturbations. The typical temporal impairments caused by the encoding process
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

come from temporal downsampling which can be performed uniformly or non-uniformly, depending on different underlying reasons. The impairments generated by network perturbations come from delay or packet loss [70]. These different impairments can be categorized as the following [3, 4, 71, 72]:

- **Jerkiness**: non-fluent and non-smooth presentation of frames as a result of temporal downsampling
- **Frame freeze**: frame halts as a result of unavailability of new frames to present due to network congestion or packet loss etc.
- **Jitter**: perceived as unnatural motion due to variations in transmission delay as a result of, e.g., fluctuations in the available bandwidth or network congestion
- **Flickering**: noticeable discontinuity between consecutive frames as a result of a too-low frame rate together with high texture, coding artifacts, or motion content
- **Mosquito noise**: appears as temporal shimmering seen mostly in smooth textured areas produced by ringing and prediction error due to motion compensation mismatch

Jerkiness is the impairment perceived by the user, while jitter and frame freezes are the technical artifacts which produce jerkiness. Figure 2.6 presents an overview of how temporal impairments are computed in most of the contemporary methods. Generally, the first step is to compute the inter-frame difference of pixel intensities (usually the luminance channel only) and the obtained value can be used as it is or a mean square value can be calculated. Afterwards, various techniques can be applied to determine the location and possibility of frame freeze or frame drops. Some kind of thresholding is then useful to obtain more information about the occurrence of a potential temporal artifact. Finally, a suitable pooling mechanism is used to compute an overall value of the artifact under consideration.

Borer [71] presented a model based on the mean square difference (MSD) of frames for measuring jerkiness (both frame jitter and frame freeze) which proved its potential for quality assessment of videos with resolution ranging from QCIF up to HD. This model calculates jerkiness as an accumulative result of multiplication of three functions called relative display time, a monotonic function of display time, and motion intensity of all frames. The display time
2.1. Review

Computation of Interframe Difference of Video Frames

Input Video

Computation of Frame Freeze/Drops

Frame Freeze/Drops Thresholding

Temporal Quality Assessment

Figure 2.6: A basic scheme for NR-P-based assessment of temporal artifacts.

and motion intensity values are parameterized through a mapping S-shaped function, which is equivalent to a sigmoid function. Besides the fact that the proposed model has reasonable correlation with MOS, it does not take into account the value of the motion intensity at the start of a freezing interval.

An earlier proposed temporal quality method which is centered around measuring the annoyance of frame freeze duration is given in [72]. This method uses MSD value to mark freeze events and builds a mapping function based on such durations of freeze to estimate the subjective MOS. The method is a part of ITU-T Recommendation J.247 Annex C [73] for the objective perceptual quality measurement of video. Although the quality method has not been compared for performance against other methods, it has promising values of correlation with the subjective scores. However, the blind frame freeze detection system proposed in [74] claims to outperform the model [72] in terms of precision of correctly signaling a zero MSD event as a frame freeze event or not. They
have presented an algorithm for thresholding such as zero MSD events to be classified as frame freeze events or the absence of it. The proposed method is reported to be equally good in performance for videos encoded using low or high quantization parameter (QP) values.

Wolf proposed an approach to accurately detect video frame dropping in [75]. One of the salient features of this approach is its use in a RR method where an adaptive threshold value is determined to avoid detection of very low amount of motion (e.g., lips movement) as a potential frame drop event. Similar to the temporal-artifact-based methods discussed before, this method also derives its computations from the basic difference in pixel values between frames to check for possible frame drops.

A method for visual quality distortion due to arbitrary frame freeze is discussed in [76]. It recursively aggregates arbitrary freeze distortions in the video under test using a method which they proposed earlier in [77]. Essentially, the approach presented in [77] replaces the sum of various freeze distortions with an equivalent single freeze length for predicting the video quality.

Yang et al. targeted their research to assess both consistent and inconsistent frame drops as a measure of perceptual quality in their contribution found in [78]. The constituents of the quality method are the amount of frame drops, motion of various objects in a video, and localized contrast of temporal quality. Instead of relying on frame rate to be used as a basis for temporal quality method value, the event length of frame losses has been used. The proposed method correlates well with subjective MOS for test sequences with a range of frame rates and a variety of motion contents.

A rather general model was proposed in [79] under several fluidity break conditions: isolated, regular, irregular, sporadic, and varying discontinuity durations with different distributions and densities. Similarly, the temporal quality method proposed in [80] accounts for the impact of various frame dropping situations and spatio-temporal luminance variations due to motion.

In [81], the authors have shared their preliminary findings on estimation of the effects of lost frames on visual quality by analyzing the inter-frame correlation present at the output of the rendering application. As the lost frames are replaced by a repetition of the previous frame, this results in high temporal correlation at those locations. Analysis of this correlation results in temporal and distortion maps.
2.1. Review

Discussion

Except for temporal impairments, most of the methods reviewed in this section have been proposed and tested for images and not for videos. For example, blockiness is a common artifact at high compression rates and some coding standards such as H.264/AVC include the use of a deblocking filter while the videos are being processed by the codec. The blockiness methods proposed for images can be used in the case of videos as well where a suitable temporal pooling scheme needs to be used. We understand that development and testing of more NR methods of blockiness estimation for videos would be beneficial. For the case of spatial-artifacts-based methods, it is evident that most of the research focus has been aimed at the development of techniques that are based on a specific coding technique or image compression standard. This fact necessitates the focus towards unraveling cross-encoder methodologies. Considering the available methods related to the quantification of perceptual impacts of various temporal artifacts, it is noted that more diverse methods are required in this area that can be applied for a variety of video resolutions and frame rates. It has also been observed that many methods employ some commonly used test database of images and videos which in turn gives an opportunity to compare the performance of competitive methods on the common benchmarks of quality. One important strength of the methods that are tested for the performance using test databases such as LIVE (image or video) is their higher applicability because the media present in such databases have been assessed for overall perceptual quality and not for a particular artifact. However, the test databases should be enriched with new demanding areas such as higher-resolution images and videos (HD and above). Besides declaring the performance of the proposed methods, finding some common approaches for reporting the computational complexity would be interesting. Table 2.1 presents a summary of the methods discussed in the subsections regarding blurring, blocking, ringing, and temporal artifacts. It is noted that a very low number of methods have been tested for HD resolution images. Competitive methods can be seen at a glance by observing the significantly high values of the performance indicators.

2.1.4 Multiple artifacts NR-P-based methods

Various artifacts found in images and videos, incurred due to compression or other reasons, can be combined to predict the overall perceived quality. As shown in Figure 2.7, an image or video can be processed for the extraction of features relevant to different artifacts. A suitable pooling mechanism can be employed to combine the results of different artifact measurements, to make an
estimate of overall perceptual quality. Blurring and ringing are the main associated degradations when JPEG2000 coding is operated at low bitrate conditions. The quality method proposed in [19] predicts quality of JPEG2000 coded images by combining blur and ringing assessment methods. Based on the local image structures, a gradient profile sharpness histogram is calculated for evaluation of a blur estimation method, and a ringing method is generated from regions associated with gradients profiles. Here, a gradient profile is essentially the distribution of the gradient magnitude along the gradient direction. It has been argued that the underlying proposed blur method is insensitive to the inherent blur found in natural images, e.g., out-of-focus blur. The performance of the method is similar to or better than a number of competitive methods while tested on LIVE JPEG2000 and TID2008 datasets.

A rule-based VQA method given in [82] relies on a group of pixel domain features of a video. It includes blockiness and blurriness as well as spatial
activity, temporal predictability, edge continuity, motion continuity, and color continuity. The authors have used already available methods to measure the first five features and have proposed their own methods for the estimation of the motion continuity and color continuity. A multivariate data analysis technique has been used to combine all the features for computing a single quality score. The earliest mentioned three features (blockiness, blurriness, and spatial activity) are measured on a single frame and the rest is calculated on an inter-frame basis. The used approach is to segregate the given set of videos into one of the given feature models and then estimate an initial prediction of the quality measure. After that, using a low-quality version of the video, a correction value is added to the initial quality estimate of the original videos. The authors claim that at the time of publication, this is the first reference-free quality method for H.264/AVC encoded videos which have been tested on a relatively large test database.

Noise is an artifact found in images in the form of a random variation of brightness and color information (see Section 2.1.3 for more details on noise). An empirical formulation of the objective measure of image quality based on blur and noise has been proposed in [83]. The method is based on the level of
image intensity variations around its edges. The authors argue that in modern digital cameras, the image signal processor (ISP) enhances the image by removing noise but in doing so, it may deteriorate the image texture. Hence, there is a need of finding a trade-off between noise and blur and it provides a rationale for combining the estimation of noise and blur in the same method. Specifically, this method considers simulated conditions of white noise as source of the noise artifact in the test stimuli.

Another joint method for noise and blur estimation is found in [84]. It estimates the amount of degradations introduced by additive white noise, Gaussian blur, and defocus blur on the quality of an image. Given the fact that noise disturbs virtually all the spatial frequencies of an image and causes an adverse rise in the higher frequencies while blur has attenuation effect on them, it is justified to study the joint impact of noise and blur on the perceptual quality. The authors have evaluated the impact of noise in both spatial and frequency domain while only the frequency domain is used for blur estimation. The central idea is to influence the power spectra of the image in order to highlight the impact of the distortions on the spectrum properties. The source of noise in the test stimuli used in this work is also white noise. The proposed method has not been tested for its correlation with subjective assessment but it has a competitive performance in comparison with a contemporary method [85] of blind image quality evaluation.

In [86], a sharpness method is presented which is sensitive to the prevailing blur and noise in an image. The mathematical formulation of the method is based on image gradients computed through singular value decomposition (SVD) rather than edge detection as commonly found in contemporary pixel-based structure measures. However, it requires a prior estimate of noise variance. This issue has been resolved in the authors’ later contribution [87]. Simulations on realistic noise data have substantiated the potential usage of this method in parameter optimization issues of image restoration such as applications for denoising. The support vector regression (SVR)-based method reported in [88] uses singular vectors from the SVD data instead of using singular values as in [87]. Various artifacts would modify the singular vectors, and hence the geometry of the distorted image will be changed leading to visual annoyance as perceived by humans. The usefulness of the method was tested on multiple image databases with a variety of artifacts. The results were found to be in accordance with subjective ratings.

Another quality method based on gradient statistics of JPEG and JPEG2000...
images, degraded by blockiness and blur, is presented in [89]. This method differs from the methods given above in one way that it does not combine the estimated amount of artifacts to yield a single quality score. Instead, it uses the same approach of calculation of local features in gradient domain for both of JPEG and JPEG2000 images and then estimates the quality of the two sets separately. The obtained results lie in accordance with some contemporary methods of blocking estimation in JPEG images and blur estimation in JPEG2000 images. Further, an artificial neural network has been used in [90] to combine a blocking method, a blurring method, and a ringing method to estimate the overall quality of an image. Quality estimators targeted for images encoded by JPEG2000 usually quantify ringing only, but such images may contain blur as well. The method proposed in [91] first determines the type of distortion by using an ANN classifier and then, depending on these results, either uses a ringing [92] or blur [43] method for the quality assessment.

Different from the aforementioned IQA methods, another example of a composite method has been proposed for videos [93]. This method is based on blocking and flickering measure of H.264/AVC encoded videos. It correlates well with subjective quality assessment and also with the structural similarity (SSIM) index [94].

Most of the VQA methods are processed in the luminance plane only to simplify the computational complexity. However, the method proposed in [95] computes three artifacts both in the luminance and chrominance planes of a video. In this method, they compute the significance of the direction in which an artifact is calculated for determining its contribution to perceptual quality assessment. Hence, for example, the value of blur in vertical direction has been given more weighting than the same in horizontal direction. In this method, blocking is measured by computing boundary smoothness between $8 \times 8$ blocks and block visibility detection. The third impairment which is considered is jitter/jerkiness. Finally, a multiple regression scheme is employed for weighted integration of the six feature values towards the corresponding quality value. The suggested quality predictor bears competitive correlation with subjective MOS when compared with some contemporary methods as tested on standard-definition television (SDTV) sequences found in VQEG Phase 1 database.

A modular method of combining artifacts both from spatial and temporal domain for quality estimation has been proposed in [80]. The method accounts for frame freeze/jerkiness and clearness/sharpness in MPEG-4 encoded videos. It has been claimed that the combined model is an estimator of global
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

visual quality.

Discussion

Given the fact that a certain type of processing, e.g., JPEG2000 coding, can introduce more than one kind of artifacts, it is imperative to have quality estimators that can assess the impact of more than one artifact. The application of the estimation of multiple artifacts becomes even more interesting when a certain processing that involves removal of an artifact, such as denoising, can produce another artifact due to its underlying methodology. The popularity of digital cameras in the recent years increases the demand of a quality estimation mechanism to compute multiple artifacts that can be used as an aid to improve the photography experience. Global visual quality estimators such as in [80] are a useful contribution towards making an overall assessment of a video signal as it can be impaired by spatial artifacts like blurring and temporal artifacts like jerkiness at the same time. Table 2.2 presents a summary of some of the existing methods of quality assessment that are based on the estimation of multiple artifacts. Overall, it is noted that these methods should be tested on higher-resolution images/videos to account for the requirements of the new display devices with capability of presenting resolutions of HD and above.

2.1.5 Features measures-based methods

An image or video signal can be decomposed to obtain various features that may be used in the process of estimating the perceptual quality of an image or a video. Generally, such features can represent a particular aspect of the visual signal and its relation to the corresponding perceptual quality. Depending upon the nature of the feature with regards to its relation to perceptual quality, a certain feature can be a desired or an unwanted component of an image or video. For instance, the presence of sharpness in an image can be perceptually preferred in many cases and hence it may be considered as a wanted feature. On the other hand, an image with pixel distortions could be considered as of low quality. In addition, certain features represent different characteristics of an image or video and can be used as complementary information besides other features for making an estimate of quality. For example, the amount of spatio-temporal information content of a video can be used to characterize the masking effect on various artifacts that may be present in the signal. More examples of visual quality relevant features include local contrast, brightness, colorfulness, and structural activity [96] [97].
2.1. Review

Table 2.2: Characteristic summary of multiple-artifacts-based and features measures-based metrics

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
<th>Processing</th>
<th>Resolution</th>
<th>Test data</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple artifacts</td>
<td>Liang et al. [19]</td>
<td>JPEG2000</td>
<td>768 × 512</td>
<td>LIVE, TID2008</td>
<td>PC = 0.92, SC = 0.94</td>
</tr>
<tr>
<td></td>
<td>Pastrana et al. [80]</td>
<td>MPEG-4</td>
<td>QCIF, CIF</td>
<td>6 SRCs</td>
<td>PC = 0.9, SC = 0.9</td>
</tr>
<tr>
<td></td>
<td>Oelbaum et al. [82]</td>
<td>H.264/AVC</td>
<td>CIF</td>
<td>300 Test videos</td>
<td>PC = 0.82, SC = 0.75</td>
</tr>
<tr>
<td></td>
<td>Choi et al. [83]</td>
<td>JPEG2000, noise</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.91</td>
</tr>
<tr>
<td></td>
<td>Cohen et al. [84]</td>
<td>Noise and blur</td>
<td>256 × 256</td>
<td>75 Test images</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Zhu et al. [86]</td>
<td>JPEG2000</td>
<td>512 × 768</td>
<td>LIVE image</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Narwaria et al. [88]</td>
<td>Multiple</td>
<td>Multiple</td>
<td>(LIVE)</td>
<td>PC = 0.8894</td>
</tr>
<tr>
<td></td>
<td>Liu et al. [89]</td>
<td>JPEG, JPEG2000</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.92</td>
</tr>
<tr>
<td>Natural statistics</td>
<td>Zhou et al. [99]</td>
<td>JPEG2000</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Lu et al. [100]</td>
<td>Multiple</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>Multiple</td>
</tr>
<tr>
<td></td>
<td>Shen et al. [102]</td>
<td>Multiple</td>
<td>512 × 512</td>
<td>LIVE, 260 test images</td>
<td>Multiple</td>
</tr>
<tr>
<td></td>
<td>Moorthy et al. [103]</td>
<td>Multiple</td>
<td>768 × 512</td>
<td>LIVE image,</td>
<td>Multiple</td>
</tr>
<tr>
<td></td>
<td>−</td>
<td>−</td>
<td>512 × 384</td>
<td>TID2008</td>
<td>Multiple</td>
</tr>
<tr>
<td>Pixel-based features</td>
<td>Gastaldo et al. [111]</td>
<td>JPEG</td>
<td>480 × 720</td>
<td>LIVE image</td>
<td>PC = 0.94</td>
</tr>
<tr>
<td></td>
<td>Li et al. [112]</td>
<td>Multiple</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.87, SC = 0.87</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [113]</td>
<td>JPEG2000</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.93, SC = 0.92</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [97]</td>
<td>JPEG, JPEG2000</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.92, SC = 0.92</td>
</tr>
<tr>
<td></td>
<td>Yao et al. [119]</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>Ries et al. [121]</td>
<td>H.264/AVC</td>
<td>SIF</td>
<td>10 SRCs</td>
<td>PC = 0.93</td>
</tr>
<tr>
<td></td>
<td>Ries et al. [122]</td>
<td>H.264/AVC</td>
<td>SIF</td>
<td>10 SRCs</td>
<td>PC = 0.93</td>
</tr>
<tr>
<td>Pixel-based features</td>
<td>Sazzad et al. [114]</td>
<td>JPEG2000</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.93, SC = 0.96</td>
</tr>
<tr>
<td></td>
<td>Jiang et al. [127]</td>
<td>MPEG-2</td>
<td>HD</td>
<td>72 Test images</td>
<td>PC = 0.91</td>
</tr>
<tr>
<td></td>
<td>Keimel et al. [129]</td>
<td>H.264/AVC</td>
<td>HD</td>
<td>7 SRC videos</td>
<td>PC = 0.86, SC = 0.85</td>
</tr>
<tr>
<td></td>
<td>Sazzad et al. [130]</td>
<td>JPEG</td>
<td>640 × 480</td>
<td>490 Test pairs</td>
<td>PC = 0.93</td>
</tr>
</tbody>
</table>

aThe used test database; bperformance denotes the correlation with subjective assessment, unless stated otherwise; SRC, source sequence; PC, Pearson’s correlation coefficient; SC, Spearman’s rank order correlation coefficient.

Moreover, it has been described in [98] that natural images possess a common statistical behavior. This behavior has been termed as NSS, and it has been found to be a useful feature for the description of image quality. There have been numerous applications of NSS including image segmentation, denoising, and texture analysis and synthesis. Although it was concluded in [98] that the major usage of scene statistics would be in the investigation of visual sensory processing, these have recently been proved to be quite useful in the design of no-reference quality methods. It has been found that such common statistical characteristics get distorted by image processing applications like image compression, and a quantitative measure of this distortion can yield the relevant variations in the image quality. Thus, an NSS-driven NR quality assessment method would provide the measure of the unnaturalness introduced into the natural scene statistics under the effect of image distortions. Figure 2.8 shows a basic schematic block diagram of feature-based methods. We have divided the review of such methods into three subsections: (i) Natural scene statistics, (ii) Pixel-based features, and (iii) Pixel-based features and artifacts.
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

Natural scene statistics

It has been claimed in [92] that the distortion introduced in the nonlinear dependencies found in natural images can be quantified for making an estimate of human perceptual quality. Based on that notion, the authors presented an NSS-driven approach for quality assessment of images processed by wavelet-based compression standards like JPEG2000.

Similarly, the NSS-based image quality prediction approach presented in [99] is also limited to be applicable only to JPEG2000. The authors have used a neural network to regress between inputs from NSS-based spectral amplitude fall-off curves in combination with positional similarity measure of wavelet coefficients and the corresponding quality value.

Harnessed by the measures to keep the model attributes unaffected by image content variations, the method proposed in [100] uses a contourlet transform [101] to quantify the degradations incurred on NSS. The authors show that wavelet transform does not completely exhibit the artifacts present in the image and the effect of degradations is visible in all the subbands of the con-
tourlet domain. Hence, the contourlet domain can be more effective in image quality assessment. The proposed method has a clear advantage in precisely predicting the image quality while tested for images degraded by JPEG2000 and JPEG compression and distortions like Gaussian blur, fast fading channel, and white noise. Similarly, a statistical relationship between the characteristics of NSS in images and the corresponding quality values was studied in [102] to engineer a reference-free quality method. In order to provide the quality ranking of the filtered natural images, a histogram of a combination of image transforms, namely, curvelet, wavelet, and cosine transform is computed. The considered distortions include noise, blur, and artifacts introduced by compression using JPEG2000 and JPEG. As the authors pointed out, this is one of the few quality methods which can quantify the perceptual impact of such a broad range of degradation types. The additional advantage of this method is its ability to classify images on the basis of the presence of one or more of these artifacts. The proposed method was tested on a large set of images from the LIVE image database as well as authors’ own test set of images. As a result, a promising level of correlation with subjective quality assessment was obtained.

The distortion identification-based image quality estimation method proposed in [103] offers an NSS-based approach of image quality prediction framework and algorithm. Firstly, the pertinent distortion is identified. Then, NSS features are used to quantify the relevant quality value which is largely independent of the distortion type present in the image. The used feature set describes (1) scale and orientation selective statistics, (2) orientation selective statistics, (3) correlations across scales, (4) spatial correlation, and (5) across orientation statistics. Support vector regression is used to train the model, and the proposed method is proved to be comparable in precision of assessment to full reference methods such as peak signal-to-noise ratio (PSNR) and SSIM. The method was evaluated on images found in TID2008 and LIVE databases. It was found quite closely correlated to subjective assessment of image quality and hence proved itself to be test-set independent.

The idea of the impact of distortions on NSS has been used in [104] for prediction of video quality where each frame of the video is decomposed into a Laplacian pyramid of a number of subbands. Intra-subband statistics including mean, variance, skewness, kurtosis, energy, entropy, and inter-subband statistics, namely, Jensen Shannon divergence, SSIM, and smoothness are computed. A Minkowski pooling scheme is adopted to yield a single value out of the aforementioned statistics. The proposed method is reported to perform better than some FR metrics while tested on the LIVE video quality database.
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

Similar to NSS, a basic model is presented in [105] to develop an NR quality method based on temporal statistics of videos called as natural motion statistics (NMS). The theory of independent component analysis (ICA) has been applied in order to compute NMS. The authors have shown that independent components calculated from the optical flow vectors of a video signal follow the Laplacian distribution. Consequently, it has been observed that the root mean square (RMS) error of the fit between the extracted independent components and Laplacian distribution can be used as an indicator of video quality.

Saad et al. have presented their DCT statistics-based image integrity index in [106]. The central idea is to track the change in particular statistics of an image while it traverses from being original to a distorted one. The proposed framework is mainly DCT based. Owing to the perceptual relevance, some features representing structural information and contrast of an image have been extracted from the DCT values at two levels of spatial sampling. An improved version of this approach is found in [107] where the impact of NSS features for various perceptual degrees of degradation has been added.

In contrast to most of the approaches mentioned before that involve transformation of an image into another domain such as DCT, the NSS-based quality estimator presented in [108] performs in the spatial domain. Locally normalized luminance and its products-based empirical distribution is used to compute quality relevant features for building a spatial NSS model. The performance of the proposed method has been found to be better than FR methods such as PSNR and SSIM. The authors have validated the NR application of this method by employing it in an image denoising system. A similar approach has been adopted in [109] to define latent quality factors that were used to estimate the image quality.

The idea of NSS features-based quality estimator has been used in the case of stereoscopic images as well. In reference [110], 2D- and 3D-based statistical features are extracted from stereopsis to estimate the image quality. A support vector machine model has been trained using these features, and the model has been tested using the LIVE 3D database.

**Pixel-based features**

There are some methods of no-reference quality estimation which rely on certain statistics, mainly spatial features, derived from pixels of an image or video to perform the corresponding perceptual quality evaluation. In [111], the au-
thors present an example where they have used objective features related to energy, entropy, homogeneousness, and contrast from the color correlogram of an image.

These features have been used to train an ANN which serves as a prediction model. Li et al. [112] have also deployed an ANN-based model to devise a quality estimation method using perceptually relevant image features including phase congruency, entropy of the degraded image, and gradient of the degraded image. The importance of phase of an image for its fidelity representation is well known, and the gradient of an image is an implication of changes in the luminance of an image. An ANN model is also used in the image semantic quality method presented in [96] where a variety of quality descriptive features have been used. The authors argue that the overall visual quality can be seen in terms of the usefulness and naturalness of an image. Sharpness and clarity are considered as the representatives of usefulness of an image, whereas brightness and colorfulness represent naturalness. These four representations of usefulness and brightness are further branched into a large set of pixel-based features; edge pixel distribution, contrast, mean brightness, and color dispersion are a few of the used 14 features. The advantage of using higher number of features has been shown by better performance of the predictor.

Compared to the aforementioned methods that rely on the process of training a particular model by using an extracted set of features, the pixel-activity-based method proposed in [113] does not use such methodology. The focus here is on the activity map of an image, essentially controlled by features, namely, monotone-changing, zero-crossing (ZC), and the existence of inactive pixel, which are calculated for non-overlapping image blocks. The concept of ZC has been used to refer to the places in the Laplacian of an image where the value of the Laplacian passes through zero, i.e., the points where the Laplacian changes sign. Such points often occur at edges in an image. The use of ZC as a constituent of an activity map is justified as the method was proposed for JPEG2000 encoded images; and ringing, which can be caused by JPEG2000-based compression, has the potential of generating ZC around contours. Moreover, spatial features consisting of edge information and pixel distortion have been used to predict quality of JPEG2000 encoded images in [114]. Pixel distortion is computed using standard deviation of a central pixel and a measure of difference between central pixel and its closest neighbor pixels. Edge information relies on zero-crossing rate and a histogram measure. Particle swarm optimization has been employed to integrate these features into a single quality index. The authors have presented a similar method in their contribution [115].
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

The notion of quality estimation with regards to structural changes in images as a result of distortions has gained widespread attention. The FR method SSIM [94] is a commonly used representative method of this area. Zhang et al. [97] have put forward a similar approach of quality estimation based on structural changes. However, the nature of the particular distortion should be known beforehand. This method can be used to evaluate degradation caused by the following artifacts but one set at a time: (i) Gaussian blur and noise, (ii) blur and ringing, and (iii) blockiness. In a nutshell, local structural activity is taken in the form of direction spread whereas structural activity weight is computed through a measure of structural strength and zero-crossing activity.

Some feature-based methods make use of the properties of HVS to govern the performance of the method for better correlation with subjective assessment. A 3D multispectral wavelet transform-based method of NR quality estimation for color video has been given in [116]. Various channels of the HVS have been represented by wavelet decomposition of the video. To invoke the impact of the HVS, a perceptual mask of sensitivity with integrated impacts of spatio-temporal contrast and luminance has been applied to all wavelet bands. The final step is to draw a perceptual mask weighted flow tensor between successive frames to define the method. An ANN has been used in [117] with an extreme learning machine (ELM) algorithm for determining the relationship between spatial statistics of an image and its corresponding quality. These statistics are mainly HVS-inspired features, namely, edge amplitude, edge length, background activity, and background luminance of an image. As the proposed method is basically targeted at JPEG encoded images, some of the underlying methodologies which help calculate these features are focused on computation of blockiness. Since DCT coding is used in video coding also, the proposed algorithm can be generalized to be workable for video quality assessment.

In the experiments on determining the visual interest for different objects and locations inside an image, it has been found that HVS perception is not spatially uniform. Instead, there are specific areas called region of interest (ROI), which draw more attention and hence contribute more towards overall quality assessment of an image. Treisman et al. [118] observed that the visual system notices and codes different features in parallel channels before the observer actually recognizes the objects in an image. Features such as color, brightness, and orientation can be pooled together to form a unique entity to be observed. Based on this observation, there exist IQA methods which assess perceptual quality of an image by focusing mainly on those ROIs. One such
method is proposed in [119] where the impact of importance of various ROIs in a video frame has been integrated into a wavelet-based just noticeable difference (JND) profile of visual perception. The proposed method works better than some contemporary methods when it was tested on the VQEG Phase I database.

In order to estimate the impact of packet loss impairments on video quality, a method based on edge strength around macroblock boundaries is proposed in [51]. Edge strength values are processed through a low-pass filter, and a threshold value is applied to compute the edge maps of adjacent rows. Finally, the impact of packet loss is computed through a difference between these edge maps.

In order to quantify the quality of enhanced images, the method given in [120] divides an image into smooth and textured areas. A JND formulation of perception is derived based on the local average brightness and local spatial frequency. The effect of enhancement is monitored through a comparison of local brightness and a JND threshold. The performance of the proposed method is reported to be better than that of conventional average local variance-based methods.

Features-based assessment of the content of an image or video can be used in the estimation of perceptual quality. Ries et al. have shown the relevance of the content class of videos in the process of determination of the visual quality in [121]. The authors classify a given set of videos into five groups based on the content. One of such group, called class here, contains videos which have a small moving ROI with a still scene in the background. Another content class has videos with huge spread of angle of movie capturing device and is called panorama. These content classes are created based on the statistics that are mainly related to motion dynamics of a video. Values of zero motion vector ratio, mean size of motion vector, uniformity of the movement, horizontalness of movement, and greenness are the classification parameters which are used to segregate the set of videos into different content classes. The central idea of the method is to first check the content class of a video and then estimate the visual quality based on bitrate and frame rate. The authors continued working on the same idea in their contribution found in [122] where they have presented a method aimed at the most common content classes of videos for handheld devices. Khan et al. have proposed a content-based method to combine encoding and transmission level parameters to predict video quality in [123]. Based on spatio-temporal features, the videos are first divided into content-based groups.
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

using cluster analysis. Adaptive network-based fuzzy inference system (ANFIS) and a regression model have been used separately to estimate the quality score. As per their results, transmission parameters like packet error rate have more impact on the quality than the compression parameters such as frame rate etc. The underlying techniques of ANFIS model and content clustering have been used in the authors’ other contributions as given in [124] and [125].

**Pixel-based features and artifacts**

Some of the existing no-reference perceptual quality assessment methods are composed of a set of spatial features combined with some measurement of artifacts. A set of spatial artifacts has been combined with some spatial image features to estimate perceptual image quality in [126]. An ANN model was trained with these features for the quality prediction. Working on a similar approach, the method presented in [127] integrates spatial features such as picture entropy (represents the amount of information of a picture) and frequency energy (distribution of frequency energy in images) with artifacts, namely, blur and blockiness. The proposed method seems prominent because of its use of the chrominance information also while most of the contemporary quality measures are based on statistics from the luminance channel only. In this contribution, it has been shown that extraction of these features from ROI further improves the value of correlation with subjective scores. Five features of quality significance have been used to model an ANN-based quality predictor in [128] where the features set constitutes a measure of artifacts such as blocking and ringing and spatial statistics such as zero-crossing rate, edge activity measure, and z-score. Another method built on similar principle is found in [129] where the amount of blurring and blocking has been combined with spatial activity in an image and predictability of an image. A partial least square regression (PLSR) approach has been used to determine the function between these features and the quality value.

The approach given in [130] uses local segmented features related to degradation and dissimilarity for quality estimation of 3D images. In fact, the essential methodology used in [114] for 2D images have been extended to be employed for 3D images in [130]. One of the key means used to check disparity in left and right images of a stereoscopic image is the block-based edge information measure.

The authors in [131] propose a method for the assessment of facial image quality. Eye-detection, sharpness, noise, contrast, and luminance values of a
test image are calculated. A weighted sum of these quantities constitutes the value of the quality method. In view of the discussion presented in [132], relatively more weighting has been given to sharpness and eye-detection as they are more important for determining facial image quality.

In [133], a set of artifacts, namely, blocking, ringing, truncation of the number of bits for image values, and noise is combined with a set of features including contrast and sharpness for designing a video quality prediction method. Each of these parameters is fitted separately in a functional relationship with subjective assessment of quality such that the correlation between the parameter values and subjective scores is maximized. Subsequently, these individual fitting functions are merged together to form a joint relationship with the perceptual quality. The data used for training includes original videos as well as different combinations of sharpness-enhanced and noise-contaminated videos. The trained model is tested on another data set which reveals a promising correlation with subjective scores.

Unlike the aforementioned NR-P-based artifacts or features measures-based methods, the mean square error distortion due to network impairments for a H.264/AVC encoded video is computed in [134]. An estimate of MSE is computed using the pattern of lost macroblocks due to an erroneous transmission of the video. Information about the lost macroblocks is estimated through the traces of the error concealment process. The same methodology has been enhanced in [135] for more general application scenarios such as no assumption is done about a certain error concealment algorithm and it does not require the knowledge of exact slicing structure.

Discussion

From the review of the features measures-based methods, we can make some general observations. The approach of estimating visual quality by quantifying the impact of distortions on natural scene statistics has gained a wide interest to gauge degradations due to different image processing techniques including compression. However, more of such approaches should be tested in the case of videos as well. Moreover, assessment of quality degradation due to network impairments using NSS-based approaches could be useful. The pixel-based and features-based approaches can be seen as composed of techniques that rely on a variety of spatial features including those related to edges, contrast, and some measures of structural information. The performance of these approaches can be enhanced by adapting the computational procedure with regards to
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

the input of HVS preferences. Additionally, including the impact of mostly occurring artifacts such as blurring, blocking, or noise could be an advantage. We observe that most of the pixel domain features-based approaches have been designed for images and it is desirable to generalize the relevant methods for applications in the case of videos. Temporal pooling quality methods such as Minkowski summation or other methods such as adaptive to perceptual distortion [136] can be used for this purpose. Table 2.2 presents a summary of some of the methods discussed in this section. It is evident that most of the methods in this category exhibit very promising performance, with correlation coefficient values equal to or higher than 0.85.

### 2.1.6 Bitstream-based methods

An estimate of the quality of an encoded video can be made by parsing the coded bitstream to deliver readily available features such as encoding parameters and network quality of service (QoS)-related parameters. The methods that adopt the usage of the bitstream data for quality estimation avoid the computational complexity of processing the full video data, as full decoding of the input video is not usually required in the case of bitstream-based methods. Another advantage of this type of method is the use of readily available information from the bitstream that is significant for the quality estimation, for example, the motion vectors, coding modes, and quantization parameter values. However, these methods are inherently coding standard specific as different encoders have different formats of bitstream. There is a range of quality relevant features that can be extracted by partial decoding or primary analysis of the bitstream data. The performance of such methods significantly depends upon the level of access to the bitstream [137].

A block diagram of general framework in bitstream-based methods is given in Figure 2.9. We have divided the discussion of these methods into three categories based on the level of information used for processing, in accordance with the standardized models recommended by telecommunication standardization sector of International Telecommunication Union (ITU-T), as discussed in [138,139]. This includes parametric models (parametric planning model and parametric packet-layer model) and bitstream layer model. In the former type, extrinsic features of a video that are of parametric nature such as bitrate, frame rate, and packet loss rate are used. Bitstream layer models have detailed access to the payload and intrinsic features related to a video such as coding modes, quantization parameter, and DCT coefficients. The standardization of these
models includes the methods designed for estimation of audio quality as well, but our discussion is limited to video quality only.

![Figure 2.9: A basic scheme used for video quality assessment methods based on bitstream-based features.](image)

**Parametric planning model**

The parametric planning models have rather low complexity as they do not access the bitstream and utilize bitrate, codec type, and packet loss rate for making a crude estimation of video quality. The work item related to this category in ITU-T is known as Opinion model for video-telephony applications, G.1070 [140]. ITU-T Recommendation G.1070 proposes a method for the assessment of videophone quality, based on speech and video parameters, that can be used by the network performance planners for ensuring the given level of end-to-end quality of the service. A quality prediction model for MPEG-2 and H.264/AVC encoded videos for IPTV is presented in [141]. The model takes some parameters related to encoding information, packet information and client information to assess the overall quality. In reference [142], a parametric model is proposed that is based on a simple method of estimating MSE that occurs due to a given pattern of packet loss. The authors derived a relationship between average motion vector length and MSE and this relation gives a fair estimate of the actual MSE.
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

**Parametric packet-layer model**

The packet layer models have access to the packet header of the bitstream and can extract a limited set of parameters including bitrate on sequence or frame level, frame rate and type, and packet loss rate. Parametric packet-layer models are also known as QoS-based methods. The work item related to this category in ITU-T is known as non-intrusive parametric model for the assessment of performance of multimedia streaming (PNAMS) [143]. The visual quality estimation method proposed in [144] presents an approach where it is not required to decode the video at any level, suitable for situations where the encoded video is encrypted. Given the observation that error concealment is more effective when there is less motion in the video, an estimation of the motion dynamics of a particular video is required to assess the effectiveness of an error concealment strategy. In this method, the ratio of the average of the B (bi-predictive coded) frame data size to the average of the size of all frames is compared with a predetermined threshold to adjust the value of the video quality score. The results obtained from the effectiveness of error concealment are refined by adjusting the values in accordance with the importance of the region in which the error has occurred.

The models in [141, 145] are designed for H.264/AVC coded SD and HD videos where a support vector machine (SVM) classifier has been used to assess the video quality based on the visibility of packet loss. By the same research group, the packet layer model presented in [146] uses video resolution, bitrate, packet loss rate, and some information of the codec settings to design a quality estimator for H.264/AVC- and MPEG-2-based encoded videos. An improvement on such statistical parameters-based models is found in [147] where temporal and spatial characteristics of a video are estimated from the packet header to build a content-adaptive model for quality assessment. The no-reference method presented in [148] is based on a nonlinear relationship between an objective quality metric and the quality-related parameters. To make it computationally simple, the authors have used only two parameters, namely, packet loss rate and the value of the interval between intra-frames of a video.

In [149], the authors have presented preliminary results of their investigation into streamlining the impacts of different conditions of packet loss over visible degradation to classify packet loss as visible or invisible. The parameters used in the decision making are extracted from the encoded bitstream. This model was tested for SD resolution H.264/AVC coded videos. If 25% or less subjects perceived an artifact, such a packet loss event was classified as invisi-
ble. If 75% or more subjects perceived an artifact, the corresponding packet loss event was classified as visible. In this case, the artifacts perceived by subjects between 25% and 75% were not accounted for at all. This issue was addressed in the authors’ later contribution [150] where all artifacts perceived by less than 75% subjects were classified as invisible. Moreover, they extended the model by including more quality-relevant parameters and generalized it by testing it on HD videos. The authors applied the same model for High Efficiency Video Coding (HEVC) encoded videos to examine its cross-standard performance, as reported in [151]. It was observed that the artifact visibility slightly increases while changing from H.264/AVC to HEVC-based video coding.

**Bitstream layer model**

In the bitstream-based methods, bitstream layer models have access to most of the data that can be used for the video quality estimation. The work item parametric non-intrusive bitstream assessment of video media streaming quality (P.NBAMS) [152] in its mode 1 (Parsing mode) is related to the bitstream layer models. In this mode, it is allowed to do any kind of analysis of the bitstream except the usage of the pixel data. The input information includes parameters extracted from the packet header and payload. Besides the parameters included in the parametric models, this model uses QP, DCT coefficients of the coded video, and pixel information. This makes the model comparatively more complex but it generally offers better performance. A low-complexity solution of video quality prediction based on bitstream extracted parameters is found in [153]. The features used are mainly related to the encoding parameters and are taken on sequence level. Low complexity has been achieved by using a simple multilinear regression system for building the relationship between the parameters and quality values. An improvement of this approach is presented in [154] where the required number of parameters has been reduced for computational efficiency and the prediction accuracy has been improved by the virtue of the usage of an ANN. A further improvement is found in [155] where a larger features set is used and the prediction of subjective MOS is also performed. A set of 48 bitstream parameters related to slice coding type, coding modes, various statistics of motion vectors, and QP value was used in [156] to predict the quality of high-definition television (HDTV) video encoded by H.264/AVC. PLSR was used as tool for learning by regression between the feature set and subjective assessment. This method outperformed the authors’ earlier contribution [129] and some contemporary objective methods of visual quality assessment.
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

H.264/AVC employs an in-loop filter to suppress blocking, and this filter has a specific parameter called boundary strength (BS) assigned to transform blocks. Statistics of BS combined with QP and average bitrate has been used in [157] to predict quality of H.264/AVC encoded videos. The proposed method formulates a linear combination of these parameters and a linear regression was conducted to determine its relationship with the subjective assessment scores. A motion-based visual quality estimation method was proposed in [158] for H.264/AVC encoded videos. In this method, some statistical features related to motion vectors along with bitrate and frame rate are calculated. PCA is used to identify the parameters most influential in the estimation of video quality value. Finally, the selected set of features is fed to an equation of quality computation. The inclusion of motion features into the reference-free quality assessment is justified by the fact that the reduction in visual quality is less for a certain level of compression when the motion is low, for example, the case of videos with static scenes.

A PSNR estimator for H.264/AVC encoded video is presented in [159] where bitrate, QP value, and coding mode are used as the features for quality prediction. The method given in [160] uses QP and block coding mode parameters for quality prediction of H.264/AVC encoded videos.

Based on an opinion model from ITU-T [140], an automatic QoE monitoring method is proposed in [161]. It depends on the network level information derived from packet loss pattern and loss rank of a frame in a group of pictures (GOP) and a measure of motion vectors to represent motion activity to train an ANN model against subjective scores of expert viewers.

In [162], the authors proposed a framework for quality estimation where a QoS parameter, packet loss rate, is combined with spatial and temporal complexities of a video. Usually, a complete decoding of the video is required to estimate its spatial and temporal complexity as these complexity values are generally obtained by an average measure of the pixel variance of codeblocks in a frame. However, the authors have proposed a method of estimating spatial and temporal complexity from the bitstream information only. Specifically, they have developed a rate-distortion model for QP value and bitrate which helps in estimating the complexity measure. Combining this complexity estimate with effects of packet loss delivers a measure of frame quality. Temporal domain quality degradation is computed through occurrences of frame freeze or frame loss. An overall estimate of the video quality is made by a pooling scheme which integrates the spatial and temporal quality indicators. The au-
thors have argued that the suggested method can be used for real-time video services due to its fair accuracy and efficiency in computational cost.

In [163], the impact of compression on quality estimated through MSE prediction using DCT coefficients data [164] is combined with (i) a packet loss model similar to the one presented in ITU-T Recommendation G. 1070 [140], (ii) a frame type-dependent packet loss model, and (iii) a frame type- and error pattern-dependent model separately. It was concluded from the obtained results that a combination of [164] and (iii) offers the best prediction of visual quality of these three combinations.

Bitstream layer methods can also utilize the DCT coefficients data of the encoded image or video, as it can be obtained by partial decoding [138]. There are several such methods which make a quality estimate based on the statistics of the DCT coefficient values. Eden [165] has proposed an algorithm for estimation of PSNR using the assumption that the probability distribution function (pdf) of DCT coefficients follows Laplacian distribution for H.264/AVC encoded videos. A modified Laplacian model for estimation of DCT coefficients distribution has been presented in [166] for JPEG images. The authors proposed to use maximum likelihood with linear prediction estimates to compute the parameter \( \lambda \) (lambda) of the Laplacian pdf, where \( \lambda \) is a parameter of the distribution. Investigation of the correlation between distribution parameters at adjoining frequencies and integration of the prediction results using maximum-likelihood parameters are the key components of this method. They have also used Watson's model [167] for perceptual weighting of local error estimates in an image. The method given in [166] has been upgraded to be workable for videos in [168]. Here, the video quality predictor has a local error assessment unit, besides having statistics from motion vectors. These values are passed to a perceptual spatio-temporal model that incorporates the HVS sensitivity to produce the visual quality score. Two more methods based on the similar approach from these authors are PSNR estimation for H.264/AVC encoded videos [169] and PSNR estimation for MPEG-2 encoded videos as given in [170].

Contrary to the assumption of Laplacian distribution to model DCT coefficients, it has been argued in [171] that a Cauchy distribution better suits the H.264/AVC encoded data in the process of quality estimation. The proposed approach has been found to be better than the Laplacian distribution [165] in terms of bias between the actual and estimated values of PSNR.
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

The authors in [172] have used DCT basis functions to evaluate kurtosis measure of images for quality assessment. Three different kinds of kurtosis measures have been made, namely, 1D kurtosis based on frequency band, basis function-based 1D kurtosis, and 2D kurtosis. However, the proposed scheme is meant only for images degraded by blur and it has been tested on LIVE [14] data set JPEG2000 encoded images.

Nishikawa et al. presented a PSNR estimation method of JPEG2000 coded videos in [173] which is actually a no-reference version of their earlier article that needed reference information [174]. The method estimates the PSNR by using wavelet coefficients from the neighboring frames of the frame which has lost some compressed codeblocks. It is assumed that the effect of packet loss upon codeblocks is possible to compute at the receiver end, given that only packet loss occurs and no bit errors exist.

**Discussion**

Bitstream-based methods of VQA have recently received a significant attention for their computational simplicity and applications in the online quality monitoring. Potentially, the main advantage of these methods is the variety in choice of the features which can be used for quality estimation that in turn means the privilege of adapting to the desired level of complexity. As compared to pixel-based processing, the bitstream-based methods have special advantage of having access to readily available information such as bitrate, frame rate, QP, motion vectors, and various types of information regarding the impacts of network impairments. However, these methods are coding scheme specific that makes them less generally applicable. In the case of parametric planning models, the performance of quality estimation remains limited due to the constraints of the information that can be obtained from the allowed level of access to the bitstream. Packet layer models have better performance with popular application in intermediate nodes of a network as they do not need complex processing and decryption of the data. Bitstream layer models are superior in the performance and the complexity can be flexible depending upon the desired level of accuracy. For possible future works in this area, some comparative performance reports of various models, such as the ones presented in [139,175] would be useful to further accelerate the research in designing better bitstream-based VQA approaches. As we notice in the summary of bitstream-based methods in Table 2.3, the research community has mostly embraced H.264/AVC-based coding for the design of such methods. It would be advantageous to develop
such methods for other popular coding standards as well. Moreover, analysis of the features relevant for quality estimation for the recently approved ITU-T standard of video coding, namely, H.265/HEVC [176] would be useful. For example, in [177], it has been shown that the existing methods of MSE estimation are not feasible for HEVC as it has significantly different coding structure as compared to the previous standards.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
<th>Processing</th>
<th>Resolution</th>
<th>Test data$^a$</th>
<th>Performance$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bitstream-based</td>
<td>Saad et al. [106]</td>
<td>Multiple</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>SC = 0.8</td>
</tr>
<tr>
<td></td>
<td>Saad et al. [107]</td>
<td>Multiple</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.93, SC = 0.93</td>
</tr>
<tr>
<td></td>
<td>Yamada et al. [144]</td>
<td>H.264/AVC</td>
<td>1,440 × 1,080</td>
<td>-</td>
<td>PC = 0.85</td>
</tr>
<tr>
<td></td>
<td>Rossholm et al. [153]</td>
<td>H.264/AVC</td>
<td>QCIF</td>
<td>288 Test videos</td>
<td>PC (PEVQ) = 0.95</td>
</tr>
<tr>
<td></td>
<td>Shahid et al. [154]</td>
<td>H.264/AVC</td>
<td>QCIF</td>
<td>288 Test videos</td>
<td>PC (PEVQ) = 0.98</td>
</tr>
<tr>
<td></td>
<td>Shahid et al. [155]</td>
<td>H.264/AVC</td>
<td>QCIF, CIF</td>
<td>120 Test videos</td>
<td>PC = 0.98</td>
</tr>
<tr>
<td></td>
<td>Keimel et al. [156]</td>
<td>H.264/AVC</td>
<td>HD</td>
<td>-</td>
<td>PC = 0.93</td>
</tr>
<tr>
<td></td>
<td>Lee [157]</td>
<td>H.264/AVC</td>
<td>QCIF</td>
<td>13 SRCs</td>
<td>PC = 0.9</td>
</tr>
<tr>
<td></td>
<td>Ries et al. [158]</td>
<td>H.264/AVC</td>
<td>QCIF, SIF, CIF</td>
<td>-</td>
<td>PC = 0.80</td>
</tr>
<tr>
<td></td>
<td>Yang et al. [162]</td>
<td>MPEG-4</td>
<td>QCIF</td>
<td>-</td>
<td>PC = 0.93, SC = 0.93</td>
</tr>
<tr>
<td></td>
<td>Eden [165]</td>
<td>H.264/AVC</td>
<td>HD</td>
<td>5 SRCs</td>
<td>PC (PSNR) = 0.99</td>
</tr>
<tr>
<td></td>
<td>Brandão et al. [166]</td>
<td>JPEG</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.97, SC = 0.97</td>
</tr>
<tr>
<td></td>
<td>Brandão et al. [168]</td>
<td>H.264</td>
<td>-</td>
<td>50 Test videos</td>
<td>PC = 0.93, SC = 0.95</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [172]</td>
<td>JPEG2000</td>
<td>768 × 512</td>
<td>LIVE image</td>
<td>PC = 0.9</td>
</tr>
<tr>
<td></td>
<td>Ichigaya et al. [191]</td>
<td>MPEG-2</td>
<td>SD, HD</td>
<td>26 Test videos</td>
<td>PC (PSNR) = 0.95</td>
</tr>
<tr>
<td>Hybrid of NR-P and NR-B</td>
<td>Farias et al. [181]</td>
<td>H.264/AVC</td>
<td>CIF</td>
<td>5 SRCs</td>
<td>Regression (PSNR) = 0.79</td>
</tr>
<tr>
<td></td>
<td>Shanableh [182]</td>
<td>MPEG-2</td>
<td>CIF</td>
<td>-</td>
<td>PC = 0.93</td>
</tr>
<tr>
<td></td>
<td>Shanableh [183]</td>
<td>MPEG-2</td>
<td>CIF</td>
<td>-</td>
<td>NRMSE = 0.046</td>
</tr>
<tr>
<td></td>
<td>Davis et al. [184]</td>
<td>H.264/AVC</td>
<td>-</td>
<td>18 Test videos</td>
<td>PC = 0.91</td>
</tr>
<tr>
<td></td>
<td>Yamada et al. [186]</td>
<td>H.264/AVC</td>
<td>SD</td>
<td>-</td>
<td>PC = 0.95</td>
</tr>
<tr>
<td></td>
<td>Sugimoto et al. [189]</td>
<td>H.264/AVC</td>
<td>HD</td>
<td>14 SRCs</td>
<td>PC = 0.91</td>
</tr>
</tbody>
</table>

$^a$The used test database; $^b$performance denotes the correlation with subjective assessment, unless stated otherwise. SRC, source sequence; NRMSE, normalized root mean square error; PC, Pearson’s correlation coefficient; SC, Spearman’s rank order correlation coefficient.

### 2.1.7 Hybrid of NR-P and NR-B methods

There are no-reference visual quality estimation methods which combine features from the coded bitstream and some statistics from the decoded media. This type of methods inherits the simplicity of computation from the bitstream-based approaches, and further accuracy in quality estimation is achieved by adding input from the pixel-based approaches. Therefore, such methods can avoid some of the difficulties involved in the pixel and bitstream-based methods [178]. One such example is the fusion of artifacts like blocking or blurring with parameters derived from motion vectors to build up a quality estimation method. The work item PNBAMS [152] in its mode 2 (full decoding mode) is related to the hybrid models where the information from the coded bitstream as well as reconstructed video can be used. Figure 2.10 gives an overview of...
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

the methodology used in this type of methods. Essentially, the choice of the features for extraction from bitstream or pixel domain depends on the design requirements of a method, the availability of a certain type of data for quality estimation, and the encoding scheme. The discussion on this class of methods is divided into two categories, namely, pixel-based and bitstream-based features or artifacts, and statistics of transform coefficients.

![Diagram](image)

**Figure 2.10:** A basic scheme for quality assessment methods based on hybrid of NR-P- and NR-B-based approaches.

**Pixel-based and bitstream-based features or artifacts**

Video quality-related features and measures of artifacts can be computed both from the pixel and bitstream data and can be pooled for an overall quality estimate. One such method which focuses on quantifying the perceptual quality of H.264/AVC encoded videos degraded by loss of packets in the IP networks is presented in [179]. The error incurred due to packet loss becomes propagative due to the two types of coding predictions involved in H.264/AVC encoders, namely, intra-prediction (spatial) and inter-prediction (temporal) at the encoder end. Even more errors can be introduced while the decoder tries to conceal for the prediction residuals and/or motion vectors lost due to missing packets in the IP bitstream. For simulating the packet loss conditions, a packet loss rate in the range [0.1, 20]% with error patterns generated using a two-state Gilbert model set for average burst length of three packets was used. Quantitatively, the measures involved in the modeling of the proposed method encompass the impact of errors due to concealment, errors propagated due to loss of reference MBs, and the channel-induced degradation due to H.264/AVC-specific coding
2.1. Review

techniques. The calculations of these distortions are done on the macroblock level, and the resulting values are summed up to frame and sequence levels. It has been observed that the proposed method yields results which bear good correlation with SSIM [94]. Another method was presented by the same authors in their earlier published contribution [180] where the effects of loss of motion vector information and prediction residuals were incorporated for quality estimation. A method in which transmission and compression artifacts are integrated for VQA is presented in [181]. The constituents of the method are estimations of blockiness, blurring, and packet loss ratio.

Two MPEG-4 encoded video quality prediction methods based on several MB level statistics, derived from bitstream and reconstructed videos, are reported in [182] for PSNR and in [183] for SSIM. A plethora of bitstream-based and pixel-based features at macroblock level have been used in these two methods. One of the distinctive aspects of these two contributions is the usage of different models for system identification between the parameters and the corresponding quality index. In the method targeted for PSNR estimation, spectral regression and reduced model polynomial network have been employed. A multipass prediction system based on stepwise regression has been used in the estimation of SSIM method. The statistical features in both of the methods constitute mainly the coding information of an MB, some relative measures of motion vector of neighboring MBs, and some numerical values related to the texture of an MB.

Average QP values were combined with pixel difference contrast measure values to offer a visual quality method in [184]. The authors have shown that the method outperforms PSNR for a wide range of bitrates of H.264/AVC encoded videos. Similarly, two parametric models have been combined in [185] to design a hybrid model of perceptual quality for H.264/AVC encoded videos. This method uses average value of QP and an average measure of contrast from the decoded video, besides having input from noise masking property of the video content.

A hybrid of bitstream-based and pixel domain quality estimator is proposed in [186]. It has been argued that a video quality estimation merely based on the amount of impaired macroblocks could be erroneous as, in modern video decoders, some error concealment methods are applied to cure the impaired macroblocks and this concealment is not accounted for in such estimations. As the error concealment may not always be effective, the proposed method uses motion intensity and luminance discontinuity measures to esti-
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

mate the number of impaired macroblocks for which error concealment remains ineffective. In essence, the visual quality, in terms of MSE, is estimated directly based on the macroblocks for which the error concealment could not perform well. The same authors have generalized this approach for three methods of error concealment and a different value of packet length in [187].

In order to estimate the impact of transmission errors on the quality of H.264/AVC encoded videos, a saliency map-based method is proposed in [188]. Color, contrast, and luminance information has been used to compute spatial saliency map, while motion vector information, readily available in coded bitstream of a video, has been exploited for the computation of temporal saliency maps. A squared sum of spatial and temporal saliency maps has been used to pool them together for computing the overall spatio-temporal map. Accordingly, this map is used for weighting of an error map for each video frame to calculate the value of the proposed model.

Another hybrid method of perceptual quality measurement, which is based on information from the bitstream and spatio-temporal image features, is presented in [189]. The weighted Minkowski method is employed to integrate the average quantization scale with their proposed measures of flickering and blocking for H.264/AVC encoded videos.

A framework for a hybrid method for videos transmitted over long term evolution (LTE) networks is proposed in [190]. It suggests to include parameters from packet layer (packet loss rate, packet size), bitstream layer (frame error, frame duration), and media layer (blurring, blocking) for estimation of the quality. However, a suitable pooling scheme to integrate these parameters into a quality indication value remains as a future work.

**Statistics of transform coefficients**

In some cases, the transform coefficients can be obtained through partial decoding of the coded bitstream data and features from bitstream as well as pixel domain can be combined for the quality estimation. One such example is found in [191] where an estimate of PSNR has been computed for MPEG-2 coded videos using DCT coefficients. This is actually an improved version of the authors' earlier contribution [192] in which they modeled the distribution of DCT coefficients as a Laplacian pdf to calculate PSNR of the video frames one-by-one for all types, i.e., I, P, and B frames. However, it lacks in accuracy of assessment for B frames. Therefore, the authors conjectured that this happens as a result
of fall in the amount of DCT coefficients information which is available for B frames due to processes of rate control and motion compensation. Henceforth, a hybrid approach to resolve this issue has been found in [191] where picture energy has also been used in addition to DCT coefficients. There is a significant improvement of correlation with estimated and actual PSNR, in the case when the proposed method was tested on SDTV and HDTV sequences.

Discussion

The hybrid methods use not only pixel-based information but also bitstream-based information, which in turn makes the hybrid framework having a potential of being the most accurate quality estimator as compared to the other approaches [193]. Thus, the importance of careful combination of the features from pixel and bitstream domains is evident. Further studies are needed to investigate the interaction among various types of artifacts due to compression and transmission and the joint impact towards the video quality.

Various approaches exist on how to combine the impact of various sources of degradation into one representative value for all the artifacts under consideration. In the recommendation ITU-T P.1202.1 that presents a complementary algorithm of NR quality assessment for the recommendation P.NBAMS [152], four types of degradations are sorted with respect to their impact on quality. The values of the two most significant artifact types are pooled together through a linear combination. A higher weighting is applied to the artifact value that is found to be the most significant out of the four types. As different artifact types can exist in different range of values, it is important that all of them are aligned to the same scale before the sorting is applied. Besides linear combination, some contributions [189] adopt the Minkowski metric [3] for pooling the values of different artifacts into a single quantity.

With regards to the preference on which a pooling strategy should be used, it may depend on many factors including relative severity of different artifacts, spatio-temporal characteristics of the contents, and the presence of masking effects. Linear combination is more valid if the constituents can be related to the quality value through a linear regression. While combining different artifacts through a linear relation, different artifacts can be given different significance. For example, more weight is given to the impact of bitstream layer features than to media-layer features in the hybrid model given in [194]. On the other hand, the Minkowski metric of summation has its roots in additive properties
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

of low-level vision. Therefore, it is required to find the suitable value of its exponent through measurements.

Most of the aforementioned hybrid methods make an assessment of quality in terms of MSE but this measure of quality is known to be rather inaccurate in representing the perceptual quality [195]. This fact necessitates the desire of enhancing such methods for better correlation with subjective quality assessment. As can be seen from the summary of hybrid methods in Table 2.3, the main focus of the development of hybrid methods has been on videos.

2.2 Conclusions

Motivated by the growing interest in NR methods of quality assessment for images and videos, we have presented a classification and review of recent approaches proposed in this research area. The available contributions have been classified into different categories based on the methodologies used in the design. Recognized classifications and standardizations in the area have been extrapolated to introduce our approach of classification. The new classification enabled us to present a review of a large amount of recently published work. On the highest tier, three categories have been identified to group the existing NR methods. The NR methods that employ pixel domain approaches for quality assessment are called NR-P-based methods, and the methods that employ encoded bitstream and parametric information of the media signal are called NR-B-based methods. The third category is called hybrid methods which are designed by a composite of NR-P- and NR-B-based methods. A further subcategorization has been presented to organize the discussion of the review.

It is observed that the majority of the publications introduce methods that are processed in the pixel domain. This trend can be attributed to the rich heritage of work in the image processing area. In most cases, pixel-based methods require more processing power than bitstream-based methods. NR quality estimation is a widely adopted application in the area of online quality monitoring. It is thus required to employ computationally less complex methods. This fact necessitates to focus towards designing bitstream-based or hybrid methods. The distortions present in a network can introduce a variety of temporal perturbations in a video transmitted through it. Such perturbations have to be monitored by service providers to ensure a given threshold of visual quality at the end users' premises. This can be performed using NR metrics which estimate the impact of degradation in the temporal domain. Unfortunately, most
of the existing methods are designed to account for a single or a limited set of degradations. Therefore, it is not easy to make an estimate of the overall loss of visual quality. Hence, methods which can make a complete assessment of the quality are desirable. Similarly, attention can be drawn towards designing less complex methods which are not application specific, such as the methods that are not limited to a particular coding standard.

In the context of the reviewed methods, it is interesting to compare the approaches adopted in case of IQA and VQA methods. In case of IQA, the main focus has been on addressing the most occurring spatial artifacts such as blocking, blurring, and ringing as a result of popular compression techniques such as JPEG and JPEG2000. Besides many methods that are specifically designed for particular artifacts and hence have limited application, it is exciting to see many methods that are not restricted to a specific artifact type and have a wider application area. In such methods of global application, the mostly adopted approach is based on computing the impact of distortions on natural scene statistics of natural images. This also suggests that such approaches may not be applied to artificial images (such as animations and cartoons). This issue can be considered as a challenge for future work in IQA. More focus has been seen on the development of bitstream-based approaches in the case of VQA methods. This is of advantage in the sense that bitstream-based approaches have relatively low computational complexity. However, they face the drawbacks of being coding scheme specific and sometimes less accurate. We believe that the development of more robust approaches based on hybrids of NR-P and NR-B methods may be beneficial to meet these challenges associated with the NR VQA area.

We observe that many of the existing contributions in NR IQA and VQA have reported the results of the proposed methods therein by doing the relevant performance tests on the publicly available test databases. This is useful for independent benchmarking and performance comparison tests of these methods by other researchers. Therefore, more variety in the content and resolution of the media available through public test databases would be of great value. On the other hand, one general drawback of many existing methods of NR quality assessment lies in the limited use of the test data, as the data used for the designing or training of a metric is often also used for its performance verification. This drawback actually does not allow to draw meaningful results from such studies. Also, it has been observed that most of the existing methods for video quality assessment are limited to one encoder implementation or rather one particular setting of an encoder. Hence, cross-encoder design of VQA met-
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

metrics would be a useful contribution. Moreover, we enlist the following trends and directions towards future research in the area of NR quality assessment:

- The trend of contribution in the NR quality estimation has been settling towards finding approaches of lesser complexity as shown by the growing interest in the bitstream-based methods. However, bitstream-based methods face the challenge of being limited for a specific codec. Given the fact that such methods have been shown to have promising performance by having reasonable values of correlation with subjective quality assessment, it would be advantageous to generalize the methodologies of these methods for diverse coding schemes.

- The performance of the bitstream-based methods has been found to be largely content dependent, as the spatio-temporal perceptual complexities vary with varying content, and, in turn, the nature of the features used for quality estimation also changes. However, in the case of pixel-based methods, it is relatively easier to differentiate the content characteristics. Thus, it is required for bitstream-based models to be trained on a sufficiently high variety in content, enabling them to be used in practice. Future inventions can focus on the development of methods that can be applied in more general scenarios with the desired amount of variety in content.

- The existing NR methods are usually designed and tested for cases where the quality difference is beyond the threshold of perceiving of an artifact, i.e., rather clearly visible. However, attention needs to be paid to scenarios where the test stimuli may already be of high quality. Future developments should therefore envisage the degradations that are considered in the category of subthreshold artifacts. The need of such methods becomes even more important with regards to the newly approved HEVC standard that supports ultra-high definition video resolutions of 4K and beyond.

- It has been observed that emphasis is being put towards making the quality estimation more in line with the perceived quality by HVS. In the future, the NR quality assessment methods should continue to adapt for HVS parameters and further advancements in the understanding of HVS, such as attention-driven foveated quality assessment models [196] should be taken into consideration.

- A robust combination of audio-visual quality estimators can be devised for designing scenario-dependent models. For example, in quality mon-
2.2. Conclusions

itoring of sports video, more emphasis can be put on the visual component than audio as the viewers might be interested more in video. For example, a video of a football match would draw more focus towards visual scene than audio, as compared to news or head and shoulder scenario. Moreover, audio-visual quality estimation is challenging due to the complex interplay of HVS preferences. In terms of the mutual impact of audio-visual subjective quality, some studies report an average cross-modal interaction level of 0.5 MOS [197] to above 2 MOS points [198] on a scale of 1 to 5 quality rating.

- Given the presented comprehensive literature review, it has been observed that developments of NR methods that consider visual attention are rather limited, especially, in the case of videos. As noted in [199], visual attention models can be integrated into existing NR methods to make them more robust. Generally, the advantage of including visual attention-based modeling appears to be larger for methods designed for video quality assessment than for image quality assessment methods. Visual attention becomes more significant in scenarios of audio-visual stimuli, as it is required to account for the cues from visual channels as well as auditory channels.

- To make the quality estimation closer to the subjective assessment, intelligent methods are needed that consider the content preference and expectations of humans in a given scenario. For example, the subjective quality assessment results mentioned in [200] indicate that desirable content is rated significantly higher than the undesirable and neutral contents.

- The task of finding the optimal trade-off between temporal and spatial resolution, and the level of quantization for its impacts on the perceptual quality in different scenarios of application, is challenging [201]. This issue should be taken into consideration for the future development of NR methods.

- In order to combine independent and isolated approaches for the development of hybrid NR VQA methods, a five-point agenda has been identified by joint effort group (JEG) at VQEG [202]. We believe that such collaborative works will be instrumental in paving the ways of NR VQA towards a measurable evolution.

We believe that our contribution in this article can be utilized and extended in various ways. One can use this review as a systematic literature review to
2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**

perform comparisons on a class of NR methods using the same image or video test database to highlight the state-of-the-art. Furthermore, this review can be very useful for the beginner researchers in this area to achieve a concise yet comprehensive overview of the field. This way, we expect this contribution to be instrumental for future research and development in the area of NR visual quality assessment. Moreover, a possible future work is to survey the contributions for audio-visual quality assessment based on NR paradigm, similar to [203] that deals with FR methods of audio-visual quality assessment.

### 2.3 Bibliography


2.3. Bibliography


2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches


[31] NG Sadaka, LJ Karam, R Ferzli, GP Abousleman, A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling,
2.3. Bibliography


2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches


[44] Z Hua, Z Wei, C Yaowu, A no-reference perceptual blur metric by using OLS-RBF network, in Pacific-Asia Workshop on Computational Intelligence and Industrial Application, (Wuhan, China 19-20 December, 2008) pp. 1007–1011


2.3. Bibliography


2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**


2.3. Bibliography


[86] X Zhu, P Milanfar, A no-reference sharpness metric sensitive to blur and noise, in International Workshop on Quality of Multimedia Experience, (San Diego, CA, USA 29-31 July, 2009) pp. 64–69

2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches


2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches


2.3. Bibliography


[122] M Ries, O Nemethova, M Rupp, Performance evaluation of mobile video quality estimators, in European Signal Processing Conference, (Poznan, Poland 3-7 September, 2007)


2. **No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches**


2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches


2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

[170] T Brandão, PQ Maria, Blind PSNR estimation of video sequences using quantized DCT coefficient data, in Picture Coding Symposium, (Lausanne, Switzerland 12-13 November, 2007)


108
2.3. Bibliography


[190] S Zhao, H Jiang, Q Cai, S Sherif, A Tarraf, Hybrid framework for no-reference video quality indication over LTE networks, in *23rd Wireless*
2. No-Reference Image and Video Quality Assessment: A Classification and Review of Recent Approaches

*and Optical Communication Conference (WOCC), (Newark, NJ, USA 9-10 May, 2014) pp. 1–5*


2.3. Bibliography


Part II
Part II

On Objective Methods of No-Reference and Reduced-Reference Video Quality Assessment
Three

A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

This chapter has been published as:

A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

Muhammad Shahid, Andreas Rossholm, and Benny Lövström

Abstract

There is a growing need for robust methods for reference free perceptual quality measurements due to the increasing use of video in hand-held multimedia devices. These methods are supposed to consider pertinent artifacts introduced by the compression algorithm selected for source coding. This paper proposes a model that uses readily available encoder parameters as input to an artificial neural network to predict objective quality metrics for compressed video without using any reference and without need for decoding. The results verify its robustness for prediction of objective quality metrics in general and for PEVQ and PSNR in particular. The paper also focuses on reducing the complexity of the neural network.

3.1 Introduction

There has been a huge growth over the last years in multimedia applications for portable devices like mobile phones. A variety of methods for lossy compression for videos has been developed to manage bandwidth and memory usage, introducing specific artifacts impairing the visual quality as perceived by the end user. Therefore, video quality assessment has become important for many stake-holders like the mobile phone industry, network operators and video chatting application developers. The degradation of the video content is measured through various metrics of quality indicators in order to quantify the introduced artifacts, a way forward to improve the visual quality or to compare the competing methods and devices. Traditionally, objective metrics like signal-to-noise ratio (SNR), mean-square-error (MSE) and peak signal-to-noise ratio (PSNR) has been used to measure such distortions. However, end-user perception of the visual quality may not necessarily fall in line with the results of these measures [1]. Subjective testing for the quality evaluation requires human assessors and can provide the actual quality estimation but this method
3. A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

is quite expensive and not applicable in many situations. In order to achieve objective metrics closer to the human perception, Video Quality Experts Group (VQEG) started performing a comprehensive standardization of quality metrics and reported their results in [2] and [3]. Furthermore, the development of quality metrics like SSIM [4], its video adapted version VSSIM [5], VQM [6] and Opticom’s PEVQ included in ITU-T Rec. J.247 [7] was witnessed. The aforementioned metrics calculate the quality measure by comparing the original frame with the distorted version of it, a full reference (FR) approach. In real-time applications like video streaming and video chat, access to the original frame is unlikely. Thus, the need of quality assessors which can work in the absence of any reference frame is a natural demand. Some of the recently introduced no-reference quality assessment methods include single feature based prediction [8] or an ensemble of objective features based prediction [9] [10] and [11] enlists many no-reference quality assessment metrics. These techniques involve appropriate processing of the received frame which makes them less usable for real-time applications. Rossholm et. al. [12] proposed a method of predicting video quality using coded bitstream information without decoding the video which makes the method useful for mobiles with limited processing power. The method uses multi-linear regression to map the encoder parameters as inputs for targets of video quality metrics. However, some of the used parameters exhibit non-linear relationship with the quality metrics. Artificial neural networks (ANN) are well known for their ability of handling non-linear problems in general and have been successfully used in the area of video quality assessment [8] [9] [10]. In this paper, we propose a no-reference ANN based video quality predictor which outperforms [12] in several prediction statistics. Video sequences used in this work were encoded by state-of-the-art H.264/AVC codec and have QCIF resolution. The proposed method has been found to be robust, fast and quite precise in terms of the statistics of its results. We think that it can be employed for monitoring visual quality in a network, e.g. 3G/4G cellular networks.

This paper is structured as follows: Section 3.2 gives information about the features used as input for the proposed model and the corresponding target quality metrics. The proposed model is discussed in Section 3.3. Section 3.4 provides a discussion on the experimentation and results of this work. Finally, Section 3.5 draws conclusive remarks about this paper.
3.2. The inputs and targets for the model

This section gives insight into the choice of encoder parameters used in the proposed method and presents in brief the quality metrics which are predicted using the proposed ANN model.

### 3.2.1 Selection of the parameters

While encoding a video, the coding process has to tune and calibrate various parameters to match the demand of specifications like bit-rate, frame rate and the pixel density. Such a parameter setting may be fixed from within a part of a frame to the entire frame and ultimately for the whole video sequence under consideration. The coded bitstream carries motion vector arrays, quantized residual coefficients and header information. The idea used in building up this quality predictor is to extract and calculate certain parameters from this bitstream to be deployed as input for the ANN model. Among the many parameters which may be obtained and used for this purpose, the potentially most contributing ones for the quality prediction were selected in [12] and are given in table 3.2.

However, the list of table 3.2 can be further simplified with a minor trade-off in the performance. The objective is to decrease the computational load of the ANN quality predictor. By observing the cross-correlation in table 3.1 between different parameters and the Pearson correlation coefficient values of prediction by the proposed model shown in Fig. 3.1, the following analysis is performed in order to reduce the number of parameters in a controlled manner.

- Fig. 3.1 shows that Bits/frame, P4x4, P8x8 and average QP are the most...
3. A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

### Table 3.2: The Used Parameters List

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bits/Frame</td>
<td>Bits per Frame</td>
</tr>
<tr>
<td>2</td>
<td>P16x16[+]</td>
<td>Percentage of inter blocks of size 16x16</td>
</tr>
<tr>
<td>3</td>
<td>P4x4[+]</td>
<td>Percentage of inter blocks of size 4x4</td>
</tr>
<tr>
<td>4</td>
<td>P8x8[+]</td>
<td>Percentage of inter blocks of size 8x8</td>
</tr>
<tr>
<td>5</td>
<td>Avg MV</td>
<td>Average motion vector length</td>
</tr>
<tr>
<td>6</td>
<td>Avg QP</td>
<td>Average quantization parameter (QP) value</td>
</tr>
</tbody>
</table>

contributing parameters when each of them is used alone for quality prediction.

- The parameters to be employed for prediction should ideally represent different aspects of the coded video. These aspects could be the motion content dynamics of the video (avg MV), structural formation of the video frames (macroblock size in inter-coded frames) and bitrate (Bits/frame, avg QP) related information. Bits/frame and avg QP are hence chosen for their high individual contribution towards quality prediction (Fig. 3.1) and to have appropriate input regarding the bitrate information for the given resolution of the video under test. Several H.264 codec applications are not using the macroblocks of size 4x4. Hence, P8x8 is selected from this group.

- Avg MV has a low contribution to the quality prediction when used alone but it has a very low cross-correlation with the parameters avg QP and Bits/frame (table 3.1). But, it is desirable to have motion content dynamics information to make the quality prediction process more robust.

- These arguments lead to the selection of Bit/frame, P 8x8, avg MV and avg QP as the simplified list of parameters to be used for a lesser complex video quality predictor.

#### 3.2.2 The quality metrics

Experiments have been performed using the proposed quality predictor to assess the following quality metrics.
3.2. The inputs and targets for the model

![Graph showing Pearson correlation coefficient for individual parameter use.]

Figure 3.1: Pearson correlation coefficient for individual parameter use.

![Diagram showing the general framework of the proposed model.]

Figure 3.2: The general framework of the proposed model

1. PSNR. This metric is a quite widely used measure of quality of reconstruction of lossy compression and it is based on the mean-square error of the luminance values of the two frames under comparison.

2. PEVQ. Perceptual Evaluation of Video Quality estimates mean opinion scores (MOS) scores of the video quality by modelling the behaviour of the human visual tract. After successful benchmarks by the VQEG, PEVQ has become part of ITU-T Recommendation J.247 (2008) [7].
3. **A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor**

![Figure 3.3: The ANN architecture](image)

3. SSIM. Structural Similarity Index is a technique of measuring the structural similarity between two frames [4]. SSIM is a still used alternative way to evaluate perceptual quality.

### 3.3 The Proposed Quality predictor Model

The problem at hand may be addressed by a mapping function \( Y = F(X) \), with \( X = [x_1, x_2, x_3...x_n] \) where \( x_1, x_2, x_3...x_n \) are the parameter vectors obtained from the coded bitstream of the videos under test produced by different codec settings. The vector \( Y = [y_1, y_2, y_3...y_n] \) holds the corresponding video quality metric values \( y_1, y_2, y_3...y_n \). So, building up such a function boils down to constructing a model which can regress between the set of parameters obtained from the coded bitstream and the related set of values of the quality metrics. The model should yield a pertinent value from vector \( y \) when a specific \( x \) vector is injected into it.

The desired mapping model may be designed using one of the many available methods like multi-linear regression in [12] or machine learning methods like artificial neural network [9] or support vector machine [13]. Each approach has its inherited advantages or disadvantages with regards to complexity and performance and any approach may has to trade off one advantage for the other. Linear regression is simple but unable to deal with the possible non-linearities, and machine learning methods are quite intelligent in handling complicated problems but they may become too complex to be implemented in devices with limited processing capacity. In this paper, an ANN based quality predictor is proposed which is relatively simple yet has proven its robustness in measuring video quality. The general framework of the proposed model is as depicted in the Fig. 3.2.
3.3. The Proposed Quality predictor Model

Table 3.3: The available sample space of videos

<table>
<thead>
<tr>
<th>Session</th>
<th>Training (168 videos)</th>
<th>Testing (120 videos)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video sequence</td>
<td>Foreman, Cart, Mobile, Shine, Cropped 3G Fish, Soccer goal, Car phone</td>
<td></td>
</tr>
<tr>
<td>bitrate (kbps)</td>
<td>30, 40, 50, 100, 150, 200</td>
<td></td>
</tr>
<tr>
<td>frame rate (fps)</td>
<td>7.5, 10, 15, 30</td>
<td></td>
</tr>
</tbody>
</table>

3.3.1 The architecture of the model

Contrary to the contemporary networks like [9] or [10] used for video quality prediction with no-reference, a reasonably simpler architecture for the ANN model is proposed here, using a two-layer feed-forward network with 10 or 6 sigmoid hidden neurons and one linear output neuron, depicted in Fig. 3.3. As will be shown in section IV, the work presented has aimed at decreasing the required number of input parameters compared to the number used in [12]. The proposed model performs best with 10 hidden neurons when it uses all the listed parameters and it uses only 6 neuron when it is employed with lesser number of input parameters.

The network is trained with Levenberg-Marquardt backpropagation algorithm which is considered to be a fast method. The performance of the model is evaluated using various error statistics and regression analysis. The already stated number of neurons in the hidden layer has been fixed as a result of experiments under various number of hidden neurons to achieve an efficient system with high performance and no over-fitting. A properly trained network gives reasonable answers when tested on unseen inputs. Typically, a new input leads to an accurate output, if the new input is similar to inputs used in the training set. To this end, the network should be trained to have generalization property. Usually, generalization is achieved by regularization or early stopping. Experiments have been conducted in this regard and it was found out that early stopping makes the training process faster than regularization and hence early stopping was used in this work. The training data is further divided into training set and validation set. The error on the validation set is monitored during training. The validation error normally decreases during the initial phase of training, as does the training set error [14]. However, when the network begins to over-fit the data, the error on the validation set typically begins to increase, and the training is stopped when the validation increases over a fixed number of iterations and the weights and biases at the minimum
3. A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

Table 3.4: An average of the results obtained by using the proposed predictor

<table>
<thead>
<tr>
<th>Quality metric</th>
<th>PSNR</th>
<th>PEVQ</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Stats</td>
<td>ANN</td>
<td>ANNs</td>
</tr>
<tr>
<td>PCC</td>
<td>0.99</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>MSE</td>
<td>0.91</td>
<td>2.3</td>
<td>4.74</td>
</tr>
<tr>
<td>SD</td>
<td>0.88</td>
<td>1.28</td>
<td>1.43</td>
</tr>
<tr>
<td>ME</td>
<td>0.37</td>
<td>0.8</td>
<td>-1.65</td>
</tr>
<tr>
<td>MAE</td>
<td>0.61</td>
<td>0.96</td>
<td>1.77</td>
</tr>
</tbody>
</table>

of the validation error are used for testing on a different data set.

3.3.2 The video sequences data set

The proposed predictor has been tested and trained on a wide range of bit-rates and frame rates of video sequences with a variety of motion content dynamics. The video sequences used in this work were approximately 3 seconds long (90, 45, 30, and 23 frames) in QCIF resolution and they were encoded with the H.264/MPEG-4 AVC reference software, version 12.2 generated by JVT [15] using the baseline profile. In table 3.3, the given 7 video sequences were encoded using 6 different bit-rates and 4 different frame rates as stated in the table 3.3, and this resulted into 7x6x4 = 168 videos for training. These were further divided into actual training set and validation set to have better generalization of the neural network. The trained network was tested on 120 video sequences which were produced by clipping five sequences from different parts of the 3G video sequence, cropping to QCIF resolution and encoding according to the same bit-rates and frame rates as for training set and hence resulted into 5x6x4 = 120 videos. The five parts of the 3G video sequence had different characteristics and motion dynamics and the original letter box aspect ratio of 3G video sequence was not maintained.

3.4 Experiments and Results

The proposed model was trained with the training set for various quality metrics using the parameters listed in table 3.2 with 10 neurons in the hidden layer and using the simplified list (Bit/frame, P8x8, avg MV and avg QP value)
with 6 neurons in the hidden layer. The network was trained on the training data set and was then tested on the test set as provided in Table 3.3. For each quality metric, the experiment was run five times and then an average measure was recorded. Table 3.4 shows the performance of the proposed method and also offers a comparative view with the results obtained by the multi-linear regression (MLR) method. Fig. 3.4 shows a sample regression graph for PSNR prediction with 6 neurons in the hidden layer using the simplified list of four parameters as input. Similar plots are presented in Fig. 3.5 and Fig. 3.6 for PEVQ and SSIM assessment respectively. As the results show, the proposed model is capable of predicting the video quality with a considerable high precision and accuracy. The error statistics and Pearson correlation coefficient values as given in Table 3.4 motivate why this model could be preferred over the MLR method for quality prediction. A considerable care has been taken to
3. A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

![Regression plot for PEVQ assessment](image)

**Figure 3.5: Regression plot for PEVQ assessment**

avoid over-fitting the network and early stopping approach worked well in this regard. Also, the number of neurons in the hidden layer is a crucial parameter to be aware of and it was optimized on the basis of tests performed in the work. Table 3.4 also shows results under column $\text{ANNsimp}$ obtained from the proposed model using only 6 neurons in its hidden layer and taking simplified list of parameters as its input for quality prediction. These results further substantiate the motivation of using the proposed ANN predictor for quality prediction as even after using 4 out of the 7 parameters used in MLR approach, the ANN predictor works better.

The proposed model is designed to measure compression artifacts but this method can be extended to be employed for other issues like transmission and network error if the model is expanded with related features. The important consideration in this regard is to train the network on a reasonably large data
3.5 Conclusion

A robust, reference free method to predict mostly used perceptual quality metrics for coded video sequences has been proposed, using only features readily available in the coded bitstream. The method predicts the PSNR and PEVQ metrics quite accurately, while the SSIM metric is predicted with less accuracy. For all three metrics, the proposed method performs better than, or equal to, the earlier proposed method in [12] in all the presented statistics. The proposed method performs better than [12] even when it is used in its reduced complexity form of ANN architecture using lesser number of input parame-
3. A Reduced Complexity No-Reference Artificial Neural Network Based Video Quality Predictor

ters, except for the correlation coefficient for PSNR. The main reason why the proposed ANN based model outperforms the linear regression approach is the possible non-linear dependency of the input decoder parameters and the output target quality metric values. These promising results even with a reduced complexity ANN architecture encourage continued development of this neural network based predictor. The accuracy and precision of results obtained in this work for predicting the objective mean opinion scores (MOS) for PEVQ, which has been accepted by ITU-T to be rather close to the human perceptual assessment, show the potential of this predictor to be employed in prediction of MOS of subjective tests. The future work should focus on developing the same model for subjective MOS.

3.6 Bibliography


3.6. Bibliography


Four

A No-Reference Machine Learning Based Video Quality Predictor

This chapter has been published as:

A No-Reference Machine Learning Based Video Quality Predictor

Muhammad Shahid, Andreas Rossholm, and Benny Lövström

Abstract

The growing need of quick and online estimation of video quality necessitates the study of new frontiers in the area of no-reference visual quality assessment. Bitstream-layer model based video quality predictors use certain visual quality relevant features from the encoded video bitstream to estimate the quality. Contemporary techniques vary in the number and nature of features employed and the use of prediction model. This paper proposes a prediction model with a concise set of bitstream based features and a machine learning based quality predictor. Several full reference quality metrics are predicted using the proposed model with reasonably good levels of accuracy, monotonicity and consistency.

4.1 Introduction

The recent advancement in digital imaging technology and availability of efficient transmission systems have resulted in a proliferation of videos more than ever before. Videos transmitted to and from mobile devices will account for 66% of the global mobile data traffic by 2017 as per forecasts [1]. Video services that have gained wide interest are many and television broadcast, DVD, Blu-Ray, Mobile TV, Web TV etc. are some to name. One of the key characteristics of video services is perceptual aspect of the quality of experience (QoE) as observed by the end user. Quality of visual media can get degraded during capture, storage, transmission, reproduction and display due to distortions which might occur at any of these stages. The true judges of video quality are humans as end users of the video services. The scientific process of evaluation of video quality by humans is called subjective quality assessment and it is usually given as mean opinion score (MOS) value. However, subjective evaluation is often too inconvenient, time-consuming, expensive and it has to be done by following special recommendations in order to yield reproducible
standard results. These reasons give rise to the need of some intelligent ways of automatically estimating the perceived quality that can be performed swiftly and economically. This non-subjective way of quality evaluation is known as objective quality assessment. Due to the overwhelming number of approaches on which objective methods of video quality assessment (VQA) are based, there are many ways we can classify them. For example, data metrics such as PSNR and MSE consider only the fidelity of the signal and remain agnostic to the visual content. On the other hand, picture metrics do take into account the visual information present in the signal. Moreover, objective metrics can also be divided on the basis of the reference information used for the quality estimation. Such metrics fall in one of the three categories namely full-reference (FR), reduced-reference (RR) and no-reference (NR) metrics [2].

In the area of video quality, Video Quality Expert Group (VQEG) has the paramount role of validation of objective metrics for International Telecommunication Union (ITU) [3]. ITU has been the cardinal body for issuing recommendations and performing standardization activities for objective methods of video quality. Based on the normative reports from ITU, objective metrics have been classified into the following five groups in terms of the type of data used.

1. Media-layer models: Models found in this category estimate quality value using the video signal. Applications such as codec comparison and optimization are common uses of these models.

2. Parametric packet-layer models: These models do not access the media signal but employ the packet-header information to estimate visual quality.

3. Parametric planning models: In these models, the quality planning parameters for network and terminals are employed for the assessment of quality.

4. Bitstream-layer models: These models use the encoded bitstream and packet-layer information for the quality assessment.

5. Hybrid models: A composite of two or more models mentioned earlier belongs to this category.

PSNR and MSE have been the traditional pixel by pixel comparison objective metrics which, however, don’t correlate well with perceptual assessment of humans [4]. It is desirable to have perceptually relevant objective metrics, hence Structural Similarity Index (SSIM) and Perceptual Evaluation of Video Quality
4.2. Related work

(PEVQ) included in ITU-T Rec. J.247 are preferred.

Our contribution in this paper builds on our previous works [5] and [6]. The use of neural network for image quality prediction has given results better than the use of linear regression technique [6]. However, neural network techniques often face the problem of overfitting and the computational complexity of neural network based methods can grow with increase in dimension of input space. In order to circumvent such issues, support vector machine technique is used [7]. Mainly, this paper presents: carefully performed subjective tests of quality assessment on a varying content of video sequences, a rational selection of the bitstream features based on their contribution towards quality estimation, improved results by using a better regression technique and a comprehensive comparison of results using commonly used FR quality metrics such as PSNR, SSIM, MSSIM and PEVQ. The remaining part of the document is organized as follows. A description of the test media generation and its subjective quality assessment, used in this work, is provided in Section 4.3. It is followed by the details of the proposed model of quality prediction in Section 4.4. A comparison of the proposed method with similar techniques is presented in Section 4.5. Finally, some conclusive remarks about this work are drawn in Section 4.6.

4.2. Related work

With regards to the use of quality metrics in practice, such as in-service monitoring, only NR metrics are suitable as availability of a reference video signal (FR) or some features of the reference video (RR) is not possible mostly. Bitstream-layer models provide a low complexity solution for NR quality assessment. The parameters used in these methods can be obtained without fully decoding a video. A low complexity solution of video quality prediction based on bitstream-layer parameters is found in [5]. Low complexity has been achieved by using a simple multi-linear regression system for estimating a relationship between the parameters and quality values. An improvement of this approach was presented in [6] where the required number of parameters have been reduced for computational efficiency and the prediction accuracy has been improved by the virtue of the usage of artificial neural network (ANN). The results presented in both of these methods were not, however, validated by subjective scores. A set of 48 bitstream parameters related to slice coding type, coding modes, various statistics of motion vectors and QP value was used in [8] to predict quality of HDTV videos. Partial least squares method was used as a tool for regression between the feature set and subjective MOS. However,
4. A No-Reference Machine Learning Based Video Quality Predictor

![Figure 4.1: A snapshot of the test video sequences](image)

the computational complexity for processing such a huge set of parameters on slice level can be a point of concern. Two methods tested on MPEG-4 encoded videos and based on several macroblock (MB) level statistics, derived from bitstream and reconstructed videos, were reported in [9] for PSNR and in [10] for SSIM. In the metric targeted for for PSNR estimation, Shanableh [9] has employed spectral regression and reduced model polynomial network. A multi-pass prediction system based on stepwise regression has been used in [10]. The video features, in both of these metrics, constitute mainly coding information of an MB, some relative measures of motion vectors of neighboring MBs and a couple of numerical values related to texture of an MB. Similarly, a PSNR estimator for H.264/AVC encoded video is presented in [11] where bitrate, QP value and coding mode are used as the features for quality prediction. Prediction of perceptual quality (subjective MOS) can be an applicable improvement of these methods.

4.3 Subjective Video Quality Assessment

Subjective evaluation of video quality is the fundamental ground truth to test and compare the performance of objective methods of quality assessment. It involves setting up an experimental environment according to specific recommendations where a prescribed number of people watch the test set of video sequences and rate the videos in terms of the visual quality. International Telecommunication Union (ITU) has two sub organizations namely Radiocommunication Unit (ITU-R) and Telecommunication Unit (ITU-T) which are responsible for standardizing such recommendations.

4.3.1 Test Video Sequences

Six different video sequences of CIF and QCIF spatial resolutions were selected in raw progressive format with different amount of motion contents and texture complexity. A method put forward by ITU-R P.910 [12] recommends an
4.3. Subjective Video Quality Assessment

Figure 4.2: SI and TI plot computed for luminance component of selected CIF and QCIF videos

approach to characterize videos on the basis of their spatial perceptual information (SI) and temporal perceptual information (TI). SI and TI are calculated in the luminance plane of a video. The test videos used in this work have been selected to represent both of these two parameters from low to high range and Fig. 4.2 shows a plot for the chosen videos, displayed in Fig. 4.1. The selected videos were encoded at two frame rates and five different bitrates, resulting into a test set of 120 videos, Table 4.1 shows these values in detail. The encoding was done using JM reference software (http://iphome.hhi.de/suehring/tml) based on H.264/AVC standard in baseline profile mode. This encoding approach is specifically useful in applications that require low delay (such as video conferencing) as bi-predictive coding (B frames) are not used in it and for the hand-held devices with low computational power. Each of 120 video sequences were of 20 seconds in length. The first frame is coded as intra (I) frame and has comparatively high quality which is supposed to reduce for upcoming successive frames due to predictive coding. In order to have relatively uniform quality of the whole test sequence, last 10 seconds of each video were
4. A No-Reference Machine Learning Based Video Quality Predictor

Table 4.1: The Test Video Sequences

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Video Sequence</th>
<th>Bitrate (kbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIF @ 30,15 fps</td>
<td>Akiyo</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>News</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Foreman</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Soccer</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Football</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td>QCIF @ 30,15 fps</td>
<td>Akiyo</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>News</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Foreman</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Soccer</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Football</td>
<td>100,200,300,400,500</td>
</tr>
</tbody>
</table>

selected for the subjective quality assessment to remove the I frame effect and to avoid forgiveness effect [13].

4.3.2 Subjective Quality Assessment Setup

We considered the recommendations given by ITU-R BT 500-12 [14] for the lab setup of our experiments. Particularly, the method followed was single stimulus quality evaluation where a test video sequence is shown once without the presence of any explicit reference, corresponding to the reality where users see only the processed version of videos [15]. A detailed account of these subjective tests is presented in [16]. In order to obtain reliable results out of raw subjective scores, screening of the observers was employed to discard observers that are considered as outliers. The algorithmic details of these steps are reported in Annex 2 of [14]. After performing the refining process, mean opinion score (MOS) was calculated and used in this work.

4.4 The Proposed Method

The proposed method consists of extraction of visual quality relevant bitstream parameters/features and building of a machine learning based model for quality prediction. These parameters were selected carefully to keep the complexity
4.4. The Proposed Method

in control and to get reasonable coding information which can represent the degradations caused on visual quality by compression. Here, we describe the extracted parameters and the rationale of including a feature in the proposed model. Note that the terms parameter and feature are used alternatively to mean the same in this context. We broadly divide the chosen bitstream parameters into two types namely Coding Distortion and Content Structural Information as given below:

- Coding Distortion: Quantization is the main source of distortion introduced during the video encoding and the Quantization Parameters (QP) value is used to steer the scale of quantization. The QP value has its impact on the bitrate and it is observed that a video sequence coded at higher bitrate has higher visual quality and the quality is degraded by lowering the coding bitrate when the encoding settings is fixed.

- Content Structural Information: In H.264/AVC, the 16x16 macro block (MB) can be sub-partitioned into blocks of sizes 4x4, 4x8, 8x4 or 8x8, 8x16, 16x8 depending on the coding mode being chosen for the sake of minimal error in prediction [17]. The relative selection of these alternatives during encoding provides an estimate of the structural information of the content present in a video. The ratio of intra blocks out of the total blocks in an inter frame would give an estimate of temporal complexity present in a video sequence. Motion Vector (MV) based statistics can be used to characterize the motion contents of a video.

As the rate control mechanism in the used JM reference software is driven to find the optimum coding modes based on a requested bitrate, the above mentioned two types have an aggregate impact on the visual quality and it differs depending on the content of a sequence.

4.4.1 Feature Extraction of H.264/AVC Bitstream Data

In the light of the discussion given before, following features were extracted from the H.264/AVC coded bitstream data of the test video sequences. 1. Average bitrate. 2. Percentage of Intra coded macroblocks. 3. Percentage of Inter coded macroblocks. 4. Percentage of skipped macroblocks. 5. Percentage of inter macroblocks of size 16x16. 6. Percentage of inter macroblocks of size 8x8, 16x8, and 8x16. 7. Percentage of inter macroblocks of size 4x4, 8x4, and 4x8. 8. Average of motion vector difference lengths in X direction. 9. Average of
4. A No-Reference Machine Learning Based Video Quality Predictor

motion vector difference lengths in Y direction. 10. Average of absolute motion
vector lengths in X direction. 11. Average of absolute motion vector lengths
in Y direction. 12. Percentage of macroblocks with absolute motion vector of
length zero. 13. Percentage of macroblocks with motion vector difference of
length zero. 14. Motion Intensity (I) defined by:

$$\sum_{i=1}^{N} \sqrt{MVX_i^2 + MVY_i^2}$$ (4.1)

where N is the total number of macroblocks in each frame and MVX_i and
MVY_i are the absolute motion vector values of the i-th macroblock in X and Y
directions respectively. 15. Motion Intensity (II) defined by:

$$\sqrt{MVX^2 + MVY^2}$$ (4.2)

where MVX and MVY are the average of absolute motion vectors in each frame
in X and Y directions respectively. 16. Percentage of Intra coded macroblocks
of size 4x4 in I frames. 17. Percentage of Intra coded macroblocks in P frames.
18. Average quantization parameter.

The features were computed on frame level and an average value of each fea-
ture was obtained for a video. In this work, the coded sequences had only one
I frame (1st one), so the feature number 16 was not included in our current
model.

4.4.2 Proposed Prediction Model

For VQA, in order to predict the quality value based on an objective model, first
we need to train such a model on a set of video features with known quality
value. Given a matrix X of m features of n videos with the corresponding n
values of the quality in vector y, where m and n are real positive numbers, as
expressed by the following:

$$X = \{x_1, x_2, x_3, ..., x_m\} \quad X \epsilon R^{m \times n}$$ (4.3)

$$y = \{y_1, y_2, y_3, ..., y_n\} \quad y \epsilon R^{n \times 1}$$ (4.4)

A given prediction model is trained using this m-dimensional feature set for
target values of quality measure given in y. This training process can be per-
formed by building a functional model between the video features set and the
corresponding quality values set. The trained model is used to predict quality
measure of another set of m-dimensional feature set of any k test videos, for
which it was not trained with. The accuracy of the prediction can be evaluated
by using the actual quality measure of the $k$ videos.

One of the regression analysis techniques relies on kernel based learning
methods. These methods solve any problem by mapping the input data set
into high dimensional feature space via linear or nonlinear mapping. Support
vector machine (SVM) is a supervised kernel based learning algorithm and it
is commonly used for classification and regression analysis [18]. Regression
computations in SVM are done based on structural risk minimization (SRM)
principle which employs capacity control to avoid over-fitting issue. In order
to simplify the implementation of SVM, in which a solution of inequality con-
straints is sought, Suykens et al. [19] developed a variant of SVM called least
square support vector machines (LS-SVM). It reformulates the standard SVM to
solve linear Karush-Kuhn-Tucker equation systems and it helps decrease com-
putational complexity. In essence, LS-SVM transforms quadratic programming
into a set of linear equations which are easier to solve. In our work, we adopted
LS-SVM algorithm for regression analysis due to its known ability of handling
non-linear data and the simplicity of computation.

Radial Basis Function (RBF) was selected as the kernel function for realiza-
tion of implicit mapping of given input data into higher dimensional feature
kernel space which results in obtaining better training and testing errors. An
optimization algorithm was employed for tuning the hyper parameters in order
to achieve better generalization performance and prediction accuracy. There are
three optimization algorithms: simplex which can be employed for all kernels,
grid search which is limited to 2-dimensional tuning parameter optimization
and line search which can only be employed to linear kernel. We need to tune
the hyper parameters for building LS-SVM model, so we employed simplex
optimization algorithm. The mechanism of the tuning process is operated by
coupled simulated annealing (CSA) which is better than multi start gradient de-
scent optimization [20]. We considered leave one out cross validation (LOOCV)
as a cost function for estimating the performance of LS-SVM model. This al-
gorithm is controlled by the performance metric mean square error (MSE) and
offers good generalization ability [21]. The performance and accuracy of LS-
SVM model depends on setting of $(\sigma^2, \gamma)$, where $\sigma^2$ is width of kernel and $\gamma$
is regularization parameter. For each pair of hyper parameters $(\sigma^2,\gamma)$, LOOCV
method is performed on training set to estimate the prediction error and thus a
robust model is obtained by selecting those optimal pair of hyper parameters.
Interested reader is referred to [19] for a detailed treat of these concepts.
4. A No-Reference Machine Learning Based Video Quality Predictor

Table 4.2: Comparison of LS-SVM with MLR [5] and ANN [6] based methods

<table>
<thead>
<tr>
<th>VQM</th>
<th>SSIM</th>
<th>MOS</th>
<th>PEVQ</th>
<th>PSNR</th>
<th>MSSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/S</td>
<td>SVM</td>
<td>ANN</td>
<td>MLR</td>
<td>SVM</td>
<td>ANN</td>
</tr>
<tr>
<td>PCC</td>
<td>0.99</td>
<td>0.98</td>
<td>0.95</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>SRC</td>
<td>0.99</td>
<td>0.98</td>
<td>0.95</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>OR</td>
<td>0.05</td>
<td>0.13</td>
<td>0.20</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>MSE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>9.79</td>
<td>14.30</td>
</tr>
</tbody>
</table>

4.4.3 Benchmark Measurements

According to VQEG phase II [22], performance of an objective quality prediction model can be evaluated by three parameters which describe prediction accuracy using Pearson linear correlation coefficient (PCC), monotonicity using Spearman rank order correlation coefficient (SRC) and consistency using outlier ratio (OR). We have used these metrics to present the results of this work.

4.4.4 Simulation Results

The set of the extracted parameters of 120 video sequences was divided into a training and a test set. The training data was obtained from a set of 80 videos selected randomly from the 120 videos, that makes two third of the total. The proposed LS-SVM model was trained on the training set for a variety of VQA metrics including subjective MOS. The training was done separately for each quality metric and the training data was expected to be different for each due to random selection. The proposed method has been tested for its prediction ability of subjective MOS, PEVQ, PSNR, SSIM and MSSIM. As can be seen in the regression plot given in Figure 4.3, the proposed model is able to predict subjective MOS with reasonably high accuracy and a similar trend is observed for other quality metrics also. Magnitude of the error bars of PCC values are below 0.005 that shows a high accuracy level in the prediction.

4.5 Comparison with Contemporary Techniques

Table 4.2 reports the statistical results achieved after using the features listed in Section 4.4.1. Results are shown for up to two degrees of precision in most cases. Under each quality metric column, values of the considered performance
4.5. Comparison with Contemporary Techniques

Figure 4.3: Prediction Accuracy and Monotonicity Coefficients Values of the Proposed Model, under 95% CI

Table 4.3: Comparison of LS-SVM with MLR [5] and ANN [6] based methods (Reduced set of features)

<table>
<thead>
<tr>
<th>VQM</th>
<th>SSIM</th>
<th>MOS</th>
<th>PEVQ</th>
<th>PSNR</th>
<th>MSSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>M/S SVM ANN MLR SVM ANN MLR SVM ANN MLR SVM ANN MLR SVM ANN MLR SVM ANN MLR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCC</td>
<td>0.98</td>
<td>0.97</td>
<td>0.90</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td>SRC</td>
<td>0.98</td>
<td>0.96</td>
<td>0.85</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td>OR</td>
<td>0.10</td>
<td>0.10</td>
<td>0.64</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>MSE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.19</td>
<td><strong>10.50</strong></td>
<td>14.28</td>
</tr>
</tbody>
</table>

measures are presented for the three methods. For all of the quality metrics, LS-SVM has outperformed the ANN and MLR methods in all the considered performance measures. Specifically, SSIM and MSSIM have been predicted with comparatively high accuracy, monotonicity and consistency. This result is in-line with findings related to structure based quality metrics for their ability of estimating the visual quality. That explains the widespread adaptation of these metrics for a variety of image processing applications such as image restor-
4. A No-Reference Machine Learning Based Video Quality Predictor

ition, contrast enhancement, watermarking, and rate-distortion optimization in standard video compression [23]. Performance results on subjective MOS are also promising except relatively high outlier ratio for all three methods. That might occur due to relatively high variability in the subjective scores, which is inherent in subjective assessment. It was observed that certain features can be omitted from the prediction model without any significant loss in performance. A discussion on the rationale of deletion of features was provided in [6]. Consequently, feature number 2, 3, 4, 6, 16, 17 were deleted to form a reduced set of features. Table 4.3 provided a comparative performance of the proposed method when tested on the reduced set. The simplification of the model for a lesser number of parameters is valid as indicated by the performance measure values. The performance of LS-SVM based model remains almost the same even with the reduced set of features, in contrast to the competitor models. It is ascertained from the results that the use of support vector machine is advantageous in learning the bitstream based features of an encoded video for its quality prediction.

4.6 Conclusion

We have presented a model of no-reference video quality estimation based on a well-known machine learning method known as LS-SVM. Selection of the quality-relevant features was done based on the rationale of encoding fidelity, structural information of contents, motion information, coding distortion and spatio-temporal complexity. The proposed method was tested for its prediction ability of quality metrics PSNR, PEVQ, SSIM, MSSIM and subjective MOS. Its performance was evaluated using three standard methods of performance measure and reasonably acceptable results were obtained. Moreover, a comparative analysis with two similar methods has been performed to prove its application. The contribution of this paper is two-folds; introduction of a visual quality estimation model with bitstream-based features and comparison of contemporary techniques used for regression. The LS-SVM based predictor performs marginally better than ANN based, and reasonably better than linear-regression based, predictors for accuracy, consistency, monotonicity and error in prediction. The performance of the proposed LS-SVM model remained substantially high while a reduced set of features was used for the quality prediction. This gain in performance can be attributed to the fact that SVM is less prone to overfitting and hence has better generalization capabilities. Future work in this area is to build a model that can account for the impact of network distortions on the visual quality.
4.7 Bibliography


4. A No-Reference Machine Learning Based Video Quality Predictor

Speech and Image Processing, Multimedia Communications and Services, pages 114–117, Jun. 2007.


Five

Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

This chapter has been submitted as:

Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

Muhammad Shahid, Katerina Pandremmenou, Lisimachos P. Kondi, Andreas Rossholm, and Benny Lövström

Abstract

In view of the growing demand for perceptual quality assessment due to the increased usage of video services, this article proposes Reduced-Reference (RR) and No-Reference (NR) models for video quality estimation. A variety of perceptually-motivated features are examined to account for the impact of coding artifacts, packet losses, and video content characteristics. The proposed RR and NR sets of features have not been used in the literature and some of the features are used for the first time in this study. These features are employed for estimating video quality using the Least Absolute Shrinkage and Selection Operator (LASSO) regression. This regression technique utilizes a subset of the input features, by selecting only those that have relatively higher impact in the process of video quality estimation. It is demonstrated that the selected features using LASSO are able to estimate the perceptual quality with reasonably high accuracy, comparable to a baseline performance provided by Ridge regression that employs all the input features. Performance measures as recommended by Video Quality Experts Group (VQEG) are utilized in order to gauge the effectiveness of the proposed techniques. Additionally, comparisons with competitive works verify the superiority of our models in all examined performance measures.

5.1 Introduction

The video portion of the global mobile data traffic has increased tremendously and it is estimated to exceed 67% by 2018, from being 53% in 2013 [1]. Therefore, with this growing usage of videos, it is believed that the end-users are
5. **Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models**

becoming more aware of the perceptual quality characteristics of video services. A required amount of compression of the raw (original) videos has to be performed in order to meet the practical limits of data storage devices and transmission channels. Depending upon its intensity, the compression can introduce different visual artifacts in a video that may decrease its perceptual quality as compared to its original version.

Besides compression, video quality can also suffer from degradations due to transmission over lossy networks. Losses of video data in a network can occur for various reasons such as network fluctuations, buffer overflows, and any operational management procedures. However, there is a growing trend of video communications through reliable transmission methods, where losses can be recovered through retransmissions, though it might be difficult to avoid all packet losses in the case of varying network characteristics. Moreover, real-time communications such as video-conferencing and other low-delay demanding video services may suffer from packet losses, as the underlying transport mechanisms generally do not apply retransmissions, e.g., User Datagram Protocol (UDP), Real-time Transport Protocol (RTP) etc. A parameter that is commonly used by service providers in order to evaluate the quality of service for an end-user is the Packet Loss Rate (PLR), which is generally considered as a useful measure for quantifying losses in a network. Hence, a study based on PLR to evaluate the performance of video communications in lossy networks can be quite useful.

In most scenarios of processing or transmission of visual information, the ultimate judges of quality are human observers. Despite the fact that many evaluation methods of objective performance have been developed, subjective assessment is the most valid solution, since it provides the ground truth of quality. The recommended procedures for subjective Video Quality Assessment (VQA) involve the collection of quality scores from a viewers’ panel, under a controlled laboratory environment. The product of such assessments is typically a Mean Opinion Score (MOS) [2] for each test sample, which corresponds to the average value of the scores given by the panel. Crowdsourcing-based subjective VQA is an emerging technique, where test material is transferred to the viewers’ premises through the Internet and the quality scores are collected through a loosely controlled environment [3]. However, subjective VQA is rather tedious and time-consuming, and hence, it is impractical to incorporate it in real-time applications.

In the last two decades, many modern models/metrics of perceptual VQA
have been developed and they can be computed automatically, based on quality-
relevant features of a video. The goal of such objective metrics is the compu-
tation of a perceptual quality estimate that correlates well with the results of
subjective assessment. A classification of the objective metrics can be made
on the basis of the reference information used for quality estimation [4]. Given
that “original” refers to the unprocessed pristine video and “impaired” refers to
its processed version (coding and/or transmission losses); Full-Reference (FR)
metrics have full access to both the original and impaired videos, Reduced-
Reference (RR) metrics have access to some suitable features transmitted from
the server’s side and full access to the impaired video, and No-Reference (NR)
metrics have access only to the impaired video.

It is generally believed that FR metrics have the capacity to provide the most
accurate estimations of video quality, since they use input information from
both the original and impaired videos. However, because of the dependence on
the original video, FR metrics are mostly suitable for offline applications, such
as encoder performance comparisons. In addition to the processed video, RR
metrics can also access selected features of the original video. These features
can be sent to the receiver through an ancillary channel [5] or alternatively,
they can be embedded in the video content itself, by using techniques such as
watermarking [6]. For the same purpose of quality estimation, NR metrics
make use of either the bitstream of the impaired video or the decoded pixels
of it, or both. Because of the limited or no dependence on the original video,
RR and NR metrics are suitable for real-time applications and online quality
monitoring of the streaming videos [7].

In this article, we present a study on design, implementation, and evalua-
tion of RR and NR models, which employ two different sets of perceptually-
motivated features. These features are extracted from H.264 encoded videos,
impaired with different amounts of packet losses. The use of Least Absolute
Shrinkage and Selection Operator Regression (LASSO) [8, 9, 10] is proposed
aiming at the dual goal of feature selection and MOS estimation. As a base-
line, Ridge regression [11, 12, 13] is applied, which performs MOS estimation
without any feature selection.

The rest of this article is organized as follows: Section 5.2 presents an
overview of the related work and concludes with a summary of the key points
and contributions of this paper. A discussion on employed features that are
related to perceptual video quality as well as a study about their suitability
takes place in Section 5.3. The problem of video quality estimation based on a
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

set of quality-relevant features and its solution using linear regression models is presented in Section 5.4. In the same section, the procedure followed for the models’ development is also described. The employed measures of performance, the provided experimental results and their analysis, and even the performance comparison of our proposed models with related works are given in Section 5.5. Finally, conclusive remarks on this work and an outlook of future research directions is given in Section 5.6.

5.2 Related Work

During the last years, a considerable part of the scientific community has focused its interest on efforts for the development of objective video quality metrics that target at reliable and accurate modeling of subjective VQA. Some RR metrics of VQA are presented in [14,15,16]. In [14] the authors designed an RR metric that is targeted for applications related to wireless communications. It is built based on the principle that humans tend to have different impairment perceptibility based on the spatial and temporal affected regions of a video sequence. The work in [15] presented a family of RR VQA models that differ in the amount of reference information required for video quality measurement, while [16] proposed a wavelet-based video distortion metric that can operate in FR or RR mode, as required. Actually, RR metrics can be an alternative to FR metrics when the original video is not accessible. However, in some cases, the cost of maintaining an ancillary channel may be high for an RR approach, while such metrics may not meet the requirements of quality estimation in the event of a failure in RR data delivery to the receiver’s end.

For these reasons, NR metrics are the most broadly applicable solution for VQA, though quality estimation with limited available input information can be challenging [17]. A NR metric tested on MPEG-4 compressed video that estimates the Peak Signal to Noise Ratio (PSNR) on MacroBlock (MB) level was proposed in [18] and a similar method that estimates the Structural SIMilarity (SSIM) index was introduced in [19]. The study presented in [20] described a PSNR estimator that considers only the compressed bitstream of an H.264/AVC coded video. However, the estimation of perceptual quality in terms of MOS could be an applicable improvement for the works presented in [18,19,20].

A set of bitstream-based features related to slice coding type, coding modes, various statistics of motion vectors, and Quantization Parameter (QP) value were employed in [21] with the goal of quality estimation of high definition
5.2. Related Work

television video, encoded by H.264/AVC. For the same purpose, statistics of boundary strength values of the deblurring filter, QP, and average bitrates were used in [22] for H.264/AVC encoded videos. Also, a motion-based quality metric was explored in [23] for H.264/AVC encoded videos as well. For this metric, some statistical features related to motion vectors along with the bitrate and frame rate were calculated, and the principal component analysis method was used to identify the parameters that can be the most influential in quality value. Similarly, a low complexity solution of VQA based on bitstream features was proposed in [24]. An improvement of this approach was included in [25], where the required number of features was reduced so as to promote computational efficiency. In that work, an improvement was noted in estimation accuracy by the virtue of the usage of an artificial neural network. A further improvement of [25] can be found in [26], where a larger set of parameters was used and the estimation of subjective MOS was also considered. However, the models built in [21, 22, 23, 24, 25, 26] are oriented towards capturing distortions due to lossy source coding only, and thus, they cannot be applied in the case of packet-loss impaired videos.

In reference [27], the authors extracted a set of features from the MPEG-2 bitstream and proposed two different modeling approaches: i) a tree classifier to decide if a packet loss is visible or invisible and ii) a Generalized Linear Model (GLM) to estimate the probability that a packet loss is visible. In [28] the GLM approach was extended for H.264/AVC bitstreams to model the visibility of individual and multiple packet losses. An application of the proposed GLM scheme to packet prioritization of a video stream, considering factors not only within a packet but also in its vicinity was suggested in [29]. The visual effect of whole B-frame losses was investigated in [30]. For this purpose, a GLM was used to estimate probability of the visibility of a B-frame loss and a router was able to decide about which frames to drop in a video transmission scenario, where the incoming bitrate was higher than the outgoing rate. However, methods presented in references [27, 28, 29, 30, 31] classified packets in a binary mode as visible or invisible based on the viewers’ responses to the glitches they spotted. For example, a packet loss was assumed to be visible when the percentage of the viewers that identified an impairment was over a threshold and invisible when this percentage was under a threshold. On the contrary, [32] introduced a NR bitstream-based model that predicts continuous estimates for the visibility of packet losses, and the impact of the lost packets on perceptual video quality was also studied. However, most of these metrics mainly target only the visibility of packet losses and a direct estimation of perceptual quality
5. **Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models**

is not made by including also the features related to video coding.

The NR method presented in [33] estimates the quality of videos transmitted over wireless networks using information from MBs of inter-frame encoded pictures of a video. The proposed method analyzes the impact of both encoding and channel conditions to the video quality degradation by using motion vectors and residual error from the received P-frame and/or B-frame. In addition, in [34] a Quality of Experience (QoE) evaluation model was proposed to estimate the end-users’ perception on a video streaming service considering different video content types. This QoE model extracts key parameter information directly from degraded video frames in order to estimate the video QoE. A similar NR quality metric for networked video was introduced in [35] using information extracted from the compressed bitstream only. This metric accounts for picture distortion caused by quantization, quality degradation due to packet losses and error propagation, and temporal effects of the human visual system.

Moreover, we have compared the performance of proposed models with the following related work. The work presented in [36] proposed an FR method that uses both singular values and singular vectors as visual features, and a machine learning technique for feature pooling was also introduced. The work presented in [37] proposed an RR metric that compares the phase and magnitude of the 2-D discrete Fourier transform of the reference and distorted images in order to compute visual quality. A NR bitstream-based quality metric that considers both the effects of lossy H.264/AVC video encoding and packet losses over Internet Protocol (IP) networks was proposed in [38]. In [39], an NR video quality metric for H.264/AVC video transmissions in packet-based networks was introduced, which used features from the headers that encapsulate compressed video data. Similarly, in [40] an enhanced algorithm based on the G.1070 model [41] was developed that compensates for the impact of varying video content characteristics on encoding bitrate. Lastly, genetic programming-based symbolic regression is used in [42] in order to build a bitstream-based NR model. The used features characterize encoding settings, parameters related to network distortions and video content.

Accordingly, in the context of the aforementioned related works, we propose an approach that directly estimates video quality by employing perceptually-motivated video features, by extending our previous studies presented in [24, 25, 26]. The key points of this article as well as the contributions it brings are summarized as follows:
5.3. Features Related to Perceptual Video Quality

1. We propose RR and NR models in order to estimate the perceptual quality of H.264/AVC video sequences, which are affected by packet losses.

2. A variety of features that may have an effect on perceptual video quality are collected in order to be used for building the proposed models using regression techniques. The suitability of the proposed RR and NR datasets are examined for redundancy and goodness-of-fit [43] with perceptual quality. It is worth mentioning that the constitution of the RR set of features as a whole and that of the NR set of features as a whole are employed for the first time, while this study also introduces the utilization of 11 new features.

3. LASSO regression [8,9,10] is utilized in order to i) indicate the most useful features for video quality estimations, and ii) calculate MOS. To the best of our knowledge, this is the first time that LASSO is employed in video quality estimation problems. From the obtained experimental results we confirmed that this is a very efficient tool for feature selection, while producing accurate quality estimations through the use of sparse models, at the same time. As a baseline, Ridge regression [11,12,13] is used, which offers similar performance as LASSO, but requires a significantly larger number of features.

4. The proposed models exhibit considerably high performance as gauged by different statistical measures. Particularly, they offer impressively high accuracy, nearly perfect monotonicity, and very low estimation errors. In addition, a performance comparison of the proposed approaches is made with a number of related studies. Moreover, the performance statistics for two FR metrics that are oriented towards measuring video quality of digital video systems are explored, namely Perceptual Evaluation of Video Quality (PEVQ) [44] and Video Quality Metric (VQM) [45] are also presented for comparison. A close inspection of the results reveals that our proposed models offer competitive performance with that of FR metrics, while the comparative advantage over the related works is apparent in terms of all used performance measures.

5.3 Features Related to Perceptual Video Quality

In this section, we describe the video features that we have used in order to model the impact of various impairments on video quality. These features are related to video content characteristics, signal factors, error factors, motion factors, as well as to the effectiveness of the error concealment. Besides description
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

of these features with respect to their selection, we make an analysis about their suitability in estimating perceptual video quality and check for any redundancies in them. Table 5.1 summarizes the features, their type, and the attributes through which they can be related to video quality estimation.

5.3.1 Examined Features

Table 5.1: Description of the examined features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Type</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Intra[%]</td>
<td>The percentage of I coded MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>2.</td>
<td>16 × 16 MB[%,]</td>
<td>The percentage of MBs of size 16 × 16 in an I slice.</td>
<td>NR</td>
</tr>
<tr>
<td>3.</td>
<td>16 × 4 MB[%,]</td>
<td>The percentage of MBs of size 16 × 4 in an I slice.</td>
<td>NR</td>
</tr>
<tr>
<td>4.</td>
<td>InSlice[%]</td>
<td>The percentage of I coded MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>5.</td>
<td>PP[%]</td>
<td>The percentage of P coded MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>6.</td>
<td>PSkip[%,]</td>
<td>The percentage of P MBs coded as PSkip in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>7.</td>
<td>P16 × 16[%,]</td>
<td>The percentage of P MBs coded with no sub-partition of MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>8.</td>
<td>P8 × 8[%,]</td>
<td>The percentage of P MBs coded with 8 × 8 partition of MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>9.</td>
<td>P8 × 8b[%,]</td>
<td>The percentage of P MBs coded with 8 × 8 in a sub-partition of MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>10.</td>
<td>P4 × 4[%,]</td>
<td>The percentage of P MBs coded with 4 × 4 sub-partition of MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>11.</td>
<td>P4 × 8[%,]</td>
<td>The percentage of P MBs coded with 4 × 8 or 8 × 4 sub-partition of MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>12.</td>
<td>P4 × 8b[%,]</td>
<td>The percentage of P MBs coded with 4 × 8 or 8 × 4 in a sub-partition of MBs in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>13-20.</td>
<td>B modes</td>
<td>B modes that correspond to the same features as given in features 5 to 12, but for B coded MBs.</td>
<td>NR</td>
</tr>
<tr>
<td>21.</td>
<td>ΔMVx</td>
<td>The average measures of motion vector difference values for x direction in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>22.</td>
<td>ΔMVy</td>
<td>The average measures of motion vector difference values for y direction in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>23.</td>
<td>avg(MVx),</td>
<td>The average measures of motion vector values for x direction in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>24.</td>
<td>avg(MVy),</td>
<td>The average measures of motion vector values for y direction in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>25.</td>
<td>MVx[a,</td>
<td></td>
<td>The percentage of motion vector values equal to zero for x direction in a slice.</td>
</tr>
<tr>
<td>26.</td>
<td>MVy[a,</td>
<td></td>
<td>The percentage of motion vector values equal to zero for y direction in a slice.</td>
</tr>
<tr>
<td>27.</td>
<td>MotionIntensity 1 Defined as:</td>
<td>Defined as:</td>
<td>NR</td>
</tr>
<tr>
<td>28.</td>
<td>MotionIntensity 2 Defined as:</td>
<td>Defined as:</td>
<td>NR</td>
</tr>
<tr>
<td>29.</td>
<td>avg(MVx),</td>
<td>The average measures of absolute value of motion vector for x direction in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>30.</td>
<td>avg(MVy),</td>
<td>The average measures of absolute value of motion vector for y direction in a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>31.</td>
<td>MotionIntensity 3 Defined as:</td>
<td>Defined as:</td>
<td>NR</td>
</tr>
<tr>
<td>32.</td>
<td>MotionIntensity 4 Defined as:</td>
<td>Defined as:</td>
<td>NR</td>
</tr>
<tr>
<td>33.</td>
<td>NonStill</td>
<td>Boolean. True, if a slice includes motion.</td>
<td>NR</td>
</tr>
<tr>
<td>34.</td>
<td>HighMot</td>
<td>Boolean. True, if a slice includes high motion level.</td>
<td>NR</td>
</tr>
<tr>
<td>35.</td>
<td>MaxResEngy</td>
<td>The maximum residual energy over all the MBs of a slice.</td>
<td>NR</td>
</tr>
<tr>
<td>36.</td>
<td>LostSinFrm</td>
<td>Number of lost slices in a frame.</td>
<td>NR</td>
</tr>
<tr>
<td>37.</td>
<td>Height</td>
<td>Vertical location of the lost slice within a frame.</td>
<td>NR</td>
</tr>
<tr>
<td>38.</td>
<td>SpatialExtend</td>
<td>Number of consecutive lost slices in a frame.</td>
<td>NR</td>
</tr>
<tr>
<td>39.</td>
<td>SpatialExtend2</td>
<td>Boolean. True, if SpatialExtend = 2.</td>
<td>NR</td>
</tr>
<tr>
<td>40.</td>
<td>ErrorFrm</td>
<td>Boolean. True, if TMDDR = 1.</td>
<td>NR</td>
</tr>
<tr>
<td>41.</td>
<td>DistToRef</td>
<td>Distance in frames between the current frame and the reference frame used for concealment. Based on the GOP pattern, P frames are concealed using images 3 frames ago, while I frames and B frames are concealed using images 1 frame ago.</td>
<td>NR</td>
</tr>
<tr>
<td>42.</td>
<td>FarConceal</td>
<td>Boolean. True if [DistToRef] ≥ 3.</td>
<td>NR</td>
</tr>
<tr>
<td>43.</td>
<td>SigMean</td>
<td>The mean of the slice luminance.</td>
<td>RR</td>
</tr>
<tr>
<td>44.</td>
<td>MeanMSE</td>
<td>The mean MSE, over all MBs of a slice.</td>
<td>RR</td>
</tr>
<tr>
<td>45.</td>
<td>MaxMSE</td>
<td>The maximum MSE, over all MBs of a slice.</td>
<td>RR</td>
</tr>
<tr>
<td>46.</td>
<td>MeanSSIM</td>
<td>The mean SSIM, over all MBs of a slice.</td>
<td>RR</td>
</tr>
<tr>
<td>47.</td>
<td>MinSSIM</td>
<td>The minimum SSIM, over all MBs of a slice.</td>
<td>RR</td>
</tr>
</tbody>
</table>

In H.264/AVC based coding, several coding modes are typically dependent
5.3. Features Related to Perceptual Video Quality

on the content of a video. Mainly, the coding starts with the prediction of one part (block) of a video frame from its adjacent frames so as to eliminate any temporal redundancies. The first frame is intra (I) coded, followed by a predetermined sequence of forward predictive (P) and bi-directional predictive (B) frames, with a periodic recurrence of I frames if required. These predictions can be applied on an MB, i.e., a $16 \times 16$ block of pixels, or on its sub-sized blocks. The available information regarding these coding modes provides an estimation of the structural content of a video. The features that we computed from the lossy bitstream and can be grouped in this category are listed from 1 to 20 in Table 5.1. In addition, inter frame prediction, which takes advantage of the temporal redundancy between neighboring frames, involves the determination of motion vector information. This information can be used to estimate the relative motion found in the blocks of different frames of a video. Besides using the absolute values of the motion vectors, a number of related statistics were computed so as to represent the motion content of a video, as listed from 21 to 32 in Table 5.1. Except for the newly proposed features 13 – 20 and 31 – 32, others were inspired from the study presented in [26].

Driven by the fact that a packet loss is significantly less visible in still video scenes [29], we propose the use of feature 33, in order to define if a video slice includes motion or not. Using the motion vector magnitude values as they were computed from feature 28, we assume that a slice includes motion (NotStill=1), if its magnitude value is greater than $1/10^{th}$ of the highest magnitude value of all slices. Similarly, we assume that a slice includes high levels of motion (feature 34) if its magnitude value is greater than $8/10^{th}$ of the highest magnitude value of all slices. Additionally, features 35-36 represent the maximum and mean residual energy over all the MBs of a slice, where the residual energy for an MB is computed as the sum of squares of its transform coefficients.

Continuing with the features 37-45 of Table 5.1, they capture the effect of a packet loss in a video sequence, under various aspects. They are all computed from the lossy bitstream, except for the feature 47, which is calculated from the reconstructed video sequence after error concealment. Specifically, the proposed feature 37 and features 38-42 [29] model the impact of a packet loss based on its frequency, location, duration etc. These features can capture various aspects of slice loss, such as the vertical location of the lost slice in a frame, represented by feature 38. Its use for quality estimation is motivated by the fact that a lost slice in the middle of a frame can have different perceptual impact as compared to a lost slice in the top or bottom of a video frame. Moreover, features 43-44 are related to the concealment strategy applied to the decoder and
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

particularly, they deal with the distance from the frame that is used as reference for the concealment of a frame impaired with a slice loss. Thus, these features take into account the considered Group Of Pictures (GOP) structure and size. The last feature of this category is the Slice Boundary Mismatch (SBM), that is feature 45 of Table 5.1. Having detected the location of the lost slice, we applied the metric as it is described in [46] with the goal of capturing the mismatch on the boundaries between correctly received and concealed slices in the decoded frames, on a pixel-by-pixel basis. Thus, SBM captures the error introduced after applying an error concealment at the decoder.

Lastly, features 46-51 of Table 5.1 are calculated on a pixel-by-pixel basis, by using the compression-impaired and compression-and-network-impaired versions of a video. Features 48-51 model the Mean Squared Error (MSE) and SSIM metrics, which are commonly used in order to characterize the error amplitude and perceptual quality. In the current study, we precompute the MSE and SSIM values for each MB at the server side and keep the maximum MSE and minimum SSIM values over all MBs in a slice. Afterwards, all MSE and SSIM values are averaged to obtain a representative value for each of them, over each slice, and the resulting values are next sent to the client’s side. Thus, once it is known which slices are actually lost, we are able to know the corresponding MSE and SSIM values. This process of precomputing and transmitting the values from the server to the client renders these features of RR type [29].

With regard to comparison of the proposed approach of RR-based VQA, our method has some advantages over the standardized RR model called ITU-T J.342 [47] in the following ways. J.342 is based on the edge PSNR measurement, which is performed on the edge pixels of the video being transmitted over the ancillary channel. In our case, it is required to compute MSE and SSIM values for a sequence and hence, it may require less bandwidth. Our RR features are not dependent on the video content; on the other hand, edge pixels may vary for different contents (spatial details and frame resolution etc.), requiring less or more bandwidth.

5.3.2 Suitability of the Features

The set of features described above was extracted from the test-stimuli of the Ecole Polytechnique Fédérale de Lausanne (EPFL) and Politecnico di Milano (PoliMi) database [48]. The original SouRCe (SRC) videos were selected for the representation of a variety of spatiotemporal perceptual information, as suggested by ITU-T Rec. P.910 [49]. The selected SRCs were in raw progres-
5.3. Features Related to Perceptual Video Quality

The video sequences comprising the EPFL-PoliMi's database of Common Intermediate Format (CIF) resolution are "Mother", "Foreman", "Paris", "News", "Mobile", "Hall", each of 298 frames at 30 frames per second (fps) and of 4CIF resolution are "Harbour" and "Soccer" of 298 frames at 30 fps, "Parkjoy", "Crowdrun" and "Ducktakeoff" of 250 frames at 25 fps and "Ice" of 238 frames at 30 fps. For each video sequence, a full row of MBs was coded as a separate slice, while the bitstreams of the coded videos were impaired by a PLR of 0.1%, 0.4%, 1%, 3%, 5% and 10%. At the decoder, Motion Compensated Error Concealment (MCEC) was applied. It should be noted that this database also includes the MOS values as they were collected after subjective experiments separately conducted at EPFL and PoliMi. Further details on the generation of this dataset as well as the testing conditions can be found in [48].

It is worth mentioning that the features employed in this study were examined for possible redundancies. To this end, we present Fig. 5.1, which includes the cross-correlation of the features listed in Table 5.1. The line of white rectangles on the main diagonal represents the correlation of a feature with itself and an inspection of either upper or lower diagonal is adequate to study the cross-correlation of a feature with all the other features. Interestingly, most of the features are largely uncorrelated with the others; only a few of them are considerably correlated because of the similarities in the method used for their computation. For example, the motion intensity-based features appear to have relatively high correlation with the feature "HighMot" and it is expected, because they have been computed using the values of motion vectors. However, as discussed in [25], it might be important to keep some features, even if they are correlated, due to their relation to perceptual quality. Therefore, we chose to do not eliminate any feature from our dataset, as the issue of multicollinearity can be dealt with using the regression techniques employed in this study. Nonetheless, we preferred to perform feature selection based on their impact on MOS estimations.

Besides the aforementioned motivations on the selection of features in terms of their relevance towards perceptual quality, we also performed a statistical analysis to determine the suitability of these feature for developing an estima-
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

Figure 5.1: Cross-correlation of the examined features as mentioned in Table 5.1

Table 5.2: Linear fitting of the examined video features with subjective MOS.

<table>
<thead>
<tr>
<th>Fitting Target</th>
<th>Reduced-Reference Features</th>
<th>No-Reference Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>R² = 0.965, RMSE = 0.279, p-value = 5.1e-53</td>
<td>R² = 0.884, RMSE = 0.497, p-value = 2.59e-32</td>
</tr>
</tbody>
</table>

To this end, we constructed a least-squares fit (fitlm function in Matlab) between the feature values of all the examined videos and the subjective MOS. With reference to Table 5.1 for the specifications of features as of RR or NR type, Table 5.2 presents the fitting statistics. It can be seen that the coefficient of determination (R²) and the Root Mean Squared Error (RMSE) are extremely good, both for the RR and NR models. The p-values for each case is far below 0.05, which is a threshold value used to indicate that the model features are statistically significant within 95% confidence interval. Overall, the statistical evidence provided in Table 5.2 confirms the suitability of the considered features for building RR and NR video quality estimations models.
5.3.3 Pre-processing of the Features

In the used test-stimuli, a slice of a video frame corresponds to a packet. Therefore, considering the impact of a packet loss in terms of data loss on the test-stimuli, it is noted that an integral number of slices are also lost as a result of a packet loss event. In light of this, in this study, the features that are related to the occurrence of a packet loss were computed at slice level. On the other hand, some features, such as related to motion vectors, are more suitably computed at MB level. Hence, we found it reasonable to follow a bottom-up approach for computing most of these features at MB level and subsequently, an average value was obtained at slice level. Henceforth, we computed the average values of the slice level features to obtain their values at frame level.

Moreover, the frame-level feature values were averaged further to obtain their values at video sequence level. Instead of using simple arithmetic mean for frame-level data to video-level data conversion, we used Minkowski summation [51], as we investigated that it promotes overall performance of the estimation models. It is computed using the following expression:

\[
Minkowski_{\text{summation}} = \sqrt[p]{\frac{\sum f_i^p}{N}}, \tag{5.1}
\]

where \(N\) is the total number of frames in a video, \(f_i\) represents the feature value for the \(i^{th}\) frame, and \(p\) is the empirically determined Minkowski exponent. It can be noted that for an exponent value of \(p = 1\), the arithmetic mean is obtained.

5.4 Video Quality Estimation Using Linear Regression Models

The problem of perceptual video quality estimation based on a set of quality-relevant features is solved by building computational models that take the given set of feature values as input and produce appropriate quality estimates. The choice of a particular solution to be used for regression depends upon the requirements of the problem under consideration as well as the tradeoff preferences between the complexity and performance of a method. However, the theory associated with linear regression is well-understood and allows for the construction of different types of easily-interpretable, stable and sparse regression models. In the current study, we propose the use of LASSO [8, 9, 10], due
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

to its ability to perform feature selection and, based on the reduced set of features, to provide accurate MOS estimations. For the sake of comparison with the MOS estimations given by LASSO, we apply Ridge regression [11, 12, 13] that does not perform any kind of feature selection. Ridge is an extension of the Ordinary Least Squares (OLS) regression method and is able to improve the OLS estimates by allowing a little bias in order to reduce the variance of the estimated values, offering a good generalization capability to unseen data. Practically, it solves the following minimization problem:

$$\min_{\mathbf{w}} \left( \frac{1}{2} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \phi(x_i))^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right).$$

(5.2)

In this equation, the vector $\mathbf{y}$ includes the measured quality values for all $n$ observations, which is the total number of videos of the test-stimuli, and $\mathbf{w}$ is an $m$-by-1 vector of regression coefficients, including the intercept. The basis function $\phi(x_i)$ is an $m$-by-1 vector at observation $x_i$, which includes the values for all examined features for a particular video sequence. The shrinkage parameter $\lambda$ is a nonnegative regularization parameter; it shrinks regression coefficients values towards zero. Therefore, Ridge attempts to tradeoff the goodness-of-fit, as it is described by the first term of Eq. (5.2), and the penalty, as it is described by the second term of the same equation.

In a more compact form, Eq. (5.2) can be written as:

$$\mathbf{w} = (\Phi^\top \Phi + \lambda \mathbf{I})^{-1} \Phi^\top \mathbf{y},$$

(5.3)

where $\Phi$ is the matrix of observations and $\mathbf{I}$ is the identity matrix. Therefore, a number of biased estimators $\mathbf{w}$ are obtained by augmenting the diagonal matrix $\Phi^\top \Phi$ with a small positive quantity, as indicated by the $\lambda$ value. In this way, the system behaves more like an orthogonal system and regression coefficient values with smaller MSEs are obtained.

5.4.1 Least Absolute Shrinkage and Selection Operator

Feature selection is useful when a collection of input features is available, from which we expect to select a small subset for efficient estimation of a response variable, e.g., the perceptual quality of a video. Nonetheless, Ridge does not perform any feature selection (no regression coefficient is set to zero), implicating the risk of harming the estimations, when irrelevant or noisy features
5.4. Video Quality Estimation Using Linear Regression Models

are employed. In this context, LASSO can be used for both feature selection and estimation of regression coefficients. Similarly to Ridge, LASSO is able to effectively address with possible issues that arise when $\Phi^\top \Phi$ is not of full rank and thus, it is infeasible to be inverted using the OLS method. However, unlike Ridge, it has the benefit of not only shrinking some coefficients close to zero, but also setting some others equal to zero, offering feature selection and producing interpretable models. Thus, LASSO combines the stability of Ridge and interpretability of subset selection.

Practically, it minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Specifically, for a given nonnegative $\lambda$ value, it solves the following minimization problem:

$$\min_w \left( \frac{1}{2} \sum_{i=1}^{n} (y_i - w^\top \phi(x_i))^2 + \frac{\lambda}{2} \sum_{j=1}^{m} |w_j| \right). \tag{5.4}$$

Unlike Ridge, the regression coefficients $w$ for the LASSO methodology have no closed form and the solution involves quadratic programming techniques using convex optimization. The tuning parameter $\lambda$ controls the amount of regularization. As $\lambda$ is increased, an increasing number of regression coefficients become equal to zero, while for $\lambda = 0$, no shrinkage is obtained.

5.4.2 Model Learning

We standardized the values of the input features by calculating their $z$-score values; that is we subtracted the mean from each feature vector and the obtained values were divided by the standard deviation of the particular feature vector. The available data used for the model training and testing consisted of 144 sequences. Specifically, for each of the 12 SRC videos, 12 different realizations of a packet-loss environment were simulated.

Before validating the robustness of our estimation models, we ensured a clear distinction between the training and test data such that no content is common in both of the sets. The test set comprised all distorted versions of one SRC sequence (12 sequences) and the training set comprised of the data of test-stimuli generated from the distorted versions of 11 SRCs (132 sequences). This process of splitting the dataset into training and test sets was iterated so that each impaired sequence set from each SRC takes its place on the test set. This procedure is the well-known $k$-fold Cross-Validation (CV) [52] (in our case 12-fold CV), where the data is partitioned into $k$ equally sized subsets and an
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

iterative procedure is repeated $k$ times such that $k - 1$ subsets are used for training and the remaining one subset is used for testing (validation).

Once the training data was selected, the next important step was to find the optimal value of the regularization parameter $\lambda$ in Eqs. (5.2) and (5.4). For this purpose, we applied nested 10-fold CV on the training data. From all nine sets included in the training set, the regression coefficient values were calculated and they were applied to the one set left for validation. A number of 100 different values for the $\lambda$ parameters of Eqs. (5.2) and (5.4) were tested and the MSE between the subjective and estimated MOS values were calculated. It is worth noting that the same 100 $\lambda$ values were tested in each fold of both of the Ridge and LASSO methods. Therefore, we obtained an array of $100 \times 10$ values of MSE, where each row corresponds to MSE values for different $\lambda$ setups and each column to each different fold. Afterwards, we computed the average MSE values for each considered $\lambda$ over all folds and a vector of $100 \times 1$ MSE values emerged. Based on the minimum MSE value of this vector, we selected the $\lambda$ value among all 100 $\lambda$ values with the same index. Thus, using the chosen $\lambda$ values for Ridge and LASSO, we trained our models and obtained values for the regression coefficients.

Algorithm 1 Model Development

```
loop
  if (a SRC is not tested) then
    Split the dataset in training set and test set, s.t. the test set includes all impaired versions of the same SRC.
    Execute exclusively on the training set.
    b. Determine the optimal $\lambda$ values of Eqs. (5.2) and (5.4).
    c. Using the optimal $\lambda$ values, train the whole training set of Ridge and LASSO models.
    d. Get the regression coefficient estimates.
    Apply the regression coefficient estimates on the test set.
    Get video quality estimations.
    Evaluate Ridge and LASSO model performance.
  end if
end loop
```

The obtained regression coefficient values were applied to the data of the test set in order to get the MOS estimations. Using the estimated values, we were able to evaluate the model's performance in comparison with the subjective MOS values. Algorithm 1 summarizes the methodology adopted in order to develop our proposed models.
5.5 Experimental Results

This section presents the results obtained after evaluating the performance of the proposed models and a comparison with related approaches is also reported. In order to calculate the performance, the following measures, as recommended by Video Quality Experts Group (VQEG) [53] were used.

- **Accuracy**: The Pearson linear correlation coefficient (PCC) is used to describe the accuracy of the estimation.

- **Monotonicity**: The Spearman Rank Order Correlation Coefficient (SROCC) is used to describe the monotonicity of the estimation.

- **Estimation Error**: The RMSE is used to describe the error of the estimation.

The values of PCC and SROCC lie in the range $[-1, 1]$ where values closer to 1 represent high positive correlation. Regarding Minkowski summation, there is no general rule for a suitable value of the exponent $p$ of Eq. (5.1). In light of this, we evaluated a vector of $p$ values that lies in the range $[0.1, 15]$ with intervals of 0.2. Based on the observed performance with respect to PCC for each exponent, the optimal value of $p$ was used in the simulations. In addition, as it was also referred in the previous section, in order to determine the optimal $\lambda$ values of Eqs. (5.2) and (5.4), we performed 10-fold CV, using a set of 100 possible $\lambda$, different for each tested content. For both regression methods, we tried the same 100 values of $\lambda$ for the same test content. All tested $\lambda$ values were in the range of slightly above 0 and slightly above 1.

5.5.1 Results

As a result of the test setup described in Section 5.3, a number of simulations were performed for the aforementioned set of video features, using the MOS values collected by both EPFL and PoliMi. It holds that subjective MOS values are usually compressed at the ends of the rating scale (0 and 5), while this is not the case for objective video quality models that are unable to mimic this weakness of subjective data. Therefore, following the VQEG report on validation of objective video quality models [54], a third order monotonic mapping function was applied on the estimated values of our models before the computation of the performance measures.
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

Table 5.3 presents the obtained results for both RR and NR models using Ridge and LASSO, when the PoliMi MOS values are used. Besides the performance measures, the optimal \( \lambda \) values and the number of used features for each proposed model are also mentioned. Each cell of the table cites the results when a specific SRC sequence is used to generate its 12 impaired versions and the bottom cell shows the arithmetic mean (average) of the performance over all the SRCs (of all cells) with resolution CIF and 4CIF separately. In the same table, related performance of PEVQ and VQM metrics is also mentioned. PEVQ is an FR metric that is a part of the ITU-T Recommendation J.247 [44]. VQM is also an FR metric that has been largely adopted in research community for taking quality estimates [45]. In an effort to be fair when comparing the estimated and subjective MOS, we scaled PEVQ and VQM values in the range [0, 5]. In addition, since for the VQM the smaller the value the better the video quality, we “reversed” these values to follow the trend of MOS. It holds that comparing RR and NR models against FR metrics is a challenging task, as FR metrics have far more data to process for estimating quality. Nonetheless, as it turns out, the proposed models perform equally well, or somewhat better than the considered FR metrics. The advantage of our proposed models is more evident mainly in terms of fairly lower values of estimation error.

One salient aspect of comparing the performance of Ridge and LASSO is the level of accuracy and sparsity offered by each solution. Observing the performance results of Table 5.3, we do not confirm an advantage of a particular regression method over the other, since their statistics is very similar in both CIF and 4CIF resolutions. However, for all examined cases, it can be construed that LASSO models use far less than half of the features employed by Ridge for making quality estimations, while this claim is more evident in the NR case. Hence, LASSO regression can be used as an efficient solution for video quality estimation as well as feature selection, maintaining impressively good performance statistics.

On the basis of statistics, shown in Table 5.2, we expected that RR models may perform better than NR models. In fact, from the individual results for each SRC as well as the average results (bottom cell of Table 5.3), it is observed that RR models have slightly better performance than the corresponding NR models, as indicated through the correlation coefficients as well as the estimation errors, regardless of the regression method used. The superiority of RR models is more prominent in the case of some SRCs including “Mobile”, “Mother”, and “Parkjoy”, while this is not the case for the “Hall” and “Ice” SRCs, where NR Ridge keeps better statistics compared to RR Ridge. This is
5.5. Experimental Results

Table 5.3: Performance of the proposed models and reference FR metrics for MOS collected by PoliMi [48].

<table>
<thead>
<tr>
<th>CIF</th>
<th>Method</th>
<th>SRC</th>
<th># Features</th>
<th>PCC</th>
<th>SROCC</th>
<th>RMSE</th>
<th>SRC</th>
<th># Features</th>
<th>PCC</th>
<th>SROCC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RR Ridge</td>
<td>Foreman</td>
<td>0.0717</td>
<td>45</td>
<td>0.987</td>
<td>0.979</td>
<td>0.221</td>
<td>0.0149</td>
<td>15</td>
<td>0.990</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.983</td>
<td>0.963</td>
<td>0.792</td>
<td>–</td>
<td>–</td>
<td>0.966</td>
<td>0.986</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.971</td>
<td>0.979</td>
<td>1.548</td>
<td>–</td>
<td>–</td>
<td>0.991</td>
<td>0.986</td>
<td>0.343</td>
</tr>
<tr>
<td></td>
<td>RR Ridge</td>
<td>Hall</td>
<td>0.1417</td>
<td>45</td>
<td>0.940</td>
<td>0.965</td>
<td>0.392</td>
<td>0.0011</td>
<td>29</td>
<td>0.987</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.944</td>
<td>0.818</td>
<td>0.573</td>
<td>–</td>
<td>–</td>
<td>0.942</td>
<td>0.930</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.940</td>
<td>0.895</td>
<td>1.064</td>
<td>–</td>
<td>–</td>
<td>0.988</td>
<td>0.996</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>NR Ridge</td>
<td>Foreman</td>
<td>1.0797</td>
<td>45</td>
<td>0.991</td>
<td>0.986</td>
<td>0.562</td>
<td>1.0973</td>
<td>45</td>
<td>0.954</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.983</td>
<td>0.979</td>
<td>2.108</td>
<td>–</td>
<td>–</td>
<td>0.978</td>
<td>0.972</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.980</td>
<td>0.979</td>
<td>1.044</td>
<td>–</td>
<td>–</td>
<td>0.983</td>
<td>0.975</td>
<td>0.343</td>
</tr>
<tr>
<td></td>
<td>NR Ridge</td>
<td>Hall</td>
<td>1.0591</td>
<td>45</td>
<td>0.991</td>
<td>0.986</td>
<td>0.598</td>
<td>1.0968</td>
<td>45</td>
<td>0.974</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.944</td>
<td>0.818</td>
<td>0.573</td>
<td>–</td>
<td>–</td>
<td>0.942</td>
<td>0.930</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.940</td>
<td>0.895</td>
<td>1.064</td>
<td>–</td>
<td>–</td>
<td>0.988</td>
<td>0.996</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>NR Ridge</td>
<td>Mobile</td>
<td>0.0224</td>
<td>9</td>
<td>0.941</td>
<td>0.965</td>
<td>0.427</td>
<td>0.0213</td>
<td>23</td>
<td>0.978</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.969</td>
<td>0.937</td>
<td>0.404</td>
<td>–</td>
<td>–</td>
<td>0.960</td>
<td>0.930</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.940</td>
<td>0.895</td>
<td>1.064</td>
<td>–</td>
<td>–</td>
<td>0.988</td>
<td>0.996</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>NR Ridge</td>
<td>Mobile</td>
<td>0.0577</td>
<td>51</td>
<td>0.993</td>
<td>0.986</td>
<td>0.145</td>
<td>0.0013</td>
<td>29</td>
<td>0.979</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.944</td>
<td>0.818</td>
<td>0.573</td>
<td>–</td>
<td>–</td>
<td>0.942</td>
<td>0.930</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.940</td>
<td>0.895</td>
<td>1.064</td>
<td>–</td>
<td>–</td>
<td>0.988</td>
<td>0.996</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>NR Ridge</td>
<td>Mobile</td>
<td>0.0141</td>
<td>45</td>
<td>0.935</td>
<td>0.944</td>
<td>0.344</td>
<td>0.0175</td>
<td>45</td>
<td>0.970</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.944</td>
<td>0.444</td>
<td>0.361</td>
<td>–</td>
<td>–</td>
<td>0.933</td>
<td>0.930</td>
<td>0.446</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.940</td>
<td>0.895</td>
<td>1.064</td>
<td>–</td>
<td>–</td>
<td>0.988</td>
<td>0.996</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>NR Ridge</td>
<td>Mobile</td>
<td>0.0224</td>
<td>9</td>
<td>0.928</td>
<td>0.944</td>
<td>0.361</td>
<td>0.0007</td>
<td>23</td>
<td>0.969</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.969</td>
<td>0.937</td>
<td>0.404</td>
<td>–</td>
<td>–</td>
<td>0.960</td>
<td>0.930</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.940</td>
<td>0.895</td>
<td>1.064</td>
<td>–</td>
<td>–</td>
<td>0.988</td>
<td>0.996</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>NR Ridge</td>
<td>Mobile</td>
<td>0.0250</td>
<td>12</td>
<td>0.938</td>
<td>0.965</td>
<td>0.276</td>
<td>0.0008</td>
<td>15</td>
<td>0.979</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>PEVQ</td>
<td>–</td>
<td>–</td>
<td>0.969</td>
<td>0.937</td>
<td>0.404</td>
<td>–</td>
<td>–</td>
<td>0.960</td>
<td>0.930</td>
<td>0.419</td>
</tr>
<tr>
<td></td>
<td>VQM</td>
<td>–</td>
<td>–</td>
<td>0.940</td>
<td>0.895</td>
<td>1.064</td>
<td>–</td>
<td>–</td>
<td>0.988</td>
<td>0.996</td>
<td>0.489</td>
</tr>
</tbody>
</table>

probably attributed to the fact that some of the employed features used in the RR case may be irrelevant or noise that harm the quality estimations. In contrast, when LASSO regression based model is applied for feature selection and MOS estimation, we do not observe such irregular cases. Therefore, summarizing the comparison between RR and NR models, we have to mention that the estimation accuracy in NR cases is very promising as well and hence, the proposed list of NR features (see Table 5.1) presents an acceptable solution for reference-free quality estimation for video transmissions over lossy networks.

In addition to the earlier mentioned detailed statistics, the overall performance of RR LASSO and NR LASSO models is shown as scatter plots in Fig. 5.2, for both CIF and 4CIF resolutions. The values of the “overal” performance as indicated in these plots are obtained by comparing the values of estimated and subjective MOS, when all examined sequences of a specific spatial resolution
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

are considered as a whole. The scatter plots not only indicate a very high overall performance in each case, but they also show that the quality estimation for “Mobile” CIF sequence seems to be difficult for NR LASSO, while RR LASSO manages its estimation in a more efficient way. Similarly, “Ice”, “Parkjoy”, and “Ducktakeoff” sequences of 4CIF resolution are a big challenge in the case of NR LASSO-based estimation. However, as it can be seen in Fig. 5.2 (d), for the case of RR LASSO, the MOS estimation is more precise for these sequences. Thus, the advantage of using an RR model over an NR model is more obvious from these plots.

5.5.2 Feature Selection in LASSO

Following the previous discussion about the feature selection capability of LASSO, it is intriguing to investigate which features are actually selected using the particular regression method. The values of the regression coefficients are considered as an indication of feature selection or not. Particularly, if the coefficient associated with a certain feature acquires a zero value, this means that the specific feature is excluded from the estimation process. On the contrary, a feature is selected, if it is assigned a non-zero regression coefficient value. As the values of the input features are normalized to the same scale, a higher value of a coefficient implies higher significance of the related feature, and vice versa. Moreover, the features that are associated with positive-signed coefficients are considered to cause an increase in quality if their values are increased. In contrast, the features associated with negative-signed coefficients are considered to decrease quality if their values are increased. Table 5.3 provides information on the number of selected features, while in order to save space we omit the presentation of the regression coefficient values. In the following, we examine feature selection with regard to the following two cases:

1) Mostly selected features: We consider that a feature is mostly selected, if it is chosen for more than half of the 12 different training sets considered in this study.

In the case of NR LASSO, the list of such features is: 8, 13, 22-24, 39, and 40 of Table 5.1. By examining the regression coefficient values of these features, it appears that partitioning of MBs into 8 × 16 or 16 × 8 sum-MBs can be relatively more significant for estimating the quality of a video. Moreover, we observe that the regression coefficient assigned to feature 8 is positive in all considered training sets. Hence, a preference for this partitioning can be favorable for perceptual quality. The percentage of bi-predictive coded blocks (feature 13) can
5.5. Experimental Results

<table>
<thead>
<tr>
<th>Subjective MOS</th>
<th>Estimated MOS</th>
<th>Overall Pearson Correlation Coefficient: 0.969</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit</td>
<td>45-degree</td>
<td>Foreman</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hall</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mobile</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mother</td>
</tr>
<tr>
<td></td>
<td></td>
<td>News</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Paris</td>
</tr>
</tbody>
</table>

(a) NR LASSO

<table>
<thead>
<tr>
<th>Subjective MOS</th>
<th>Estimated MOS</th>
<th>Overall Pearson Correlation Coefficient: 0.984</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit</td>
<td>45-degree</td>
<td>Foreman</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hall</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mobile</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mother</td>
</tr>
<tr>
<td></td>
<td></td>
<td>News</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Paris</td>
</tr>
</tbody>
</table>

(b) RR LASSO

<table>
<thead>
<tr>
<th>Subjective MOS</th>
<th>Estimated MOS</th>
<th>Overall Pearson Correlation Coefficient: 0.958</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit</td>
<td>45-degree</td>
<td>Crowdrun</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Duckstakeoff</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harbour</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Parkjoy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Soccer</td>
</tr>
</tbody>
</table>

(c) NR LASSO

<table>
<thead>
<tr>
<th>Subjective MOS</th>
<th>Estimated MOS</th>
<th>Overall Pearson Correlation Coefficient: 0.976</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit</td>
<td>45-degree</td>
<td>Crowdrun</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Duckstakeoff</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Harbour</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Parkjoy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Soccer</td>
</tr>
</tbody>
</table>

(d) RR LASSO

Figure 5.2: Overall performance of the proposed NR and RR LASSO models for CIF resolution in (a), (b) and for 4CIF resolution in (c), (d).
5. **Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models**

have a negative impact on the perceptual quality (in the case of CIF videos), as indicated by the sign of the associated regression coefficient. The importance of motion-related features (22-24) confirms many earlier studies (e.g., see [23]) that suggest a relationship between video motion characteristics and perceptual quality. Interestingly, features 22-24 have non-zero regression coefficient values in the vast majority of the different training sets. Lastly, features 39 and 40 account for the impact of packet losses and are very crucial in estimating perceptual quality. In fact, they keep negative values in all 12 training sets.

In the case of RR LASSO, the list of such features is: 8, 23, 26, 39, 40, 45, 49, and 51 of Table 5.1. The reappearance of some of the features (8, 23, 39, 40) that were selected in the case of NR LASSO as well, further highlights their significance. However, the relative importance of these features appears to have changed with the inclusion of the RR features in the model. SBM (feature 45) is assigned a negative-signed regression coefficient in all different training sets, confirming that slice-boundary-mismatch has a negative impact on video quality. Out of the considered six RR features, maximum MSE and particularly minimum SSIM are found to be more crucial as compared to the other features, based on the magnitude of their regression coefficients.

2) *Features selected from all training sets:* In this case, we enlist the selected features, after performing a union operation on the lists of the selected features for all 12 training sets. In the case of NR LASSO, the list of such features is: 1, 3, 4, 5, 7-9, 12-15, 20-24, 28, 30, 33, 34, 36, 38-43, and 45 of Table 5.1. This list contains 28 features being selected out of the total 45 features. Besides the significance of the selected features, this list also shows that some features, including the percentage of blocks of size 4 × 4 (feature 2), the number of lost slices in a frame (feature 37), and the distance from the reference frame used for concealment being greater than 3 (feature 44) have a negligible impact on perceptual quality.

In the case of RR LASSO, the list of such features has a large overlap with the one of the NR case and the selected features are 34 (out of 51 in total). Specifically, this list includes the features: 1, 3, 5, 7-9, 12-15, 19, 20, 22-27, 33, 34, 36-41, 43, 45-51 of Table 5.1. It is instructive to note that the percentage of MBs coded as PSkip (feature 6), block sizes around 8 × 8 sub-partitioning for P and B modes (feature 10 and 18), and distance from the reference frame used for concealment being greater than 3 (feature 44) are not selected in either of the RR and NR LASSO models. Also, it worths mentioning that the motion-related features (motion vectors and motion intensity) are selectively employed
5.5. Experimental Results

by these models as some of them may be conveying redundant information (see Fig. 5.1 and related discussion).

### 5.5.3 Comparison with Related Work

In this part of the article, we compare the results produced in this study, with the results of existing publications that address video quality estimation problems in FR, RR, and NR modes. It is to be noted that Table 5.4 includes the results of the performance measures as they were calculated in “overall” fashion, that is considering estimated and subjective MOS values for all examined sequences as a whole.

From Fig. 8 (b) of the work presented in [36], we observe that the PCC and SROCC values of the EPFL-PoliMi video database [48] are between 0.85 and 0.95, for all proposed Q-mentioned FR metrics (Qvector, Qcsiq, Qtid, Qlive). These results were generated by training the aforementioned models on the CSIQ [55], TID [56] and LIVE [57] databases, while the same models were tested on the CIF sequences of EPFL-PoliMi database. Therefore, due to the fact that the models of [36] are of FR type and they have been trained on image databases only, the comparison with the results of that article is not completely fair. Nonetheless, for paper’s completeness we point out that, despite the fact that we propose RR and NR models, and thus the task of making estimations is more challenging as compared to a FR model, from Table 5.4 we infer that our NR models offer PCC values equal to or higher than 0.968 and SROCC values equal to or higher than 0.970 and our RR models offer 0.978 and 0.971 as the least values of PCC and SROCC.

Moreover, the performance of Fourier transform-based RR model proposed in [37] for its variant $Q_{\text{combined}}$ that offers the best results is compared against our proposed models in the same table (Table 5.4). We realize that the model presented in [37] can be considered more general as it is trained and tested on different media contents in contrast to the models of our work where test-stimuli from only one source was employed. Thus, this implies that the worse performance of the models in [37] for the CIF resolution sequences of the EPFL-PoliMi’s database [48] may be justified as they use a broader dataset for models’ development. However, from the provided experimental results, we confirm the superiority of both the RR and NR models that we propose, in all examined measures of performance. Interestingly, the estimation error of our models is nearly equal to half of the corresponding amount of $Q_{\text{combined}}$. 

173
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

Table 5.4: Comparison of the overall performance of the proposed models with related works.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Based on MOS values by PoliMi</th>
<th>Based on MOS values by EPFL</th>
<th>CIF resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NR Ridge</td>
<td>NR LASSO</td>
<td>FE-PLM [38]</td>
</tr>
<tr>
<td>PCC</td>
<td>0.974</td>
<td>0.970</td>
<td>0.95</td>
</tr>
<tr>
<td>SROCC</td>
<td>0.968</td>
<td>0.970</td>
<td>0.95</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.288</td>
<td>0.318</td>
<td>0.43</td>
</tr>
</tbody>
</table>

4CIF resolution

<table>
<thead>
<tr>
<th>Metric</th>
<th>Based on MOS values by PoliMi</th>
<th>Based on MOS values by EPFL</th>
<th>G.1070E [40]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NR Ridge</td>
<td>NR LASSO</td>
<td>SLR_{IP} + SLR_{B} [39]</td>
</tr>
<tr>
<td>PCC</td>
<td>0.960</td>
<td>0.954</td>
<td>0.91</td>
</tr>
<tr>
<td>SROCC</td>
<td>0.959</td>
<td>0.954</td>
<td>0.91</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.326</td>
<td>0.333</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Furthermore, another work that utilizes the EPFL-PoliMi database [48] and specifically, the CIF resolution sequences to assess the performance of the proposed NR metric is the one presented in [38]. In that work, the MOS collected by PoliMi are used, while the best proposed model is called “Frame-type and Error pattern dependent Packet-Loss model”, as denoted by “FE-PLM” in Table 5.4. It can be seen that our proposed NR models are better in terms of all examined statistics compared to those of [38]. One of the underlying reasons behind this difference in performance can be the fact that we use a variety of features to capture various characteristics of a video including the impact of packet losses. On the other hand, the models in [38] are based on the assumption that visual quality can always be exponentially related to the PLR which, in practice, may not hold in varying bitrates and different contents [58].

Similarly, the NR model presented in [39] was evaluated using MOS values collected at PoliMi [48]. In order to design the model, the authors assumed that PLR and MOS can be characterized by a two region piecewise linear relationship. Based on this assumption, a number of variants of the basic NR model were proposed, which differ mainly on the type of data used for estimating losses introduced by the network. The results that we considered from [39] are based on the quality estimation using the $SLR_{IP} + SLR_{B}$ model variant (based on slice loss rate of I/P slices and B slices) that offers the best results. The conclusion derived after looking at the results is that our models achieve competitive performance in terms of PCC and slightly lower RMSE compared to the model proposed in [39], despite the fact that a two region piecewise linear model is employed in [39] as opposed to the one region linear models proposed in this study.

Moreover, an enhanced version of the ITU-T Recommendation G.1070: Opinion model for video-telephony applications [41], called G1070E can be found
5.6 Conclusions

The estimation accuracy of the G1070E model is validated using the 4CIF resolution sequences of the database presented in [48], with the MOS data collected from the subjective tests, conducted both at the EPFL and PoliMi institutions. Specifically, in [40] the estimations models were trained on a large variety of CIF resolution sequences, other than those included in the EPFL-PoliMi’s database, which were compressed at various bitrates and were impaired with different PLRs. Comparing the performance of our proposed NR models with the G1070E [40] model, when the PoliMi MOS values are used, we easily perceive a clear advantage of our proposed models and considerably better performance in terms of all presented measures of performance. However, talking about models’ evaluation using the EPFL MOS values, we conclude that our models achieve a lower RMSE, although they keep slightly lower PCC and SROCC values.

Lastly, we studied the performance achieved by a genetic programming-based NR regression model presented in [42]. In that work, the authors validated the performance of their proposed model, by considering the video quality estimates and subjective MOS values together, for both the CIF and 4CIF resolution sequences of the EPFL-PoliMi’s database [48]. Accordingly in this case, our NR LASSO model offers PCC and SROCC values equal to 0.964 and 0.963 respectively, as compared to the corresponding values offered by [42], which are equal to 0.881 and 0.883, respectively.

5.6 Conclusions

In this study, we investigated a fairly large variety of video features for estimating video quality. These features include different attributes related to perceptual quality and encompass the impacts of coding and network impairments on H.264/AVC encoded sequences. Most of these features can be computed without any access to the original video and hence, they are applicable to design a NR model of quality estimation. The rest of the features can be pre-computed and sent to the client’s end for providing RR information of the original video.

Based on these features, we propose RR and NR models of quality estimation, by employing linear regression techniques namely Ridge and LASSO. Ridge was used to achieve a baseline performance when all the examined features are used and LASSO was investigated for its capability to offer similar performance as Ridge, but by employing a smaller number of features. The simulation results reveal that our proposed models are able to estimate sub-
5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models

jective MOS with reasonably high accuracy and fairly low errors in quality estimations, while they outperform many existing FR, RR, and NR techniques, used for video quality estimation.

An insight into the process of feature selection using LASSO is also provided. It is shown that certain features of NR type are relatively more significant for quality estimation, as they are selected in the cases of using both RR LASSO and NR LASSO models. It is also observed that a preference of a specific block partitioning type can be favorable for perceptual quality. The selection of features that are related to motion vectors is in line with many previous studies, where it is claimed that the motion contents of video sequences have an impact on perceptual quality.

In the future, this study can be extended to test the proposed models on higher resolution test-stimuli with more variety in compression and network-related features. In addition, in view of the recently approved standard of video compression, i.e., High Efficiency Video Coding (HEVC), the proposed models can also be applied on the HEVC encoded test-stimuli.

5.7 Bibliography


5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models


5. Perceptual Quality Estimation of H.264/AVC Videos Using Reduced-Reference and No-Reference Models


5.7. Bibliography


Part III
Part III

On Subjective Methods of Video Quality Assessment
Six

Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

This chapter has been published as:

Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

Muhammad Shahid, Amitesh Kumar Singam, Andreas Rossholm, and Benny Lövström

Abstract

Advancements in the video processing area have been proliferated by services that require low delay. Such services involve applications being offered at various temporal and spatial resolutions. It necessitates to study the impacts of related video coding conditions upon perceptual quality. But most of studies concerned with quality assessment of videos affected by coding distortions lack in variety of spatio-temporal resolutions. This paper presents a work done on quality assessment of videos encoded by state-of-the-art H.264/AVC standard at different bitrates and frame rates. Overall, 120 test scenarios for video sequences having different spatial and temporal spectral information were studied. The used coded bitstreams in this work and the corresponding subjective assessment scores have been made public for the research community to facilitate further studies.

6.1 Introduction

The recent advancement in digital imaging technology and availability of efficient transmission systems have resulted in a proliferation of videos more than ever before. Videos transmitted to and from mobile devices will account for 66% of the global mobile data traffic by 2014 as per forecasts [1]. Video services that have gained wide interest are so many and television broadcast, DVD, Blu-Ray, Mobile TV, Web TV etc. are some to name. One of the key characteristics of video services is the quality of experience (QoE) as observed by the end user. Quality of visual media can get degraded while capturing, storing, transmission, reproduction and display due to the distortions which might occur at any of these stages. Although many automatic (objective) methods of visual quality assessment (VQA) have been proposed and are in-use but the true judges of
6. **Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos**

the quality are humans as end users. Subjective assessment of video quality is done by following standardized recommendations for experimental set-up, lab environment, stimuli characteristics and the number of viewers.

Video Quality Expert Group (VQEG), formed in 1997, has been the principal body for conducting comparative study of objective methods by performing subjective assessment tests. The full reference television (FR-TV) project report of phase I [2] concluded that no objective measure of VQA can replace subjective VQA. The test data and the corresponding subjective mean opinion score (MOS) was released to public for facilitating research on VQA. In the phase II campaign of subjective tests [3], it was found that some of the candidates of objective VQA performed better than PSNR, the conventional quality measure still in use. However, the MOS and test data was not made accessible for everyone, hence researchers in the field of development of objective VQA can not use the VQEG phase II data for the verification of their algorithms. Even the data which was shared earlier constitutes video sequences of high resolutions like 720x576 @50 fps and 720x486 @60 fps tested in PAL and NTSC TV formats respectively. Some independent efforts of subjective VQA have also been made such as [4, 5, 6] but unfortunately the test stimuli and subjective scores are not shared to public.

Tremendous efforts have been made to develop objective methods of VQA which are easy to repeat and not so time consuming as subjective assessment. Objective methods are categorized as full reference (FR), reduced reference (RR) and no reference (NR) methods based on the reference information required. Examples of each type include [7], [8], [9] for FR, RR and NR respectively. Most of the objective methods estimate visual quality by quantifying spatial (intra frame) degradations such as blocking, blurring, ringing and temporal (inter frame) degradations such as jitter [10]. High amount of quantization and frame dropping or resolution reduction to meet rate requirements of limited capacity transmission sources generate the aforementioned artifacts. However, the objective methods of VQA have to validated through subjective VQA for the appraisal of their performance. To study the effect of varying bitrate and

![Figure 6.1: A snapshot of the test video sequences](image-url)

190
Figure 6.2: SI and TI plot computed for luminance component of selected videos

spatio-temporal conditions on the video quality, only limited work has been performed and as mentioned before, the shared data lacks in variety of resolution. To the best of authors’ knowledge, publically available databases of subjective VQA are only found from:

- [11], where the test stimuli for various coding conditions of H.264/AVC is limited.
- [12], that doesn’t address H.264/AVC coded videos.
- [6], that has shared MOS values for H.264/AVC coded data but the stimuli was not made public.
- [13], that has shared MOS values for H.264/AVC coded data but the shared stimuli was impaired by transmission error only.

Moreover, in most of the reported work, either the coding details are not shared completely or the video encoding is done using bi-predictive coding
6. Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

which makes the decoding process slower. Real time applications like videoconferencing and popular video services such as videos viewed on handheld devices require low delay in the coding process. In this paper, we present a study of subjective VQA on QCIF (176x144) and CIF (352x288) resolution videos encoded by H.264/AVC standard at various bitrate and frame rate conditions. To be applicable for low delay services, bi-predictive coding is not preferred. We believe that sharing this kind of data has fundamental role in facilitating comparison and benchmarking of objective methods of VQA. In the sequel, Section 6.2 presents details on the generation of test stimuli and the subjective assessment. Finally, Section 6.3 presents conclusive discussion on this work.

6.2 Subjective Assessment of the Test Stimuli

For a purposeful subjective VQA database, the test stimuli should constitute videos with varying amounts of spatial and temporal spectral information and the tests should be conducted by following standard recommendations. This section provides details of our approach on both of these considerations.

6.2.1 Generation of Test Stimuli

In order to characterize the spectral contents of a video sequence, spatial spectral information (SI) and temporal spectral information (TI) indices are used as suggested in the ITU T-recommendations [14]. SI and TI are calculated in the luminance plane of a video and the used formulae are given below.

\[ SI = \max_{\text{time}}[\text{std}_{\text{space}}[\text{sobel}(f_n)]] \]  \hspace{1cm} (6.1)

- \( \text{sobel}(f_n) \) is each frame at time n filtered with Sobel filter.
- \( \text{std}_{\text{space}} \) is standard deviation over pixels.
- \( \max_{\text{time}} \) is maximum value in time series.

\[ TI = \max_{\text{time}}[\text{std}_{\text{space}}[M(i,j)]] \]  \hspace{1cm} (6.2)

Where \( M(i,j) \) is difference of motion between pixel values in space for sequential frames in the luminance plane and \( (i,j) \) represent the pixel location. High textured videos have higher values of SI and higher motion content in a video sequence leads to higher TI value. Videos sequences selected for this work cover
6.2. Subjective Assessment of the Test Stimuli

![Graph showing MOS values for Akiyo and Soccer](image)

Figure 6.3: MOS Values for Akiyo and Soccer

a wide range in both of these indices as depicted in Figure 6.2 namely *Akiyo, News, Foreman, Crew, Soccer, and Football* without any audio signal. One selected frame from each video is shown in Figure 6.1. These videos were in raw (YUV) progressive format with 4:2:0 color space having 300 frames in QCIF and CIF resolutions. Thus using these 12 source (Source Reference Circuit, the SRC) files, 120 test sequences (Hypothetical Reference Circuit, the HRC) were generated at ten different bitrates and two different frame rates with JM reference software for H.264/AVC, available online at [15]. In particular, the baseline profile was employed that is suitable for low delay applications. Table 6.1 presents details on the generated test sequences. For subjective VQA, all 120 HRCs were scaled to CIF size at frame rate of 30 fps. *Bicubic interpolation* and *repeat frame* methods were used for up-scaling QCIF videos to CIF and 15 fps to 30 fps respectively.
6. **Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos**

Figure 6.4: Overall Average Score by Each Subject
6.2. Subjective Assessment of the Test Stimuli

Table 6.1: Description of Used Videos

<table>
<thead>
<tr>
<th>Spatial Resolutions</th>
<th>Video Sequences</th>
<th>Bit rates(kbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIF @ 30,15 frame rates</td>
<td>Akiyo</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>News</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Foreman</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Soccer</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td></td>
<td>Football</td>
<td>200,400,600,800,1000</td>
</tr>
<tr>
<td>QCIF @ 30,15 frame rates</td>
<td>Akiyo</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>News</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Foreman</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Crew</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Soccer</td>
<td>100,200,300,400,500</td>
</tr>
<tr>
<td></td>
<td>Football</td>
<td>100,200,300,400,500</td>
</tr>
</tbody>
</table>

Table 6.2: Pointers used in Fig. 6.5

<table>
<thead>
<tr>
<th>Sequence/Coding Condition</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akiyo</td>
<td>▽</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>Crew</td>
<td>▽</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>Football</td>
<td>▽</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>Foreman</td>
<td>▽</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>News</td>
<td>▽</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>Soccer</td>
<td>▽</td>
<td>×</td>
<td>○</td>
</tr>
</tbody>
</table>

6.2.2 Subjective Assessment

We have followed the recommendations given by ITU-R BT 500-12 [16] for performing our experiments of subjective VQA. Particularly, the method followed was single stimulus quality evaluation where a test video sequence is shown once without presence of any explicit reference, corresponds to the reality where users see only the processed version of a video [17] [18]. A flat LCD screen with non-glare surface treatment was used for displaying the video sequences. The used monitor had resolution 1440x900 with 5 ms response time and its color temperature was set at 6500K in sRGB mode. Other hardware includes a desktop HP system having 3 GHz AMD processor and 4 GB RAM. A comfortable seating arrangement was made for the subjects at a distance of three to four times the height of the display size of a video. A software tool
developed at the department was used to automate the process of presenting the videos in the center of the screen. Videos were played in a random order for each subject with insertion of the standard intervals (10 sec.) in between for grading. Viewers were not given the privilege to repeat any video and software front end had no controls available for the subjects to alter the intended processes in anyway. The used grading scale was 0-100 to have scores in a continuous manner. The software automatically stored the results in an excel sheet. In total, 21 non-expert subjects were invited to participate in the tests. The subjects were international students at different master programmes offered at the university and some staff members also took part in the grading campaign, both male and female. The viewers were introduced to the tests by dictating a common text saying that they are supposed to grade a set of videos on visual quality basis. To avoid any viewer fatigue, the test sessions were kept around half an hour length. Only one subject viewed the test stimuli at one time and the lab environment was kept silent to avoid any distractions for the subjects.

In order to obtain reliable results out of raw subjective scores, a two step filtering method was employed to refine the results. The first step was to detect and discard the scores from viewers who exhibited large change of votes.
6.2. Subjective Assessment of the Test Stimuli

MOS Values for Some Test CIF Videos

MOS Values for Some Test QCIF Videos

Figure 6.6: MOS versus bitrate plots
6. Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

compared to the average scores. The second step involved the screening of inconsistent observers without any thought of systematic change. The algorithmic details of these steps are reported in Annex 2 of [16]. After performing the refining process, the outliers were removed and we were left with scores by 18 subjects. Mean opinion score (MOS) was calculated from the scores of these subjects for each test condition \( k \) as following, where \( x \) represents score given by one subject and \( N \) is the number of subjects after outlier removal.

\[
MOS_k = \frac{1}{N} \sum_{n=1}^{N} x_k(n)
\]

(6.3)

As the number of the subjects in the refined scores is not large, we can assume that our data follows Students \( t \) distribution. Thus confidence interval (CI) estimates for each test condition is calculated using the following expression, where \( \alpha \) is taken equal to 0.05 for having 95\% CI, \( N \) is the number of subjects after outliers removal and \( \sigma_k \) is the standard deviation of a test condition \( k \) across \( N \) subjects.

\[
CI_k = t(1 - \alpha / 2, N)\sigma_k\frac{1}{\sqrt{N}}
\]

(6.4)

Figure 6.4 shows average score graded by each subject for all the videos before and after the refining process, within 95\% CI. It is observed that variability of the scores among the subjects has been reduced by the refining process. Figure 6.3 presents the MOS values of Akiyo and Soccer sequences at 5 bitrate conditions for QCIF resolution videos as given in table 6.1. First 5 values are obtained at frame rate of 30 fps and the next 5 values are obtained at frame rate of 15 fps. Overall, MOS values have an increasing trend with increase of bitrate at both of the frame rates.

MOS values for coding condition 1: 600 kbps for CIF and 300 kbps for QCIF, coding condition 2: 400 kbps for CIF and 200 kbps for QCIF, coding condition 3: 200 kbps for CIF and 100 kbps for QCIF are plotted against frame rates and frame resolution in Figure 6.5. The plotted values conform to the commonly known tendencies of perceptual quality at varying bitrates and frame rates. For further illustration, MOS values for a set of test videos are plotted in Figure 6.6 for CIF and QCIF resolutions. Table 6.3 and 6.4 presents the complete list of MOS values for CIF and QCIF test videos.
### 6.2. Subjective Assessment of the Test Stimuli

Table 6.3: MOS Values for CIF Resolution Videos

<table>
<thead>
<tr>
<th>Test Sequence</th>
<th>Frame rate</th>
<th>Bitrate</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Akiyo</strong></td>
<td>30</td>
<td>200</td>
<td>78.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>83.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>74.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>76.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>76.55</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>200</td>
<td>75.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>76.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>73.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>73.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>68.55</td>
</tr>
<tr>
<td><strong>Crew</strong></td>
<td>30</td>
<td>200</td>
<td>33.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>58.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>71.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>69.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>76.27</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>200</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>62.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>67.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>71.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>74.38</td>
</tr>
<tr>
<td><strong>Football</strong></td>
<td>30</td>
<td>200</td>
<td>21.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>40.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>62.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>65.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>70.66</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>200</td>
<td>25.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>45.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>61.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>67.77</td>
</tr>
<tr>
<td><strong>Foreman</strong></td>
<td>30</td>
<td>200</td>
<td>45.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>64.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>75.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>75.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>77.05</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>200</td>
<td>55.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>67.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>68.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>68.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>73.66</td>
</tr>
<tr>
<td><strong>News</strong></td>
<td>30</td>
<td>200</td>
<td>72.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>76.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>75.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>75.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>200</td>
<td>72.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>72.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>70.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>73.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>71.83</td>
</tr>
<tr>
<td><strong>Soccer</strong></td>
<td>30</td>
<td>200</td>
<td>46.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>64.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>74.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>73.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>78.05</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>200</td>
<td>51.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>60.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>600</td>
<td>67.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>800</td>
<td>64.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>70.44</td>
</tr>
</tbody>
</table>
## 6. Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos

Table 6.4: MOS Values for QCIF Resolution Videos

<table>
<thead>
<tr>
<th>Test Sequence</th>
<th>Frame rate</th>
<th>Bitrate</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akiyo</td>
<td>30</td>
<td>100</td>
<td>49.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>50.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>49.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>46.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>50.22</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>100</td>
<td>47.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>49.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>44.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>49.94</td>
</tr>
<tr>
<td>Crew</td>
<td>30</td>
<td>100</td>
<td>18.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>35.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>42.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>44.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>44.66</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>100</td>
<td>24.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>43.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>46.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>49.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>46.61</td>
</tr>
<tr>
<td>Football</td>
<td>30</td>
<td>100</td>
<td>19.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>33.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>45.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>42.88</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>100</td>
<td>23.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>33.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>38.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>44.77</td>
</tr>
<tr>
<td>Foreman</td>
<td>30</td>
<td>100</td>
<td>26.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>42.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>46.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>47.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>49.83</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>100</td>
<td>33.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>43.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>46.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>47.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>49</td>
</tr>
<tr>
<td>News</td>
<td>30</td>
<td>100</td>
<td>39.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>36.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>41.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>42.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>43.16</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>100</td>
<td>43.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>40.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>41.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>42.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>42.94</td>
</tr>
<tr>
<td>Soccer</td>
<td>30</td>
<td>100</td>
<td>28.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>45.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>49.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>53.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>52.94</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>100</td>
<td>26.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>200</td>
<td>41.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>300</td>
<td>44.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>400</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>53.66</td>
</tr>
</tbody>
</table>
6.3 Conclusion

This paper provides details on a subjective VQA of 120 video sequences coded by following H.264/AVC standard at different bitrates, frame rates and frame resolutions. Selection of the test stimuli was performed based on the values of spatial and temporal spectral information. The subjective tests were performed in accordance with standard recommendations of ITU-T. The MOS values and the compressed bitstreams are shared online at: http://www.bth.se/ing/shm.nsf/pages/research-resources. Our motive is to contribute to the field by sharing the complete test database with the fellow researchers. Related future works includes performing similar experiments on other frame resolutions and frame rates also. Moreover, the impact of other distortions can also be included such as packet loss and delay in the IP networks.

6.4 Bibliography


6. **Subjective Quality Assessment of H.264/AVC Encoded Low Resolution Videos**


Seven

Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

This chapter has been published as:

Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

Andreas Rossholm, Muhammad Shahid, and Benny Lövström

Abstract

The growing consumer interest in video communication has increased the users’ awareness in the visual quality of the delivered media. This in turn increases, at the service provider end, the need for intelligent methodologies of optimal techniques for adapting to varying network conditions. Recent studies show that constraints on the bandwidth of transmission media should not always be translated to an increase in compression ratio to lower the bitrate of the video. Instead, a suitable option for adaptive streaming is to scale down the video temporally or spatially before encoding to maintain a desirable level of perceptual quality, while the viewing resolution is constant. Most of the existing studies to examine these scenarios are either limited to low resolution videos or lack in provisioning of subjective assessment of quality. We present here the results of our campaign of subjective quality assessment experiments done on a range of spatial and temporal resolutions, up to VGA and 30 frames per second respectively, under a number of bitrate conditions. The analysis shows, among other things, that keeping the spatial resolution is perceptually preferred among the three parameters that have impact on the video quality, even in the case with high temporal activity.

7.1 Introduction

As video communications constantly continues to grow both regarding its share of all data traffic and the amount of data in absolute terms, the consumers demand on perceived quality also increases. Also, new cellular wireless technology evolves and an increasing share of all data communication will be wireless. This results in many new scenarios with different services and
requirements where the provider want to optimize the perceived quality or quality of experience (QoE). One new challenge with new mobile networks like 3G and 4G is that even if high peak link rates are possible the cellular wireless networks experience rapid link rate variation and occasional long delays in one or both direction. This requires either long receiver buffers, resulting in long end-to-end delay, or fast adaptation, resulting in need for the possibility to change used bandwidth [1]. In this context the need of optimizing the delivered quality of experience by a service provider is raised. To this end, one significant issue to be resolved is finding the best trade-off among spatial resolution, temporal resolution, and quantization level, giving the optimal value of QoE in a given scenario. In practice, this includes applications such as adaptive streaming [2], [3], as well as different real time video communication services where maintaining the desired level of perceived quality is required in fluctuating network conditions. Also, there is a growing demand for objective quality measurement or monitoring techniques estimating perceived video quality in these scenarios, especially to be able to compare different spatial and temporal resolutions. An overview of various types of contemporary objective Video Quality Assessment (VQA) is presented in [4].

The quest of discovering the optimal trade-off has been the subject of video scalability for assuring stipulated level of visual quality. For service providers, it is useful to ascertain the best QoE of a video at a given bandwidth capacity. In order to optimally address any fluctuations in the transmission network, it becomes pertinent to determine the parameter that can be scaled up or down with minimal deviation in the level of delivered visual quality. To serve this matter, a number of studies have been made that focus on examining the impacts of changes in the aforementioned three parameters of a video. Subjective quality assessment of low resolution, QCIF (176x144) and CIF (352x288), videos encoded using H.264/AVC has been reported in [5] for 150 test scenarios. Under low bitrate conditions, it was concluded that small frame size is mostly preferred. For CIF resolution or high temporal (30 fps) resolution at low bitrates, it was found that it was most efficient to reduce quantization except for video sequences containing very low spatial activity. It was also pointed out that a minimum threshold value of 0.1 Bits Per Pixel (BPP) is required to achieve a good or excellent perceptual quality.

Subjective experiments conducted using low resolution videos, CIF, in [6] show that frame rate can be compromised to maintain the perceptual quality by keeping the compression ratio at low value. Similar results can be observed in the study reported in reference [7]. Impact of encoding strategy on the qual-
ity of MPEG-2 encoded videos, QCIF and CIF, while transmitted over lossy network has been investigated in [8]. It has there been observed that videos with high spatial activity are perceptually preferred with higher spatial resolution, and videos with higher temporal activity are preferred in full frame-rates. The validity of these results needs to verified in the case of videos encoded by H.264/AVC. Considering the case of high resolution video conferencing applications, video scalability has been tested for high definition videos (1920x1080) in [9]. It was observed that the quality level can be maintained by decreasing frame rate and frame resolution to cater the constraints of the transmission bandwidth. Hence, high compression rates can be avoided. Moreover, as the bandwidth begins to grow, it is perceptually preferred to increase the frame rate up to a certain higher level first and the frame resolution can be increased afterwards. Unfortunately, these conclusions have been drawn only from the results of objective metrics, with no subjective assessment to support the results. A detailed discussion and a review of the studies performed on the video scalability for quality can be found in [10] and the references therein.

Table 7.1: The original frame rate and a brief description of the SRCs used in the experiment

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Frame rate [fps]</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>30</td>
<td>Two children sitting on the floor, slowly moving, low SI and low TI</td>
</tr>
<tr>
<td>City</td>
<td>25</td>
<td>Panning view over a city from an airplane, high SI and medium TI</td>
</tr>
<tr>
<td>Elisa</td>
<td>30</td>
<td>Head and shoulder of a talking woman, medium SI and low TI</td>
</tr>
<tr>
<td>Ice</td>
<td>25</td>
<td>Several persons skating on white ice, low SI and high TI</td>
</tr>
<tr>
<td>Soccer</td>
<td>25</td>
<td>Close up view of soccer game, panning, low SI and high TI</td>
</tr>
</tbody>
</table>

Figure 7.1: The SRCs used for generation of PVSs
7. Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

Table 7.2: The PVS combinations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MVGA</td>
<td>10</td>
<td>50</td>
<td>MVGA</td>
<td>8.33</td>
<td>50</td>
</tr>
<tr>
<td>MVGA</td>
<td>10</td>
<td>150</td>
<td>MVGA</td>
<td>8.33</td>
<td>150</td>
</tr>
<tr>
<td>MVGA</td>
<td>10</td>
<td>300</td>
<td>MVGA</td>
<td>8.33</td>
<td>300</td>
</tr>
<tr>
<td>MVGA</td>
<td>15</td>
<td>50</td>
<td>MVGA</td>
<td>12.5</td>
<td>50</td>
</tr>
<tr>
<td>MVGA</td>
<td>15</td>
<td>150</td>
<td>MVGA</td>
<td>12.5</td>
<td>150</td>
</tr>
<tr>
<td>MVGA</td>
<td>15</td>
<td>300</td>
<td>MVGA</td>
<td>12.5</td>
<td>300</td>
</tr>
<tr>
<td>MVGA</td>
<td>30</td>
<td>150</td>
<td>MVGA</td>
<td>25</td>
<td>150</td>
</tr>
<tr>
<td>MVGA</td>
<td>30</td>
<td>300</td>
<td>MVGA</td>
<td>25</td>
<td>300</td>
</tr>
<tr>
<td>QVGA</td>
<td>10</td>
<td>50</td>
<td>QVGA</td>
<td>8.33</td>
<td>50</td>
</tr>
<tr>
<td>QVGA</td>
<td>10</td>
<td>150</td>
<td>QVGA</td>
<td>8.33</td>
<td>150</td>
</tr>
<tr>
<td>QVGA</td>
<td>10</td>
<td>300</td>
<td>QVGA</td>
<td>8.33</td>
<td>300</td>
</tr>
<tr>
<td>QVGA</td>
<td>10</td>
<td>600</td>
<td>QVGA</td>
<td>8.33</td>
<td>600</td>
</tr>
<tr>
<td>QVGA</td>
<td>15</td>
<td>50</td>
<td>QVGA</td>
<td>12.5</td>
<td>50</td>
</tr>
<tr>
<td>QVGA</td>
<td>15</td>
<td>150</td>
<td>QVGA</td>
<td>12.5</td>
<td>150</td>
</tr>
<tr>
<td>QVGA</td>
<td>15</td>
<td>300</td>
<td>QVGA</td>
<td>12.5</td>
<td>300</td>
</tr>
<tr>
<td>QVGA</td>
<td>15</td>
<td>600</td>
<td>QVGA</td>
<td>12.5</td>
<td>600</td>
</tr>
<tr>
<td>QVGA</td>
<td>30</td>
<td>150</td>
<td>QVGA</td>
<td>25</td>
<td>150</td>
</tr>
<tr>
<td>QVGA</td>
<td>30</td>
<td>300</td>
<td>QVGA</td>
<td>25</td>
<td>300</td>
</tr>
<tr>
<td>QVGA</td>
<td>30</td>
<td>600</td>
<td>QVGA</td>
<td>25</td>
<td>600</td>
</tr>
<tr>
<td>HVGA</td>
<td>10</td>
<td>150</td>
<td>HVGA</td>
<td>8.33</td>
<td>150</td>
</tr>
<tr>
<td>HVGA</td>
<td>10</td>
<td>300</td>
<td>HVGA</td>
<td>8.33</td>
<td>300</td>
</tr>
<tr>
<td>HVGA</td>
<td>10</td>
<td>600</td>
<td>HVGA</td>
<td>8.33</td>
<td>600</td>
</tr>
<tr>
<td>HVGA</td>
<td>10</td>
<td>900</td>
<td>HVGA</td>
<td>8.33</td>
<td>900</td>
</tr>
<tr>
<td>HVGA</td>
<td>15</td>
<td>150</td>
<td>HVGA</td>
<td>12.5</td>
<td>150</td>
</tr>
<tr>
<td>HVGA</td>
<td>15</td>
<td>300</td>
<td>HVGA</td>
<td>12.5</td>
<td>300</td>
</tr>
<tr>
<td>HVGA</td>
<td>15</td>
<td>600</td>
<td>HVGA</td>
<td>12.5</td>
<td>600</td>
</tr>
<tr>
<td>HVGA</td>
<td>15</td>
<td>900</td>
<td>HVGA</td>
<td>12.5</td>
<td>900</td>
</tr>
<tr>
<td>HVGA</td>
<td>30</td>
<td>300</td>
<td>HVGA</td>
<td>25</td>
<td>300</td>
</tr>
<tr>
<td>HVGA</td>
<td>30</td>
<td>600</td>
<td>HVGA</td>
<td>25</td>
<td>600</td>
</tr>
<tr>
<td>HVGA</td>
<td>30</td>
<td>900</td>
<td>HVGA</td>
<td>25</td>
<td>900</td>
</tr>
<tr>
<td>VGA</td>
<td>10</td>
<td>300</td>
<td>VGA</td>
<td>8.33</td>
<td>300</td>
</tr>
<tr>
<td>VGA</td>
<td>10</td>
<td>600</td>
<td>VGA</td>
<td>8.33</td>
<td>600</td>
</tr>
<tr>
<td>VGA</td>
<td>10</td>
<td>900</td>
<td>VGA</td>
<td>8.33</td>
<td>900</td>
</tr>
<tr>
<td>VGA</td>
<td>15</td>
<td>300</td>
<td>VGA</td>
<td>12.5</td>
<td>300</td>
</tr>
<tr>
<td>VGA</td>
<td>15</td>
<td>600</td>
<td>VGA</td>
<td>12.5</td>
<td>600</td>
</tr>
<tr>
<td>VGA</td>
<td>15</td>
<td>900</td>
<td>VGA</td>
<td>12.5</td>
<td>900</td>
</tr>
<tr>
<td>VGA</td>
<td>30</td>
<td>600</td>
<td>VGA</td>
<td>25</td>
<td>600</td>
</tr>
<tr>
<td>VGA</td>
<td>30</td>
<td>900</td>
<td>VGA</td>
<td>25</td>
<td>900</td>
</tr>
</tbody>
</table>
By examining the existing drives to investigate the impacts of three basic parameters of video encoding, the requirement of a comprehensive study on a wider range of videos, in a highly interesting bandwidth range, supported by subjective assessment of quality becomes evident. Therefore, we present here the details of an extensive campaign of subjective quality assessment experiments of videos encoded using combinations of multiple levels of the bitrate, frame rate and resolution. This enables examinations of e.g. the perceptual trade off between spatial and temporal resolution at a certain bitrate. The rest of this paper is organized as follows. In Section 7.2 the video sequences used in the test are described, encoding configuration, as well as the subjective assessment setup. Also the pre-processing of the sequences before the assessment is described. In Section 7.3 the findings from the subjective tests are given, and finally in Section 7.4 conclusions are drawn.

Figure 7.2: MOS vs. bitrate for different frame rate and resolutions where Children is characterized to have low SI and low TI.
7. Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

![Graph showing MOS vs. bitrate for different frame rate and resolutions. City is characterized to have high SI and medium TI.](image)

**7.2 Test Stimuli and Subjective Video Quality Assessment**

To perform a comprehensive subjective quality assessment that can be used to infer useful conclusions, it is imperative to select the SouRCE sequences (SRCs) carefully. Such SRCs should possess a variety of spatio-temporal characteristics to be representative of most commonly used videos. To this end, we followed the ITU recommendation P.910 [11] for the selection of SRCs based on spatial perceptual information (SI) and temporal perceptual information (TI). The SI and TI values are calculated in the luminance plane of a video. The five SRCs used in this study are Children, City, Elisa, Ice, and Soccer, all of 10 s duration. The starting frame of each of the sequences is shown in Fig. 7.1, and table 7.1 gives a short description of the content and lists the original frame rate of the sequences as well as their SI and TI characteristics.
Figure 7.4: MOS vs. bitrate for different frame rate and resolutions where Elisa is characterized to have medium SI and low TI.

### 7.2.1 Encoding configuration

The SRCs have been encoded following the standard H.264/AVC using the JM reference software to produce Processed Video Sequences (PVSs). For the encoding of the PVSs a number of combinations of resolutions (Res), frame rates (FR) and bitrates (BR) have been used, based on several considerations. For the bitrates, the band width fluctuation and the built in limitations running realtime communication over cellular wireless network was taken into count. Based on this and the requirements of a realistic BPP value, and also de facto configurations from industry, the resolution and frame rate was limited, as shown below.

- **BR:** 50, 150, 300, 600, and 900 kbps
- **Res:** VGA = 640 × 480, HVGA (Half VGA) = 480 × 320, QVGA (Quarter VGA) = 320 × 240, and MVGA (mobile VGA) = 192 × 144
Figure 7.5: MOS vs. bitrate for different frame rate and resolutions where Ice is characterized to have low SI and high TI.

- FR: A: 30, 15, 10 fps, and B: 25, 12.5, 8.33 fps

All used combinations of bitrate, resolution, and frame rate are shown in Table 7.2, where columns A and B shows the different combinations for the videos with original frame rates 30 fps and 25 fps, respectively. It can be seen in table 7.2 that it results in 38 combination for every SRC. To conduct a suitable subjective test the combination of resolution, frame rate and bitrate is based on realistic combinations used in practice, which means that combinations with too low or very high BPP are excluded.

### 7.2.2 Pre-processing the test sequences

Before executing the subjective assessment all the processed video sequences (PVSs) are pre-processed. The reason for this is to enable a more realistic test scenario as in streaming or realtime video applications, where the viewing resolution is usually fixed even if the source data is changed, e.g. down sampled, in
Figure 7.6: MOS vs. bitrate for different frame rate and resolutions where Soccer is characterized to have low SI and high TI.

In the context of adapting to fluctuating bandwidth. Therefore all PVSs with spatial resolution MVGA, QVGA, and HVGA were up-scaled to VGA (640x480), performed with bicubic filtering as it produces sufficient quality and does not require too much of processing power. Also, all files with sub-sampled temporal resolution from the original 25fps or 30fps were up-sampled to the original frame rate by frame repetition. This was performed to limit difference in play out during the subjective assessment between the PVSs.

### 7.2.3 Subjective Video Quality Assessment Setup

The subjective quality assessment has been performed on 32 test subjects with video sequences described in the previous subsection. Since not all combinations of bitrates and frame rates are used, this results in a total of 190 sequences being used in the test. The setup of the subjective quality assessment follows...
7. **Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality**

Table 7.3: **ANOVA applied to the sequences. The "*" marks the most significant variable.**

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Variable</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>FR</td>
<td>7.55e-11</td>
</tr>
<tr>
<td></td>
<td>Res</td>
<td>5.81e-33*</td>
</tr>
<tr>
<td></td>
<td>BPP</td>
<td>1.062e-19</td>
</tr>
<tr>
<td>Children</td>
<td>FR</td>
<td>4.22e-06</td>
</tr>
<tr>
<td></td>
<td>Res</td>
<td>1.51e-09*</td>
</tr>
<tr>
<td></td>
<td>BPP</td>
<td>9.75e-07</td>
</tr>
<tr>
<td>City</td>
<td>FR</td>
<td>0.0026</td>
</tr>
<tr>
<td></td>
<td>Res</td>
<td>0*</td>
</tr>
<tr>
<td></td>
<td>BPP</td>
<td>0.0035</td>
</tr>
<tr>
<td>Elisa</td>
<td>FR</td>
<td>0.0013</td>
</tr>
<tr>
<td></td>
<td>Res</td>
<td>0*</td>
</tr>
<tr>
<td></td>
<td>BPP</td>
<td>0.0308</td>
</tr>
<tr>
<td>Ice</td>
<td>FR</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>Res</td>
<td>0*</td>
</tr>
<tr>
<td></td>
<td>BPP</td>
<td>0.0001</td>
</tr>
<tr>
<td>Soccer</td>
<td>FR</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>Res</td>
<td>0*</td>
</tr>
<tr>
<td></td>
<td>BPP</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

ITU recommendations as given by ITU-R BT 500-12 [12] for the lab setup of our experiments. Particularly, the method followed was the single stimulus quality evaluation where a test video sequence is shown once without the presence of any explicit reference, corresponding to the reality where users see only the processed version of the video. Overall, the adopted methodology and lab setup has been summarized in [7]. The subjects who participated in the tests were of both genders, mainly students at the university and some staff members, and all of them were considered to be non-expert in the area of video quality assessment. In order to obtain reliable results out of the raw subjective scores on the quality scale of 1 to 100, a screening of the observers scores was employed to discard observers that are considered as outliers. The algorithmic details of these steps are reported in Annex 2 of [12]. After screening of our data no subject had to be rejected. Finally, the mean opinion score (MOS) was calculated and used in this work.
7.3 Results

In the context of adaptive streaming, an estimation of the variable bandwidth over a channel is used as a restriction for available bitrates to use for the video codec. With this in mind, the MOS results for the five test sequences with their 38 combinations are presented in Fig. 7.2-7.6 where the MOS scores are plotted versus the bitrate. To be able to identify different resolutions and frame rates in the figures, the resolution is color coded and the frame rate is marked by different symbols. In these figures, besides the obvious symbols in the legends, "s" denotes square and "d" denotes diagonal.

7.3.1 Observations from the MOS results

Some observations can be made directly by studying the MOS results. It can be seen that the sequences with lowest temporal information (TI), Elisa, City, and Children, have clear differentiation between the different resolutions, indicating that the resolution has high significance. There is though a difference regarding City versus Elisa and Children, where the later have the highest spatial information, that for City it clearly differentiates between resolutions even for VGA and HVGA which is not the case for Elisa and Children even if highest resolution is always preferred. It can also be seen for these sequences that higher resolution over increased frame rate is always preferred. For the two sequences with highest temporal information (TI), Soccer and Ice, the tendency is the same but not to the same extent. It can be seen that for increased spatial resolution at lower frame rate is preferred over increase of frame rate but keeping the spatial resolution.

7.3.2 Analysis of Variance Based Comparison

To further evaluate the MOS scores ANalysis Of VAriance (ANOVA) [13] was used. ANOVA analysis is used to determine whether or not different factors or variables are statistically significant. We considered Res, FR, and BPP for this analysis to see their statistical significance on the MOS scores, where BPP can be seen as an indicator of level of compression. To resolve the relative importance of these variables the multiway ANOVA technique was used and the variables are stated significant if the p-value was below 0.05. We used the Matlab function anovan for this purpose. The result from the ANOVA comparison is shown in Table 7.3. It can be seen in Table 7.3 that for all the cases resolution (Res) has the highest significance which was also confirmed in the evaluations.
7. Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

of the MOS scores illustrated in Figs. 7.2-7.6. Further the result indicates that BPP is the second most important variable, i.e. the compression level, except for Soccer and Ice which are the two sequences with highest temporal information where the frame rate (FR) has the same or higher significance.

7.4 Conclusion

In this paper we have addressed the increasing interest of video communication and its attempt to maximize the perceptual quality during fluctuating bandwidth conditions. In many scenarios of streaming, realtime video communication, or other video applications, adaptive streaming is used to handle fluctuating network bandwidths. A suitable option for adaptive streaming is to scale down the video temporally or spatially before encoding to maintain a desirable level of perceptual quality while viewing resolution is constant. In a subjective assessment with five original sequences, 38 different combination of bitrate, frame rate, and resolution, 32 subjects were used. Both direct and statistical evaluation was made of the MOS scores, where MOS scores were plotted versus the bitrate, and ANOVA was used for statistical analysis. The result shows that preserving the spatial resolution throughout the process has the highest significance even in the scenarios with high temporal information. In comparison, in most studies when increasing the bitrate for a sequence with high SI this results in a preference for increased resolution or decreased quantization, while for sequences with high TI it results in a preference for increased frame rate. One of the reasons to this could be that all sequences were assessed at the same or limited number of different spatial resolutions. In our study, however, four different spatial resolutions were used, and all sequences were assessed at a fixed spatial viewing resolution. Future work planned includes using the presented results to develop a bit-stream based no-reference quality metric, as well as conducting a subjective study using higher resolutions to investigate the same parameters of video coding in other user scenarios.

7.5 Bibliography


7. Analysis of the Impact of Temporal, Spatial, and Quantization Variations on Perceptual Video Quality

Eight

Crowdsourcing Based subjective quality assessment of adaptive video streaming

This chapter has been published as:

Crowdsourcing Based Subjective Quality Assessment of Adaptive Video Streaming

Muhammad Shahid, Jacob Søgaard, Jeevan Pokhrel, Kjell Brunström, Kun Wang, Samira Tavakoli, and Narciso Garcia

Abstract

In order to cater for user’s quality of experience (QoE) requirements, HTTP adaptive streaming (HAS) based solutions of video services have become popular recently. User QoE feedback can be instrumental in improving the capabilities of such services. Perceptual quality experiments that involve humans are considered to be the most valid method of the assessment of QoE. Besides lab-based subjective experiments, crowdsourcing based subjective assessment of video quality is gaining popularity as an alternative method. This paper presents insights into a study that investigates perceptual preferences of various adaptive video streaming scenarios through crowdsourcing based subjective quality assessment.

8.1 Introduction

Interest in quality of experience (QoE) of video services is growing due to increasing usage of videos over networks, such as the portion of video data in mobile networks is expected to exceed 67% by 2018 [1]. Hypertext Transfer Protocol (HTTP) based video streaming has been greatly adopted to avoid network distortions such as packet-loss. Subjective experiments are considered to be the most valid methodology to assess the QoE. Subjective experiments are typically conducted in a controlled laboratory environment. Objective or computer software assisted methods have been largely seen as an alternative approach, to get around the complications involved in the lab-based subjective experiments. However, the objective methods even with state-of-the-art performance are generally considered far from universal acceptance. Crowdsourcing based subjective experiments have gained attention to replace needs of lab-based tests and these experiments offer promising correlation with the later [2]. This methodology mainly involves collecting subjective assessment
8. CROWDSOURCING BASED SUBJECTIVE QUALITY ASSESSMENT OF ADAPTIVE VIDEO STREAMING

of quality through ubiquitous streaming via the Internet. This enables the investigator to receive opinion from a vast variety of subjects; in a time-flexible, test-data size scalable, and swift manner.

This paper includes an insight into a crowdsourcing based subjective perceptual preference of various adaptation scenarios investigated earlier in [3] and additional buffering scenarios. In the following, we present the related details on our experiments and the obtained results thereof.

8.2 Test Background

The videos for our subjective test are originally from the subjective lab experiment detailed in [3]. The original videos were all in 1280x720 resolution with a frame rate of 24 and encoded using the high profile for H.264/AVC at 4 different bitrates: 600 kbps, 1 Mbps, 3 Mbps, and 5 Mbps. Seven different sources were used; three sources were taken from entertainment movies and the rest was content from: a soccer match, a sports documentary, a newscast, and a concert. The subjective lab experiment was carried out at Acreo lab in a test room compliant with the ITU-R BT.500 [4]. Several adaptation scenarios for the videos were produced in the original experiment, such as going from a high to a low bit rate in a stepwise manner. In our subjective experiment we used the following scenarios from the original experiment: Gradual Decreasing (GD), Rapid Decreasing (RD), constant 600 kbps (N600), constant 1 Mbps (N1), constant 3 Mbps (N3), and constant 5 Mbps (N5). Additionally, we introduced new buffering scenarios to test the quality perception in relation to the aforementioned scenarios. The buffering scenarios include: 1 Freezing event lasting for 2 seconds in the constant 3 Mbps video (1F3M), 2 Freezing events lasting for 1 second each in the constant 3 Mbps video (2F3M), and 1 Freezing event lasting for 2 seconds in the constant 1 Mbps video (1F1M). In total 9 different scenarios were used, resulting in a total of 63 stimuli.

Crowdsourcing experiments should be as simple as possible for the subject, therefore we chose to follow the Paired Comparison (PC) methodology [5]. We used the optimized square design [6] based on our assumptions of the quality levels to get reliable measurements and reduce the number of pairings. Using this method, our test set consisted of a total of 126 pairings. These pairings were divided into 14 tasks with 3 videos from 3 different contents, i.e., 9 videos for each task. We used screentests [7] prior to the subjective test to filter out potential malicious workers. In total, 215 workers participated in the
experiment where almost one quarter of the participants were from the US while the rest were mainly from European countries.

![Figure 8.1: Comparison between lab-based and crowdsourcing subjective experiments.](image)

8.3 Results

To analyze the results, we applied the Bradley-Terry (BT) model to obtain quality scores and corresponding confidence intervals from the preference matrices of PC data [6]. The subjects were filtered by excluding workers with too many unlikely preferences. We define an unlikely preference as a preference where the corresponding probability in the BT model is lower than a threshold $\theta$. In our test, we allowed only 2 out of 9 unlikely preferences and therefore we set $\theta = 0.25$. With this approach 6 workers were excluded from the final results. In order to validate the results obtained from the crowdsourcing experiment, we compared the mean opinion scores (MOS) obtained from the lab experiment to the crowdsourcing experiment as shown in Fig. 8.1. The results show that the opinion scores obtained from both the experiments are strongly corre-
8. Crowdsourcing Based Subjective Quality Assessment of Adaptive Video Streaming

lated, though not to the degree one would expect of a repetition of the lab test. This can be due to the differences in the test setup, such as evaluation method, viewing environment and the introduction of new distortions. Our experiment verify the results from earlier studies, e.g., [8], that buffering events has a high impact on the QoE. Due to this, users generally prefer viewing videos at lower bitrates than experiencing buffering events in videos at higher bitrates.

The quality of the videos can also be compared against the average bitrate of the videos. This has been illustrated in Fig. 8.2, where the mean of the subjective scores has been calculated over the video contents. Generally, users prefer videos at higher bitrates, i.e., 3 or 5 Mbps and the difference between them is probably more due to the difference in content than the difference in compression levels. Users dislike buffering events and it seems that the frequency is more important than the total duration of these events (both videos at 3 Mbps with buffering have a total buffering time of 2s), which is in line with earlier studies e.g. [9]. But if the bitrate is high enough and the frequency of the buffering events is low enough, e.g., the 1F3M video, this seems to be a viable alternative to decreasing the bitrate of the video or having a constant low bitrate, e.g., 600 kbps or 1 Mbps.

---

Figure 8.2: Opinion Scores (BT model) versus the average bitrate.
8.4 Conclusion

The subjective experiment conducted in a crowdsourcing environment verifies the results of earlier studies of adaptation scenarios, including the effect of buffering events. Also, our study suggests that in a network environment with fluctuations in the bandwidth, a medium or low video bitrate which can be kept constant is the best approach. Moreover, if there are only a few drops in bandwidth, one can choose a medium or high bitrate with a single or few buffering events.

8.5 Bibliography


Part IV
Part IV

On Speech Enhancement in Modulation Frequency Domain
Nine

Modulation Domain Adaptive Gain Equalizer for Speech Enhancement

This chapter has been published as:

9.1 Introduction

Modulation Domain Adaptive Gain Equalizer for Speech Enhancement

Muhammad Shahid, Rizwan Ishaq, Benny Sällberg, Nedelko Grbic, Benny Lövström and Ingvar Claesson

Abstract

This paper evaluates speech enhancement by filtering in the modulation frequency domain, as an alternative to filtering in conventional frequency domain. Adaptive Gain Equalizer (AGE) is a commonly used single-channel speech enhancement algorithm. A recently introduced class of signal transformations called modulation transform has successfully made its place alongside classical time/frequency representations. This paper presents an implementation of AGE within modulation system, for the purpose of enhancing the speech signal. The successful implementation of the proposed system has been validated with various performance measurements, i.e., Signal to Noise Ratio Improvement (SNRI), Mean Opinion Score (MOS) and Spectral Distortion (SD). A spectrogram analysis is also presented to further substantiate the performance of this work.

9.1 Introduction

Speech as the main part of the communication systems, is usually degraded during the transmission by different types of noise, e.g., Gaussian noise, engine noise, periodic noise and other interferences. There are a variety of methods for reduction of noise from speech signal, e.g., spectral subtraction (frequently used for noise reduction) [1] and optimum Wiener filtering [2]. Adaptive Gain Equalizer (AGE) [3] is a noise reduction method that focuses on enhancing the speech signal instead of suppressing the noise. The speech enhancement is carried out by weighting the sub-bands in time-frequency domain according to an estimate of the Signal-to-Noise Ratio (SNR). This method offers better result in terms of low complexity, low delay, low distortion and there is no need for Voice Activity Detector (VAD).
The modulation system assumes that a speech signal is composed of a modulator and a carrier. The signal is represented by,

\[ x(t) = m(t)c(t) \]  \hspace{1cm} (9.1)

where \( m(t) \) denotes the low frequency part of the signal, called modulator, and it modulates a high frequency carrier \( c(t) \). Studies have shown that the modulators of speech signal are most important for the intelligibility of the speech signal. The importance of the modulator in speech signals brought the attention of many researchers.

AGE implementation has been intended so far in time-frequency domain, but here an implementation of AGE in a modulation system is proposed. Modulation systems which are based on sub-band modulators, perfectly fit the AGE system which works on the sub-bands of the signal.

### 9.1.1 Literature Survey

Zadeh [4] is considered to be the pioneer of the field of modulation domain who suggested a two dimensional bi-frequency system, where time variation of the acoustic frequency is the second dimension of frequency. Atlas et al. used the concept of coherent modulation for the target talker enhancement in speech enhancement [5]. They proved that modulation domain moderately increases the speech intelligibility. Coherent modulation using the frequency reassignment has been used for speech enhancement and for demodulation of a signal into modulator and carrier [6]. Li et al. described the theory behind modulation filtering which offers a new approach to modifying non-stationary signals e.g., speech. They presented the coherent modulation analysis based on instantaneous frequency estimation using conditional mean frequency. In addition, they showed that the proposed method accurately estimates the carriers and modulators of the signals [7]. Speech polluted by wind noise has been enhanced by using coherent modulation comb filtering by King et al. [8]. Although the modulation filtering has mostly been used for the purpose of speech enhancement, Vinton et al. also used it for audio compression. They showed that a 32 kb/s/channel outperformed MPEG-1 coded at 56 kb/s/channel (both at 44.1 kHz), using the modulation technique [9]. The concept of homomorphic demultiplication is connected to the modulation spectral analysis/synthesis and it was outlined by Atlas et al. in [10]. Clark et al. showed in [11] the effectiveness of modulation filtering by measuring the empirical modulation frequency response and got a near-ideal response performance, and 25 dB improvement has been shown for suppressing undesired modulation frequencies.
over incoherent modulation. Clark presented the Center of Gravity (COG) method for decomposition of a sub-band signal, and he used coherent modulation filtering for the interpolation of long gaps in acoustic signals [12]. The concept of AGE for the reduction of noise in speech signals, has shown its success in real time and proven to be a low complexity system [3]. The method used an FIR filter bank to get the required results and it was also shown that the system adapted itself for different types of noise. The proposed AGE method using the mixed analog and digital hybrid approach yielded around 13 dB speech enhancement [13]. The AGE was originally intended for the digital domain, but [14] provides an analog implementation which does not use quantization and digitization and it is also best fitted for battery powered applications. A hybrid solution to overcome problems related to a digital and an analog implementation of the AGE is found in [15].

9.1.2 Main Contributions

The main contribution of this paper is to combine the AGE and modulation system domain for speech enhancement. Hence, the advantage of benefits from both of the fields has been taken to build up a new system. This approach has proven to be robust, flexible in implementation and has been validated by performance measures like Signal to Noise Ratio Improvement (SNRI), Mean Opinion Score (MOS) and Spectral Distortion (SD). Section 9.2 briefly introduces the modulation system, Section 9.3 introduces the concept of AGE and its operation in the modulation frequency domain and Section 9.4 evaluates the proposed system. Section 9.5 concludes this work with a summary and future research directions in the area.

9.2 Modulation System

A modulation domain spectrum is obtained from a certain acoustic spectrum by taking short-time Fourier transform (STFT) of the speech signal at the given acoustic frequency. The speech signal modulators are the most important components for speech intelligibility. Shamma [16] reported that auditory cortex neurons possibly decompose the acoustic contents into spectro-temporal modulation contents. It has been found that if the modulators of the speech signal are replaced by constant amplitude modulators, while carriers are preserved, speech is not intelligible. However when the modulators are preserved but carriers are altered, the speech is intelligible [17]. Modulation domain actually

decomposes the speech, or other natural signals, into modulators and carriers whereafter the modulators of the signals are analyzed. A modulation frequency system is described by the following steps:

- Filter bank to get sub-band signals
- Demodulation i.e., decomposition of each sub-band signal into a modulator and a carrier.
- Analysis of the modulators of the sub-band signals by discrete Fourier transform of each modulators
- Modification of the modulators (e.g. linear filtering)
- Re-modulation (recombination of modified modulators with original carriers)
- Synthesis of signals

The modulation system filter bank divides the wide-band signal into K narrow-band sub-bands. The signal $x(t)$ is passed through the filter bank’s set of band-pass filters $h_k$, which renders the sub-band signals $x_k(t)$.

$$x_k(t) = h_k * x(t)$$  \hspace{1cm} (9.2)

where $*$ denotes the convolution operator. The demodulation process decomposes the sub-band signal into its envelope and carrier. Its efficient to decimate the sub-band signals so that the redundant samples may be removed. Modification of the modulators is done by the modulation filtering which mostly uses linear time invariant filters $g(t)$, i.e., $\tilde{m}_k(t) = m_k(t)g(t)$. A modulation spectrogram and modulation analysis can be done by computing the Fourier transform along the time-axis of the spectrogram (magnitude) or by utilizing the spectrum of the envelop signals, which gives the modulation frequency along horizontal axis and acoustic frequency along vertical axis. Re-modulation is the process in which modified modulators $\tilde{m}_k(t)$ are combined with the original carriers, obtained in the process of demodulation, to get the modified sub-band signals $\tilde{x}_k(t)$. The synthesis process reconstructs the modified signal $\tilde{x}(t)$ using the modified sub-band signals $\tilde{x}_k(t)$, according to the following equation. Interpolation must be performed prior to this stage if decimation was done before.

$$\tilde{x}(t) = \sum_{k=1}^{K} \tilde{x}_k(t)$$  \hspace{1cm} (9.3)
9.2. Modulation System

Envelope detection is used for demodulation of a signal and it is the most important part of the modulation frequency system. There are two types of envelope detectors mostly used, coherent envelope detection and incoherent envelope detection. Magnitude, or magnitude-like, operations are used to estimate modulators in incoherent detection, while coherent detection use the carrier estimate operations. Incoherent envelope detection detects the envelope and carrier independently and coherent detection uses the carrier estimation for the calculation of the envelope. Following is a brief description about one of the methods used for coherent carrier detection which is used in this work.

9.2.1 Spectral Center of Gravity Carrier Estimation

In this recently introduced method of the center-of-gravity approach, instantaneous frequency $\omega_k(n)$ is defined as instantaneous spectrum average frequency of $x_k(t)$ at time $t$ [18]. An instantaneous spectrum with short-time Fourier transform is computed as,

$$S_k(\omega, t) = \sum_p g(p)x_k(t + p)e^{-j\omega p}$$

(9.4)

where $g(p)$ is a short spectral-estimation window. The instantaneous frequency $\omega_k(t)$ of the sub-band signal $x_k(t)$ is estimated as,

$$\omega_k(t) = \frac{\int_{-\pi}^{\pi} \omega |S_k(\omega, t)|^2 d\omega}{\int_{-\pi}^{\pi} |S_k(\omega, t)|^2 d\omega}$$

(9.5)

The phase $\phi_k(t)$ of the carrier is computed as follows

$$\phi_k(t) = \sum_{p=0}^{t} \omega_k(p)$$

(9.6)

The carrier $c_k$ is

$$c_k(t) = e^{j\phi_k(t)}$$

(9.7)

and the complex valued modulator $m_k(t)$ is given by

$$m_k(t) = x_k(t)c_k^*(t)$$

(9.8)
9.3 Adaptive Gain Equalizer System

As discussed in [3], the AGE consists of a filter bank with different band-pass filters. Each sub-band is weighted by a gain function which amplifies the signal when speech is present and keeps the noisy part of the signal, where no speech is present, to unity. A filter bank of K bandpass filters divides the input signal $x(n)$ into K sub-bands $x_k(n)$.

$$x_k(n) = h_k * x(n) \quad (9.9)$$

Here $h_k$ is the impulse response of the filter bank sub-band k and $*$ denotes the convolution. The time domain signal is modeled as a sum of sub-band signals, according to:

$$x(n) = \sum_{k=1}^{K} x_k(n) = \sum_{k=1}^{K} (s_k(n) + w_k(n)) \quad (9.10)$$

where $s_k(n)$ is the desired speech signal related to $k^{th}$ sub-band, while $w_k(n)$ is the additive noise in the sub-band k. The output signal $\hat{x}(n)$, with the amplified speech signal, is computed as

$$\hat{x}(n) = \sum_{k=1}^{K} G_k(n)x_k(n) \quad (9.11)$$

where $G_k(n)$ is the AGE weighting function which amplifies the signal when speech is active.

9.3.1 Gain Function

Two terms used for the calculation of the gain function are; a long term (slow) average $A_{s,k}(n)$ and the short term (fast) average $A_{f,k}(n)$. The short term average, for sub-band k, $A_{f,k}(n)$ is calculated as

$$A_{f,k}(n) = a_k A_{f,k}(n-1) + (1 - a_k) | x_k(n) | \quad (9.12)$$
9.3. Adaptive Gain Equalizer System

where \( \alpha_k \) is a small positive constant, given by

\[
\alpha_k = \frac{1}{T_{s,k} F_s}
\]

where \( F_s \) is the sampling frequency in Hz and \( T_{s,k} \) is a time constant in seconds. In the same manner, a slow average is computed as

\[
A_{s,k}(n) = \beta_k A_{s,k}(n-1) + (1 - \beta_k) | x_k(n) |
\]

if \( A_{s,k}(n-1) \leq A_{f,k}(n) \),

\[
A_{s,k}(n) = A_{f,k}(n)
\]

if \( A_{s,k}(n-1) > A_{f,k}(n) \)
where \( \beta_k \) is a small positive constant. The AGE gain function is computed as:

\[
G_k(n) = \left( \frac{A_{f,k}(n)}{A_{s,k}(n)} \right)^{p_k}
\]

where \( p_k \geq 0 \), and \( A_{s,k}(n) > 0 \).

9.3.2 Modulation Domain AGE

The functionality of the AGE has been extended to work in the modulation domain for speech enhancement. Modulation domain separates each sub-band

signal into a carrier and a modulator. While only modulators are considered here, the AGE is implemented on each modulator to enhance the speech. The system is shown in figure 9.1. The mathematics for AGE in the modulation domain is the same as for AGE in the sub-band domain, the long term average and the short term average are calculated for each sub-band modulator, instead of the sub-band itself. The gain function is multiplied with the modulator of the sub-band to yield a modified modulator \( \tilde{m}_k(n) \) which is then used with the carrier in the reconstruction stage of the modulation system.

\[
\tilde{m}_k(n) = m_k(n)G_k \quad (9.17)
\]

\[
\tilde{x}_k(n) = c_k(n)\tilde{m}_k(n) \quad (9.18)
\]

The synthesized signal \( y(n) \) is finally calculated by adding up all the components.

\[
\tilde{x}(n) = \sum_{k=1}^{K} \tilde{x}_k(n). \quad (9.19)
\]

The gain function \( G_k \) is given by

\[
G_k = \min \left( L, \frac{A_{f,k}}{L_{opt}A_{s,k} + \epsilon} \right) \quad (9.20)
\]

where \( A_{f,k} \) denotes short term average and \( A_{s,k} \) denotes the long term average, \( L \) is a limiting threshold which limits the gain function’s value and \( L_{opt} \) is an optimum level of control on the value of the gain function. The averages are computed as:

\[
A_{f,k}(n) = \alpha_f A_{f,k}(n-1) + (1 - \alpha_f) | m(n) | \quad (9.21)
\]

\[
A_{s,k}(n) = \alpha_s A_{s,k}(n-1) + (1 - \alpha_s) | m(n) | \quad (9.22)
\]

\[
A_{s,k}(n) = \min (A_{s,k}(n), A_{f,k}(n)) \quad (9.23)
\]

where \( \alpha_f \) and \( \alpha_s \) are time constants of the short term and long term averages, respectively.

9.4 Evaluation of The Proposed System

Figure 9.2 shows the experimental setup, where \( s(n) \) is the clean speech signal, \( v(n) \) is a noise signal and \( x(n) \) is the sum of speech and noise signals
9.4. Evaluation of The Proposed System

Table 9.1: Spectral distortion (SD) results

<table>
<thead>
<tr>
<th>Noise SNR</th>
<th>0 dB</th>
<th>10 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 to 5</td>
<td>5 to 20</td>
</tr>
<tr>
<td>(L_{opt}) range</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>SD for FN [dB]</td>
<td>-18 to -4</td>
<td>-4 to -2</td>
</tr>
<tr>
<td>SD for IN [dB]</td>
<td>-18 to -4</td>
<td>-4 to -2</td>
</tr>
<tr>
<td>SD for TN [dB]</td>
<td>-18 to -6</td>
<td>-4 to -2</td>
</tr>
</tbody>
</table>

\(s(n) + 10^{-\frac{\text{SNR}}{20}} v(n)\) scaled by desired level of Signal to Noise Ratio (SNR). \(m_s, c_s, m_x, c_x, m_v\) and \(c_v\) are the signal matrices of modulators and carriers for \(s(n), x(n)\) and \(v(n)\) respectively. The gain matrix \(G\) is calculated by passing \(m_x\) through AGE system. This \(G\) is then multiplied with the \(m_x, m_s\) and \(m_v\), whereafter the re-modulation and the synthesis processes generate the output signals \(\tilde{x}(n), \tilde{s}(n), \tilde{v}(n)\), as depicted in figure 9.2. The system was evaluated with the following parameter settings. \(L = 1, L_{opt} = 1\) to 20, \(T_s = 4s\) and \(T_f = 0.04s\). The speech signals comprise male \(F_s=16\) kHz and female \(F_s=16\) kHz speech signals and the noise signals are scaled so as to have 10 dB, 5 dB, 0 dB and -5 dB SNR. Noise signals used were Engine Noise (EN), Factory Noise (FN), Gaussian Noise (GN), Tonal Noise (TN) and Impulse Noise (IN). The performance measurement was evaluated by the Signal to Noise Ratio Improvement (SNRI), Perceptual Evaluation of Speech Quality (PESQ) and Spectral Distortion (SD). SNRI of male speech signal for TN at 0 dB SNR with \(L_{opt} = 20\) was around 10 dB and for other noises was between 4 dB to 6 dB. The female speech signal also had SNRI of 9 dB for TN and around 3 to 5 dB for EN, FN, GN, IN at 0 dB SNR. PESQ has been calculated by comparing \(s(n)\) and \(y_s(n)\) which gives an objective measure of how much degradation the system has introduced on the speech signal due to introducing the AGE gain function. The objective Mean Opinion Score (MOS) as computed by the PESQ for most of the tests given above was 3, which is considered fair for speech signals. Experiments have been performed to find out the optimal value on the critical system parameter \(L_{opt}\), for different noise cases and for different speaker situations. Figure 9.3 shows the MOS values for both male and female speech signals at 10 dB of noise SNR. It is interesting to note that female speech has higher values of

Figure 9.3: MOS for the processed male speech signal (upper) and female speech signal (lower) with noise at 10 dB SNR

MOS than male speech under similar conditions. This observation is attributed to the fact that female speech with higher pitch is less affected by some noises. Moreover, the SD is very low for $L_{opt} < 5$ and then increases rapidly with increasing $L_{opt}$ values for all tests. For male speech signal, the SD at $L_{opt} = 20$ is around -2 dB and -4 dB for FN, GN, TN and IN and some of them are shown in table 1. The female speech signal has different behavior than the male speech
9.4. Evaluation of The Proposed System

The proposed method was also compared against the speech enhancement method by AGE proposed in [3]. It was observed that the proposed method has better performance than the reference method of [3]. One such comparison is shown in figure 9.4 where a male speech signal having mixed with 5 dB SNR factory noise is enhanced by two methods and the proposed method clearly outperforms its counterpart in [3].

Figure 9.4: SNRI plots of two speech enhancement methods

signal on SD. For female speech, SD is found to be -2 dB for EN, GN, IN and -4 dB for FN and -8 dB for TN at the $L_{opt}$=20.

9.4.1 Spectrogram Analysis

The spectrogram of a male speech signal that has been mixed with gaussian noise at 10 dB SNR and the spectrogram after enhancement with the proposed AGE system, are given in figure 9.5. The AGE algorithm converges after 0.2 seconds for all test cases, whereafter it may be observed that the disturbing noise is reduced while the formants of the speech are maintained. Enhanced signal $y_x(n)$ has shown the formants very clearly after the processing. Although the Gaussian noise is spread throughout the frequency plane, the AGE works very efficiently, but a little bit speech signal energy has also been lost. The spectrogram of male speech signal mixed with tonal noise at 0 dB SNR and enhanced male speech signal by the AGE was also observed. The tonal noise which had

Figure 9.5: Spectrogram of noisy male speech (upper) having Gaussian noise at 10 dB SNR and the enhanced signal by the proposed method (lower)

all of its energy around 1 kHz has been reduced by the AGE, i.e., reduced its energy, while maintaining the formants of speech. Moreover, the impulse noise at 0 dB SNR, which is similar to gaussian noise in spreading its energy through all the frequencies, has been successfully eliminated.

9.5 Conclusion

A novel approach of speech enhancement in modulation frequency domain has been explored and the promising results obtained by using the proposed method have been presented in this paper. The adaptive gain equalizer (AGE), which has shown its advantages already in digital, analog and hybrid domains by its simplicity, low complexity for being robust to different noisy environments, has been implemented in the modulation frequency domain in this paper. The detailed analysis of the system has put light on its advantages and disadvantages, i.e. where the evaluation section highlights the compromise between low SD and high SNRI. The system provides good improvement on the female speech signal, with better SNRI, low SD, fair MOS, and output speech signal sounds good. The maximum SNRI obtained for the female speech signal analysis was approximately 9 dB and SD of female speech for some noise has been shown 0 dB.
9.5. Conclusion

The spectrogram analysis provides another view of these results. The AGE gain function adapts during the first 0.2 seconds. This start-up time can be reduced by varying the integration time, but changing the integration time has obvious consequences on the signal integrity and the noise reduction performance. Moreover, the proposed method has shown its potential as a better alternative to the traditional methods of speech enhancement. Future work is to implement this system in real time and other speech enhancement methods may also be tried in modulation domain.

References


9. **Modulation Domain Adaptive Gain Equalizer for Speech Enhancement**


Ten

Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization

This chapter has been published as:

Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization

Rizwan Ishaq, Muhammad Shahid, Benny Lövström, Begoña García Zapirain, and Ingvar Claesson

Abstract

Adaptive gain equalizer (AGE) is a commonly used single-channel speech enhancement algorithm. AGE and its variants have been widely used for speech enhancement applications. There are two broad categories of these variants. The first deals with its improvement in time-frequency domain with readjustment of the used parameters and the second one deals with performing the main filtering operation in modulation frequency domain. This paper evaluates the working of AGE in modulation frequency domain with the use of a demodulation technique which solves the demodulation process as a convex optimization problem. The performance of the modified AGE is compared with the traditional AGE and another modulation frequency domain AGE based on demodulation using the spectral center-of-gravity. These used performance measures are Signal to Noise Ratio Improvement (SNRI), Spectral Distortion (SD) and Mean Option Score (MOS).

10.1 Introduction

Different types of background noise corrupts the otherwise clean speech signals in everyday communication. A phone call can be disturbed by a variety of noises present nearby ranging from computer fan noise to factory noise. There have been a variety of methods for reducing noise from speech signal, e.g., spectral subtraction [1] and optimum Wiener filtering [2]. The commonly used method for reducing noise is spectral subtraction but it has an inherent problem of generating musical noise due to spectral flooring [3]. There have also been some efforts to reduce this musical noise such as [4] but this improvement has the tendency of producing audible-distortion causing listening discomfort.
10. Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization

even compared to the unprocessed signal [5]. Reducing noise without generating artifacts was proposed in [6] but this method fails to address unvoiced speech.

The Adaptive gain equalizer (AGE) is a time domain speech enhancement algorithm in which the speech signal is amplified based on signal-to-noise (SNR) estimates in subbands. A signal is divided into subbands for calculation of a gain which is independent for each band. The algorithm has shown advantages over contemporary techniques because of its low complexity implementation, no requirement of voice activity detector (VAD) and has no presence of musical noise as a result of controlled gains [7]. Additionally, hardware implementations of AGE [8] indicate its importance in speech processing applications.

As an alternative to time domain processing, an implementation of AGE in the modulation domain was presented in a recent study [9]. This method was mainly inspired by the performance advantages of splitting the signal into its frequency bands. The modulation system assumes a speech signal as composed of a modulator and a carrier. Thus the signal is represented by

\[ x(t) = m(t)c(t) \]  

(10.1)

where \( m(t) \) denotes the low frequency part of the signal, called the modulator, that modulates a high frequency carrier \( c(t) \). Studies have shown that the modulators of a speech signal are more important for the intelligibility of the speech signals than their counterpart carriers [10]. Modulation systems are based on sub-band modulators and hence perfectly fit the AGE system which works on the sub-bands of the signal. Besides the fact that the study in [9] has reported improvement in performance measures in speech enhancement in comparison to time-domain AGE, the proposed center of gravity (COG) demodulation does not involve an optimization step, the need of which we state in the following.

In this work, we consider AGE in modulation domain by demodulation process as a convex optimization problem presented in [11]. The reason of adaptation of this technique for AGE in modulation domain is mainly the ambiguity associated with the demodulation process of having unlimited number of possible modulator-carrier pairs. Moreover, proven ability of this method for efficiently demodulating a variety of carriers such as harmonic, stochastic and time-varying ones further justifies its usage.

An account of related work in modulation domain and a brief introduction of AGE is provided in Section 10.2. Section 10.3 describes a modulation system, a summary of a demodulation technique called spectral center of gravity that used in AGE implementation given in [9]. Section 10.4 starts with an introduction of solving demodulation as an optimization problem and completes with
the description of the proposed model of AGE. A comparison of performance of the proposed model is presented in Section 10.5 with its time-domain and modulation domain counterparts. Finally, some conclusive remarks about this work are drawn in Section 10.6 with an outline of possible future works in the area.

10.2 Background

AGE can attenuate noise in speech signals in real time with low computational complexity [12]. It uses an FIR filter bank to divide a speech signal into sub-bands where speech in each subband is amplified independently. It was also shown that the system can adopt itself for different types of noise. The proposed AGE method using the mixed analog and digital hybrid approach yield around 13 dB speech enhancement [13]. The AGE was originally intended for the digital domain, but [13] provides an analog implementation which does not use quantization and digitization and is best suited for battery powered applications. A hybrid solution to overcome problems related to a digital and an analog implementation of the AGE is found in [14].

Zadeh [15] introduced the modulation domain as a two dimensional bi-frequency system, where time variation of the ordinary frequency is the second dimension. Since then, there have been reasonably large interest in this field for various tasks related to speech processing. Atlas et al. used the concept of coherent modulation for the target talker enhancement in speech enhancement [16]. They proved that working in modulation domain can increase the speech intelligibility. Coherent modulation using the frequency reassignment has been used for speech enhancement and for demodulation of a signal into modulator and carrier [17]. Speech polluted by wind noise has been enhanced by using coherent modulation comb filtering as reported in [18]. Although the modulation filtering has mostly been used for the purpose of speech enhancement, we find some of its applications in audio compression as well [19]. It was showed that a 32 kb/s/channel outperformed MPEG-1 coded at 56 kb/s/channel (both at 44.1 kHz), using the modulation technique.

10.3 Modulation Domain and AGE

An acoustic spectrum is transformed by short-time Fourier transform into the modulation domain spectrum at a particular acoustic frequency. It has been observed that speech intelligibility can be altered by operating on modulator
10. Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization

part of the signal. Shamma [20] reported that auditory cortex neurons possibly decompose the acoustic contents into spectro-temporal modulation contents. It has been found that if the modulators of the speech signal are replaced by constant amplitude modulators, while carriers are preserved, speech does not remain intelligible anymore. However, when the modulators are preserved but carriers are altered, the speech is intelligible [10]. A modulation frequency system is described by the following steps:

- Filter bank to get sub-band signals
- Demodulation i.e., decomposition of each sub-band signal into a modulator and a carrier.
- Analysis of the modulators of the sub-band signals by discrete Fourier transform of each modulators
- Modification of the modulators (e.g. linear filtering)
- Re-modulation (recombination of modified modulators with original carriers)
- Synthesis of signals

The modulation system’s filter bank divides the wide-band signal into K narrow-band sub-bands. The signal \( x(t) \) is passed through the filter bank set of band-pass filters \( h_k \), which renders the sub-band signals \( x_k(t) \).

\[
x_k(t) = h_k(t) * x(t)
\]  

(10.2)

where \( * \) is convolution operator. The demodulation process decomposes the sub-band signal into its envelope and carrier. It is efficient to decimate the sub-band signals so that the redundant samples may be removed. Modification of the modulators is done by the modulation filtering \( g(t) \), i.e., \( \tilde{m}_k(t) = m_k(t)g(t) \).

A modulation spectrogram and modulation analysis can be done by computing the Fourier transform along the time-axis of the spectrogram (magnitude) or by utilizing the spectrum of the envelop signals, which gives the modulation frequency along the horizontal axis and acoustic frequency along the vertical axis. Re-modulation is the process in which modified modulators \( \tilde{m}_k(t) \) are combined with the original carriers, obtained in the process of demodulation, to get the modified sub-band signals \( \tilde{x}_k(t) \). The synthesis process reconstructs the modified signal \( \tilde{x}(t) \) using the modified sub-band signals \( \tilde{x}_k(t) \), according
to the following equation. Interpolation must be performed prior to this stage if decimation was done before.

\[ \tilde{x}(t) = \sum_{k=1}^{K} \tilde{x}_k(t) \]  

(10.3)

Following is a brief description on one of the methods used for coherent carrier detection which is also used in this work, apart from convex optimization demodulation process.

### 10.3.1 Spectral Center of Gravity Carrier Estimation

In the Center-of-Gravity (CoG) approach, instantaneous frequency \( \omega_k(n) \) is defined as instantaneous spectrum average frequency of \( x_k(t) \) at time \( t \) [21]. An instantaneous spectrum with short-time Fourier transform is computed as,

\[ S_k(\omega, t) = \sum_p g(p) x_k(t + p) e^{-j\omega p} \]  

(10.4)

where \( g(p) \) is a short spectral-estimation window. The instantaneous frequency \( \omega_k(t) \) of the sub-band signal \( x_k(t) \) is estimated as,

\[ \omega_k(t) = \frac{\int_{-\pi}^{\pi} \omega |S_k(\omega, t)|^2 d\omega}{\int_{-\pi}^{\pi} |S_k(\omega, t)|^2 d\omega} \]  

(10.5)

The phase \( \phi_k(t) \) of the carrier is computed as follows

\[ \phi_k(t) = \sum_{p=0}^{t} \omega_k(p) \]  

(10.6)

The carrier \( c_k \) is

\[ c_k(t) = e^{j\phi_k(t)} \]  

(10.7)

and the complex valued modulator \( m_k(t) \) is given by

\[ m_k(t) = x_k(t)c_k^*(t) \]  

(10.8)

### 10.3.2 Adaptive Gain Equalizer System

The AGE consists of a filter bank and each sub-band is weighted by a gain function which amplifies the signal when speech is present and keeps the noisy
10. MODULATION FREQUENCY DOMAIN ADAPTIVE GAIN EQUALIZER

Using Convex Optimization

Figure 10.1: Adaptive gain equalizer in modulation domain

part of the signal, where no speech is present, to unity [7]. A filter bank of K bandpass filters divides the input signal \( x(n) \) into K sub-bands \( x_k(n) \).

\[
x_k(n) = h_k(n) * x(n)
\]

(10.9)

Here \( h_k \) is the impulse response of the filter bank sub-band \( k \) and \( * \) denotes the convolution. The output signal \( \tilde{x}(n) \), with the amplified speech signal, is computed as

\[
\tilde{x}(n) = \sum_{k=1}^{K} G_k(n)x_k(n)
\]

(10.10)

where \( G_k(n) \) is the AGE weighting function which amplifies the signal when speech is active and is given by

\[
G_k(n) = \min \left\{ \left( \frac{A_k(n)}{L_{opt}B_k(n)} \right)^{p_k}, L_k \right\}
\]

(10.11)

where \( L_{opt} \) is the optimized suppression level for gain function and \( p_k \) gain rise exponent constant. \( L_k \) is a limiting threshold limiting gain function value. Fast average \( A_k(n) \) and slow average \( B_k(n) \) of sub-band \( k \) calculated according to:

\[
A_k(n) = \alpha_k A_k(n-1) + (1-\alpha_k)|x_k(n)|
\]

(10.12)

where \( \alpha_k = \frac{1}{f_s T_a} \) is forgetting factor constant and \( f_s \) is sampling frequency.

\[
B_k(n) = \begin{cases} A_k(n) & \text{if } A_k(n) \leq B_k(n-1) \\ (1+\beta_k)(B_k(n-1)) & \text{otherwise} \end{cases}
\]

(10.13)

where \( \beta_k = \frac{1}{f_s T_b} \) is a positive constant control the noise level. Based on the above mentioned principle of AGE, a speech signal modulator can also be enhanced by the equalizer. Modulation domain separates each sub-band signal...
### 10.4 Convex Optimization and the Proposed Model

into a carrier and a modulator. While only modulators are considered here, the AGE is implemented on each modulator to enhance the speech. The system is shown in figure 10.1. The mathematics for AGE in the modulation domain is the same as for AGE in the sub-band domain, the long term average and the short term average are calculated for each sub-band modulator, instead of the sub-band itself. The gain function is multiplied with the modulator of the sub-band to yield a modified modulator $\tilde{m}_k(n)$ which is then used with the carrier in the reconstruction stage of the modulation system.

$$\tilde{m}_k(n) = m_k(n)G_k$$  \hspace{1cm} (10.14)

$$\tilde{x}_k(n) = c_k(n)\tilde{m}_k(n)$$  \hspace{1cm} (10.15)

The synthesized signal $\tilde{x}(n)$ is finally calculated by adding up all the components.

$$\tilde{x}(n) = \sum_{k=1}^{K} \tilde{x}_k(n).$$  \hspace{1cm} (10.16)

#### 10.4 Convex Optimization and the Proposed Model

One inherent problem with the demodulation technique is the unfortunate presence of unlimited number of possible yet valid modulator-carrier pairs. This predicament can be understood by taking example of a sinusoidal signal that is composed of multiple frequency sinusoids. Such a signal can be decomposed into more than one legitimate modulator and carrier pairs, that are equally correct mathematically. Similar is the case with speech signals when the problem of demodulating it into modulator and carrier is dealt. Thus there is need to add some conditions to the problem which can make the algorithm result into the desired solution. A general optimization problem minimizes a given objective function while fulfilling a set of equality and inequality constraints. If the objective function and inequality constraints are all convex and the equality constraints are all affine, the problem is called a convex optimization problem [22]. Although the modulation problem of equation 1 is not convex as it is, two methods have been suggested in [11] for constraining modulation into convex restrictions. One solution is to work in logarithm domain where the optimization variables can be defined simply as the logarithm of the
squared linear optimization variables $m(t)$ and $c(t)$. A convex relation is then obtained by just summing the two logarithmic domain variables. The other method of making the problem convex is to work in linear domain where the process involves eliminating the carrier $c(t)$ and minimization of only the modulator signal is done. The final expression obtained in linear domain convex optimization is given by the following:

Minimize $C_m(m(t)) + C_c(m(t)^{-1}x(t))$

where the modulator cost function $C_m$ can be any convex function but the carrier cost function $C_c$ must be both convex and non-decreasing as a requirement of making the problem a convex one. We have followed the linear domain convex optimization method in our work. The interested reader is referred to [11] for detailed analysis of these methods.
10.5 Comparative Performance Analysis

10.5.1 Mean Opinion Score (MOS)

The Mean Opinion Score (MOS) calculated by observing the clean speech signal processed by a system to check how much it degrades the clean speech signal. Fig. 10.2 shows a female speech signal processed by a system where SNR has been set -10dB for both Engine Noise (EN) and Factory Noise (FN). The system with convex demodulation has MOS value around 3.5 for EN and 3.8 for FN which provides less degradation as compare to CoG modulation and AGE system where is average MOS observed 3, and less than 3, respectively.
### 10.5.2 Spectral Distortion

Fig. 10.3 shows the Spectral Distortion (SD) for female speech signal contaminated by EN and FN at the SNR of -10dB. The increasing value of $L_{opt}$ increases SD up to 10dB for EN when the system uses AGE while for convex demodulation average SD around 7dB and for CoG demodulation its around 9dB, but for FN, SD for all the system observed around 3dB average.

### 10.5.3 Signal to Noise Ratio Improvement (SNRI)

Fig. 10.4 shows the Signal to Noise Ratio Improvement (SNRI) for AGE, MAGE (CoG and Convex demodulation) for a female speech signal distorted by EN and FN having SNR of -10dB. The MAGE methods with convex demodula-
10.5. Comparative Performance Analysis

Figure 10.5: Spectrogram with Factory Noise (FN) (SNR=-10dB)

...portion has the highest SNRI for all the values of $L_{opt}$ and around 5dB and 8dB improvement over the AGE and MAGE (CoG) methods for EN. But for FN system show improvement after $L_{opt} = 12$. The MAGE (CoG) in start improved significantly but with increasing value of $L_{opt}$ MAGE (Convex demodulation) has better improvement.

10.5.4 Spectrogram Analysis

Fig. 10.5 and 10.6 shows spectrogram of original signal with processed signal with AGE, MAGE (convex and CoG demodulation) for FN and EN respectively. The MAGE (convex demodulation) improvement can be observed in term of speech formants being not effected, as visible in spectrogram for both EN and FN.
10. Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization

Figure 10.6: Spectrogram with Engine Noise (FN) (SNR=-10dB)

10.6 Conclusion

An alternative method of demodulation has been proposed for AGE in the modulation frequency domain. The presented method solves the demodulation process as a convex optimization problem, thereby avoiding the inherent problem of multiple solutions of a demodulation algorithm. We have tested the proposed method for various conditions and magnitudes of noise injected in a clean speech signal. The performance of our method has been validated by mean opinion score, spectral distortion, signal to noise ratio improvement and spectrogram analysis in comparison to two other techniques.
10.7 Bibliography


10. Modulation Frequency Domain Adaptive Gain Equalizer Using Convex Optimization


