Content-Aware Video Transmission in HEVC Context: Optimization of compression, of error resilience and concealment, and of visual quality

Ahmed Aldahdooh

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Ahmed ALDAHDOOH

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : Sciences et technologies de l'information, et mathématiques

Discipline : Informatique et applications
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)

Date du soutenance le 25 août 2017

Content-Aware Video Transmission in HEVC Context

Optimization of compression, of error resilience and concealment, and of visual quality

JURY

Présidente :	**M^me^ Christine GUILLEMOT**, Directeur de Recherche, INRIA
Rapporteurs :	**M^me^ Amy REIBMAN**, Professor of Electrical and Computer Engineering, Purdue University **M. François-Xavier COUDOUX**, Professeur au sein du groupe Communications Numériques de l'IEMN DOAE, Université de Valenciennes
Examinateur :	**M. David BULL**, Professor of Signal Processing, University of Bristol
Directeur de thèse :	**M. Patrick LE CALLET**, Polytech Nantes, Université de Nantes
Co-encadrant de thèse :	**M. Marcus BARKOWSKY**, Polytech Nantes, Université de Nantes

Dedication

To my parents, brothers, and sisters
To my wife, children
“I wish if you joined my happiness in the harvest day. I am optimistic, a lot of happiness come”

Acknowledgment

This work is supported by the Marie Skłodowska-Curie under the PROVISION (PeRceptually Optimised VIdeo CompresSION) project bearing Grant Number 608231 and Call Identifier: FP7-PEOPLE-2013-ITN and by UltraHD-4U project.

Special thanks to everyone helped me through my thesis. Thanks to my professors and advisors.

Background

1

Introduction

The technology evolution of cameras, mobile devices, screens, etc., leads to increase on using such electronic devices. For instance, nowadays, people prefer to use videos in their daily life activities. These activities include video calls and messages, recording videos for memories, recorded CV, etc. Hence, and according to Cisco report [1], the IP video traffic is 70% of all IP traffic in 2015 and expected to increase up to 82%. Besides, according to Cisco forecast [2], Internet video downloads and streaming are the major applications that are expected to have 80% of the bandwidth,by 2020, of all Internet traffic. It is also reported that content delivery network (CDN) is the dominant way to stream videos. 61% of all Internet video traffic crossed CDNs in 2015 and is expected to reach up to 73%. Since people are now connected to different types of networks especially the Internet using different devices (computers, TVs, portable devices, ... etc.), they are demanding for high quality videos. Moreover, the demanding includes immersive contents as well such as ultra-high definition (UHD), high dynamic range (HDR) and 360^o videos. The delivery (from capturing to the end user) of such high quality is challenging. It requires saving bandwidth by using good encoders, producing robust streams to be sent over error-prone channels, ability to recover errors, and finally good quality estimation for perceived videos.

1.1 Problem Statement

Indeed, there are different types of video sequence contents such as sporting, news, natural scene, movies, computer-generated, and cartoon sequences. Each of these content types can be divided into sub categories. Each content has its own features and underlying content characteristics that make the video content different from other contents. For instance, Figure 1.1 shows four video content sequences that cover different amounts of spatial (SI) and temporal information (TI) (perceptual information). SI is a statistical texture measure that finds edges in the frame. TI is a statistical measure that measure the amount of difference between adjacent frames.

As indicated at the beginning of this chapter, the delivery of video contents with quality that satisfies the end users is challenging. One way to improve the current state-of-the-art of video delivery chain is to take advantage of video *content characteristics*. The wide variety of video content causes a challenge for content-based research since, of course, natural scenes differ from sport scenes and sport scenes differ from cartoons, etc. Hence, considering content types and their corresponding features is very important in different aspects. First, in setting up subjective experiment; in [3–5], Video Quality Experts Group (VQEG) and Pinson et al. mention some limitations to video source selection to conduct researches. Second, in improving objective measures of video quality; in [6–8], the video content features are analysed to improve the objective video quality measure. Third, in improving video coding efficiency; it can be concluded from the subjective experiment of Pitrey et al. [9] that the video content influences the video encoding. Fourth, in designing error resilience tools: in [10–13], switch algorithms that utilize the content features are used to decide which error concealment technique should be applied.
Hence, the main focus of this dissertation is to take the advantages of content *features/indicators* to improve the video delivery chain in different aspects. The delivery chain includes, in this dissertation, pre-encoding process, error resilience, error concealment, and quality assessment.

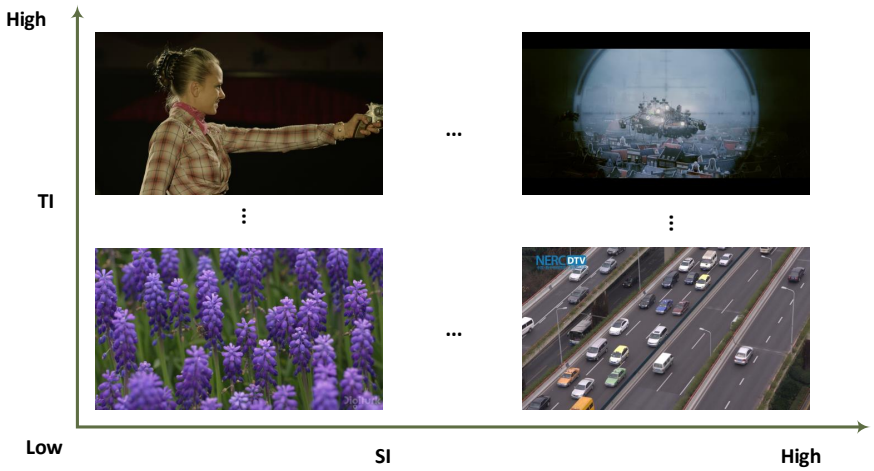


Figure 1.1 – Four sequences with different amounts of spatial and temporal information

1.1.1 Pre-encoding process

The aim of any coder is to reduce the size of the video file to target a specific bitrate budget. The latest video coding standard, High Efficiency Video Coding (HEVC) [14], is designed especially to target different types of applications and particularly high resolution video applications [15]. Quality, bitrate and complexity (encoding time) are the key elements of video coding performance evaluation. The complexity of HEVC is increased due to the new/improved coding tools. This complexity is a liability for some targeted users, for some applications, or, for some devices. Some targeted users, like content providers, may not care about the complexity since they have the power to build high performance encoders, i.e. parallel encoders. Some applications (security and safety applications) require that the captured videos need to be quickly encoded and sent. Due to the limited computational power and batteries of some devices, the complexity is an important issue. This dissertation targeted this issue and investigates the impact of using *content characteristics* to reduce the effect of complexity.

1.1.2 Error resilience

Video transmission system characteristics depend on the application type. Figure 1.2 shows the abstract overview of the video transmission/storage system layers. Real-time application uses RTP/UDP based systems since it is suitable for low-delay applications while progressive download based applications use HTTP/TCP based systems since it is a reliable protocol. In [16] the authors summarize these system technologies to:

- RTP [16], is a real-time transport protocol developed by Internet engineering task force (IETF) and is used over UDP.
- MPEG-2 Systems [17]. It is used in broadcast systems as IPTV.
- The ISO Base Media Format [18] and MPEG-DASH, Dynamic Adaptive Streaming over HTTP, [19]. They are used in VoD streaming, progressive download, and HTTP streaming over Internet. For more information about the standard principles and concepts, readers are advised to refer to [20–23].

The decoded video quality might not be satisfying if one or more packets are lost in error-prone channels. The main goal of video coding like high efficiency video coding (HEVC) [14] is to minimize the coding distortion for a target bitrate. This requires a complex prediction process to remove the redundant information in the video signal [15]. As a result, the error resilience in HEVC is decreased compared to H.264/AVC due to the increase of temporal dependency [24]. Protection methods of a compressed stream [25] can be categorized into three categories: error resilience, error control coding, and error concealment. It is stated in [25] that “Error resilience refers to schemes that introduce error resilient elements at the video compression stage, which reduce the interdependencies of the data-stream, in order to mitigate error propagation”. Error propagation is a phenomenon when successive frames are affected by single/multi transmission error. Several error resilience techniques are introduced in the literature [26–28]. This dissertation targeted this issue and investigates the impact of using *content characteristics* and the impact of using good network structure when multiple description coding is used as an error resilience tool. The target is to: 1) reduce the effect of error propagation that happens due to packet losses, and 2) to reduce the amount of redundant data that needs to be sent.

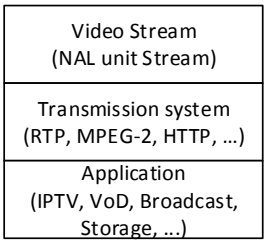


Figure 1.2 – Abstract overview of video transmission/storage system layers

1.1.3 Error concealment

Error concealment (EC) is one category of methods to protect the compressed stream. EC is a passive operation that does not place resilient element in the streams but uses the correctly received video data to recover the lost information. One objective of an error resilient tool is to make the job of error concealment easier. Spatial EC techniques [29, 30] utilize available surrounding pixels to reconstruct the missing pixels. They are not efficient for large areas, non-constant areas, and in terms of complexity. They usually reconstruct the texture (such as edges) but not the structure. Temporal EC techniques use available motion information to predict the missing motion vectors (MVs), for instance, by interpolating [31] or by selecting the MV that minimizes the side match distortion [32]. Despite providing information about whether the current area is moving or not, this technique is efficient only for low-motion and smooth sequences and for small areas since the precision of predicted MVs is not guaranteed. Thus, the structure (of copied data) is reconstructed but not the texture.
The target of any error concealment algorithms is twofold: reconstructing a satisfying reconstruction of a lost area and reducing the mismatch between the encoded and the reconstructed blocks which yield to reduce the error propagation effect. To achieve that we need to reconstruct the texture and the structure of a missing area and that can be done using inpainting techniques. This dissertation targeted this issue and investigated the impact of using *content characteristics* in improving the inpainting-based error concealment algorithms.

1.1.4 Quality assessment

Human satisfaction is one aspect quality of experience (QoE). The popular metric to measure video quality is MSE/PSNR which does not necessarily express the human satisfaction especially when parts of the video are delayed or concealed. Researchers trust observers' judgements in video quality assessment, although building subjective experiments are very expensive and cannot be incorporated in real-time systems. Many efforts have been dedicated to implement objective measurements to automatically assess the image or video quality and give results very close to what human observers give. Quality of Experience (QoE) is defined in [33] as "the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state". Subjective experiment is the accurate way to judge the quality of the perceived video. Because of its inability to be part of the video delivery and its time and money consumption, efforts have been dedicated to implement objective measurements to automatically assess the image or video quality taking into account *perceptual properties* of the content and the *human visual system properties*.
This dissertation targets different issues of video quality assessment. Firstly, how to study the agreement of different objective video quality measures and the influence of the *content types* and the encoding conditions on this agreement. Secondly, the impact of *content characteristics* are studied in order to see the correlation of content features with respect to the behaviour of full-reference objective measurements for error-free and loss-impairment videos. Thirdly, selecting encoding conditions for testing the objective measures is also investigated. Finally, to mitigate the shortcomings of current methods (correlations and mean square errors) to evaluate the performance of the video quality measure, analyses are conducted to introduce new methods.

1.2 Main research questions and contributions

As the subtitle of this manuscript indicates, the main focus of this dissertation is to *utilize the content characteristics of the video shots to improve the video delivery chain.* For each component of the video delivery chain that is mentioned in the Section 1.1, the following research questions are investigated:

— **Pre-encoding process:**

1. Can the generic/global content features be used as indicators for finding the links between the content features and the encoders parameters? If so, then building a joint content and complexity aware encoder's parameters prediction model is applicable.
   In addition to that, the following secondary research questions are investigated too:
   + How does the encoder behave in terms of complexity with different content?
   + How is the encoder complexity linked with different parameters per content?

— **Error resilience:**

1. Which content features can be used in order to take advantage of the received redundant representations/descriptions when using $n-$MDC with $n \geq 4$?
2. With these features, is the quality of experience (QoE) of the reconstructed video sequence improved?
3. Which content features would help to build an adaptive MDC scheme?
   In addition to that, the following secondary research question is investigated too:
   + Can we trade-off between quality and bitrate in MDC schemes by not always using a specific MDC scheme? In other words "Is it better to use SDC, 2-MDC, or 4-MDC for a specific content?"
4. How can the quality of temporal-MDC scheme be evaluated?
   In addition to that, the following secondary research question is investigated too:
   + How to take advantage of TCP and UDP protocols to build a good networking structure that allows to reduce the amount of redundant MDC data to be sent.

— **Error concealment:**

1. What is the information, content indicators, that needs to be considered as inputs of the inpainting-based EC algorithm?
   In addition to that, the following secondary research question is investigated too:
   + How to adapt the state-of-the-art inpainting-based EC algorithms to be suitable for low delay communication?
2. Does the observer get disturbed with the proposed inpainting-based EC algorithm? Does that correlate with DMOS?
3. Which content indicators may help in predicting this correlation?

— **Quality assessment:**

1. How do different full-reference video quality measures behave in terms of ranking for the error-free and loss-impairment sequences?
2. Characterize the behaviour of FR video quality measures at frame and sequence levels with respect to video content and coding parameters.
3. What is the impact of using pixel-based content features in building machine learning based NR VQA for error-free and loss-impairment sequences along with channel and coding parameters?
4. Can a representative subset be selected from a large-scale database such that this small-scale database can be further analysed and the conclusions drawn on the small-scale database also apply to the large scale database?
5. In case that the PLCC and RMSE cannot report the goodness of a model, what other performance measures that we need to report this goodness?
   In addition to that, the following secondary research question is investigated too:
   + How does the goodness analysis work when the content sources are different as well as the HRCs?

Research has been conducted to utilize the *content characteristics* in the above-mentioned fields. The aims, based on the raised research questions, were to develop frameworks, methods, and algorithms to integrate video content features as a main component. To reach that end, different types of analysis and experiments (including subjective experiments) have been conducted. As a result of these efforts, this dissertation presents the following contributions:

— **Contribution #1:** A framework to predict the encoder parameters at the sequence level using content indicators has been proposed: **(Part II)**

1. The primary contribution of this work is the prediction of the encoding parameter values leading to minimum complexity in terms of execution time using the underlying content features. For instance, features like cross correlation, Laplacian-based, chrominance information, and motion intensity features have a high impact in finding the links between the content features and the motion search range parameter in HM encoder. The model trades-off rate (R), distortion (D), and complexity (C). For instance, if a video sample is encoded using different configurations and the output videos are in same bitrate and distortion ranges, the configuration that achieve the minimum encoding time will be chosen. If the output videos achieve same complexity and distortion ranges, the configuration of the lowest bitrate will be chosen.

— **Contribution #2:** Introducing content awareness in MDC and subsequently introducing a quality control scheme for MDC that exploits state-of-the-art network architectures: **(Part III)**

1. A new temporal MDC scheme which is characterized by standard compatibility, redundancy tuning, lightweight complexity, and suitability for n-MDC schemes. This scheme includes the process of generating descriptions and the process of reconstructing video sequences when primary data is lost. Coding unit splitting and the temporal distance properties are used to train a weighting coefficient to reconstruct the lost primary frame from the redundant frames.
2. The subjective experiment that shows the preference of the proposed scheme against other MDC schemes is introduced.
3. An adaptive content-aware framework to predict the suitable description scheme (SDC, 2-MDC, or 4-MDC) to be transmitted over an error-prone channel in order to maximize the quality of experience. The contrast of Gray Level Co-occurrence Matrix and the ratio of entropy of Laplacian levels 4 and 5 features are used to build the adaptive MDC scheme.
4. Quality evaluation framework for temporal-MDC schemes is proposed. The framework introduces an interactive networking structure that helps reducing the amount of redundant data to be sent. During this work, all the steps of the proposed framework are done except the quality evaluation (subjectively and objectively) stage. This due to time and computing power limitations.

— **Contribution #3:** Content-aware inpainting-base EC algorithm and subjective experiment to study the subject's disruption: **(Part IV)**

1. A modified version of the inpainting-based error concealment [34] is proposed. The following improvements are achieved in the proposed algorithm:
   — The concept of motion map M_c is introduced. It includes the predicted motion vectors M_{mv}, the pixel-based motion intensity M_{pi} and the motion vector of interests (MVI) that relate to camera motion M_{cm}. It was shown that the proposed motion map improves the performance of the inpainting.
   — The algorithm is adapted to be practical for low-delay video communications.
   — An adaptive search window size for temporal and spatial inpainting is introduced.
   — Reduce the spatio-temporal artifacts using simple Poisson blending strategy with the proposed mask strategy.
2. A subjective experiment that analyses the observer disruption when loss-impairments are introduced in the video sequence. It is observed that the disruption measure has a high correlation with the perceived DMOS. In addition to that, it is shown that the inpainting-based EC technique achieves a better perceived quality with respect to one of the-state-of-the-art EC techniques.
3. Three content features are introduced to study the subject disruption as one step forward to help measure the quality of the perceived degraded videos. The features are: texture, colour, and motion entropy maps.

— **Contribution #4:** Content-aware VQA that predicts the behaviour of full-reference measures is proposed and a content-based subset selection algorithm is proposed as well: **(Part V)**

1. The agreement between the three tested measures PSNR, SSIM, and VIFP showed that the results of their predictions are similar, notably in the high and low quality range, less so in the middle range. It was further noted that the disagreement of the measurements is more pronounced in case of packet loss than for coding-only conditions which may be seen as a first step towards an automatic identification of the scope of application for objective measures. Thanks to the large size of the analysed dataset, some important effects on the characterization of the performance were highlighted that are not evident when a limited set of contents and parameters is considered.
2. The disagreement between several objective measures exist on a frame-level even if the measures agree on a sequence level. However, the particular patterns of this disagreement point to two important conclusions. The first conclusion is that the usage of one single measure may not be sufficient. In particular, it may be beneficial to analyse the usage of several complementary algorithms within the coding loop, i.e. for rate-distortion optimization. In addition, it should be noted that performance bias may occur when improvements are measured only objectively and only using one single method, thus weakening such proposals. The second conclusion is that the pronounced correlation between content characteristics and encoder parameter selection encourages further analysis, for example with respect to the efficiency of rate-control algorithms. Some coding factors are almost not influential, whereas others have a strong impact, suggesting that quality comparisons among sequences without considering the detailed behaviour of the quality over the frames in the sequence itself could be strongly misleading.
3. A content-based NR VQA is built for error-free and loss-impairment video sequences along with coding, and channel characteristics. It predicts the behaviour of the full-reference VQA. The following features are found useful for the prediction model:
   — Channel parameters: loss rate, average and burst length.

- Coding parameters: GOP size/type, intra-period, number of slices, open/close GOP, QP.
- Channel and coding parameters: number of frames hit, number of slices hit, and number of affected frames.
- Content-based features:
  + Gray-level co-occurrence matrix properties,
  + Chrominance information,
  + Spatial and temporal information,
  + Cross-correlation,
  + DCT and Laplacian based properties,
  + Motion intensity, and
  + MPEG-7 motion activity descriptor.

4. Two subset selection algorithms are proposed. They are targeting a wide range of a specific target; quality/bitrate or content targets. Specifically, a small-scale set is selected from a large-scale database such that this small-scale database can be further analysed and the conclusions drawn on the small-scale database also apply to the large scale database.
5. The following new performance measures are proposed for learning-based video quality assessment algorithms:
   - Measures depend on analysing the residual error using PCA,
   - Measures depend on analysing the confidence intervals of the predicted data, and
   - Measures depend on analysing the confidence intervals of the linear coefficients of the trained and tested models.

1.3 Dissertation structure

Box 1.1 – Note

The work that is presented in this dissertation has been published or submitted for publications in different international conferences and journals. Hence, the chapters of this dissertation provide a complete overview of these publications.

This thesis consists of six parts including the background, conclusions, and perspective future works. Figure 1.3 shows the main four parts of this thesis and how it is connected together.

1.3.1 Part I: Background

The scientific part of this manuscript begins with Chapter 2. It reviews the related works of the main parts that are mentioned in Section 1.2. Firstly, the related works of how the complexity of the encoder can be reduced are reviewed. Secondly, the definitions of the quality of service and the quality of experience are reviewed. Thirdly, since multiple description coding (MDC) is a method that falls in source coding category of error resilience, an overview of source coding technique is discussed. It explains why we conducted research efforts on MDC. Fourthly, error concealment techniques are reviewed to explain why inpainting-based error concealment technique is pursued. Finally, the research efforts to introduce the no-reference video quality measures are reviewed.
In Chapter 3, an overview of global/local content features are introduced. Some of these features are developed for this work. Besides, the procedures to select few video sequences of a big video sequences set are illustrated. Finally, the list of contents that are used in different parts of this dissertation is listed.

1.3.2 Part II: Proof-of-Concept: Role of Generic Content Characteristics in Optimizing Video Encoders; predicting the video encoder's parameters

This part consists only of one chapter, Chapter 4. This chapter deals with the complexity issue from a novel perspective. The well-known techniques to deal with this matter mainly focus on not using some of the coding tool(s) to reduce the complexity. On the other hand, the proposed framework uses the underlying content features to predict the encoder parameters that trade-off between bitrate, distortion, and complexity. The prediction process starts before the encoding. It means that the predicted parameter values are applied to the whole sequence. The prediction model is built offline. The input of the prediction model are 1) the content features as variables, 2) the preferred encoder parameter value as a class/label. This class are selected after the proposed analysis space is analysed. The framework is tested with UHD video sequences against different encoder (HM) configurations.

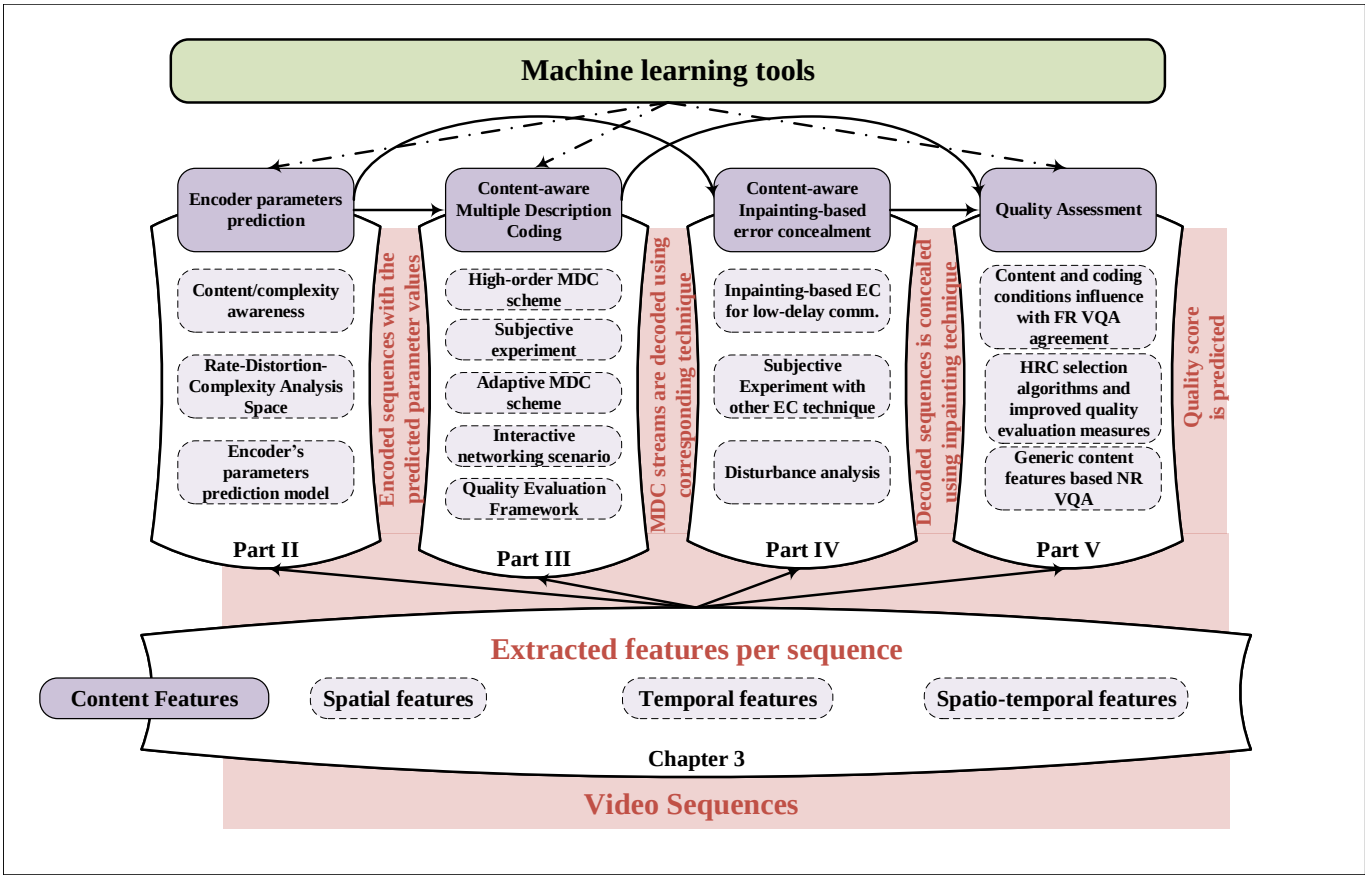


Figure 1.3 – Thesis's parts map

1.3.3 Part III: Content-aware Multiple Description Coding (MDC)

This part introduces the content-aware multiple description scheme and consists of 3 chapters. In the first chapter, Chapter 5, an MDC scheme that is suitable for high-order MDC is introduced. It also provides the recovery procedures if primary frames are lost. The scheme takes advantages of two important properties of the MDC streams in the recovery process: 1) the distance of the corresponding redundant pixels of the lost primary pixels from their reference frames that are used for predictions (three variables), 2) the coding unit (CU) area that belongs to each pixel of the redundant data (three variables). Each group of these six variables are trained with pixel values of the redundant and primary frames to provide weights that minimize the error of the primary pixel value. Then, this scheme is subjectively evaluated and it has been shown that it is preferred than other MDC schemes.
In the second chapter, Chapter 6, an important observation that results from the subjective experiment is analysed. It led to the introduction of an adaptive content-aware MDC scheme. It simply recommends the transmission mode for a specific content. Instead of always sending one type of MDC scheme, a specific type of MDC schemes can be sent according to the content features of the sequence.
Finally, in Chapter 7, it is observed that the content-aware MDC that is proposed in Chapter 6, together with a good/smart networking implementation, proposed in Chapter 7, provide a promising solution to use temporal-MDC scheme as one way to maximize the quality of experience. Quality evaluation framework for temporal-MDC schemes is proposed. The framework introduces an interactive networking structure that helps reducing the amount of redundant data to be sent.

1.3.4 Part IV: Inpainting-based error concealment (EC) technique in video communication

To reconstruct the texture and the structure of a missing area; we can use inpainting techniques. In Chapter 8, an adaptation to one of the state-of-the-art EC algorithms is proposed. The main enhancements include 1) It uses different motion maps; motion vector map, motion intensity map, and camera motion map. 2) adapted to low-delay video communications, 3) adaptive window size for different inpainting steps is proposed. The algorithms have three main steps: inpainting the moving foreground object, inpainting the stationary background temporally and spatially. In Chapter 9, observations from existing loss-impairment video datasets are analysed. Particularly, we analysed the perceived subjective quality of videos containing H.264/AVC transmission impairments, incident at various degrees

of retinal eccentricities of subjects. We relate the perceived drop in quality, to five basic types of features that are important from a perceptive standpoint: texture, colour, flicker, motion trajectory distortions and also the semantic importance of the underlying regions. Then, a subjective evaluation of inpainting-based EC is conducted to study the observer's disturbance of inpainting-based EC technique. This disturbance is correlated with different content properties such as entropy maps of motion, colour, and texture.

1.3.5 Part V: Role of measured content characteristics in quality assessment

This part contains three chapters. In the first one, Chapter 10, we are aiming to use the large-scale database in order to conduct analysis and see observations that cannot be obtained with the subjective experiments. 1) FR measure agreement for error-free and loss-impaired sequences is analysed. 2) Impact of content and coding condition in FR agreement consistency is studied frame-wise within individual sequences and across sequences.
Following the conclusions of Chapter 10, Chapter 11 utilizes and trains content features to predict the behaviour of the FR video objective measures for error-free and loss-impairment sequences.
Finally, in Chapter 12, we targeted two important issues that are existing in evaluating the objective quality measure. The first issue is that how to select a set of hypothetical reference circuits (HRCs) that is representative for the large-scale database. The second issue is what shall we do if the PLCC and RMSE cannot report the goodness of a model, what other performance measures that we need to report this goodness? A set of new performance measures are proposed to target this issue. The HRC selection algorithms and the new performance measures are tested with other randomly selected HRC sets.

1.3.6 Part VI:Conclusion and future perspectives

Simply, this part, in Chapter 13, summarizes the main conclusions of the research efforts that are conducted in this dissertation. It highlights the usefulness of the PhD work. Then, the work perspectives are discussed to highlight the room of the improvements that can be conducted for the proposed models in this dissertation.

1.4 Publications

1.4.1 Journals

— **Ahmed Aldahdooh**, Enrico Masala, Glenn Van Wallendael, Marcus Barkowsky. "Reproducible research framework for objective video quality measures using a large-scale database approach", Elsevier Digital Signal processing, **submitted**.
— **Ahmed Aldahdooh**, Enrico Masala, Glenn Van Wallendael, Marcus Barkowsky. "Framework for reproducible objective video quality research with case study on PSNR implementations", Elsevier SoftwareX, **submitted**.
— **Ahmed Aldahdooh**, Marcus Barkowsky and Patrick Le Callet, "Proof-of-Concept: Role of Generic Content Characteristics in Optimizing Video Encoders," Springer Multimedia Tools And Applications, **submitted**.

1.4.2 Conferences

— **Ahmed Aldahdooh**, Marcus Barkowsky, Patrick Le Callet, and David Bull, "Inpainting-Based Error Concealment For Low-Delay Video Communication," The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, 2017.
— **Ahmed Aldahdooh**, Enrico Masala, Glenn Van Wallendael, Marcus Barkowsky. "Comparing temporal behaviour of fast objective video quality measures on a large-scale database" 32nd Picture Coding Symposium (PCS 2016)., Nuremberg, Germany, 2016.
— Yashas Rai, **Ahmed Aldahdooh**, Suiyi Ling, Marcus Barkowsky and Patrick Le Callet, "Effect of content features on short-term video quality in the visual periphery," 2016 IEEE Workshop on Multimedia Signal Processing (MMSP 2016), Montreal, Canada, 2016.
— **Ahmed Aldahdooh**, Marcus Barkowsky and Patrick Le Callet, "Content-aware adaptive multiple description coding scheme," 2016 IEEE International Conference on Multimedia & Expo Workshops (IC-MEW), Seattle, WA, 2016, pp. 1-6. doi: 10.1109/ICMEW.2016.7574726
— **Aldahdooh, Ahmed**, Marcus Barkowsky, and Patrick Le Callet. "Spatio-temporal Error Concealment Technique for High Order Multiple Description Coding Schemes Including Subjective As-sessment" Quality of Multimedia EXperiences (QoMEX) 2016, International Conference on. IEEE, 2016.
— **Ahmed Aldahdooh**, Enrico Masala, Olivier Janssens, Glenn Van Wallendael, Marcus Barkowsky. "Comparing Simple Video Quality Measures for Loss-Impaired Video Sequences on a Large-Scale Database" Quality of Multimedia EXperiences (QoMEX) 2016, International Conference on. IEEE, 2016.

— **Aldahdooh, Ahmed**, Marcus Barkowsky, and Patrick Le Callet. "The impact of complexity in the rate-distortion optimization: A visualization tool." Systems, Signals and Image Processing (IWSSIP) 2015, International Conference on. IEEE, 2015.

2

Related works

2.1 Introduction

In this Chapter, the state of the art of different parts of the dissertation will be reviewed. The video content characteristics will be reviewed in a separate chapter, Chapter 3. Since this PhD work targets different parts of the video delivery chain, each section of this Chapter will cover one target with a specific scope. The structure of the chapter is shown in Box 2.1.

Box 2.1 – Chapter structure

This chapter will be organized and structured as shown in Figure 2.1. In Section 2.2, works that are related to the encoder parameters prediction is illustrated in Section 2.2. In Section 2.3, a brief review of quality of service (QoS) and of quality of experience (QoE) will be presented. Section 2.4 reviews the source coding technique as a way to enhance the QoE. The second way to enhance the QoE is reviewed in Section 2.5. Finally, Section 2.6 shows the state of the art of content-aware video quality and no-reference quality assessment (QA).

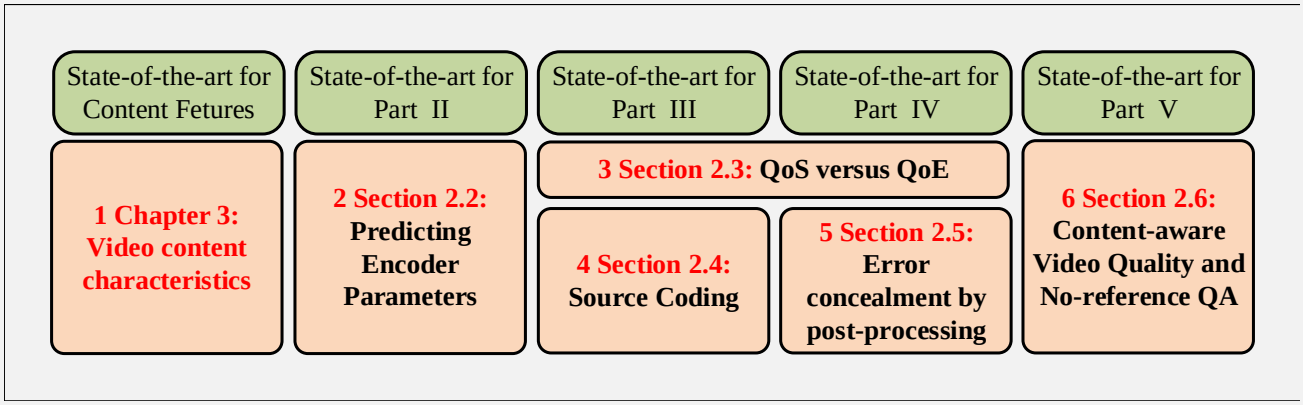


Figure 2.1 – Chapter 2 Structure

2.2 Predicting Encoder Parameters

The aim of any encoding system is to provide the best-effort video quality for the end users. Due to the limited bandwidth, coding systems since H.261 employ rate-distortion optimization (RDO) model aiming to achieve minimum degradation in video quality for a given bitrate. It is expressed mathematically as [35]:

$$min\{D\}, \quad \text{subject to} \quad R \leq R_c \tag{2.1}$$

where R_c is the given bandwidth. This is solved using Lagrangian optimization as expressed in [35]:

$$min\{J\}, \quad \text{where} \quad J = D + \lambda R, \tag{2.2}$$

where λ is a Lagrange multiplier.
When a new video coding standard is introduced, new or improved tools are introduced that increase the complexity dramatically since the RDO needs to test all possible combinations of the modes. Therefore, many research efforts have been conducted to reduce the encoder complexity [36–56]. The awareness of complexity in these algorithms came from the fact that not each coding tool needs to be utilized. In [37,38,40,41,43,49,51], the complexity control algorithms are implemented for inter frames based on fast mode decision, changing motion estimation search methods, or reducing the number of reference pictures. Going into more detail, in [37], three complexity controls are proposed; the first uses spatial and temporal blocks to determine the search range using proper threshold, the second uses the sum of absolute differences (SAD) cost with two thresholds to determine the prediction mode, finally, SAD, motion vectors and optimal reference frame are used to decide the number of reference frames. In [38], the authors proposed to use both the fractional motion estimation and fast integer motion estimation algorithms to reduce the complexity. Kim et al. [40] use the best mode information of a correlated macro-blocks (MB) in the time-successive frame to determine the search mode and use the adaptive rate distortion cost threshold for early termination process. Su et al. [41] manage the complexity by changing the motion estimation parameters and by adjusting mode decision processes using different complexity levels. In [43], the motion estimation tools are categorized into five states according to the complexity using SAD cost. Shen et al. [49] utilize inter-level correlation of quadtree structure and the spatiotemporal correlation to determine the inter mode. In [51], the rate distortion cost on reference frame is used to determine the CU splitting. Other algorithms are implemented for intra-frames [36,39,40,53,54]. In [36], partial computation of the cost function is used to determine the intra mode while in [39], the Discrete cosine transform (DCT) based dominant edge direction is used. In [40], the best inter mode is used to determine the proper intra mode. Chen et al. [53] map the edge direction to a proper prediction mode in HEVC while Zhao et al. [54] use SSIM structure similarity of neighbouring coding units to determine the intra-mode. Algorithms like [41] analyse the complexity of each coding tools and range them to provide coding levels of complexity. Some of the above-mentioned algorithms are implemented in H.264/AVC and they might be adapted in HEVC. The largest amount of complexity of HEVC is due to quadtree structure, therefore a lot of efforts have been introduced in this domain [44, 47, 48, 50, 52, 53, 55, 56]. In [44], the authors utilize the fact of correlation between consecutive frames to ignore the rarely used depth information at frame level and utilize the neighbouring and co-located blocks to determine the CU splitting while in [47], the authors extract features that are related to the content at the CU level and use them to build a prediction model to determine the CU splitting. Shen et al. [48] use Mean Absolute Deviation (MAD) to measure the texture homogeneity of the CU to early terminate the splitting. In [50], the CU depth decision is determined by utilizing the spatial correlations in the sequence frame while Nguyen et al. [52] determine the most probable CU depth ranges by utilizing temporal correlation of depth levels among CUs and the continuity of the motion vector field. Chen et al. [53] propose a bottom-up partition process by utilizing the gradient information of pixels. In [55], the back-propagation neural network (BPNN) was used to build a classifier to decide the splitting of the CU using the sum of absolute transform difference (SATD) and the coded block flag (CBF) as features while in [56] the authors use the decision trees to decide the CU splitting by utilizing the encoding information such as rate-distortion, skip merge flag, and merge flag. In [46], the authors use the max tree depth of unconstrained frame to encode a specific number of next consecutive frames while in [45], the complexity is controlled by weighting the basic operations in the reference encoder.
The above-mentioned techniques achieve significant reduction in complexity although the complexity awareness came from the fact that some of the introduced modes and tools of the encoder either are rarely used or unnecessary in some situations. Most of these algorithms employ content properties as demonstrated in the aforementioned algorithms. Properties like spatio-temporal correlation between blocks, SAD cost, MVs, RD cost, or flags are used in these algorithms. Moreover, these algorithms work on block or frame level which may yield block-to-block and frame-to-frame variations in quality. The aim of these algorithms is to reduce the complexity while the bitrate and quality are not balanced. A room of improvement can be accomplished by utilizing the underlying content features to predict the encoder's parameters at sequence level.

2.3 Quality-of-Service Versus Quality-of-Experience

The video delivery chain consists mainly of two services; compression and transmission which are the sources of distortions due to the use of quantization and best-effort networks respectively. The service quality can be judged using two factors, technical and human factors. The technical factor is named quality-of-service (QoS) while the human factor is named quality-of-experience (QoE). In the following subsections, a short review of each term will be introduced.

2.3.1 Quality of Service (QoS)

The application type has a vital role to determine a suitable error resilient tool to use to mitigate transmission loss. In [26], the author mentioned that end-to-end delay can be calculated using following factors and all factors except the transmission delay are relatively fixed since they are acceptable by the underlying application:

1. encoder processing delay (including acquiring the data and encoding),
2. encoder buffer delay (to smooth the rate variation in the compressed bit stream),
3. transmission delay (delay caused by the transmission itself, which is usually very small, and that due to queuing and perhaps retransmission in packet-based network),
4. decoder buffer (to smooth out transmission jitters), and
5. decoder processing delay (including both decoding and display buffer for constant frame play-out).

Hence, the QoS of the video transmission application is identified by the maximum transmission delay, latency, and the delay variation, jitter, allowed by the application and the probability loss rate of the underlying network.
For example, the video telephone application allows 150ms as a maximum delay and 40ms as a maximum delay of the decoder. Therefore, designing error concealment tools with low complexity is required for this type of application. A tighter delay constraint is required for ultra-low delay applications.
There are intensive efforts to enhance the QoS to stand solidly against the existing challenges like:

1. Real-time applications: this challenge still undermining the enhancement of QoE. It allows the implementing of simple error concealment mechanisms which leads to dissatisfaction and instability of the quality.
2. Thirst for bandwidth: although the network capacity is increased but, from another side, the need for high definition and even for Ultra-HD applications are increased. Relying on the compression efficiency is the promising solution.
3. Network stability: network conditions are time varying and these variations are not predictable especially the packet loss behaviour. Hence, the service is stand as best-effort service and relying on the error robustness tool is promising.

2.3.2 Quality of Experience (QoE)

In the previous Section, an overview of QoS is introduced. In this Section, the human opinion on the quality and the satisfaction is presented. This satisfaction has a term namely quality of experience (QoE). The popular metric to measure video quality is MSE/PSNR which does not necessarily express the human satisfaction especially when parts of the video are delayed or concealed. Researchers trust observers' judgments in video quality assessment despite the fact that building subjective experiments are very expensive and cannot be incorporated in real-time systems. Many efforts have been dedicated to implement objective measurements to automatically assess the image or video quality and give results very close to what human observers give.
Quality of Experience (QoE) is defined in [33] as "the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state". The reader is referred to [33] for the history of the definitions and be noticed that the above-mentioned definition is *"working definition"* since this definition might be changed in the future as QoE research efforts evolve.

2.3.2.1 Ways to Measure QoE

Subjective experiment is the accurate way to judge the quality of the video. Because it cannot be part of the video delivery and it is time and money consuming, efforts have been dedicated to implement objective measurements to automatically assess the image or video quality taking into account perceptual properties of the content and the human visual system properties.

Subjective Experiment

There were standardization efforts to formalize how to setup the experiment. ITU-T body in [57–59] standardizes subjective experiment procedures. Procedures include:
— Source signal: includes recording environment, recording systems, and scene characteristics.
— Test methods and experimental design.
— Evaluation conditions: include viewing conditions (room illumination, viewing distance, ... etc.), processing and playback systems, viewers, and instruction to them.
— Statistical analysis and results reporting.

Subjects evaluate the visual data in different ways depending on the test method used for that. For instance, absolute category rating (ACR) uses (5 down to 1) scale which reflects excellent, good, fair, poor, or bad rating respectively. Another example is Degradation category rating (DCR) test method in which the subjects report the overall quality using (5 down to 1) rating scale that reflects imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying rating respectively. For more information about test methods, readers are advised to read [58, 59].

Objective Measurements

Researchers classify video quality measurements to three categories; Full-Reference (FR), Reduced Reference (RR), and No reference (NR). The widely used technique is to evaluate frame by frame as still image independently and then one score is evaluated as global measure by using temporal pooling techniques. A review of full-reference objective measures is presented in [60]. The authors highlight that the visual characteristics and human visual system features have a high impact when designing the objective measures. Results of subjective experiments, mostly mean opinion score (MOS), are used as a ground truth dataset to evaluate the performance of the objective measure.

Traditional point-based metrics: Peak-Signal-to-Noise-Ratio (PSNR) metric

It is the most widely used metric due to it is simplicity and mathematically convenient for optimization but it is not good in the perceived visual quality as you can get two images with same PSNR but not in the same visual quality. It depends on Mean Square Error (MSE).

Natural Visual Characteristics: Natural Visual Statistics: SSIM

A new approach is introduced in [61, 62] that depends in the fact the Human Visual System is highly adapted to extract structural information from the scene. The authors extract information related to luminance, contrast, and structure measurements of the scene.

Natural Visual Characteristics: Natural Visual Features: Video Quality Model (VQM)

The VQM [63] is widely used recently as it is highly correlated to subjective experiments results. It starts with the calibration step to correct the spatial and temporal alignments. After that a set of features are extracted from original and processed video. These features characterize the perceptual changes in spatial and temporal properties. The VQM score is calculated by linearly combining the seven independent parameters calculated from extracted features.

- Four parameters are computed from Luminance (Y) component.
- Two parameters are computed from the two chrominance components (CB and CR)
- One parameter is computed from the contrast and the absolute temporal information extracted from Y components.

Perceptual (HVS): Frequency Domain: Digital Video Quality Metric (DVQ)

Watson et al. [64] use DCT to measure the human visual system aspects like light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation to evaluate the video quality.

Perceptual (HVS): Pixel Domain: Perceptual Video Quality Measure (PVQM)

The quality measure in [65] models three perceptual features, edginess, colour error, and the temporal decorrelation.

2.3.2.2 Ways to enhance the QoE

Video transmission over IP networks is challenging since it is a best effort environment. Video packets are vulnerable to loss and techniques for mitigating these errors are indeed required. Of course, enhancing the underlying network environments, i.e. enhancing QoS parameters, for video transmission is very important factor to increase the quality of experience (QoE) of the end-users. In this Section, we focus in tools that directly enhance the QoE. The well-known classification for these tools are mentioned in [26, 66, 67] and depicted in Figure 2.2. *Forward error concealment* tools refer to those tools in which the encoder plays the primary role while the decoder plays the primary role in *error concealment by post-processing or simply error concealment* tools. In *the interactive error-concealment* tools, both encoder and decoder play equally.

Channel coding is quite popular. Techniques like forward error correction (FEC) [68, 69], packets interleaving, and retransmission techniques are used in video transmission systems. In addition, joint channel-source coding might also be applied to protect some data more than others. Data Prioritization concept is introduced in [70]. Using this concept, for instance, packets that contain important data like headers or ROI data can be protected strongly to form unequal error protection.

To maximize the QoE of end users, considering the cooperation between encoder and decoder in designing robustness tools is highly recommended in types of applications. Based on the received information from the feedback channel, the encoder may change his parameters. For instance, if the packet loss probability decreased, the allocated bandwidth for FEC can also be decreased. Another way to utilize the feedback information as applied to H.263 [71] is the reference picture selection technique in which the encoder does not consider the damaged areas as references [72]. Another example is the error tracking technique [73] in which the current block will be intra coded if it affected by the lost blocks. Using feedback-based transmission system, encoders like H.264/AVC [74] and HEVC can be used. In case the

feedback is not applicable in some of the applications, other techniques like video redundancy coding can be used [75]. Several factors are considered to report the efficiency of error robustness tools: the final perceived quality, complexity, and amount of redundancy.

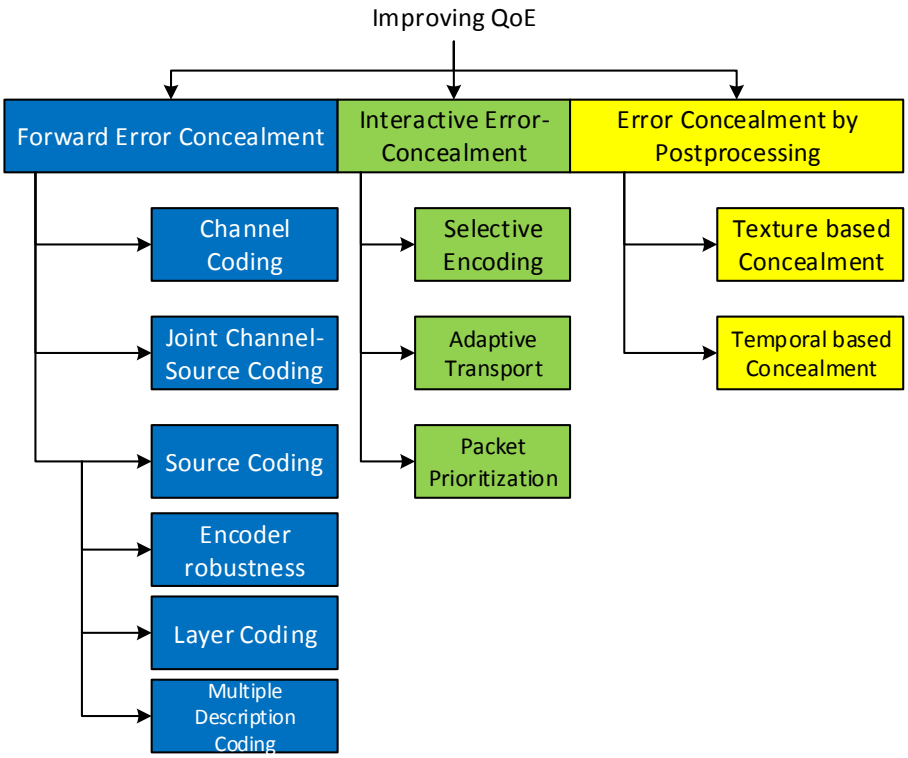


Figure 2.2 – Error robustness techniques classification

2.4 Source Coding

Despite of channel coding techniques are trying to detect or correct errors, bitstream may still contain corrupted data. Consequently, the need for other techniques that make the bitstream strong enough against channel errors. In this Section, the state-of-the-art of error robustness tools developed for the source coding will be introduced. Encoder robustness tools will firstly be visited and then the layer coding and multiple description coding.

2.4.1 Encoder Robustness

The encoding of video sequences passes through two processes, prediction and entropy coding processes. Synchronization codeword marker has to be implemented in the entropy coding to reset the entropy coder. Synchronization marker can be inserted in the block and/or frame levels. For instance, MPEG-4 [76,77] inserts the marker to separate motion and texture information. MPEG-4, in addition, uses reversible variable length coding (RVLC) to decode the bitstream backward. Later, the RVLC is not used in H.264/AVC and HEVC. Readers are advised to refer to [26,66,67] for more information about robust entropy coding and to [78,79] review entropy coding of HEVC and its comparison with respect to H.264/AVC standard.

2.4.1.1 Error robustness in prediction process

Robust bitstram contains redundancy bits to preserve video quality in presence of transmission errors and make it not optimal as standard one.

Data Partitioning/Isolation : It is one tool of the error robustness. The idea is to put coded data into different parts. For example, put header information in one part and the rest in another part. Codecs like H.263 [71,80,81], MPEG-4 [66,81], and H.264/AVC [82,83] use this technique. In H.264/AVC, the data is partitioned into three parts A, B, and C in which header information and motion vectors, intra coded blocks, and inter coded block information is stored in respectively. It is an added value to guide the error concealment in the decoder side. Wenger et al. [84]

proposed the recommended action in case one (or more) partition is lost. If, for instance, partition B and C are lost, then, MV in partition A can be used to conceal the block. Ideas like putting low frequency coefficients in partition A or putting a copy of intra block related header information to partition B is introduced in [85] and [86] respectively. In [27], data partitioning method is tested under 5%, 10%, and 20% error rates and it was shown objectively and subjectively that the data partitioning is superior for Foreman sequence and competitive for Paris sequence. Unfortunately, this technique is not supported in HEVC [28, 87].

Picture Segmentation: slice concept is introduced in different codecs H.263, MPEG-4, H.264/AVC, and H.265/HEVC. As mentioned before, the frame can be segmented to one or more slices and each slice contains blocks in sequential order. Using slice technique decrease the coding efficiency since the prediction outside the slice boundary is prohibited. In [27], the authors tested slice concept and yielded unsatisfied objective results but very satisfied subjective results. The performance increased if the slices are interleaved. Blocks can be grouped in more flexible way like chessboard fashion. Studies, like [88], showed that if the lost blocks arranged in chessboard like, the error concealment results will be enhanced. This technique is called Flexible Macroblock Ordering (FMO) and it is introduced in H.264/AVC but removed from HEVC due to the rare usage.

Redundant Slices: The idea is to have primary coded slice and one or more redundant slice representations with different qualities. In case of losing primary slice the decoder can reconstruct the lost slice with lower quality redundant slices.

Parameter Sets: This concept is introduced in H.264 and HEVC in which common parameters among video, sequence and picture are stored in sets. Parameter set [27] is not itself an error resilient tool but it can be used intelligently to maximize the error robustness by making sure that they are received reliably.

SP-/SI-Frame: The SP- and SI-Frames [89] are introduced in H.264/AVC and removed from HEVC due to rare usage. The main concept behind the SP-Frame is that it can be reconstructed even if it is predicted using two different reference frames. This concept makes it useful to be used for streams switching, splicing and random access, and error resiliency. Studies are conducted to see its benefits in the context of transmission errors [90–93]. For instance, in [91], each macroblock is additionally predicted with different reference frames and store them as SP-Frame to use them when the original block is lost and hence the error propagation is reduced, if not stopped. While in [92], Zhou et al. utilized feedback information from the decoder to re-encode the affected blocks by the errors by using concealed frame as predictor as SP-frame. I believe that removing such technique from recent video coding standard, HEVC, is not a wise decision.

Intra Refresh: The normal coding standards support enforcing the insertion of intra frame periodically to mitigate error propagation and it is usually set to one second. In inter coded frame, rate-distortion optimization in video coding [35, 94] makes its decision in the coding mode. Many enhancements are done to enforce intra coding of a block. [72, 95–103]. In [99], Haskell et al. proposed two strategies: periodic intra update and content-adaptive strategies. The periodic updates depends on the expected life of the errors while the content-adaptive leaky difference used to mitigate the error propagations and it requires to send side information. In [96, 104], a metric to determine the block sensitivity to the errors is used which depends on error probability and blocks that have high metric are encoded as intra blocks. If this idea is included in the rate-distortion model, it will give better results. In [103], Farber et al. utilized feedback information to track errors of the current block to decide if intra update or not. This system is not applicable of real-time applications. Hence, in [72], the model is adopted for low-complexity applications which makes it suitable for real-time applications. In [95], Willebeel et al. analysed the temporal dependency of the blocks in successive frames to determine the intra refresh. It is not suitable for real-time application due it is complexity. In [100, 101], the authors adopted rate distortion optimization to include the error probability and the distortion came from concealing error frames. It requires that the encoder knows the error concealment technique in the decoder side. In [98], the Recursive Optimal per-Pixel Estimation (ROPE) model is introduced to estimate the decoder distortion. It integrates with rate distortion optimization and it yields better results than [100] since the decoder estimation is measured more precisely. In [94], it changes the Lagrange multiplier and includes error-free, error-concealment, and error-propagation distortions. It was shown that this approach is promising as an effective error concealment tool, [74] since it trades-off between coding efficiency and error robustness and it robustness to mitigate the error propagation.

2.4.1.2 Layer Coding (LC) and Multiple Description Coding (MDC)

Layer Coding (LC): Layer coding, special case of scalable coding, basically encode the video sequence as a base layer with one or more enhancement layers. The temporal, spatial and quality (SNR) scaleability are implemented as an extension to different video coding standards such as [105, 106]. The idea is to allow the heterogeneous receivers to decode partial part of the compressed stream depending on their capabilities. The base layer delivers an acceptable video quality to the end-users and the highest quality achieved when all the enhancement layers are decoded. To allow the layer coding to serve as error resilience tool [26], many considerations have to be applied as:

- Unequal error protection: in which a base layer has to be protected more than the enhancement layer using FEC.
- Retransmission: in case that the base layer is lost or corrupted, retransmission is required for the base layer

since the enhancement layers are useless without the base layer.
- The inter prediction between layer has to be restricted to limit the error propagation. Hence, enhancement layers are directly predicted from the base layer so the coding efficiency is decreased.
- Common information like headers and coding information has to be copied in each layer.

Since the error-free base layer delivery is not guaranteed and the only solution for this problem is to retransmit the base layer, layer coding may not be convenient to real-time applications. Multiple description combats the drawback of the layer coding . The video sequence is encoded into two or more different bitstreams, called descriptions, and each description has acceptable quality if it is received alone without other descriptions. The highest quality will be achieved if all descriptions are received. A full review of multiple description coding can be found in [107–109]. It was shown that the multiple description coding is an effective and promising technique for error resilient tools for several reasons. First, it is suitable for real-time applications since the feedback is not required which simplifies the network design. Second, it performs better than other error resilience approaches in high error loss rates [110,111]. In the low error probability loss channels, it is advisable to use single layer coding or scalable coding with error resilience tools to save bitrate. As mentioned in [109], to judge a multiple description coding, four factors must be considered standard compatibility, redundancy tunability, complexity, and capability to increase the number of descriptions. Here, a brief classification for multiple description coding techniques will be reviewed.

Spatial domain MDC: The description is generated by downsampling the original video frame to two or more frames either by using polyphase or quincunx downsampling [112–114]. The main issue here, downsampling reduces the pixels correlation in one description. Different solutions are proposed to resolve this problem by zero padding [115], content adaptive zero padding [116,117], filtering [118,119], and duplicating data [117]. It is difficult to decide which is better since it depends on the application. For instance, zero padding and filtering is adding complexity while [114] not compatible with the standard although it achieves better objective quality in the central decoder.

Temporal domain: The descriptions are generated by downsampling the frame rate [120]. The problem is when increasing the number of description the temporal correlation between frames will be decreased especially for scenes with fast motion. Solutions is also introduced by duplicating motion vectors [121], and duplicating or dropping frames [122]. In addition to the fact that this approach is not compatible with the standard if one of the previous solution is used, this approach will increase a mount of redundant data if the number of descriptions is increased.

Frequency domain MDC: This approach can be further categorized as scalar quantization, coefficient partitioning, and transform coding multiple description coding.

Scalar quantization MDC: The descriptions are generated by using different quantization methods [123]. Methods like quantization interval-shifting [124] and tables indexing [125] are proposed. The quantizations levels for the descriptions are complementary that makes it not suitable for high number of descriptions and for high error loss rate.

Coefficient partitioning: Simply, the DCT coefficients are distributed to descriptions. How to distribute the coefficients is the problem. In [126], a threshold is used, while in [127] the high frequency components are fixed and the low frequency components are distributed among the descriptions. In [128], the distribution is block-wise. The main problem of this approach is that the side quality is (very) poor. The coefficient splitting using a threshold is not practical since it is content dependent.

Transform Coding MDC: The descriptions, in case of two descriptions, are generated by applying pairwise correlating transform to a pair of DCT coefficient to introduce correlation and put each in a description [129]. It gives better results than coefficient partitioning approach since the side decoder has more information about the coefficients and missing ones are estimated but, of course, it comes at price of complexity and redundancy.

Hybrid domain MDC: The main objective of this approach is to generate more descriptions. For example, the video can be first spatially downsampled and then temporally downsampled or first spatially downsampled and then coefficient partitioning [130]. Combining the multiple description coding with other techniques like redundant representation or scalable video coding make the error resiliency tool more robust against transmission errors. In [131], Ivana et al. use multiple description coding with redundant slices tool of H.264/AVC and propose a rate-distortion model to control the redundancy, while in [132,133] use multiple description coding with scalable video coding.

2.4.1.3 Error Resiliency Tools in HEVC

Most of the tools introduced in H.264/AVC are removed from HEVC because of the rare usage in real life. Here, I will summarize the error resiliency tools that HEVC supports [28]:

— Intra frame refresh as discussed above.
— Slices and tiles as discussed above.
— Enhanced error detection mechanism using reference picture selection (RPS) by which HEVC is able to detect losses.
— Temporal scaleability is used to limit error propagation and to guide error concealment techniques.
— Decoded picture hash SEI message that can be used to detect errors since the hash code is derived from the decoded samples.
— Structure of pictures (SOP) SEI message which describes the temporal and inter prediction structures of the stream. It can be useful to use it in the multimedia-aware networks to detect errors in the intermediate node

not in the receiver. Therefore, feedback information to retransmit the lost data can be sent to the decoder without compromising the overall transmission delay which makes it useful tool for real-time applications.
— Limit the usage of AMVP. It was shown in [24] that the error robustness in HEVC is decreased compared to H.264/AVC due to the increasing in temporal dependency. When a frame that uses AMVP is lost, the entropy decoding failure will happen with a serious quality degradation due to the error propagation. Disabling AMVP in some frames are proposed in [134–136]. In [137], a CU-level proposal has been introduced to limit the error propagation.

Errors might happen during the transmission process over lossy channels. Theses error can be bit errors or packet loss. Controlling the errors is challenging for several reasons [26, 67].

Error propagation Video coding standards use predictive and entropy coding, therefore the loss of data in one frame affecting subsequent frames that use the defected frame as reference.

Optimal method Video content is varying through time and also the network conditions, so finding the optimal solution is very difficult, if not impossible.

Complexity Implementing efficient error resilient tool in the encoder requires extra complexity. In addition, extra complexity in decoder side is also required for efficient error concealment. These extra complexity not applicable for specific types of applications.

Coding efficiency Most error resilient tools add extra redundancy the leads to bitrate increasing, so there is a trade-off between error robustness and compression efficiency.

The challenge is increased in HEVC since the prediction loop is improved. Oztas et al. [24] raised this challenge and showed that the robustness of HEVC is decreased compared with the H.264/AVC.

2.4.2 Temporal-based Multiple Description Coding

The temporal MDC schemes with their error concealment techniques are categorized into three classes; the first class is referring to the schemes that do not have any side information, the second class is referring to the schemes that introduce some additional data for each frame, while the third class is referring to the schemes that include a redundant frame for each primary frame. Table 5.1 shows the list of some MDC schemes and the corresponding hypothetical reference circuits (HRCs) as used later in this work, in Chapter 5. Apostolopoulos in [121] reviewed the first class of the schemes. All schemes in this class share the same encoding and decoding processes and differ in error concealment strategy. Suppose that an even frame is lost. Copying the previous even frame from the distorted description to replace the lost even frame in the buffer (HRC00, HRC01, HRC09), copying the previous odd frame from the undistorted description (HRC02, HRC10), averaging the previous and the next odd frames from the undistorted description (HRC03, HRC11), scaling the MVs of the next odd frame from the odd description by $\frac{1}{2}$ and use them to do the motion compensation process using the previous odd frame of the undistorted description, namely inplaceMC (HRC04, HRC12), and generating the MVs using the available previous and next odd frames, namely MCinterp (HRC05, HRC13), are the error concealment strategies that are reviewed in [121]. In the second class of schemes, a side information is introduced. This side information can be a duplicate of MVs of each frame in the description or a duplicate of I-frames (HRC06, HRC14). In [138], a different scheme is proposed in which each description contains alternatively even/odd frames and odd frames in even description are containing the motion information only predicted from the previous even frame (HRC08, HRC16). While in the third class of the temporal MDC schemes, a complete frame is used as side information. Radulovic et al. [131], suggested that each description alternatively contains a fine quantization frame (even) followed by coarse quantization frame (odd) (HRC07, HRC15).

2.5 Error Concealment by Post-Processing

All error-resilient source coding methods do not guarantee the arrival of all information to the decoder side which indeed the need for robust error concealment technique is quite important. Unfortunately, due the diversity of video content types, finding a robust error concealment tool is difficult. The decoder may need to estimate texture information, motion information, and coding mode of the missing blocks. The error concealment methods depend on the fact that the adjacent pixels, blocks, and frames are correlated and smoothly changes except in the area with edges and scene cuts. The first important step before starting the error concealment is to detect the error. In the following subsections, firstly, the error detection techniques will be reviewed and then an error concealment overview will be reviewed.

2.5.0.1 Error detection

The decoder can detect errors if abnormal syntax is detected like [139, 140].
— illegal codeword.
— Out of Range Codeword.

— Contextual Error
— unexpected value of syntax element like for instance wrong mode number.
— number of decoded DCT coefficient is more than expected.
— illegal synchronization header.
— incorrect number of stuffing bits are found

In HEVC, besides its ability to detected frame loss reliably using reference picture selection tool, using hash code SEI message is also a useful tool to detect errors in the reconstructed samples.

2.5.0.2 Error concealment categories

Video coding standards don't normalize the error concealment methods as part of the standard. The reference model JM of H.264/AVC, implements two error concealment methods, spatial and temporal [141]. HEVC reference model HM does not implement any error concealment methods. Here, spatial and temporal error concealment techniques will be reviewed. For more details, refer to [66, 142].

Spatial Error Concealment

One of the first error concealment methods is proposed by Wang et al. [143]. They utilized the smoothness property of the natural images by minimizing the variations between the damaged pixels, therefore they used boundary samples to do that. In my opinion, this technique has two main drawbacks. First, it works in smooth areas that do not contain edges. Second, it works for small regions and blurriness will occur in large missing blocks. In [144–146], the smoothness is recovered by estimating the DCT coefficients. In [147], the damaged samples are recovered using two or four neighbouring blocks. It suffers from the same drawbacks of [143]. In [88, 148],a weighted average of the neighbouring pixels are used for interpolation and this is implemented in H.264 due to its simplicity but the drawback is that it does not preserve the structure, i.e image edges. A refined proposal can be found in [149]. In [150–152], Meisinger et al. use frequency selective signal extrapolation technique to conceal missing data. The advantage of this is; it recovers the block structure since it uses FFT and not only the boundary samples of the damaged data can be used but also n-neighbouring boundary pixels can be used. The disadvantages of this technique are 1) the error threshold or number of iterations needs to be specified manually and 2) the number of neighbouring boundary pixels to be involved need to be specified which make it impractical for real-world applications.
The work in [153] classifies the edges of the neighbouring blocks using Soble operator and the dominant direction is chosen for interpolation. Since the dominant direction is only used, this method is not suitable for regions with multi-edges. In [154], a refined version is proposed. This work is further enhanced in [155] by considering only the pixels whose direction cross the missing block. Projections onto Convex Sets concept is used in [156] to conceal the damaged block. It recursively projects the damaged block with surrounding blocks to frequency domain and then applies an adaptive filter guided by block classifier as in [153] and finally, it is projected to the spatial domain. It preserves the structures since it uses the DFT. The main disadvantage of this method is its complexity since it switches between frequency domain and spatial domain many times. The results reported in [157] are interesting. Li et al. recover the missing block pixel by pixel in sequential order for eight directions and the weighted average is applied. This make it superior, but unfortunately its complexity is comparatively high. An edge-oriented spatial interpolation for consecutive block error concealment method is proposed in [158]. The edge direction is calculated using 1-D matching algorithm (MAD function) from the top and bottom boundary blocks. Its performance decreases in highly texture areas. The work in [159], does the edge-orientation based spatial pixel average. The candidate samples are extended to include n-neighbouring boundary samples to do the interpolation by which block structure is preserved.

Temporal Error Concealment

The main idea is to estimate the damaged motion vectors using available motion vectors of neighbouring or collocated blocks in current or reference frames respectively. The trivial solution is to use the zero motion vector or use the average or the median or the weighted average [142, 160, 161]. A motion vector of one of the neighbouring, which minimize the error in the block boundary, is used in the software [141]. For the whole frame loss, either freezing the previous correctly receiving frame or simple copy its motion vector and conceal it. Simple approach can be applied due to its simplicity but unfortunately, its performance is decreasing with the scenes of fast motion. In [10], the candidate list is extended to include internal and external candidates. The internal includes zero MV, the 4-neighbours, the 8-neighbours of the missing, and the average of all neighbours. While, the external includes the MV of the collocated block in the previous frame and the 8-neighbours of this collocated block. An extension to [150] temporal dimension is introduced and again this technique is not optimized and it is content dependent.

Switching Algorithms

It is highly considered to account switching algorithms in error concealment. For instance, in H.264/AVC, a switching algorithm is considered depending on the frame type. Besides, a switching between two temporal methods is also proposed depending on the motion activity. In [11], a decision tree, to be sent to the decoder, is implemented to adapt

the video content with its suitable error concealment method. Another content-adaptive spatial error concealment strategy is proposed in [12] to choose the best method depending on the analysing neighbouring blocks to smooth, edge-oriented, or texture. In [162], the authors use the proposed directional entropy to choose between their spatial error concealment implemented in [155] and the bilinear interpolation. In [10], the authors use spatial and temporal activity measures to switch between temporal and spatial error concealment methods. An enhancement to this work is introduced [13] in which the directional entropy is normalized and the threshold to switch between spatial and temporal criteria. The disadvantage of these models is that it is practical in the real world.

2.5.1 Inpainting-based Error Concealment

In post-processing techniques, the decoder utilizes the spatial and/or temporal redundancies to reconstruct the damaged/lost area in a video frame. Spatial techniques [29, 30]utilize available surrounding pixels to reconstruct the missing pixels. They are not efficient for large areas, non-constant areas, and in terms of complexity. They usually reconstruct the texture but not the structure. The work in [163] is an extension of [30] in which a spatio-temporal selective extrapolation strategy is used to reconstruct the missing area. Temporal techniques use available motion information to predict the missing motion vectors (MVs), for instance, by interpolating [31] or by selecting the MV that minimizes the side match distortion [32]. Despite providing information about whether the current area is moving or not, this technique is efficient only for low-motion and smooth sequences and for small areas since the precision of predicted MVs is not guaranteed. Thus, the structure (of copied data) is reconstructed but not the texture.
The target of any error concealment algorithms is twofold: reconstructing a satisfying reconstruction of a lost area and reducing the miss match between the encoded and the reconstructed blocks which yields reducing the error propagation effect. To achieve that we need to reconstruct the texture and the structure of a missing area and that can be done using inpainting techniques. A review of inpainting techniques can be found in [164]. In this work we focus on exemplar-based inpainting in which each lost patch is reconstructed by copying the best match from the known area. Specifically, inpainting algorithms have many target applications and in this work, we are interested in error concealment as a target application.
Inpainting-based error concealment algorithms are introduced in [34, 165]. The algorithms have three main steps: inpainting the moving foreground object, inpainting the stationary background temporally and spatially as in [166]. In the first step, inpainting the moving foreground object, the moving pixels are identified as in [34] using Bilinear Motion Field Interpolation (BMFI) [31]. Then, the best match of a moving patch is reconstructed from the neighbouring frames. In the second step, inpainting the stationary background temporally, the best match of a moving patch is reconstructed from the co-allocated patches of the neighbouring frames. The remaining pixels are reconstructed in the third step, inpainting the stationary background spatially.

2.5.2 Perceptual Effects of Packet Loss

A lot of studies are conducted to study and model the visibility of perceptual artifacts in video quality in the context of packet loss or dropping. During video transmission slice, continuous slice or the whole frame might be lost. The visibility of a loss depends on its location, the video encoding parameters, the underlying characteristics of the video signal itself, and the error concealment strategy used to recover the damaged area [167]. In [168], the individual packet loss is almost invisible for non-reference frames while losses in reference frames may last until the next synchronization signal, i.e I-frame. In [167], they model the error visibility with factors that likely affect the perceptual effects and they categorize them to content-independent and content-specific factors. Then, a subjective experiment was run to classify the perceptual effects to visible or invisible. Reibman et al. [169] used scene characteristics, scene cuts, to predict the packet-loss visibility. The study includes that study of camera motion and the effect of loss at scene cuts and before/after scene cuts. It was shown that the camera motion is a considerable factor to detect the packet loss visibility and the error at scene cut increases the error visibility more than errors before/after scene cuts. A study for burst length was done in [170], it was shown that a burst loss generally produces a larger distortion than an equal number of isolated losses. A contradicting results was shown in [171, 172]. In [172], Boulos et al. also studied the loss distribution and the percentage of error loss in the picture. They showed that “for the same loss percentage, multiple burst losses are more damaging than a single contiguous long loss”. Frame dropping technique is useful for bitrate adaptation and error concealment strategies [173]. Their perceptual effects are studied in [174–176]. In [175], they run subjective experiments to detect the threshold of temporal discontinuities that result from frame dropping. They have found that the discontinuity caused by 200ms is always detected. They also found that, in [174], regular frame dropping is less annoying than irregular frame dropping. In [177], Dai et al. studied the impact of single packet loss with different frequencies. They found that the video quality will be unaccepted if more than two times single-losses in a short period.

2.6 Content-Aware Video Quality and No-Reference QA

A lot of objective quality measurement techniques use simple spatio-temporal statistics or psychovisual-based complex experimental results. An overview of recent development in visual quality assessment can be found in [178]. It is stated that, in [179], as long as the video content and the codec type are not changed, PSNR is a valid quality measure and when the content is changed, correlation between subjective quality and PSNR is highly reduced. Le Callet et al. [180] pointed to the importance of display, resolution, content and visual attention factors in subjective quality assessment. The same conclusion is drawn in [9]. In [181] a quality metric is proposed that takes into account quantization errors, frame rate, and motion speed. In [182], the prediction model is fed with the content type, packet error rate, frame rate, and bitrate. The content type is identified by performing content classification using spatial and temporal features. A strategy to combine spatial and temporal information activity levels of the analysed video sequences with peak signal-to-noise ratio (PSNR) in order to produce more reliable estimate of the perceived subjective quality in terms of mean opinion score (MOS) in introduced in [6]. In [183], a set of spatio-temporal features derived from the encoded video, such as the AC transform coefficients, the quantization parameters and the motion vectors. These features together with the bitrate are used to build content-based video quality metric model. In [184], the author estimates the quality by using frame difference, contrast, motion vector magnitude, motion intensity, motion direction, quantization step, Gabor-based features, and spatial and temporal resolution. Recently, [7, 185], several content features are extracted and trained using machine learning approach to build a prediction model. As noticed, there are a few works that considers the content features for estimating video quality especially for loss-impaired videos.

New image or video coding standards introduce new or improved coding tools in order to improve the rate-distortion performance. Each standard may be characterized by the type and amount of degradation that is added to the encoded image or sequence [186]. A lot of efforts have been done in identifying these coding degradations for different standards [187], notably for H.264/AVC. In addition to their importance in guiding the improvements in coding standard, understanding such degradations is also important for objective quality measures especially when there is no information about the original source. This type of quality assessment is called No-Reference (NR) measurement. A classification of no-reference quality estimation models has been proposed in [188] and a variety of algorithms has been reviewed. Although H.264/AVC NR measures can be adapted to the High Efficiency Video Coding (HEVC) use case, some publications have specifically addressed HEVC. In [189], a no-Reference Pixel (NR-P) based method is proposed in which the quality estimation for loss-impaired sequences is measured by calculating the temporal variations of the power spectrum across the decoded frames. As stated in [189], the model has correlation scores between 0.7 and 0.8 and works well for low-to-medium temporal activity sequences. This calls for integrating further content characteristics, either pixel-based or bitstream features, in objective video quality measurement models. In [167, 190–192], No-Reference Bitstream (NR-B) based models are introduced. In [167], Kanumuri et al, modeled the visibility of packet-loss in MPEG-2 video using pixel and bitstream based features. In [190], the authors train a neural network using subjective scores as well as packet loss rate, frame type, GOP structure, Intraperiod, percentage of damaged frames, and percentage of frames at different temporal levels. In [192], the authors use the QP and the spatial information (SI) to introduce a two-parameter NR-B method to estimate the perceptual quality (DMOS) of encoded HEVC sequences. The SI, as in [192], is calculated as the weighted sum of the DC difference values of inconsistent transform units (TU) and their respective neighbouring TUs based on the ratio of the TU edge length. In [191], the authors rely on the bitstream features to predict the perceptual video quality.

3

Video content characteristics

3.1 Introduction

In the real world, different video classes like, natural scenes, cartoons, sports, news broadcasting and computer-generated videos exist and each class may be categorized into subclasses. The wide variety of video content causes a challenge for content-based research since, of course, natural scenes differ from sport scenes and sport scenes differ from cartoons, etc. Hence, considering content types and their corresponding features is very important in different aspects. First, in setting up subjective experiment; in [3–5], Video Quality Experts Group (VQEG) and Pinson et al. mention some limitations to video source selection to conduct researches. Second, in improving objective measures of video quality; in [6–8], the video content features are analysed to improve the objective video quality measure. Third, in improving video coding efficiency; it can be concluded from the subjective experiment of Pitrey et al. [9] that the video content influences the video encoding. In more details, one may encode a given video with several configurations and may get a similar Mean Opinion Score (MOS) for the output videos, but practically one of them spent minimal computational power. This minimum may change from one content to another depending on video content characteristics. Fourth, in designing error resilience tools: in [10–13], switch algorithms that utilize the content features are used to decide which error concealment technique should be applied. While in [193], a content-aware adaptive multiple description coding scheme is proposed.
In this chapter, different content characteristics will be listed. They cover a wide range of spatial, temporal, and spatio-temporal features. These features will be used in this dissertation in different aspects. It will be used in building the prediction models, in selecting a subset of contents to run the experiments, and in analysing results of subjective experiments. The goals of this Chapter are listed in Box 3.1 and the Chapter structure is illustrated in Box 3.2.

Box 3.1 – Goals

This chapter aims to answer the following research questions:
- Identify the global/generic content indicators/features that are going to be used in this PhD.
- How to select a set of sequences from the available content sequences?
- Identify the local content features that are going to be used in analysing the subjective experiments of the loss-impairment sequences.

3.2 Global/generic content features

3.2.1 Extracted Features

Table 3.1 shows 209 content features, listed in [8], that have been extracted from the ten original/encoded video sequences. The features cover spatial and temporal characteristics that are extracted from the luminance frame (Y), and the chrominance frames (Cb and Cr), in the spatial domain or in the frequency domain. The features are

Box 3.2 – Chapter structure

This chapter will be organized and structured as shown in Figure 3.1. Different content features that are extracted from original/distorted sequences will be listed and studied in Sections 3.2.1 and 3.3. The content selection will be demonstrated in Section 3.2.2 and Section 3.4.

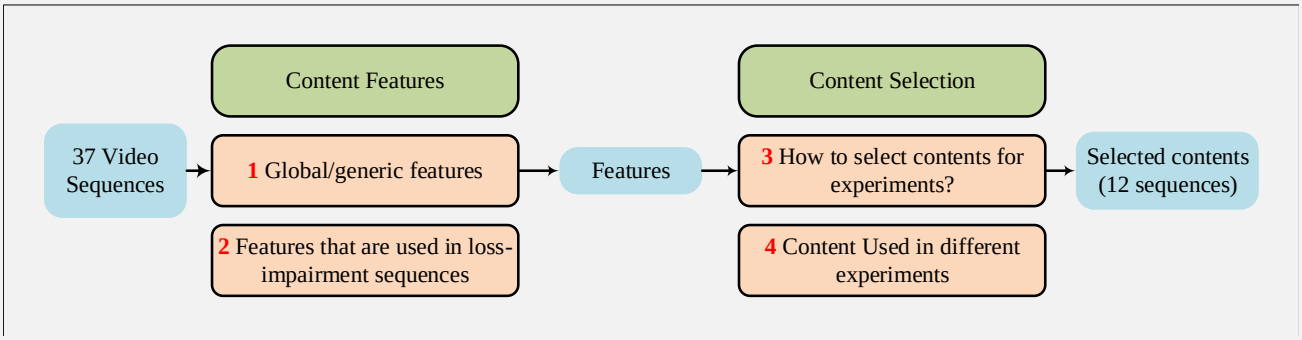


Figure 3.1 – Chapter 3 Structure

extracted on both block or frame levels. For the features that are extracted at the block level, the Minkowski sum with different power is applied to obtain a scalar value of each frame, then several statistical measures (e.g., mean, maximum, standard deviation, etc.) are applied to get a scalar value that represents the video sequence. Three spatial information features [194] are employed to measure the edge information. Twelve chrominance information features [195] are used to measure the colour information. Five contrast information features [195] are considered to measure the distribution of contrast in the frame. Twelve features that belong to spatial perceptual features [195] measure the perceptual spatial information and extract the changes in the orientation of the spatial activity. One feature measures the colourfulness of the video sequences [196]. Contrast, energy, correlation, homogeneity, and entropy of the joint probability distributions of pairs of pixels, namely Gray-Level Co-occurrence Matrix (GLCM), as in [197] and [198] on the whole frame and on 64x64 blocks are measured using four neighbouring directions (0, 45, 90, and 135 degrees). In total, 90 features have been extracted from every video sequence using GLCM. Normalized cross correlation features [199] are represented by 25 values that indicate how much the top-left 16x16 sub-block is correlated in its 64x64 block. Other 8 features are extracted from the 4x4-DCT decomposition of the luminance frame: these are kurtosis, smoothness, sharpness, similarity between different frequencies (3 features), and vertical and horizontal blockiness [200]. Features from Laplacian pyramid subband are also extracted [201]. Energy, entropy, and kurtosis are extracted from each intra-subband and the ratio between different subbands is considered as features, yielding 33 values. Other 13 features represent the inter subbands smoothness, subbands similarity and SSIM similarity. Regarding the temporal domain, 7 features are computed. Two of them directly represent the temporal information [194]. Others are computed according to the definitions the MPEG-7 motion activity descriptor [202]: they represent motion intensity, motion direction, and spatial distribution of objects. Other features are also extracted to help select the sequences for the experiment; Motion intensity maps [203], encoding bitrate, and camera motion descriptors [204].

3.2.2 Content Selection

37 UHD source videos are selected from different content providers: Shanghai Jiao Tong University (SJTU) [205], Ultra Video Group [206], Sveriges Television AB (SVT) [207], Blender Foundation [208], and MediAVentures [209]. Content features, described in Section 3.2.1, are extracted from these video sequences, Figure 3.2. Feature values are normalized linearly between [0,1]. Then, each feature is categorized into 3 or 4 classes according to their normalized values. For instance, labels of 1, 2, and 3 will be assigned to the feature's value that lies in the range [0,0.33], (0.33,0.67], and (0.67.1] respectively and labels of 1, 2, 3, and 4 will be assigned to the feature's value that lies in the range [0,0.25], (0.25,0.50], (0.50.75] and (0.75.1] respectively. Features that are represented with histograms, tree classification using Jensen-Shannon divergence as distance metric [210] is used to cluster the video samples to 3 classes. After the classification process, contents that covers different class levels for different features are selected. As a result, 12 video sequences are selected, Figure 3.24. In the following subsections, some classifications that are related to motion or spatial features are illustrated.

Table 3.1 – List of Extracted Features

Features	Formula	Count
Spatial information [194]	$SI = F_1\{F_2[Sobel(Y)]\}$ — $\{F_1, F_2\} = \{max, std\}, \{max, mean\}, \{std, mean\}$	3
Chrominance Information [195]	$C_U = F_1\{F_2\{U\}\}$ $C_V = F_1\{W_R * F_2\{V\}\}, W_R = 1.5$ — $\{F_1, F_2\} = \{mean, mean\}, \{std, mean\}, \{mean, kurt\}, \{std, kurt\}, \{max, kurt\}, \{max, max\}$	12
Contrast information [195]	$CI = F_1\{F_2[Y]\}$ — $\{F_1, F_2\} = \{mean, mean\}, \{mean, std\}, \{max, std\}, \{mean, skew\}, \{mean, kurt\}$	5
Spatial perceptual information [195]	$F_{SI13} = F_1\{F_2[SI13(Y)]\}$ $SI_{HV} = F_3\{\frac{(mean(HV[Y])\|_p)}{(mean(\overline{HV}[Y])\|_p)}\}, p(threshold) = 3$ — $\{F_1, F_2\} = \{mean, mean\}, \{mean, std\}, \{mean, skew\}, \{mean, kurt\}, \{mean, max\}, \{std, mean\}, \{max, max\}$ — F_3= $mean$, std, $skew$, $kurt$, max	12
Colourfulness [196]	$CF = mean\{CF\{YUV\}\}$	1
Gray-Level Co-occurrence Matrix (GLCM) [198]	$Contrast = F_1\{cont(GLCM)\}$, $Correlation = F_1\{corr(GLCM)\}$, $Energy = F_1\{enrg(GLCM)\}$, $Homogeneity = F_1\{homo(GLCM)\}$, $Entropy = F_1\{entropy(GLCM)\}$ — F_1 = $mean$, std, max — It is calculated per frame (5x3= 15) and per block with different Minkowski power p=(1,2,4,10, 0.1) = (5x3x5 = 75)	90
Normalized cross correlation [199]	$F_{NCC} = F_1\{NCC(block64x64)\}$ — F_1 = $mean$, max, std, $skew$, $kurt$ — It is calculated per block with different Minkowski power p=(1,2,4,10, 0.1)(5x5=25)	25
DCT based features [200]	See the reference for more details	8
Laplacian based features [201]	See the reference for more details	46
Temporal information [194]	$TI = F_1\{F_2[Y_2 - Y_1]\}$ — $\{F_1, F_2\} = \{max, std\}, \{std, max\}$	2
MPEG-7 Motion Activity [202]	See the reference for more details	5
How to read the formula: — For instance, to read the formula $SI = F_1\{F_2[Sobel(Y)]\}$ with $\{F_1, F_2\} = \{max, std\}$: - Apply the Sobel filter to each Y frame and keep the maximum value. - Calculate the standard deviation of maximum values. — Key to read abbreviations: standard deviation (std), maximum (max), skewness (skew), and kurtosis(kurt)		

3.2.2.1 Bitrate clustering

The video sequences are encoded using single description coding (SDC), and multiple description coding (MDC) (2-MDC and 4-MDC). The bitrate classification is based on the bitrate increase factor of 4-MDC with relative to SDC. The bitrate increase values are normalized from 0 to 1. Then, the value the lies in the range [0,0.25], (0.25, 0.5], (0.5, 0.75] or (0.75.1] is labelled to low, low-mid, mid-high, and high respectively. Figure 3.3 shows the classification result example.

3.2.2.2 Motion intensity clustering

The descriptor that is described in [203] represents the motion content of the video at pixel level as a Pixel Change Ratio Map (PCRM). It is claimed in [203] that the PCRM enables us to capture the intensity of motion in a video sequence and indicates the spatial location and size of the moving object. The PCRM is represented using 8/16/32-bin histogram. The PCRM is pixel-based technique in which the changes in pixel intensity over all the frames in a video segment is accumulated to generate the PCRM. Figure 3.4 shows the construction field video sequence with its corresponding PCRM. Figure 3.5 shows the 32-bin histogram of the PCRM and it is clear that the sequence contains

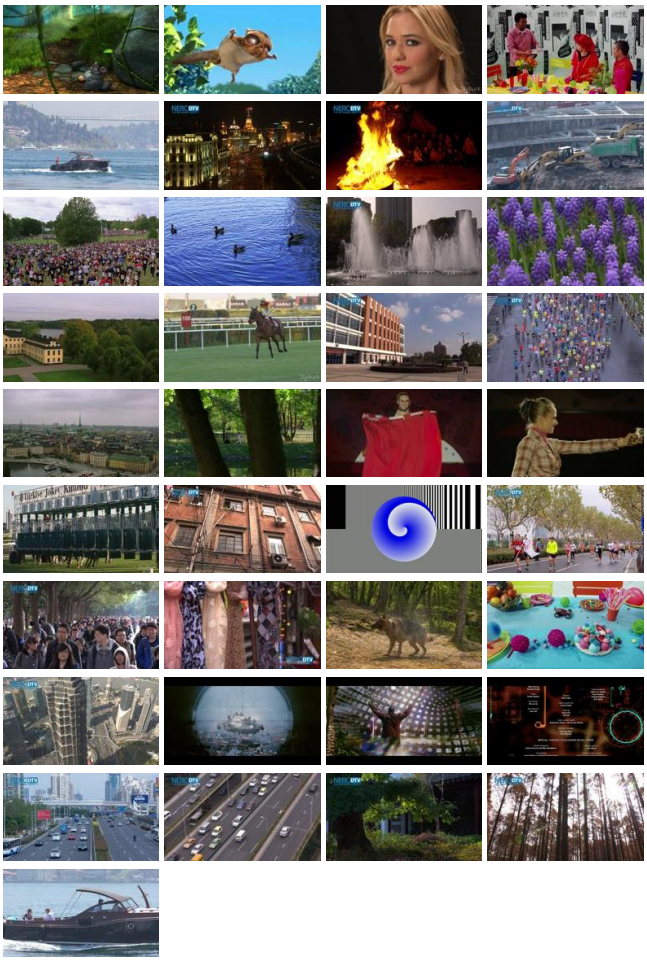


Figure 3.2 – 37 UHD video sequences from different resources.

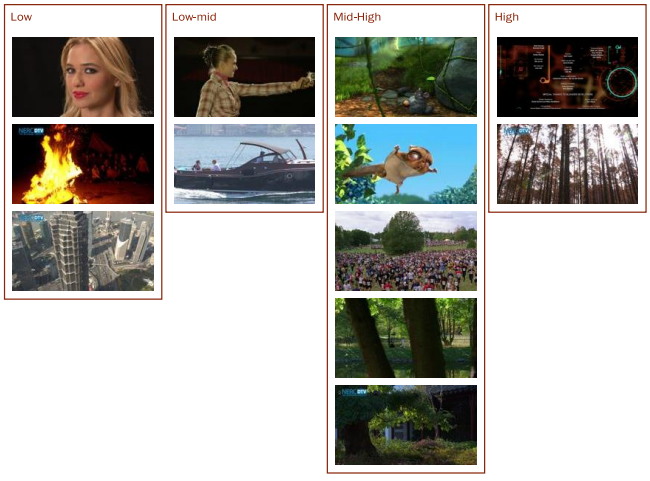


Figure 3.3 – 4-MDC Bitrate increase factor classification

high amount of low-intensity motions. Capturing the level of action and motion intensity of a video sequence are important in identifying different amount of motions in video sequences. MPEG-7 visual motion descriptors [202] also provide 5 types of motion descriptors; camera motion, motion activity, warping parameters, trajectories, and parametric motion. Motion activity descriptor captures the intensity of action; slow sequence, fast-paced sequence, and action sequence. Motion activity descriptor presents 4 attributes of video sequence. All of them depends on analysing motion vectors. First, Intensity of activity: it expresses the amount of activity from low to high using integer values (1-5). Second, direction of activity: it identifies the dominant direction of several motion objects and

it expressed with 8 equally spaced directions. Third, spatial distribution of activity: the frames are divided into regions (9 regions in this work) and then the motion intensities are computed (7 levels in this work) in each region to indicate which region is active and which region is not. Finally, temporal distribution of activity: is to express the variations in motion activity temporally . It is expressed as 5-bin histogram in which each bin represents a level of intensity. Figure 3.6 shows the BBB sequence with its corresponding temporal distribution of activity and the spatial distribution of activity respectively. The motion intensity classification is done using tree classification using Jensen

(a) Construction Field Sequence

(b) PCRM of construction Field sequence.

Figure 3.4 – The construction field video sequence with its corresponding PCRM

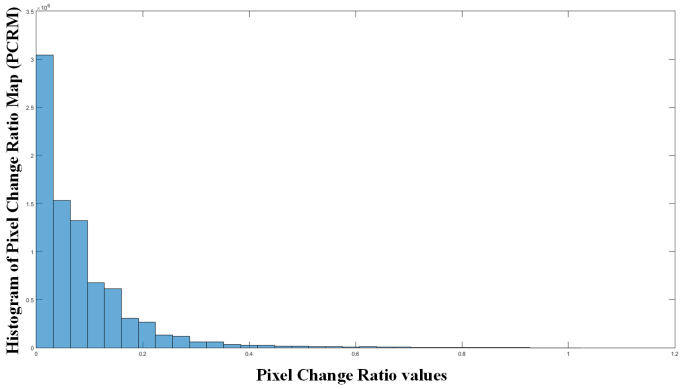
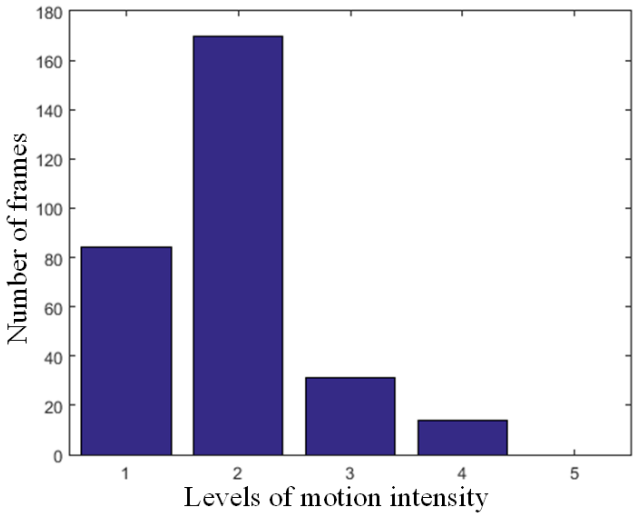


Figure 3.5 – Construction Field Sequence PCRM 32-bin Histogram. X-axis represents the pixel change ratio values. Y-axis is the count of each probability range

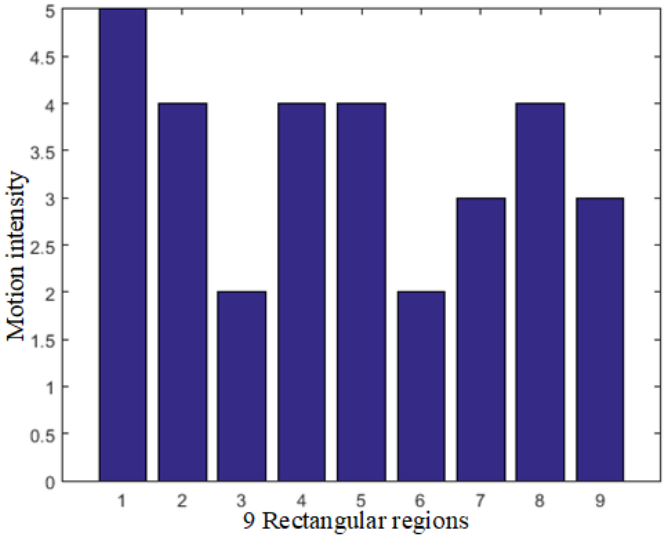
.

Shannon divergence (JSD) as a distance metric. Firstly, the 32-motion histogram is generated for each sequence [203]. Second, the distance between sequences is calculated using JSD, Figure 3.7. Third, the tree classification is used and it is adjusted for 3 classes, Figure 3.8. Figure 3.9 shows the classification result example.

(a) BBB Sequence



(b) Temporal distribution of activity of BBB Sequence. X-axis represents the 5 levels of motion intensity. Y-axis represents the count of each level in the video segment
.



(c) Spatial distribution of activity of BBB Sequence. X-axis represents the 9 rectangular regions (from left to right and top to bottom). The Y-axis represents the motion intensity level value (1-7).

Figure 3.6 – The BBB sequence with its corresponding temporal distribution of activity and the spatial distribution of activity respectively

3.2.2.3 Camera Motion clustering

A prior knowledge about camera motion in a video sequence is very important since, for instance, it helps select the suitable error concealment strategy to conceal transmission error. Camera motion histogram descriptor that is introduced in [204] analyses motion vectors to identify motion vectors of interest (MVI) that are analysed using principal component analyses to characterize the motion. The frame is divided to 9 rectangular regions. The histogram represents the dominant angle in each region. There are 13 direction identified. The output of this technique is a 117(9x13)-bin histogram. The complete description is introduced in [204]. Figure 3.10 shows the OldTownCross sequence and the corresponding camera motion histogram. It is clear that the camera motion "Zoom in" is represented well. The camera motion classification [204] is done using tree classification of camera motion descriptors using Jensen-Shannon divergence (JSD) as a distance metric. Firstly, the 117-camera histogram is generated for each sequence [204]. Second, the distance between sequences is calculated using JSD, Figure 3.11. Third, the tree classification is used, Figure 3.12. Figure 3.13 shows the classification result example.

3.2.2.4 Perceptual spatial information clustering

Using a perceptual filter, Sobel-like filter of window size of 13, the spatial information is computed to identify edges. Figure 3.14 shows the classification result example.

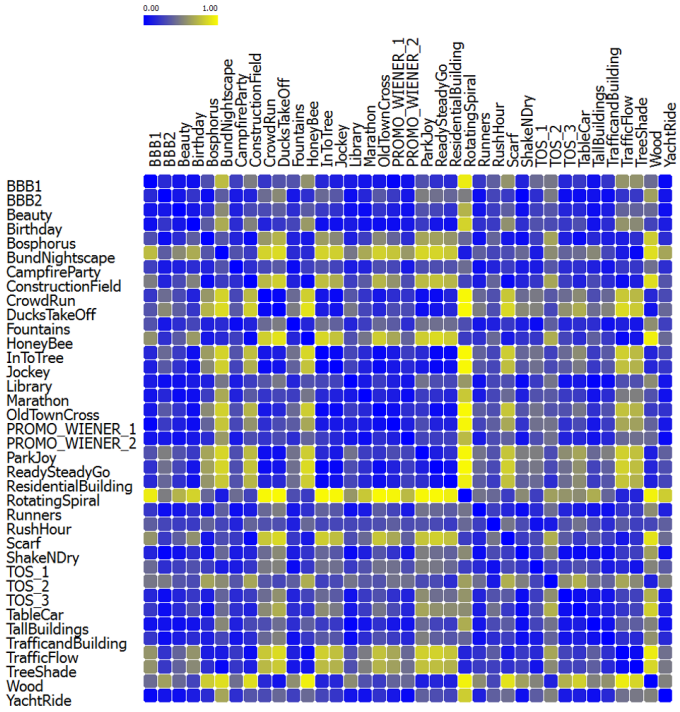


Figure 3.7 – Distance between sequences for motion intensity

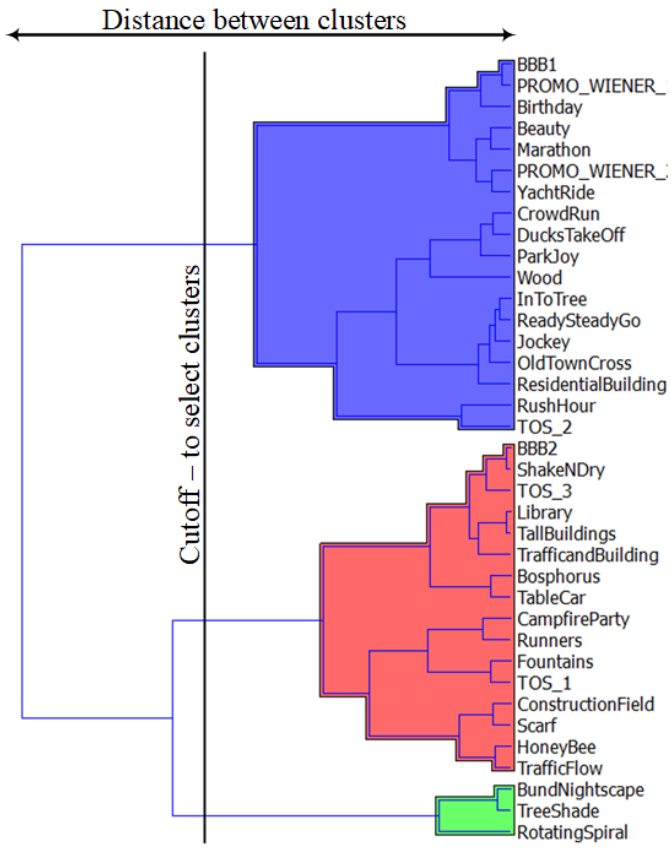


Figure 3.8 – Tree classification of motion intensity

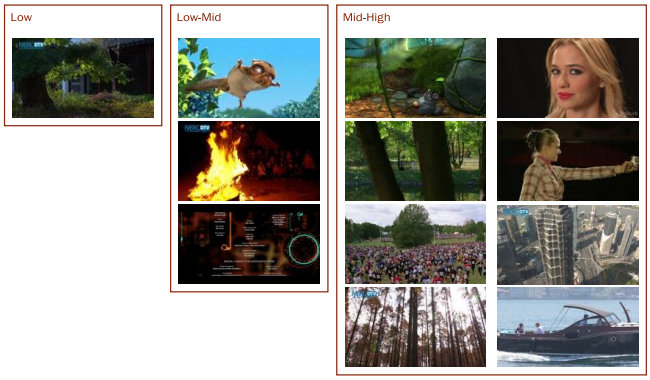


Figure 3.9 – Motion Intensity class classification

3.3 Local Video Content Features

3.3.1 Viewing Eccentricity

It has been observed in several vision studies that both: the spatial [211] and temporal [212] sensitivity of the human visual system decreases as we move away from the central region of regard. Visual processing is said to switch to a *coarser* spatial scale due to the reduced density of ganglion cells, also known in literature as *Cortical Magnification*. We measure the closest distance between the distorted region and the point of regard, see Figure 3.15, in order to calculate the viewing eccentricity. The viewing eccentricity help us to study the drop in the subjective scores of loss-impairment sequences. We analyse the effect of viewing eccentricities ranging from 0^o till 6^o of viewing angle after which the available data becomes very sparse. We analyse the fixations in the gaze data starting from the time instance when the distortion is introduced. Although a subject may possibly make a saccade as a response to the disturbance, the region of initial fixation, where he possibly perceived the change is considered for the eccentricity calculations.

3.3.2 Distortion in Texture

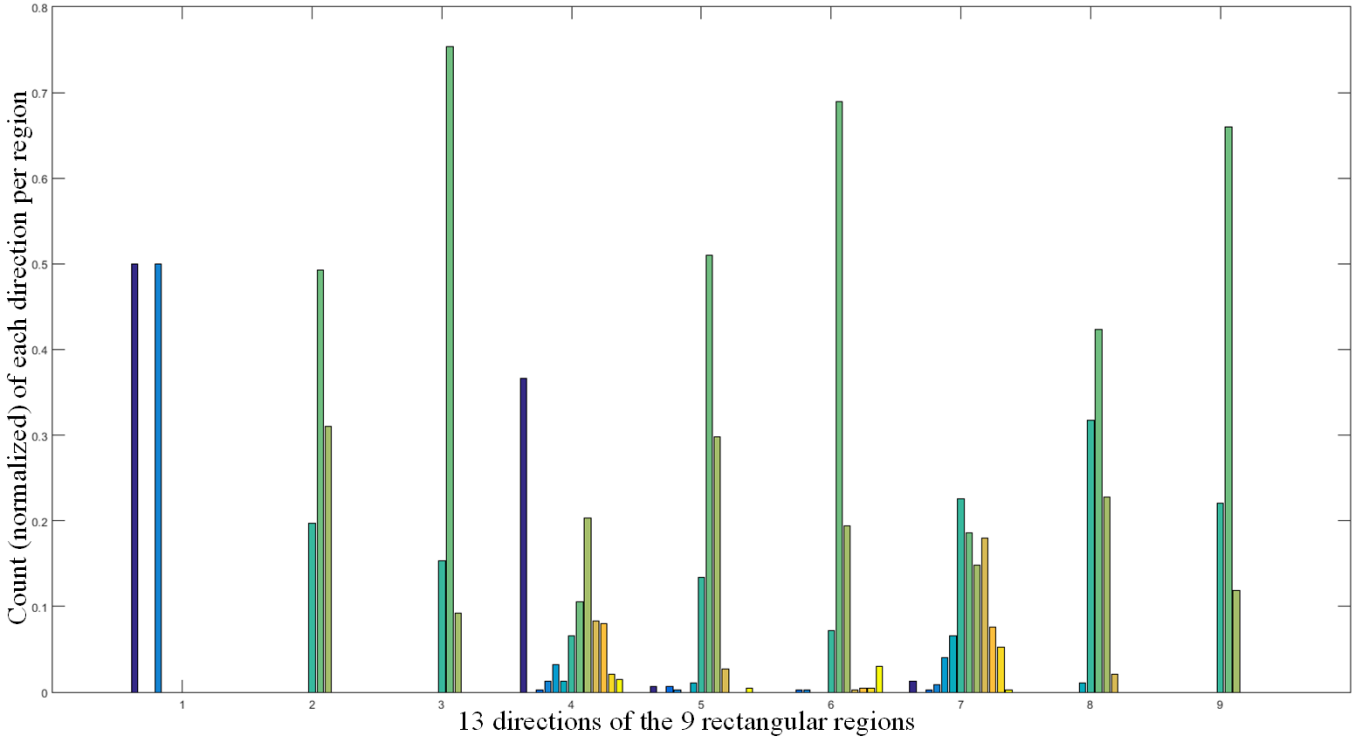
- As explained in Section 3.3.1, spatial frequency sensitivity decreases with eccentricity and visual processing changes to a *coarser* spatial scale. In addition, most modern video coders like AVC and HEVC have an effect of producing high-frequency distortions due to the aggressive quantization in these bands. Studying the effects of this frequency loss due to coding is therefore very important to determine the perceptual effects of artifacts like blurring [213, 214].
  To measure the effects of blurring, we test the strengths of three important frequencies that are deemed to be very important from former spatial contrast sensitivity studies [215]: namely 0.46 cpd, 2.8 cpd and 8.0 cpd (cycles per degree). The cpd measure is converted to a digital frequency using the conversion for viewing angle, and a complex Gabor filter centred at the three different bands (that we call Low Frequency (LF), Middle Frequency (MF) and High Frequency (HF)) and 4-orientation tuning is used to decompose the image into several bands. The responses at the four different orientations are then pooled together to produce the magnitude response for the entire image at the desired frequencies. The response obtained for one of the sequences after pooling is indicated in Figure 3.18. It is clear from Figure 3.18 that the effects of distortion are very clearly visible in the medium and high frequency bands.
  The use of Gabor filters is motivated by the fact that 1) They are optimal in space and spatial frequency in two dimensions i.e., they achieve the theoretical lower limit of joint uncertainty in space and spatial frequency; and 2) The frequency and orientation representations of Gabor filters are similar to those of human visual system [216]; 3) simple operations on Gabor filters can be established to achieve illumination, rotation, scale and translation invariance [217]. A complex Gabor filter with the response as in Equation 3.1 can be expressed as in [217].

$$\psi(x, y, f, \theta) = \frac{f^2}{\pi\gamma\nu} \exp\left(-\left(\frac{-f^2 x'^2}{\gamma^2} + \frac{-f^2 y'^2}{\nu^2} \right) + j2\pi f x' \right) \tag{3.1}$$

  where x' and y' are defined as,

$$x' = xCos(\theta) + ySin(\theta) \tag{3.2}$$

(a) OldTownCross Sequence



(b) Camera motion Histogram of OldTownCross sequence. X-axis represents the 9 rectangular regions (from left to right and top to bottom). Each region is represented with 13 directions. Y-axis represents the count of each direction per region.

Figure 3.10 – The OldTownCross video sequence with its corresponding camera motion histogram

$$y' = -xSin(\theta) + yCos(\theta) \tag{3.3}$$

f is the frequency of sinusoidal plane wave, θ is the rotation of the Gaussian envelope and the sinusoidal, γ and ν are the spatial widths of the filter along the major and the minor axis respectively. Suppose the image function is $\xi(x, y)$, the response of Gabor filter to ξ is given by the convolution between ξ and ψ as,

$$r(x, y; f, \theta) = \psi(x, y, f, \theta) * \xi(x, y) \tag{3.4}$$

The response of the Gabor filters at various scales and orientations are shown in Figure 3.16.

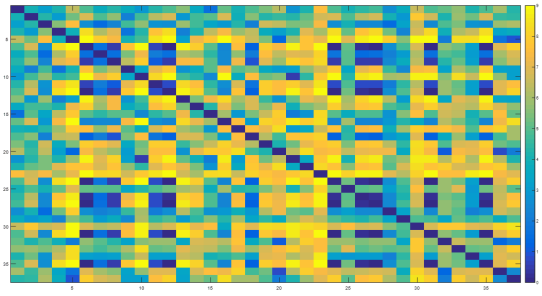


Figure 3.11 – Distance between sequences for camera motion

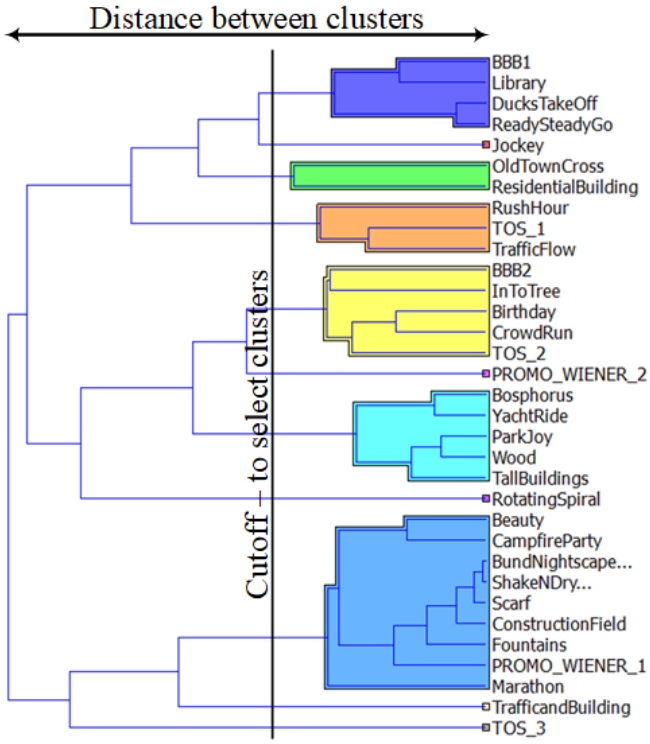


Figure 3.12 – Tree classification of camera motion

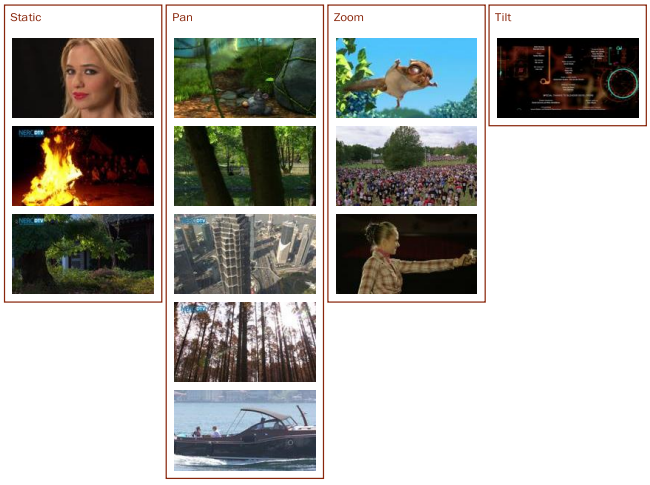


Figure 3.13 – Camera motion classification

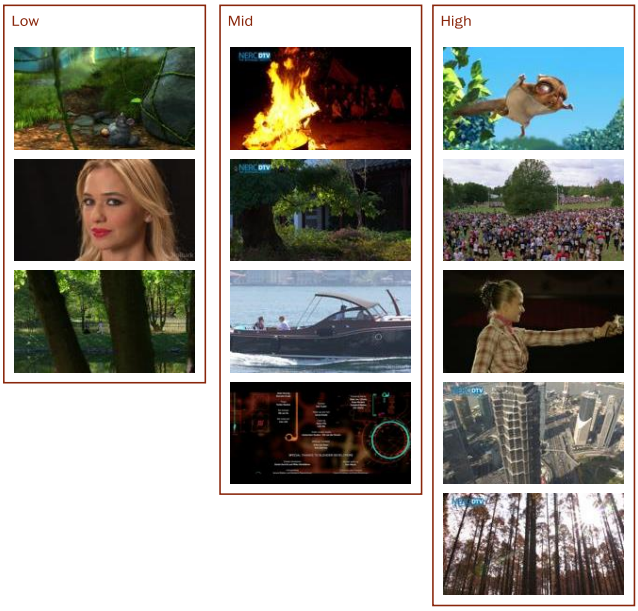


Figure 3.14 – Spatial Information classification

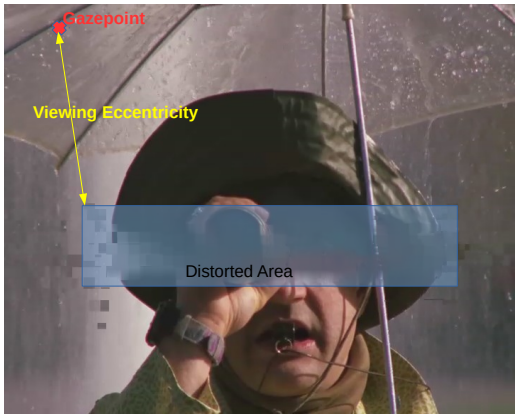


Figure 3.15 – Viewing Eccentricity definition

- The entropy map, texture entropy map, can be generated using the responses from the above step for texture clustering. Such an approach has been tried in [218]. The principle is in essentially simple : A simple square error metric is used to perform a K-Means clustering in a $M \times N$ frequency space, where M denotes the number of frequency scales and N - the orientations. Each pixel of the video is quantified by $M\ timesN$ parameters to assign it to a certain location in this multidimensional space. Afterwards the problem reduces to a simple case of clustering. The authors particularly claim that the problem has a background in the human visual system. This procedure allows us to classify a scene into a subset of regions purely based on its *texturedness*.
Calculating the entropy of such a segmented scene helps us isolate irregular or unexpected textures in a sea of rather *boring* areas. We obtain the local entropy of each region and therefore derive the amount of *unpredictability* in every local area. An output map produced by such an approach is shown in Figure 3.17.

3.3.3 Distortion in Colour

- The importance of psychophysical cone-opponent colour spaces to measure differences in colour perception, has been established in Vision Science [219, 220] and Quality evaluation [221, 222]. When light is incident on the cone receptors in the fovea, the light information is transformed into neuro-electrical signals by the three types of cone receptors (L, M, S) and are subsequently combined in an opponent manner in the cortical area V4-B of the visual system [220]. The visual periphery is said to have a deteriorated colour sensitivity as compared

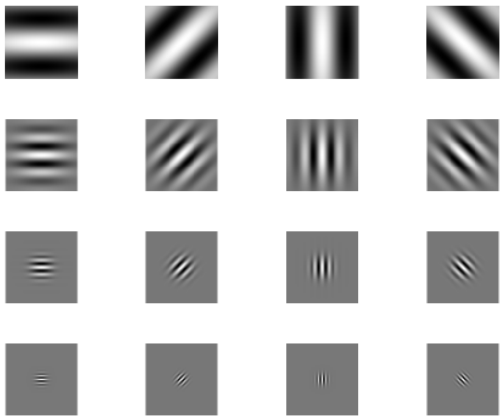


Figure 3.16 – The response of the Gabor filters at various scales and orientations

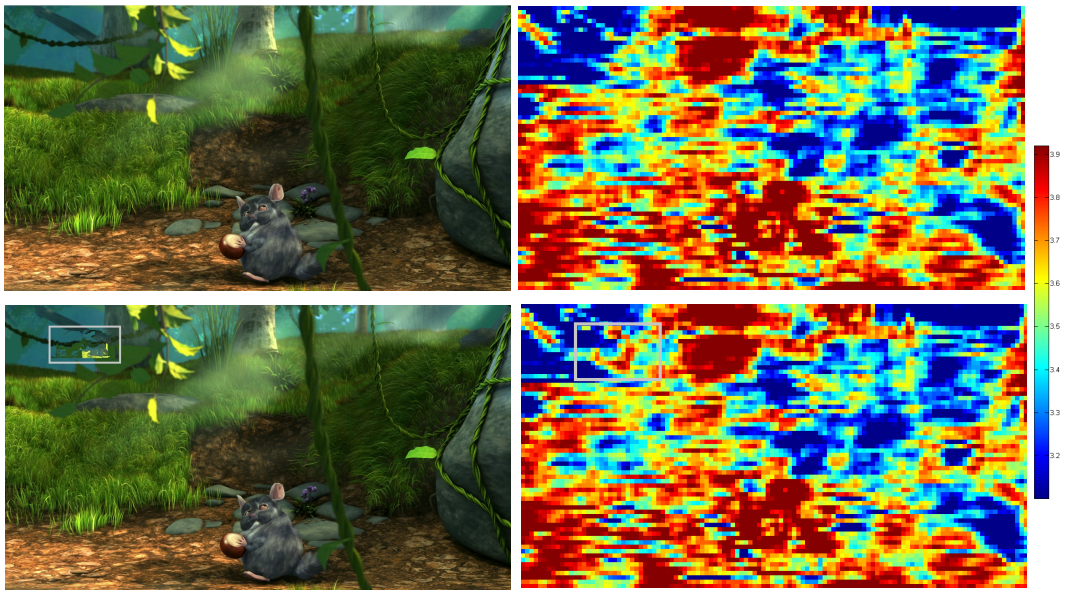


Figure 3.17 – (Left): Frames of the video without distortion and with distortion as indicated in the grey window. (Right): Responses of texture entropy for the respective cases. It is easy to notice the difference in the grey box marked area. In the other areas the responses are very similar

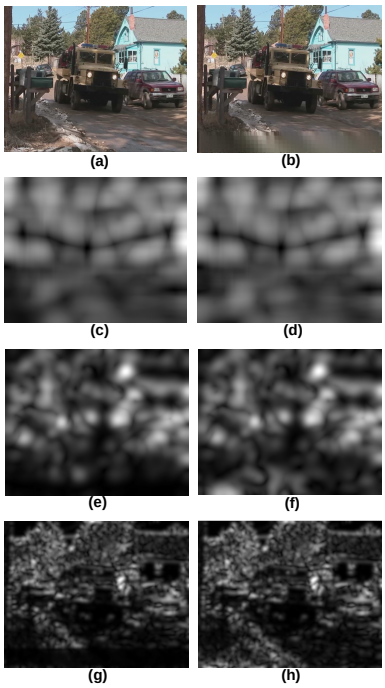


Figure 3.18 – (a,b): Original frame and a distorted frame with a distortion present in the lower part of the frame, (c,d): The responses at 0.46 cpd, (e,f): Responses at 2.8 cpd, (g,h): Responses at 8 cpd

to the fovea: more so for the red-green components than blue-yellow [219, 223].
To understand the effects of distortion in these colour channels, we use the definition of colour opposition channels from Krauskopf et. al. similar to the studies in [221, 222]. Using the display parameters like the gamma and monitor luminance, in combination with the values in the RGB colour space, we make the transformation into the Krauskopf AC_1C_2 colour space to measure the effects of distortion in C_1 and C_2 individually. However, we do not include any frequency analysis or the effects of inter/intra channel masking [222], and only analyse the influence of the channels separately.

- An entropy map, colour entropy map, can be generated using colour spaces like, for instance, luminance (Y) component of YUV colour space or lightness (L) component of Lab colour space [224]. The colour entropy map is generated as follows: firstly, for each frame, the colour space is converted to Lab colour space. Secondly, the spatio-temporal tubes are generated for each frame. Thirdly, a histogram of each spatio-temporal tube lightness is counted. Finally, the entropy value is calculated and assigned to the spatial region of the current frame. This map is generated for the original and the distorted sequences, see Figure 3.19.

3.3.4 Distortion in Motion

- Motion Trajectories: motion is a perception of an illumination that stimulates two spatially displaced photoreceptors after a specific interval of time Δt : a *percept* whose effects can be represented by a space-time receptive field which in the visual system, maps onto two separate visual path ways called the M and P pathways comprising of specific neuronal cells in the lateral-genicular nucleus(LGN) [225–228]. Experiments performed by Virsu et al [212]. in the visual periphery suggests that, although central and peripheral vision are qualitatively similar in motion perception, the quantitative differences seemed to be caused simply by the difference in spatial sampling of the retinal ganglion cells.
In the present context, motion content in the scene is represented as a series of short-term motion trajectories [229] of super-pixels. Motion vectors first created by block matching, are aggregated across each of the super-pixels, by examining the dominant direction of motion and retaining only those motion vectors lying within a certain angular range of this dominant direction. The trajectories help us construct a relatively noise free and more realistic motion representation of objects. Human observers do not deduce the motion of objects by observing merely two frames at a time, and instead have a more detailed understanding of object speeds and directions by observing them over a finite *short-term* period [229]. Any distortion in motion trajectories due to the presence of coding artifacts like in the example of Figure 3.20, therefore indicates that the naturalistic object trajectory in the scene has been affected. Examining the difference between trajectories, by computing the area between trajectory curves for example, provides a good indication of motion trajectory distortions in

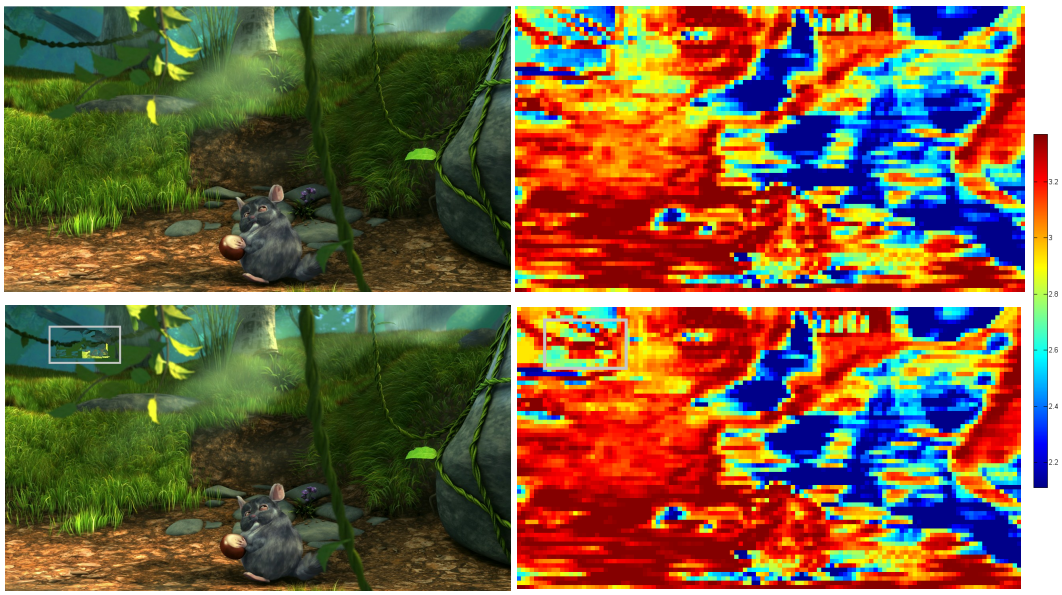


Figure 3.19 – (Left): Frames of the video without distortion and with distortion as indicated in the grey window. (Right): Responses of color entropy for the respective cases. It is easy to notice the difference in the grey box marked area. In the other areas, the responses are very similar

the scene.

- Entropy map, motion entropy map, can be generated using motion information, i.e. the motion vectors. Motion flow [230] is a very useful tool to determine the pixel-level motion between consecutive frames of the video. Because videos and natural scenes often contain discrete, smoothly moving regions, there is a spatio-temporal homogeneity in the motion map so acquired. Any perturbations or non-homogenity in the regions therefore are an indication of a localized distortion. Because motion is an important factor that often attracts attention in case of videos, motion distortion is defined as the amount of disturbance caused to an otherwise smooth motion trajectory as a result of the distortions. It is expected that the presence of the distortion causes a disturbance in the otherwise smooth motion flow.
  Soft decision: After calculating the motion flow within a certain scene, we then cluster all the pixels in the motion map in accordance to their direction and magnitude. The cluster centers are automatically chosen using a K-Means clustering for a set of frames. 8 Angular bins and 6 magnitude bins are used in the work and the assignment is done as in Equation 3.5 where $C(j)$ refer to the j cluster centers, $m(i)$ the actual motion flow value and $H(i)$ the actual hard decision for every pixel i. Such a clustering is performed separately for the magnitude and direction. Although such an approach seems like a reasonable solution to cluster the motion into separate bins, there is often a problem when we deal with motion flow values that are close to the border value in between two bins. Assigning such values to either of the two bins can have a huge impact on the final entropy score obtained.

$$H(i) = \underset{j}{\operatorname{argmin}}(m(i) - C(j)) \tag{3.5}$$

  We therefore use a *Soft Decision* process where several bins in addition to the exact cluster are incremented based on their distance to the remaining cluster centers. This approach is similar to that used in [231] where a Gaussian was used to explicitly assign the probability to each gaze-state based on its distance from each super-pixel center. The soft decision $S(i)$ for each $m(i)$ is performed as in Equation 3.6 where $\phi(x)$ indicates a Gaussian with mean value x.

$$S(1, 2 \cdots J) = \phi(m(i) - C(j)) \tag{3.6}$$

  Once the soft decision map is obtained for all the pixels in the map, an entropy filter is applied in the usual manner to identify irregularities in the motion flow.

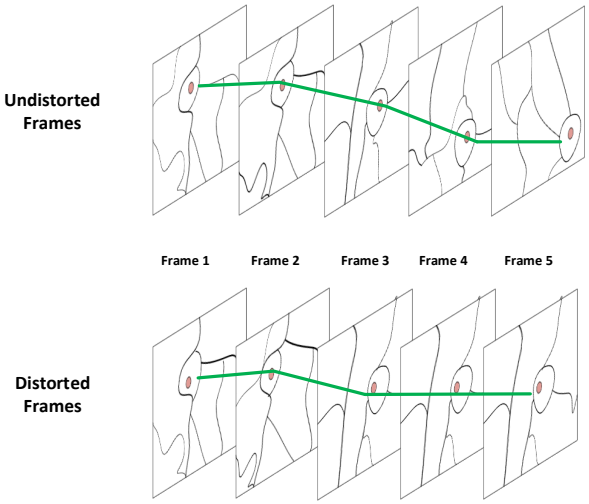


Figure 3.20 – Trajectories of a super-pixel in two different video sequences: the first, a pristine case and the second, the case when alternate frames are repeated.

3.3.5 Distortion in Temporal Harmonics(Flicker)

In the domain of video compression, it has been found that for both AVC [232, 233] and HEVC [234] video coding standards, flicker plays a very important role in deciding the overall quality of the video. Studies from Snowden et al. [235] on flicker perception in the far-periphery found that the sensitivity function is band-pass and the peak sensitivity lies at 10Hz for all eccentricities and spatial frequencies tested.
Flicker in case of [235] is defined as a sinusoidal signal moving in the temporal dimension or a *Temporal Harmonic*. To maintain a similar definition, we compare the changes in energy of the various temporal bands in the video using the scheme indicated in Figure 3.21.
The analysis begins with the motion compensation of all the forward frames in the video, in order to nullify the effects caused due to motion. We then average blocks of pixels 32×32 that roughly corresponds to a receptive field in the visual system. These super-blocks are subsequently collected from successive frames over a *short-term* and are subject to Fourier analysis as shown in Figure 3.21. Based on the studies of [235], we compare the loss/gain in temporal frequencies in the pristine reference versus the test video in three distinct bands : 0.8 to 3.9 Hz (that we call LF flicker), 4.69 to 8.6 Hz (that we call MF flicker) and 9.3 to 12.5 Hz (that we call HF flicker).

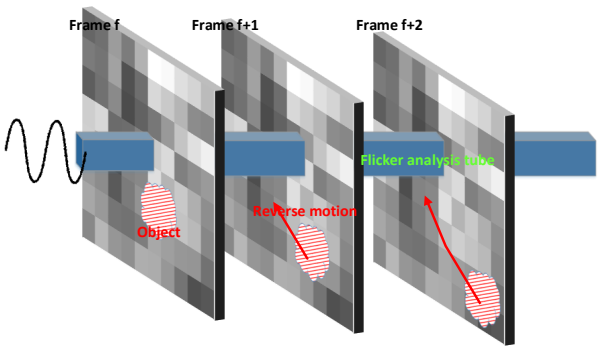


Figure 3.21 – Analysis of harmonics in a cuboidal short-term tube: Forward frames in the video are first motion-compensated and the intensity level inside each block averaged, before performing a Fourier analysis.

3.3.6 Role of Semantic Importance

It is important to understand as to, which objects human subjects regard to be the most important in a video, so that the semantic importance feature can be calculated. To measure the relative importance of objects in the scene, a ground truth knowledge of all the objects in the scene and their exact boundaries are required (sequences in a segmented form) [236].
To measure the importance of every object in the video, different set of subjects were used in order to avoid any bias due to repeated viewing. The video was played in one of the displays continuously, until the subject finished marking

the objects of primary and secondary importance in another display. A specialized tool was used for the purpose as shown in Figure 3.22. The importance score for every individual object in every video were then averaged among the subjects to obtain an average importance that is less affected by individual variations and has a better precision.
Despite their common goal of identifying the most relevant information in a visual scene, the type of relevance information that is predicted by various visual attention indicators can be very different. Rather than the salient portions of an image, users are often interested in those portions that are semantically more meaningful and convey maximum information about the scene. This perceived interest in objects is strongly driven by context and semantic information, and involves usually a voluntary control of the gaze shift. Analogous to saliency maps we define another map known as the Importance map [237] that is obtained using the procedure in section 9.2.1.4. The importance map, Figure 3.23, is a construct representing each of the objects in a scene with an importance score obtained from users.

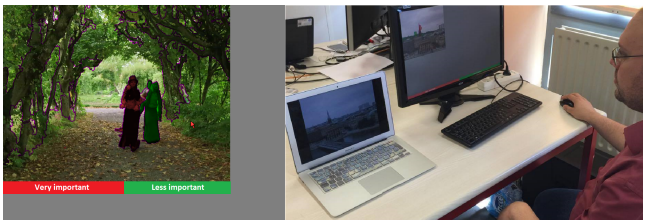


Figure 3.22 – (a) Application to mark the importance of objects. Subjects first click the red/green coloured rectangular box to select the importance level and then choose the object. (b) A subject performing the experiment by watching the video in one screen simultaneously marking the importance in the other.

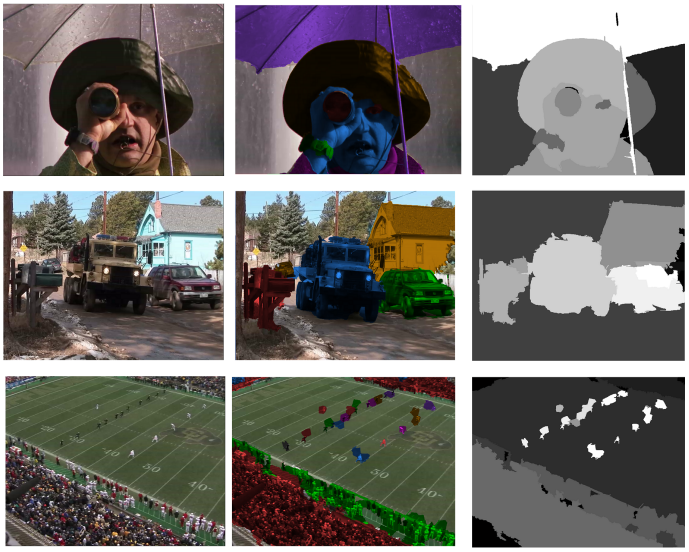


Figure 3.23 – Left Column: Frames from three different videos used for the experiment, Middle Column: Manual marking of different objects in the scene each marked with a colour, Right Column: Average importance rating of the objects from 14 different observers where white indicates high importance and black very low.

3.4 List of contents

Due to the chronological development of the PhD, different video sequences are used in the different parts of it. Therefore, the following subsections list the video sequences that are used in each experiment in the PhD.

3.4.1 Contents for Part II

In Part II, 37 UHD video sequences are used. The thumbnail of those sequences are shown in Figure 3.2. Two of them are excluded due to the licence issues. The 35 UHD video sequences that are available for research purposes are investigated. Conditions and limitations mentioned in [3] and [4] are considered.

3.4.2 Contents for Part III

The selection process that is illustrated in Section 3.2.2 is applied. As a result, 12 source sequences are selected and they are in ultra-high definition (UHD) with a resolution of 3840x2160 pixels. Figure 3.24 shows the thumbnails of the video sources. The frame rate of the video sequences varies from 25 frames per second (fps) to 120 fps. Each sequence is 10 seconds long. Video sequences cover different video properties: motion intensity, camera motion type, spatial complexity, and colours.

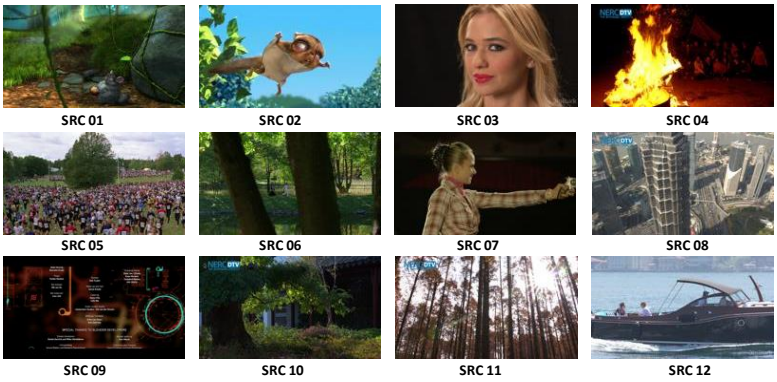


Figure 3.24 – 12 UHD video sequences that are used in Part III

3.4.3 Contents for Part IV

3.4.3.1 Contents for Chapter 8

Eight of twelve sequences that are used in Part III are selected because of the computational power issue. Besides, the contents are down sampled to the resolution of 1280×720.

Figure 3.25 – Thumbnails of the eight 1280x720 Video Sequences that are used in Chapter 8

3.4.3.2 Contents for Chapter 9

In Chapter 9, an eye-tracking subjective experiment is conducted. Hence, the twelve sequences that are used in Part III are changed, i.e. some content are removed and some are added. This is only to adapt the goal of the experiment without changing the general characteristics of the contents and to ensure that the contents have different number of objects in the scene. Since this Chapter is about content-aware disturbance analysis of the inpainting-based error concealment technique, the contents are down sampled to the resolution of 1280×720 due to the computational power issues. The thumbnails of the selected 14 sequences are shown in Figure 3.26

3.4.4 Contents for Part V

In order to achieve the goals of Part V, a large-scale database has to be used. JEG Hybrid Group, one of the Video Quality Expert Group (VQEG) projects, aims to improve quality metrics. They publish a dataset to be a reference for video quality researchers [238]. Figure 3.27 shows the thumbnails of the ten video sources, each 10 seconds long. The sequences are encoded using HM11.1 with different encoder parameters, see Figure 3.28. As a result, 5952 compression scenarios are generated. In addition, the dataset provides the results of applying different quality measures like PSNR,

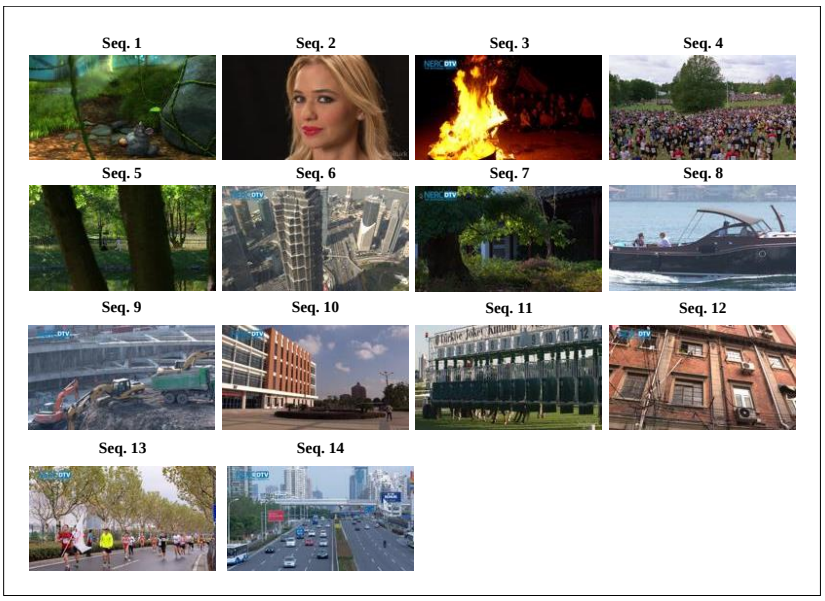


Figure 3.26 – Thumbnails of the 14 1280x720 Video Sequences that are used in Chapter 9

SSIM, VIF, and VQM. An extension to this database which adds a large number of objective quality evaluations when compressed video streams are subject to data loss. A set of 25 loss patterns has been generated by means of a 2-state Markov model [239] using loss rate values up to 1% and average burst length up to 2 slices. Applying each loss pattern to each sequence, 25 degraded bitstreams have been generated, decoded, and objective measurements have been calculated. Note that for each degraded bitstream we used the effective loss rate and average burst length, as measurable at the receiver, which may be different from the settings of the model since only part of the loss pattern has been used. Currently, due to the huge computational effort, this activity has been performed in full only for the lower-resolution set of encoded sequences (i.e., 19,840). Therefore, 496,000 combinations of encoded sequence and loss traces have been evaluated with several objective quality metrics. Depending on the position of the lost packets, determined by the 25 loss patterns, and the encoding configuration, error propagation of different duration occurs. Using a robust decoder simulation, [240], all video sequences can be decoded, there is no temporal offset, and the number of affected frames can be exactly calculated.

Figure 3.27 – 10 video sources of JEG dataset that are used in Part V

3.5 Conclusion

In this chapter, the main contributions are introduced and listed in Box 3.3. Moreover, the lists of video sequences that are used in different parts of this PhD are identified.

VBR: QP	26, 32, 38, 46
CBR: frame level	0.5, 1, 2, 4, 8, 16 Mbps
CBR: CTU level	0.5, 1, 2, 4, 8, 16 Mbps
Random access	Closed-GOP intra refresh (IDR), Open-GOP intra refresh (CRA)
Intra period	8, 16, 32, 64
Resolution	1920x1080 1280x720 960x544
Slices	Count: 1, 2, 4; Size: 1500 byte
GOP structure	GOP size 1 (IPPPPPPPPP) GOP size 2 (IBPBPBPBP) GOP size 4 (IBBBPBBBP) GOP size 8 (IBBBBBBBP)

Figure 3.28 – HM encoder parameters that are used in [238].

Box 3.3 – Contributions

— The set of global/generic content features are identified. These content features are used in different parts of this manuscript.
— A content selection process is illustrated to select a subset of video sequences to be used in the experiments that are conducted in this PhD.
— The local features that are designed, investigated, and implemented for disturbance analysis of error concealment techniques are introduced.

Proof-of-Concept: Role of Generic Content Characteristics in Optimizing Video Encoders; predicting the video encoder's parameters

4

Complexity- and Content-Aware Sequence-level Encoder Parameter Decision Framework

4.1 Introduction

The recent development in multimedia devices and mobile networks have opened the door for end users to easily capture videos with different resolutions and qualities, therefore the demand for delivering high quality immersive videos is increasing. Moreover, smart-phone applications became popular and important. On the other hand, these devices have limited computational power and batteries. The latest video coding standard, High Efficiency Video Coding (HEVC) [14], is designed especially to target different types of applications and particularly high resolution video applications [15]. Quality, bitrate and complexity (encoding time) are the key elements of video coding performance evaluation.

The complexity of HEVC is increased due to the new/improved coding tools. This complexity is a liability for some targeted users, for some applications, or, for some devices. Some targeted users, like content providers, may not care about the complexity since they have the power to build high performance encoders, i.e. parallel encoders. Some applications (security and safety applications) require that the captured videos need to be quickly encoded and sent. Due to the limited computational power and batteries of some devices, the complexity is an important issue. Therefore, tools to reduce the encoding time without compromising the coding efficiency and the perceived quality are important. There are several sources of complexity increase in video coding. First, the new or improved encoding tools that are introduced in HEVC such as new intra and inter modes, new quadtree block structure, improved motion estimation, and the number of reference frames [15]. For instance, testing all combinations of block splitting and inter modes in each reference frame will highly increase the complexity. In [241], the distribution of encoding time per operation and encoding configuration is analysed. Second, choosing encoder parameter values also trade-off the quality and the complexity. For instance, selecting a smaller motion search range value, accelerates the encoding process at the price of quality and a larger value may slow down the encoding process. Finally, many research efforts have pointed to the importance of content types and its underlying characteristics in video coding.

The existing tools, Section 2.2, are focusing in reducing complexity to a certain extent while the quality loss and bitrate decrease levels are not assured and the awareness came from the fact that some of the modes and tools of the video coding either are rarely used or unnecessary in some situations, Section 2.2. A room of improvement can be accomplished not just for complexity reduction but also to trade-off between bitrate (R), distortion (D), and complexity (C) by utilizing the underlying content features to predict the encoder parameter values. In this work, a new approach that addresses and analyses the content features to predict the encoder parameters values are demonstrated. In the analysis phase, the content features are analysed to find the content features that have an influence in deciding the appropriate encoder parameter values at sequence level and for a given QP. In order to find this relationship, R (the total bitrate of a sequence that is required for transmission/storage), D (the PSNR values or another video quality measurement), and C (the encoding time that is required to encode a sequence using specific configuration) are considered. Then, for each sample in the dataset, the content features are associated/labelled with the appropriate encoder parameter value (class). Finally, the dataset is trained using classification tree or support vector machines learning algorithms. This prediction model is used to predict the encoding parameters of the video to be encoded.

The results show, for instance, that predicting motion search range achieves complexity reduction of 36% on average when the HEVC reference software HM13 is used.
Box 4.1 shows the research questions that this chapter is trying to answer and Box 4.2 shows the structure of this chapter.

Box 4.1 – Research Questions

This chapter aims to answer the following research questions:
- Can the generic/global content features be used as indicators for finding the links between the content features and the encoders parameters? If so, then building a joint content and complexity aware encoder's parameters prediction model is applicable.
  In addition to that, the following secondary research questions are investigated too:
  + How does the encoder behave in terms of complexity with different content?
  + How is the encoder complexity linked with different parameters per content?

Box 4.2 – Chapter structure

This chapter is structured as shown in the Figure 4.1. Section 4.2 shows the observations and identifies the problem statement. Steps for the data preparation will be demonstrated in Section 4.3. Classification steps and the prediction model will be illustrated in Section 4.4. The evaluation of the proposed model will be shown in Section 4.5.

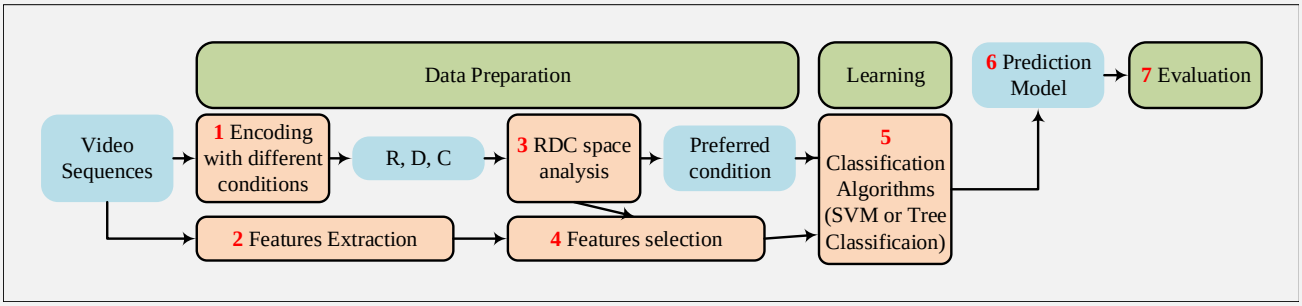


Figure 4.1 – Chapter 4 Structure

4.2 Observations and problem statement

4.2.1 Observations

In this subsection, a demonstration to show that a specific encoding configuration is not necessarily suitable for different types of contents in terms of trading-off R,D, and C. Consider a set of four video sequences that have different spatio-temporal properties $V = \{V1 = \text{TrafficFlow}, V2 = \text{HoneyBee}, V3 = \text{Jockey}, V4 = \text{CampfireParty}\}$, Figure 4.2, a subset from the 35 sequences that are used in this work. Then, encode them with different coding unit sizes and depths (CUSize/MaxDepth); 16/1, 16/2, 32/2, 32/3, 64/3, and 64/4. All other encoding parameter values are common ($QP = 32, motionSearchRange = 64$). Since only one parameter is changed, the variations in R, D, and C are due to the contents and links between content features and the varied coding parameter might be identified. Figures in the first column of Table 4.1 show the R (Kbps), D (dB) and C (hours) with respect to different sizes and maximum depth of the coding unit. It can be observed that the variations in bitrate between contents are different. For instance, the variation in V2 is low, while the variation in V3 is high. The same observation can be noticed when the motion search range (MSR) is changed to ($MSR = 16, 32, 64, 128, unrestricted(full)$) and all other parameters are common ($QP = 32, CU/depth = 64/4$), Figures in the second column of Table 4.1. This observation can be used to reduce the coding time while the bitrate and the quality are very slightly compromised. For instance, the bitrate and the quality of HonyBee sequence (V2) is very slightly compromised if the CU size/depth is set to 16/1, but the gain of complexity is very high. Moreover, Figures in the third column of Table 4.1 show all R, D, and C of all possible combination of CU size/depth and motion search range, i.e. 36 configurations. From these figures, [4.1], several observations can be concluded;

— content types/features have an impact in setting up the encoding configuration. Therefore, what is good for one content is not necessarily good for other contents. This conclusion will be clearer by the end of this subsection when trading-off between R, D, and C is conducted.
— key element of encoder configuration such as CU size/depth, motion search range are independent. This concluded observation would simplify the design of the proposed tool by predicting each parameter value independent of others.
— trading-off tool for R, D, and C is important to balance the gains and losses in terms of R, D, and C.

Table 4.1 – Bitrate, Distortion, and encoding time against (CU size/Max depth, motion search range).

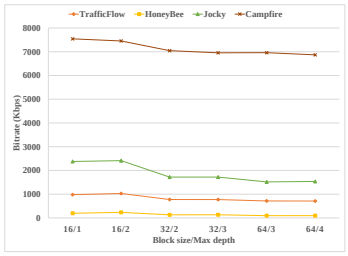
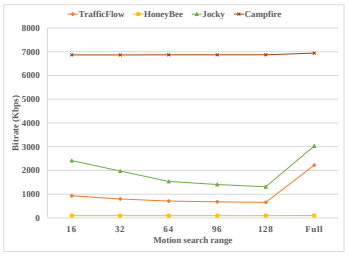
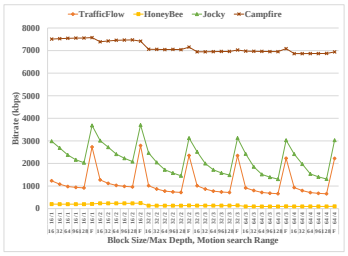
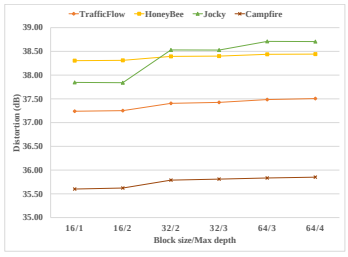
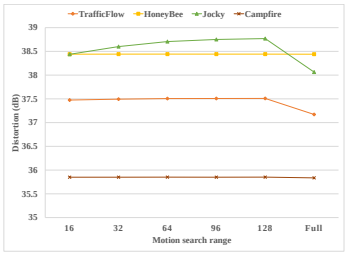
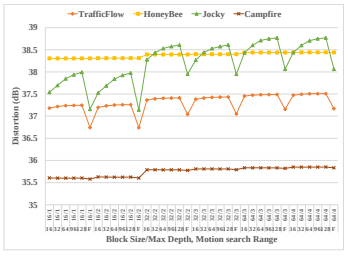
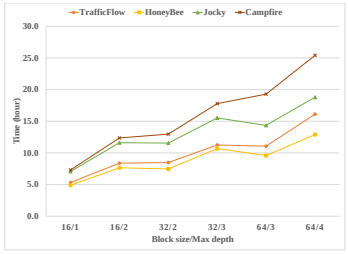
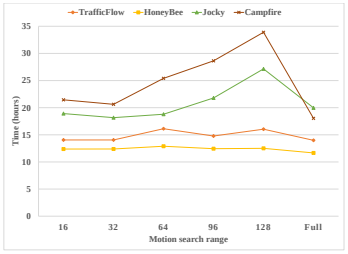
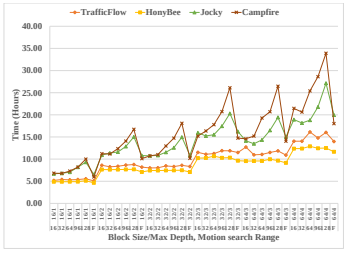
	CU size/max depth	Motion search range	Both combinations
Bitrate			
Distortion			
Time			

Figure 4.2 – Four sequences that are used for the observation, from left to right; trafficFlow, honeybee, jockey, and campfireparty

To trade-off between R, D, and C, a tool that we developed in [242] is used. The tool applies a linear optimization model and provides a visualization tool that helps select the best configuration for a specific R, D, and C point of the analysis space as shown in Figure 4.3. In [242], 13 video sequences are used to show the results of the trading-off tool, Figure 4.4. The sequences are a subset from the 35 sequences that are used in this work. These sequences are encoded with seven different configurations as shown in Table 4.2. In order to judge the configuration against others, a linear combination of the three components, bitrate saving, distortion saving, and complexity saving, is established as an optimization criterion and expressed as in Equation 4.1, where (C_{minC}) the minimum-complexity configuration.

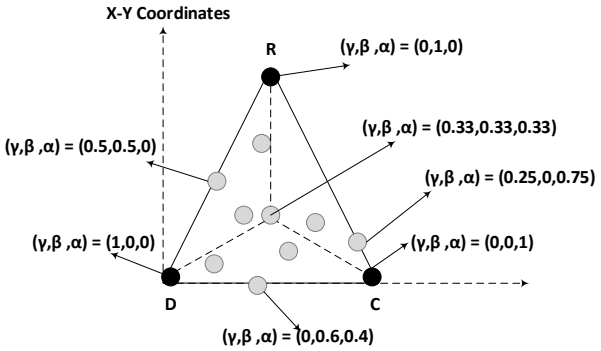


Figure 4.3 – The analysis space of the tool in [242]

Figure 4.4 – 13 sequences that are used in [242]

α, β, γ the three coefficients that need to be tuned to obtain the optimization criterion value. These coefficients are restricted to a sum of one. Each point in the analysis space represents the contribution factor of each component of the optimization criterion, i.e. bitrate, distortion, and complexity and for the visualization purposes, these points represent the best configuration.

$$O = \alpha \frac{C}{C_{minC}} + \beta \frac{R}{R_{minC}} + \gamma \frac{D}{D_{minC}} \tag{4.1}$$

An example of applying the proposed visualization tool to the 13 video sequences using configurations that are listed in Table 4.2 is shown in Figure 4.5. Each point in the analysis space represents the best configuration. The selection of the best configuration is locally optimized within a limited bitrate and distortion range since the quantization parameter is fixed. It can be concluded from Figure 4.5 that:

— Configurations 2 and 3 should not be considered for sequence #1 as no gain is obtained for any rate-distortion-

Table 4.2 – 7-different encoder configurations that are used in [242]

Parameters	Low			Medium		High	
	1	2	3	4	5	6	7
Coding Unit/Depth	16/1	16/2	32/2	32/2	32/3	64/3	64/4
Transform Unit min-max	2-2	2-3	2-3	2-3	2-4	2-4	2-5
Motion Search Range	32	32	32	64	64	Full	Full
IntraPeriod	8	8	8	16	16	32	32
- GOP is 8 (hierarchical B-Frames) with QP increased by one in each level - QP is 32 - Full: Full search mode							

complexity operation point in comparison to the other configurations.
— Configuration 1 can be considered for less complexity, configuration 4 can be considered for balancing the three components, and configurations 6 and 7 can be considered for the best quality, for sequence #1.
— For sequence #13, configurations 6 and 7 cannot be used since better results are given with 4 and 5.
— There are sequences that behave alike which is important to note as it points to content properties similarities.

In general, the optimization criterion is critical as it acts as a decision maker for which configuration should be selected. Changing this criterion alters the selected mode for one parameter, and, consequently, the video sequences properties that influence this parameter may change as well. Many changes can be done to this criterion. Firstly, a logarithmic mapping function can be applied to the bitrate (R) and to the complexity (C) due to their logarithmic behaviour. Secondly, another distortion measurement can be applied rather than the usual PSNR due to its limitations with respect to modelling the Human Visual System (HVS). Methods like SSIM [61], MS-SSIM [62], PSNR-HVS [243], PSNR-HVS-M [244], VIFp [245], and VQM [63] can be tested and it is recommended as a future work.

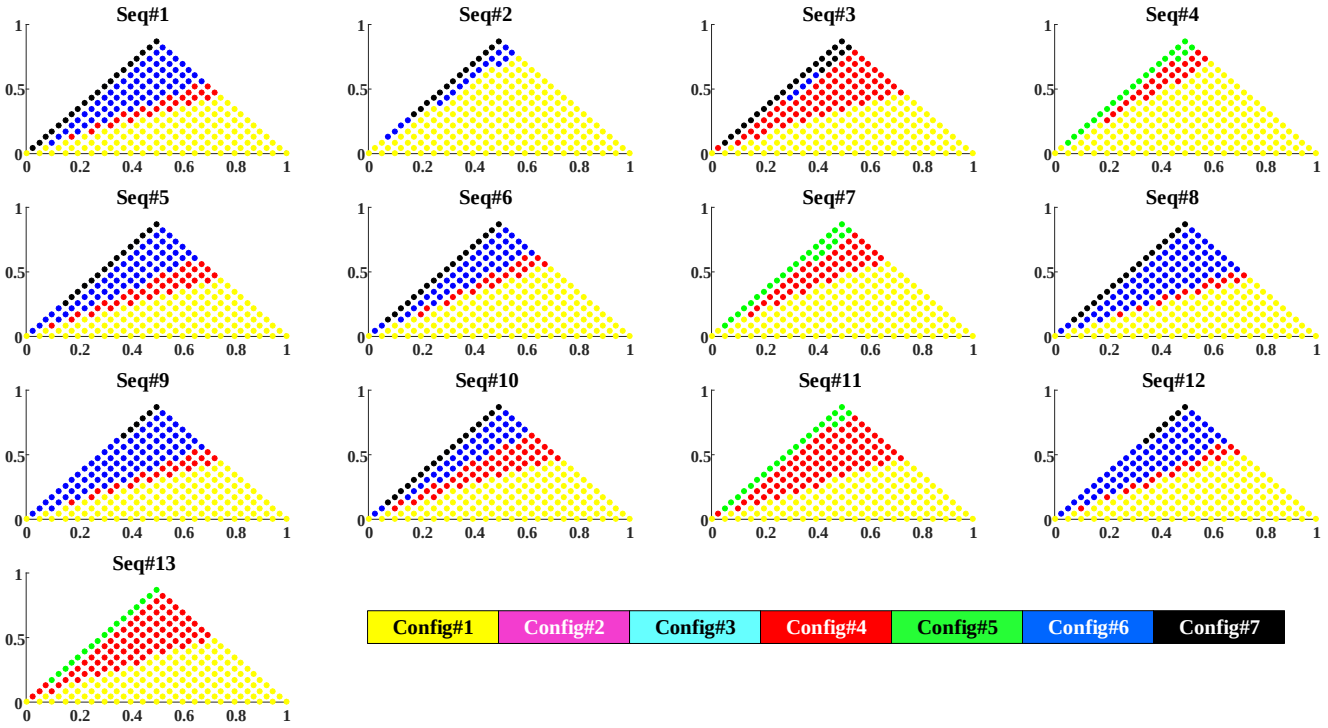


Figure 4.5 – Visual analysis of optimization criterion with 7-encoding configurations for 13 sequences [242]. The colours refer to the configurations as shown in Table 4.2

4.2.2 Problem Statement

It is observed in the previous section that in function of the content, encoder parameters lead to different results with respect to bitrate, distortion, and complexity. Moreover, some contents also behave alike which points to a fact that they share some similar features. Therefore, prior content awareness together with complexity awareness of encoding parameters helps predict the suitable encoding parameter values of a specific content type. A model that utilizes this awareness is introduced in this work. Figure 4.6 shows the proposed model. At the end, the original input video (V_I) is encoded by the encoder instance ENC using global/final decision parameters (P_{GD}) and fixed parameter(s) (P_F). The optimization process (OPT) might do many iterations until the final decision is determined (P_{GD}). The optimization process (OPT) predicts the encoder parameter values. Optimization model uses video content features (F) extracted by (EXT), the extracted features are listed and reviewed in Chapter 3. Fixed parameters (P_F) can be bitrate, distortion, quantization parameter, complexity, or any combination of them. In this work $P_F = QP$. Global parameters (P_{GD}) can be mode decision of intra or inter prediction, encoder parameters such as motion range or block size, quantization parameter, or a set of them. In this work, P_{GD} = encoding parameters that can be set in the configuration file of the encoder. When the optimization process finishes, the encoder starts encoding using P_{GD} and P_F parameters. Finally, the output video (V_O) is delivered. This model works at least on the Group of Pictures (GOP) level in order to optimize for a set of video frames. In this work, the sequence level is used in order to obtain statistically stable and relevant content features.

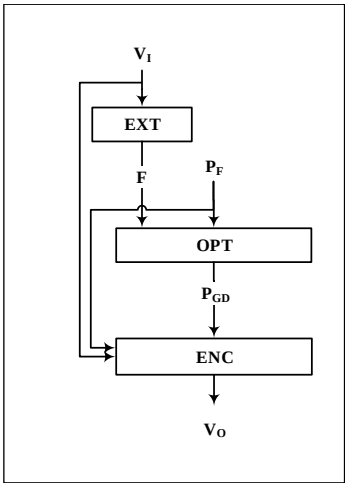


Figure 4.6 – The proposed optimization model

4.3 The Proposed RDC Optimization Model

4.3.1 Model Overview

Figure 4.7 shows how the proposed model is utilized. When starting to encode the video sequence (V) using a specific quantization parameter ($P_F = \{QP|QP \in \{0....51\}\}$), the RDC model ($OPT$), predicts the parameter values (P_{GD}), e.g. motion search range. The prediction model (OPT) is based on analysing the content properties (F) of the video sequences in order to know if these properties are correlated with encoding parameters for a given quantization parameter or not. The extracted features/properties are described in Section 3.2.1 and the offline analysis steps will be illustrated in Section 4.3.2. The model determines the encoder configuration parameters (P_{GD}). During the training phase of the model, each encoding parameter is marked as predictable or fixed. Predictable parameter means that the model found correlations between the video properties and the encoding parameter. On the other hand, the fixed parameter means that the model did not find correlation between video properties and the parameter. The optimization criterion that the model uses in the analysis steps to find the correlated properties is illustrated in [242]. This criterion considers the bitrate, complexity, and distortion and selects the best configuration. For instance, if two configurations have approximately the same bitrate and distortion, the one of the lowest complexity is selected and if they have approximately the same complexity and distortion, the one of the lowest bitrate is selected. Since the encoding parameters are independent, as noticed in observations Section 4.2, the model will start predicting the first parameter, block size, then the second parameter, intra period, and so on until all parameters are set to start encoding.

4.3.2 RDCO Model Training

The proposed system's overview is described above. In this subsection, the analysis steps are demonstrated systematically in order to build the prediction model (OPT) for each encoding parameter.
The set of video sequences will be denoted as $V = \{v_1, v_2, \ldots, v_n\}$ and each video will be encoded using the parameters $P = \{p_1, p_2, \ldots, p_m\}$ and each parameter has different values $K_{pi} = \{k_{pi1}, k_{pi2}, \ldots, k_{pil}\}$, where $1 \leq i \leq m$, l is the number of possible values of each parameter.
The flowchart, Figure 4.8, illustrates the steps that the model follows to build the ground truth dataset. Finding the links between encoding parameters and the video properties is the aim of this analysis. In the first step (1.1), different encoding configurations that range from low to high computational complexity are prepared by fixing all parameters except one (p), for instance motion range. Thereby, the variations in the encoding results are due to content properties, textures and motions. Then, encode the video sequences (V) and get the bitrate, distortion, and complexity. Complexity is measured in terms of execution time in this work˙ The second step (1.2) is to apply the optimization criterion (O) to each analysis space point. As discussed in Section 4.2, it can be said that each point in the analysis space represents the contribution factor of each component of the optimization criterion, i.e. bitrate, distortion, and complexity. By applying the optimization criterion to each point for the different encoding configurations, each point will represent the best configuration.
The goal of the third step (1.3) is to know if the parameter is influenced by the video properties (see Section 3.2.1). Therefore, for each video sample, it is assumed that the configuration that covers most of the analysis space is the winner; for that, the mode function is applied.

$$k_{pi}(v) = mode_{\alpha,\beta,\gamma}\{max\{O(p,v)\}\} \tag{4.2}$$

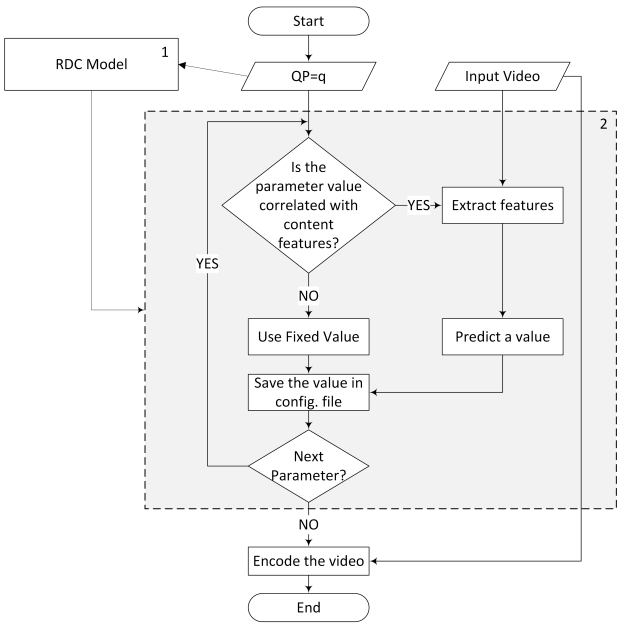


Figure 4.7 – Utilizing the proposed model. In order to simplify, the figure is restricted to a subset of parameters

.

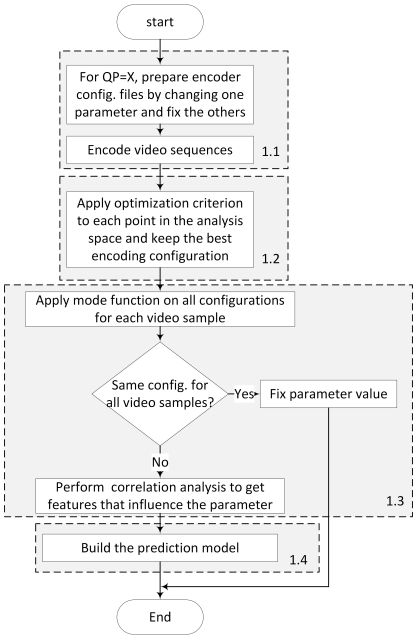


Figure 4.8 – The RDCO model training

If $k_{pi}(v)$ values of all video sequences are the same, then the model fixes the value for future use, as it is not influenced by any of the video properties. For instance, all video sequences select motion range of 64 to be the best mode. On the other hand, if $k_{pi}(v)$ values are different, then this parameter is influenced by video properties. For instance, some video sequences were best coded with motion range of 64 and others were best coded with unrestricted motion range. In this case, correlation analysis has to be conducted and features that have significant correlation on a 10% of confidence interval are selected. Figure 4.9 shows an example, the X-axis represents the selected modes (1=motion range 64, and 2=unrestricted motion range), the Y-axis represents the feature values (energy ratio of two laplacian subbands 1 and 4), and the points represent the video sample.
The fourth step is to build the prediction system. Here, two choices are available. The first one is to apply classification tree algorithm with one of the attribute selection criterion, such as information gain or gain ratio. Then the model can select up to three or four properties for future prediction. The second choice is to train a learning algorithm such as SVM or decision trees for future prediction.

4.4 Experimental Results

In this work, the model results are based on a comparably small dataset, 35-UHD video sequences, and enlarging the dataset increases the model robustness. This may change the prediction model, i.e. new correlated features may show up, for one parameter or change the fixed values. In both cases the general idea of the model can still be applied. The 35 UHD video sequences that are available for research purposes are investigated. Conditions and limitations mentioned in [3] and [4] are considered. These sequences are from;

— SJTU 4K Video Sequences [205]
— Ultra Video Group 4K sequences [206]
— Xiph.org Video Test Media 4K sequences [246]

The influence of video properties in the motion range and the block unit size parameters are investigated . HM13.0 encoder [247] is used. QP of 32 is used. For the motion range parameter, the video sequences are encoded by fixing all parameters and changing the motion range to 32, 64, 96, 128, and unrestricted search mode. The video sequences are also encoded by fixing all parameters and changing the block size and its depth to 16/1, 16/2, 32/2, 32/3, 64/3, and 64/4 for the HM encoder.

4.4.1 Study the content influence with respect to the block size parameter using the HM encoder

Using the HM encoder, the model shows that there seems to be no content influence with respect to the block size parameter. The mode of block size equal to 16/1 is selected. That means that during the analysis steps, the mode

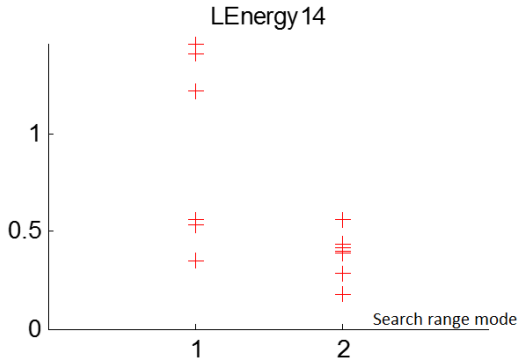


Figure 4.9 – Correlation analysis. The X-axis represents the selected mode (1=motion range (64), and 2=unrestricted motion range), the Y-axis represents the feature values (energy ratio of two laplacian subbands 1 and 4). The points represent each video sample. Here, motion search range=32 is excluded since it is not selected by any of the video sequences.

that has the majority over the video sequence set is the block size of 16/1. For now, it can be said that the block size parameters are not influenced by the video content in 4K resolution in Rate-Distortion-Complexity analysis. Please note that this conclusion may change if parameters dependency is taken into consideration.

4.4.2 Study the content influence with respect to the motion range parameter using the HM encoder

Using the HM encoder, the model shows that there is content influence with respect to the motion range. A search range of 32 and unrestricted search are the most frequent modes. According to the proposed model, it is possible to start the correlation step and categorize the video sequences into two clusters; one for motion range 32 and another for unrestricted search mode. The explanation of seeing unrestricted search mode appear is that the "fast mode" parameter of the encoder is enabled. Here, after applying the correlation step, the characteristics that are correlated on 10% confidence interval using the non-linearized model are shown in the Figure 4.10.
These features are classified using the classification tree algorithm with "information gain" as attribute selection criterion and exhaustive search for optimal split configurations, Orange software is used [248]. The aim here is to make two clusters; one for those contents that choose a motion range size of 32 and the second one for unrestricted search mode. Figure 4.11 shows the results of the analysis steps and the prune values of chosen features for the video sequence set (35 UHD) that will be used in future prediction. These features are:

— Chrominance (B) information (CBI),
— Kurtosis ratio extracted from Laplacian based features (laplacian pyramid level 4 over level 5) (LKurt45), and
— Cross-correlation (pattern=64, sub-image=128, ROI=Frame) using p=4 in Minkowski sum (CC_ALL_64_128_4).

The alternative choice is to learn a supervised learning model like SVM or classification tree. Cost-SVM with RBF kernel and automatic parameters search configuration is used. Cross-validation technique (3 folds) is applied to learn two algorithms. Table 4.3 shows the classification accuracy (CA) and area under ROC curve (AUC), Orange software is used [248].

Table 4.3 – Learning results for predicting motion search range using the HM encoder and QP=32

Method	CA	AUC
SVM	0.884	0.5788
Classification Tree	0.8611	0.6348

4.5 Performance Evaluation

This section shows the model performance evaluation and what the gain and the loss. Here, the comparison between the analysis steps results, which are called optimal configurations, and results from using the prediction model, which are called the predicted configurations, are shown. These two datasets are compared with other standalone configuration. The following steps are followed:

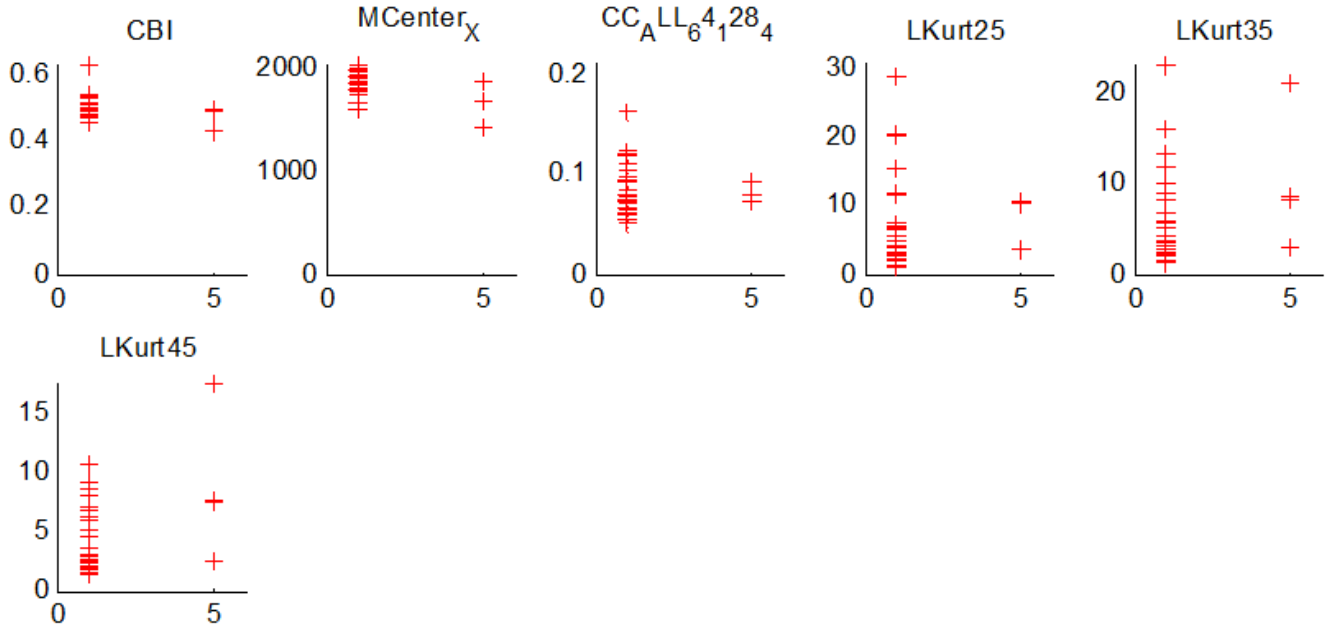


Figure 4.10 – Correlated features with motion range value using HM encoder. X-axis is the size of motion range (1:32, 2:64, 3:96, 4:128, and 5:full) and Y-axis is the feature value

.

— Get the bitrate (Kbps), complexity (sec), and distortion (MSE) on average when the 35 UHD video sequences are encoded with optimal, predicted, and standalone configurations.
— Show the gain or loss factor when using the optimal and predicted results relative to standalone results.

Remember that the prediction value of the parameter is a result of mode function. That means that there are regions in the analysis space that are not optimized. This is one of the model limitations and that explains why there are losses.

4.5.1 Evaluation of predicting motion range using HEVC HM encoder and QP=32

Table 4.4 shows the result of the predicted values against the optimal values (derived with our model) and against other configurations. For instance, if the predicted value is used rather than configuration 2 that uses motion range of 64, the loss is very small in quality and bitrate (101.73 Kbps), but the gain is 17% (213017.1 sec) in terms of complexity and 36% against configuration 4. Note that the absolute value of the complexity is the sum difference of two configurations and is not the average.

4.5.2 Evaluation of using fixed block size using HM encoder and QP=32

In this case, the analysis steps do not find links between content features and block size parameter value because the block size of 16/1 (2/2) (Block size/Depth (Transform size min/max)) is the result of *mode* function for all sequences. Table 4.5 shows the evaluation of choosing 16/1 block size mode against the others. Using this parameter value will increase the bitrate and reduce complexity. Using different distortion measurements, as mentioned in [242], may change these results. Table 4.5 shows that MS-SSIM as a quality measurement gives different results than the usual PSNR. Other quality metrics like PSNR-HVS, PSNR-HVS-M, VIFP, and VQM can be tested in future work.

4.5.3 Features Complexity

The previous results do not include the features complexity. It is clear that the complexity of the proposed model is feature dependent. Section 4.4.2 shows that there are three features that will be used for motion search range prediction in HM encoder. All content features are implemented in MATLAB, and a lot of function implementations are not optimized.

Following the classification tree results, Figure 4.11, if the three features that are used in predicting motion search range in HM encoder are used, i.e. navigate through all tree depths/levels, the added complexity is 22% with respect to the encoder complexity using predicted configuration. As shown in Figure 4.12a, with this added complexity, the model still having gain and loss against other configurations. On the other hand, if the classification tree is optimized with depth of two instead of three to reduce the complexity introduced by using the feature in the third depth, the

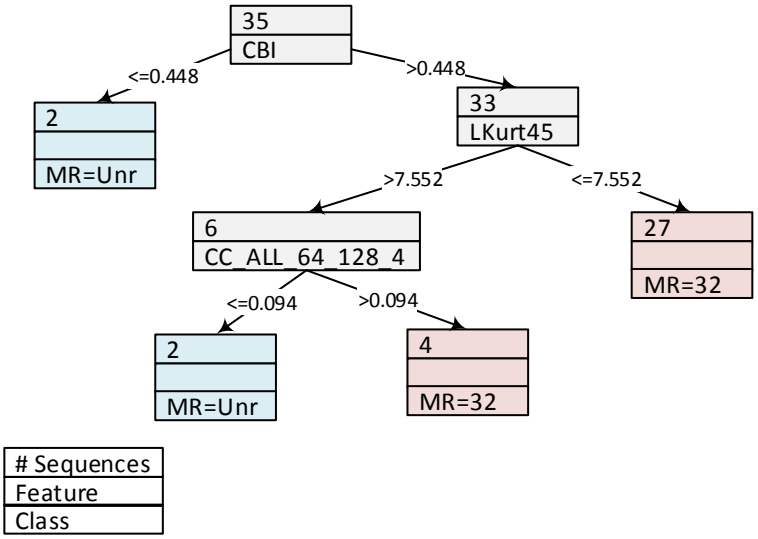


Figure 4.11 – Selected features to predict motion range using HM encoder

Table 4.4 – Performance results averaged over 35 sequences of predicting motion range using HEVC HM encoder and QP=32.

Configs.	Component	Factor compared to		Absolute value		
		Optimal	Predicted	Average	Gain/Loss	Unit
Config (1) - MR=32	Rate	1.00	1.00	7844.825	12.00	Kbps
	Distortion (MSE)	0.99	1.00	18.67218	-0.09	
	Complexity	1.00	1.00	35514	-3436.64	sec(s)
Config (2) - MR=64	Rate	0.99	0.99	7731.097	-101.73	Kbps
	Distortion (MSE)	0.99	0.99	18.58514	-0.17	
	Complexity	1.17	1.17	41698.45	213017.07	sec(s)
Config (3) - MR=96	Rate	0.99	0.98	7708.004	-124.82	Kbps
	Distortion (MSE)	0.99	0.99	18.5569	-0.2	
	Complexity	1.25	1.25	44378.98	306835.61	sec(s)
Config (4) - MR=128	Rate	0.98	0.98	7689.099	-143.73	Kbps
	Distortion (MSE)	0.99	0.99	18.5286	-0.23	
	Complexity	1.37	1.36	48520.19	451777.83	sec(s)
Config (5) - MR=unrestricted	Rate	1.30	1.30	10148.44	2315.61	Kbps
	Distortion (MSE)	1.09	1.09	20.53126	1.77	
	Complexity	1.00	1.00	35657.26	-1575.44	sec(s)
Config (Optimal)	Rate	1.00	1.00	7813.052	-19.77	Kbps
	Distortion (MSE)	1.00	1.00	18.78857	0.03	
	Complexity	1.00	1.00	35545.77	-2326.60	sec(s)

overall complexity will be reduced with no noticeable loss. The complexity of the cross correlation feature is high and not considering it in predicting motion search range will reduce the complexity overhead to 3% as shown in Figure 4.12b. One can reduce the calculation time of the feature by targeting spatial or temporal subregions. For instance, calculate the cross-correlation on 1 out of 16 blocks or for each third frame.

4.6 Conclusion

This work extends the state-of-art of optimizing video coding by analysing signal based and perceptual characteristics of video sequences. It discusses the research question listed in Box 4.1. The main contributions are listed in Box 4.3. The calculation time of some features like cross-correlation is time consuming. Hence, calculating the cross-correlation on 1 out of 16 blocks or for each third frame may significantly reduce the calculation time of the feature. This work still limited in terms of content features, enhancing the efficiency/optimization model, using other

Table 4.5 – Performance results averaged over 35 sequences of using block size (16/1) using HEVC HM encoder and QP=32. ($factor > 1$ signifies gain and $factor < 1$ signifies loss)

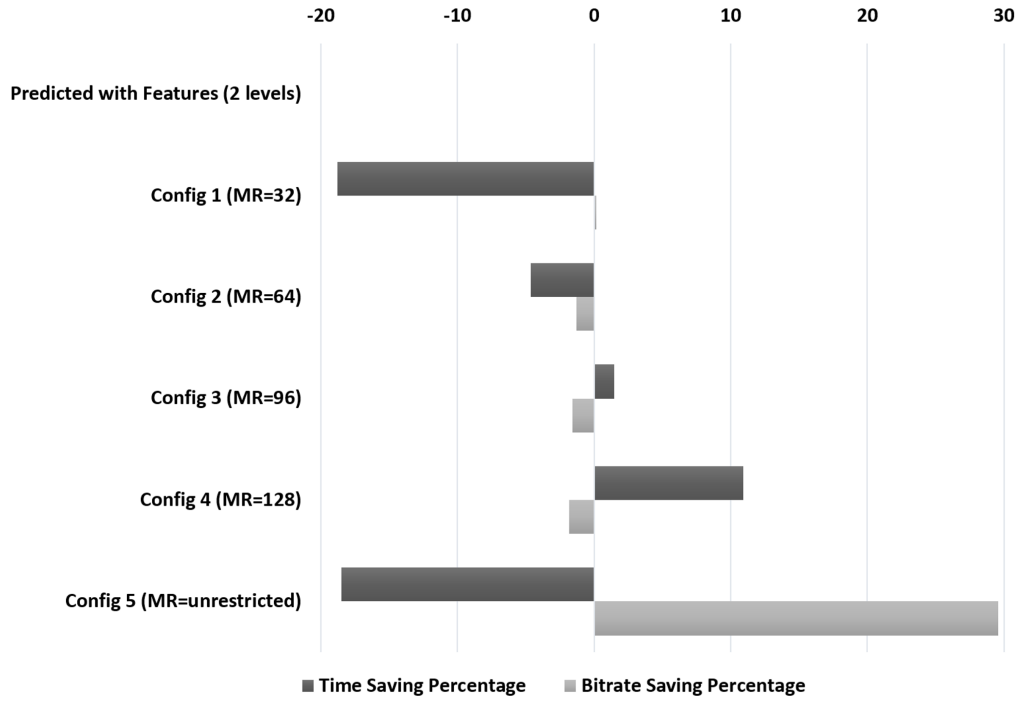
Configs.	bitrate factor	MSE factor	MS-SSIM factor	Time factor
16/1 (2/2)	1.0000	1.0000	1.0000	1.0000
16/2 (2/2)	1.0093	0.9931	0.9997	1.7281
16/2 (2/3)	1.0126	0.9202	1.0031	1.5843
32/2 (2/2)	0.9479	0.9492	1.0031	1.5449
32/2 (2/3)	0.9416	0.8880	1.0062	1.5080
32/3 (2/2)	0.9479	0.9397	1.0032	2.4753
32/3 (2/3)	0.9402	0.8765	1.0063	2.1512
32/3 (2/4)	0.9069	0.8590	1.0075	2.0399
64/3 (2/2)	0.9281	0.9291	1.0038	2.0747
64/3 (2/3)	0.9262	0.8708	1.0069	1.8789
64/3 (2/4)	0.8908	0.8531	1.0081	1.8877
64/4 (2/2)	0.9293	0.9201	1.0039	2.9215
64/4 (2/3)	0.9245	0.8598	1.0070	2.5503
64/4 (2/4)	0.8884	0.8434	1.0082	2.4302
64/4 (2/5)	0.8740	0.8355	1.0088	2.4718

distortion measurements, and setting up subjective experiments to judge the results.

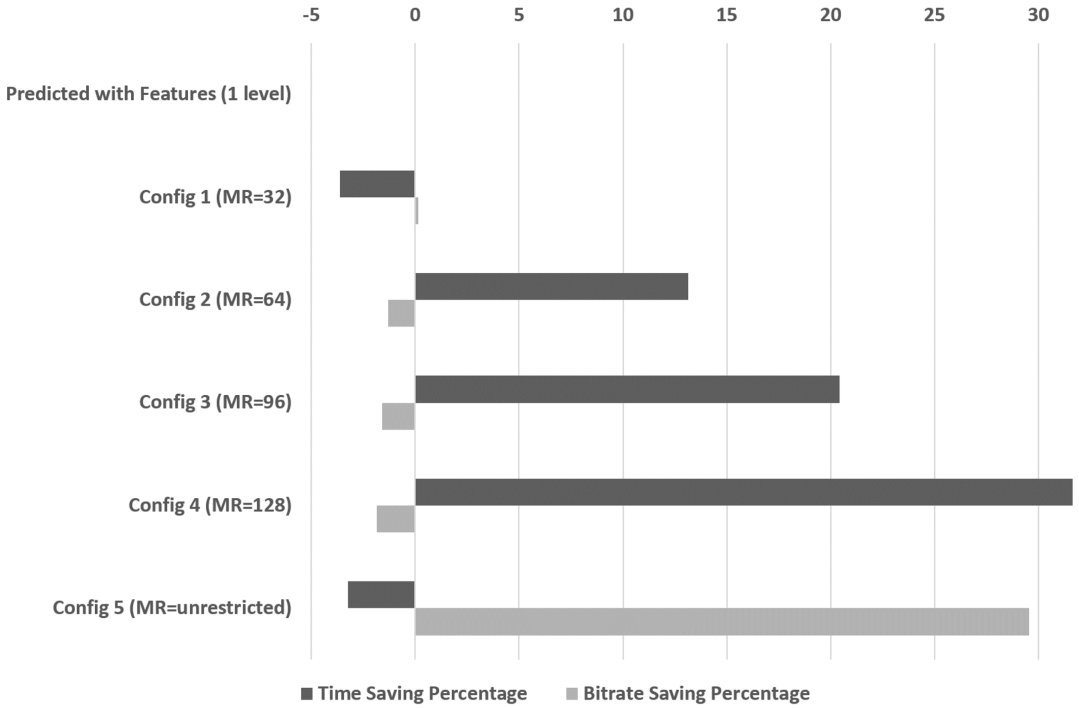
Box 4.3 – Contributions

The primary contribution of this work is the prediction of the encoding parameter values leading to minimum complexity in terms of execution time using the underlying content features. For instance, features like cross correlation, Laplacian-based, chrominance information, and motion intensity features have a high impact in finding the links between the content features and the motion search range parameter in HM encoder. The model trades off rate (R), distortion (D), and complexity (C). For instance, if a video sample is encoded using different configurations and the output videos are in same bitrate and distortion ranges, the configuration that achieve the minimum encoding time will be chosen. If the output videos achieve same complexity and distortion ranges, the configuration of the lowest bitrate will be chosen. The following properties of the proposed model can be noticed:

- The video coding tools are not changed and the candidate prediction modes are not reduced.
- Targeting global quality not local quality since block-to-block and frame-to-frame quality variations yield annoying temporal artifacts.
- The proposed model is a complementary work of other complexity reduction techniques. Suppose that there is N sequences to be encoded and there is time limitations (not necessarily power supply limitation), one possible solution is to distribute the time budget evenly. This solution might not be optimum since some of them are hard to code and some are not. Therefore, one thing to do is to map the predicted parameters values of each sequence into available budget and then use one of up-to-date algorithms of complexity reduction such as [45].



(a) Time and bitrate saving percentage (gain and loss) with added complexity using 3 features. For instance, using predicted values against Configurations 3 and 4, there is gain in complexity and loss in bitrate.



(b) Time and bitrate saving percentage (gain and loss) with added complexity using 2 features. For instance, using predicted values against Configurations 2, 3, and 4, there is gain in complexity and loss in bitrate.

Figure 4.12 – Time and bitrate saving percentage (X-axis: gain and loss) of using predicted configurations with added complexity relative to the standalone configurations (Y-axis) using (a) 3 features and (b) using 2 features

Content-aware Multiple Description Coding

5

High-order temporal-based MDC scheme

5.1 Introduction

Video applications became popular and the videos might be sent via error-prone channels. The decoded video quality might not be satisfying if one or more packets are lost. The main goal of video coding like high efficiency video coding (HEVC) [14] is to minimize the coding distortion for a target bitrate. This requires a complex prediction process to remove the redundant information in the video signal [15]. As a result, the error resilience in HEVC is decreased compared to H.264/AVC due to the increase of temporal dependency [24]. Several error resilience techniques are introduced in the literature [26–28]. Layer Coding (LC) and Multiple Description Coding (MDC) are both efficient in terms of error resilience. In LC, if the base layer is lost or corrupted and despite the presence of enhancement layers, the output video sequence will be degraded seriously. To mitigate this problem, different solutions might be applied here. One of them is to protect the base layer using forward error correction (FEC). This is useful in packet corruption with specific number of errors. Another solution is to retransmit the lost packet when feedback channels are available. The best solution is to use a hybrid scheme. Nevertheless, LC may not be convenient to real-time applications so MDC is a promising solution to deal with these drawbacks of the LC. In MDC, the video sequence is encoded into two or more different bit streams called descriptions. One of the most important design principles of MDC is that each description has to deliver videos with acceptable quality even if it is the only description received by the decoder and the highest quality will be achieved if all descriptions are received. A comprehensive review of multiple description coding can be found in [107–109]. It was shown that the multiple description coding is an effective and promising technique for error resilience for several reasons. First, it is suitable for real-time applications since feedback is not required which simplifies the network design. Second, it performs better than other error resilience approaches in high error loss rates [110,111]. In this chapter, a temporal domain multiple description coding is studied.
Each MDC scheme defines two processes, the first is how to generate the descriptions at the encoder side and how to combine them at the decoder side. The second process is how to do the error concealment when one packet, or more, of a description is lost. Some schemes introduce side information to provide additional or redundant information to help the decoder conceal the lost frame. A review of different temporal MDC schemes is discussed in Section 2.4.2. These schemes are not efficient for the following reasons. First, in schemes that do not include any side information, the error propagation will be annoying especially if the intra-period is large and if the sequence has high motion intensity. Second, in schemes that do include side information, the coding efficiency will be decreased and the error propagation will be noticeable, Figure 5.2. Third, these schemes are less efficient in n-MDC ($when\ n > 2$) since the side information is not fully utilized. For instance, in the case of 4-MDC with side information, each frame has one primary data and three redundant data and if the primary data is lost, one of redundant data will be utilized and the remaining two will not. In this chapter, a new scheme is proposed in which the redundant data is represented in a different context and a new weighted average algorithm for error concealment is also introduced in which all the redundant data is utilized if the primary data is lost. The proposed scheme, as discussed in Section 5.3, is characterized by lightweight complexity, standard compatibility, redundant data tuning, and suitability for n-MDC ($n \geq 2$). The proposed scheme along side with other schemes are tested in a subjective experiment , Section 5.4. To sum up, this chapter raises the research questions listed in Box 5.1 and the chapter structure is illustrated in Box 5.2.

Box 5.1 – Research Questions

This chapter aims to answer the following research questions:

— Which content features can be used in order to take advantage of the received redundant representations/descriptions when using $n-$MDC with $n \geq 4$?

— With these features, is the quality of experience (QoE) of the reconstructed video sequence improved?

Box 5.2 – Chapter structure

This chapter is structured as described in Figure 5.1. The generation of the descriptions is illustrated in Section 5.3.1. The extracted features from the descriptions and how the weights are generated for the recovery process is demonstrated in Section 5.3.2. The Experimental setup and the results are presented in Section 5.4. In Section 5.5, we sum up with the conclusion.

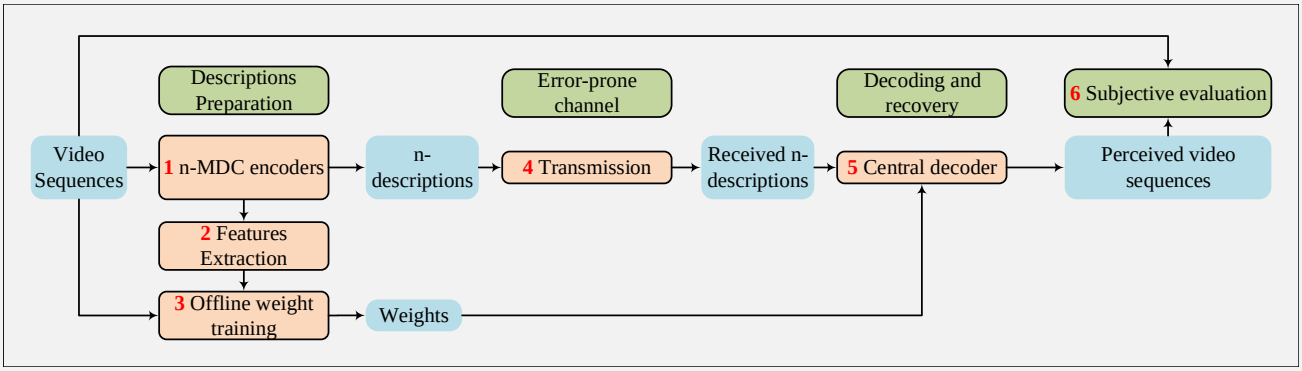


Figure 5.1 – Chapter 5 Structure

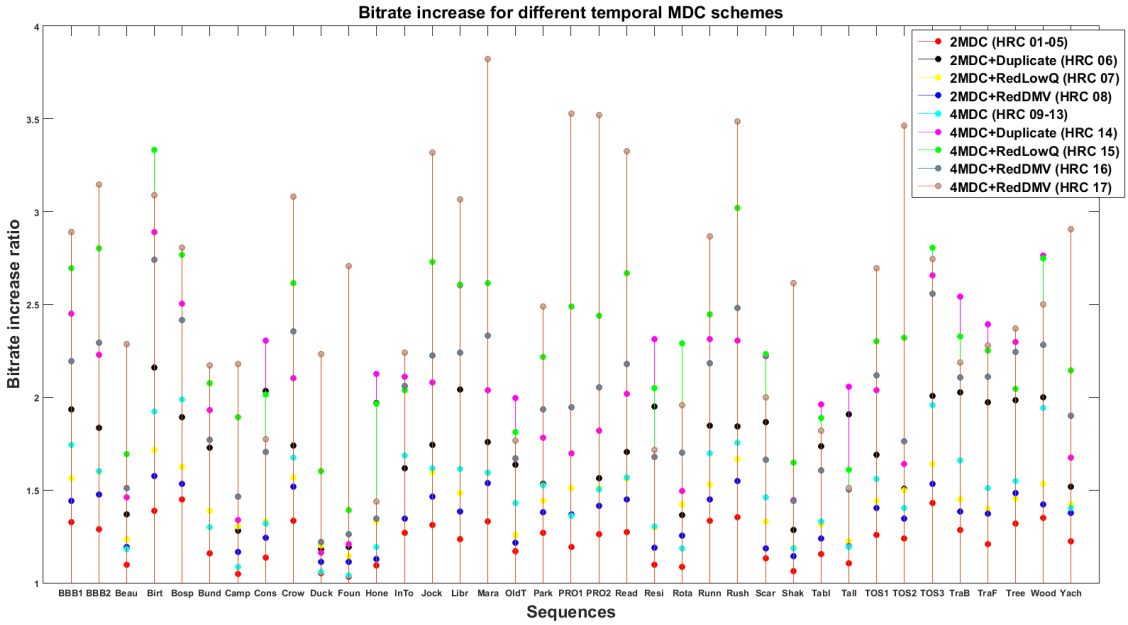


Figure 5.2 – Bitrate increase factor of all HRCs

Table 5.1 – List of hypothetical reference circuit (HRC). Check mark (✓) means that the HRC is subjectively evaluated while times mark (X) not. The dash mark (-) means that the HRC is not applicable.

EC technique	SD	2-MDC	4-MDC
Copy previous frame from the same description	✓(HRC00)	✓(HRC01)	✓(HRC09)
Copy previous frame from another description	-	✓(HRC02)	X(HRC10)
The average between the previous and next frames from another description	-	✓(HRC03)	X(HRC11)
Scale the MV of next frame from another description and use them to conceal the frame	-	X(HRC04)	✓(HRC12)
Average the two concealed frames another description using predicted MVs using Phase correlation algorithm	-	X(HRC05)	✓(HRC13)
Use the duplicate I-frames and MVs	-	✓(HRC06)	X(HRC14)
Use the duplicate degraded frames	-	✓(HRC07)	X(HRC15)
Use the MVs of redundant frames	-	X(HRC08)	✓(HRC16)
Weighted average of the concealed frames using the proposed strategy	-	-	✓(HRC17)

5.2 Problem statement

The 2-MDC scheme that is encoded in low-delay configuration (IPPPP) is used in order to provide a good illustration of different schemes. The descriptions in the 2-MDC are generated as follows; the sequence frame rate is down sampled by two to generate even/odd descriptions and each has its own encoding loop. Table 5.1 shows the list of temporal-MDC schemes that are used in this work. In HRC01/09, the distorted description will continue to decode normally, therefore, the effect of error propagation due to the correlation reduction in one description will be highly noticeable. On the other hand, in HRC02/10, the effect of error propagation will be reduced relative to the HRC01/09 respectively but still is not efficient in sequences that have large motion intensity. HRC03/11 yield a blurred concealed frame which is also not an appropriate technique to use when there are spatial and temporal variations in the sequence, while HRC04/12 work well under the assumption that the motion is completely smooth, which is not the case in most of the video sequences. HRC05/13 use the technique mentioned in [249] which employs the phase correlation motion estimation technique to conceal the lost frame. Though it adds extra complexity to calculate the MVs, it still suffers from blurriness and post-processing for the concealed frame is required. In HRC06/08/14/16, two important information are not included that have a vital impact in the concealed frame, the residual signal and the intra-block modes in inter-frames. The first class of MDC schemes perform well in terms of coding efficiency since no side information is used but it does not provide a satisfactory video quality especially if there are errors in both descriptions and if the video has high motion intensity. In the second class, a trade-off between quality and coding efficiency is achieved by including the MVs and excluding the residual signal and the intra-block modes. While in the third part, the trade-off is achieved by using the coarse frames. Unfortunately, the second and the third parts are not convenient in more than two-description schemes since not all redundant data is utilized. Therefore, these designs principles are taking into consideration in the proposed scheme. Like other schemes except HRC05/13, the complexity is lightweight since the weighted average is applied and the weights are stored in the decoding side. Standard compatibility and tuning of redundant data are preserved too. Finally, the scaling to higher number of descriptions is also considered.

5.3 The proposed MDC scheme

In this section, the proposed MDC scheme is explained. Firstly, the encoding process and the corresponding decoding process are presented, then the error concealment algorithm is elaborated. The 4-MDC is used as an example to elaborate the two processes. In the 4-MDC, the video sequence frame rate is downsampled by 4 and each description contains one fourth of the original sequence.

5.3.1 Encoding and decoding processes

Figure 5.3 depicts the encoding process. Each description contains primary frames which represent the frames of the usual 4-MDC and secondary frames which represent frames of other descriptions and located between two primary frames. The primary frames are encoded using low-delay configuration and the secondary frames are predicted from the previous primary frame in the same description. As a result, each frame of the original sequence is represented with a primary frame and three secondary frames that can be sent in-stream or out as side information. The following rules are applied in the encoding of secondary frames. First, intra blocks in inter-frames are not allowed, which enforces reconstructing the lost frame using only MVs. Second, the residual is set to zero during the encoding process and rate-distortion optimization is used to decide the best splitting in terms of distortion. Such way requires more signalling

in the stream which may increase the bitrate. Many solutions can be applied to tune the amount of redundant data (MVs) either by reducing the sub-pixel accuracy to $\frac{1}{2}$-pixel or integer-pixel accuracy or by using one of the algorithms that prioritize the MVs [250]. At the decoding side, four side decoders are used to decode the descriptions and in case of no error the central decoder assembles the primary frames from the side decoders and send them to the display buffer.

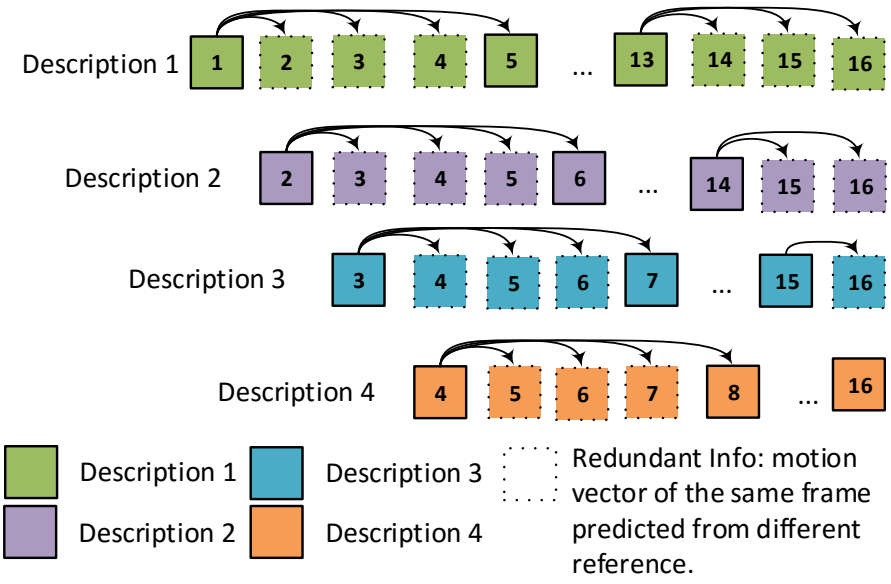


Figure 5.3 – 4-MDC with redundant data/side information. The solid-border square represents the primary frame. The dot-border square represents the redundant representations to be sent in the stream or as a side information. The arrows represent the prediction process; the redundant frames between two primary frames are predicted from the previous primary frame. Only motion vectors are transmitted. The primary frames represent the low-delay configuration.

5.3.2 Error recovery/concealment process

When one primary frame is lost the central decoder initializes the error concealment process. Figure 5.5 depicts the error concealment steps. In the first step, each secondary frame is decoded normally in the side decoders. Then, the lost primary frame is replaced with a weighted average of the three available secondary frames. The weights are applied on the pixel level and they are a function of the temporal distance (d) and the number of pixels (n) in the partition unit (PU) that the pixel belongs to. Number of pixels in the PU ranges from 4096 (64x64) down to 16 (4x4). That yields 13 different amounts of pixels in the PU. Since there are 3 redundant/secondary frames, 2197 (13^3) combinations are counted. In addition, because secondary frames are predicted from previous, frames 3 distances are counted. In total, 6591 (3x2197) combinations of (d_i, n_i) are counted. Figure 5.4 shows an example.
Using temporal distance as one factor on the weight function has already been used in the literature in other contexts [131] and it is believed that the closest frame is not always the best match for the current frame, therefore other factors may have significant influence. In HEVC, the coding unit (CU) can be split using one of the eight supported PU modes. For more details and applied constraints, please refer to [15]. In the proposed EC, for each CU, at most 3 different splitting trees are available that are optimized in terms of distortion and the authors believe that it may have an impact reducing the overall distortion and error propagation in the concealment process.

In order to train the weights, data samples are collected from video sequences. Each sample (each pixel in a primary frame) has tenth values $(d_1, d_2, d_3, n_1, n_2, n_3, p_1, p_2, p_3, p_0)$, where d_i represents the temporal distance of the current pixel, n_i represents the number of pixels in the PU that the current pixel belongs to, and p_i represents the pixel value of secondary frames and primary frame respectively. Each combination (d_i, n_i) is considered as a unique condition and enumerated with the parameter $k = [1, \ldots, 6591]$.Then the samples that share the same properties, i.e. the values of d_i and n_i, are grouped and then are split into train and validation sets to train the weights. The training

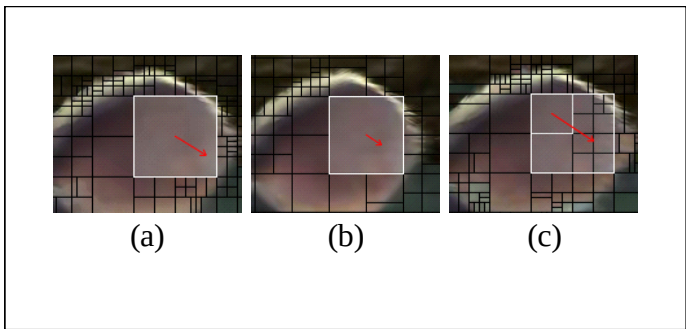


Figure 5.4 – CU partitions for the same CU with different references. (a) reference with distance 3. (b) reference with distance 2. (c) reference with distance 1. The red arrows show the direction of the motion
.

can be expressed as:

$$W^k = \underset{\{w_1^k, w_2^k, w_3^k\}}{\arg\min} ((p_0 - \sum_{i=1}^{3} w_i^k p_i)^2), \text{where, } w_1^k + w_2^k + w_3^k = 1 \tag{5.1}$$

where $W^k = \{w_1^k, w_2^k, w_3^k\}$ are the weights that minimize the error in the validation samples. After that, the weights are stored in the decoder side.

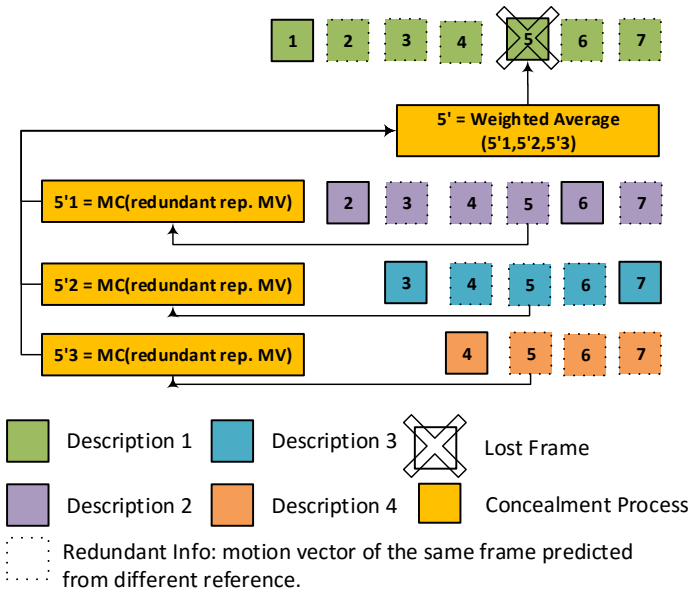


Figure 5.5 – Proposed error concealment procedures

5.4 The subjective experiment

5.4.1 Experimental setup

5.4.1.1 Source video contents

There are different types of content features. Spatial and temporal features are extracted from the luminance frame (Y), chrominance frames (Cb and Cr), from the spatial and the frequency domain. A complete list of the extracted features is listed in Chapter 3. Other features are also extracted to help select the sequences for the experiment; Motion intensity maps [203], encoding bitrate, and camera motion descriptors [204].
The above mentioned features are extracted from 37 UHD video sequences and are normalized linearly between [0,1]. Then, each feature is categorized into 3 classes according to their normalized values. For instance, labels of 1, 2, and

3 will be assigned to the feature's value that lies in the range [0,0.33],]0.33,0.67], and]0.67,1] respectively. Features that are represented with histograms, tree classification using Jensen-Shannon divergence as distance metric [210] is used to cluster the video samples to 3 classes. Finally, the contents that cover all the features classes are selected. The thumbnails of these sequences are shown in Figure 3.24.
The source videos are from different content providers: 4 from Shanghai Jiao Tong University (SJTU) [205], 2 from Ultra Video Group [206], 2 from Sveriges Television AB (SVT) [207], 2 from Blender Foundation [208], and 2 from MediAVentures [209]. The 12 source sequences are in ultra high definition (UHD) with a resolution of 3840x2160 pixels. Figure 3.24 shows the thumbnails of the video sources. The frame rate of the video sequences varies from 25 frames per second (fps) to 120 fps. Each sequence is 10 seconds long.

5.4.1.2 Hypothetical reference circuit (HRC)

In this experiment, the video sequences are encoded as single description (SD), 2 descriptions MDC, and 4 descriptions MDC. All are encoded with QP=32, intra period of 32, and motion search range of 64, 128, and 256 are used for single description, 2-MDC, and 4-MDC respectively. For HRC00/01/02/03/04/05/06/09/10/11/12/13/14, the restricted low-delay configuration is used (GOPSize=1), i.e. only the previous frame is used for prediction. For HRC07/08, the (IPP) GOP structure is used, while in HRC15/16/17, the (IPPPP) GOP structure is used, Figure 5.3. The same error pattern is inserted to all generated videos. The 34^{th}, 50^{th}, 162^{nd}, and 178^{th} NALUnits are dropped and concealed with different error concealment techniques as shown in Table 5.1. In this experiment, each NALUnit represents one frame. For encoding, HM12.1 is used while for the decoding processes, the robust decoder [251] is used and has been adapted to call the appropriate error concealment strategy. A total of 12x18=216 processed video sequences (PVS) are generated. Since this number is large for a subjective experiment, not all of them are used. Each error concealment technique is applied to either 2-MDC or 4-MDC or both as shown in Table 5.1. To sum up 12x11=132 PVS are used in the subjective experiment. Indeed, the redundancy overhead is increasing when MDC is used. In the above-mentioned HRCs, the redundancy overhead is varied from a factor of 1.1 to 1.7 in terms of bitrate in 2-MDC and from a factor of 1.5 to 3.5 in 4-MDC relative to SDC. One possible observation that might be obtained from the experiment is that when comparing two HRCs that are varied in the redundancy overhead, the HRC with lower overhead may have a better quality than the HRC of higher overhead. Output samples for Source#12 is shown for Figure 5.6.

5.4.1.3 Testing conditions

Since all processed videos were affected by error insertions, the pair comparison (PC) method from ITU-T Rec. P.910 [252] was selected to obtain the subjective scale of the experiment. Not all HRCs are involved in the experiment to reduce the number of pairs, as mentioned in the previous subsection. The optimized square design (OSD) methodology was selected to reduce the number of pairs [253] in which the ranking of the stimuli in the test is known based on pre-test results or prior knowledge. In this experiment, the 3x4 rectangular matrix was selected for 11 HRCs and the ranking of the stimuli is defined by the authors' prior knowledge as shown in the R matrix: Where the matrix on the left represents the rank of the stimuli and the matrix on the right represents the corresponding HRC. The 12^{th} cell of the matrix is filled with a repetition of the proposed error concealment strategy (HRC17). Due to OSD, the number of pairs is reduced from 11*10/2=55 to 27 pairs for each content, thus 27x12=324 in total. Unfortunately, this number of pairs is still large. In order to reduce this number, the pairs that have very close quality, ((17,7),(16,6), and (2,0)) are viewed for each observer, and the other pairs are randomly and equally distributed between the observers. Note that the pair (HRC17,HRC17) is not considered in the experiment.

5.4.1.4 Subjective assessment

For each pair, the two stimuli are viewed one after another. The replay function was supported. The observer is asked for his preference for each pair in a forced choice manner. A playlist for each observer is prepared taking into consideration that the pairs that belong to the same content are not viewed consecutively, orders of the pairs are random, and the temporal order of the pairs is also switched between the observers. The viewing distance was 1.5 times the height of the screen. The experiment was explained to the observers using a training session prior to the test session. 4 pairs are selected from the PVSs for the training session without any explicit or implicit instruction on how to choose the preference. The test duration is about 75 minutes including training and breaks. All sequences are viewed at 60 fps therefore the video of 25 or 30 frame rate are up sampled by 2, i.e. the frames are duplicated. The PVS are displayed using 3840x2140 native screen resolution with 60 fps. The screen brand is Grundig FINEARTS 55 FLX 9490 SL with a 55-inch diagonal. The ITU Recommendations BT.709-5 [254] and BT.500-13 [255] are followed to adjust the screen colour and brightness and to set up the testing room respectively. 46 non-expert observers participated in the experiment, 22 males and 24 females and the age average is 24 (18 to 38). The pairs ((17,7),(16,6), and (2,0)) are evaluated by the 46 observers and other pairs are evaluated with 11 or 12 observers. A vision check is performed before the experiment using far and colour vision tests. Any observers with normal or corrected to normal visual acuity are allowed to do the experiment.

Figure 5.6 – Output samples for Source#12

$$R = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 10 & 11 & 12 & 5 \\ 9 & 8 & 7 & 6 \end{bmatrix} \Rightarrow \begin{bmatrix} HRC09 & HRC12 & HRC03 & HRC01 \\ HRC07 & HRC17 & HRC17 & HRC13 \\ HRC16 & HRC06 & HRC02 & HRC00 \end{bmatrix}$$

5.4.2 Experimental results and discussion

In this subsection, subjective data is analysed in terms of pair comparison of raw data using Barnard's exact test [256].Fisher's exact test and Barnard's exact test are both statistical exact/significance test of contingency tables. For 2x2 contingency tables, Barnard's exact test is claimed to be more powerful than Fisher's exact test. Its powerfulness came from its unconditional rule to calculate the p-value. Suppose that $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, is the Barnard's contingency table where, for each pair (A, B), a and d are equal to the number of observers that prefer A rather than B, and b and c are equal to the number of observers that prefer B rather than A. The M matrix is the input of the Barnard's test and the output is the p-value that is calculated on the 95% confidence interval.
Table 5.2 shows the results of applying the Barnard's test for each pair per content. It shows two types of analysis. First, the significant difference for each pair per content is calculated. The first and the second columns represent the pairs while columns labelled 1 to 12 represent the content. Second, the significant difference on the pair level is calculated using two methods. The first method counts the number of sequences that have significant difference for two HRCs and is summed up either in the "#⟵" or in the "#⟶". It also counts the number of sequences that do not have significant difference for two HRCs which is recorded in the "#No" column. The significant difference between the new pair ((#⟵ or #⟶), #No) is calculated and represented in "Sig." column. The second method is to sum the votes for each pair across the video sequences and to calculate the significance of the difference between any two pairs. The result is represented in "Total Sig." column. For instance, the first row which represents the pairs that belong to HRC12 and HRC09. HRC12 significantly differs from HRC09 in 7 sequences and HRC09 significantly differs from HRC12 in one sequence while there is no significant difference in 4 sequences. In this pair we cannot apply the Barnard's test on the pair level using the first method because HRC09 is significantly preferred in one sequence, while the second method shows that there is a significance preference for HRC12. The error concealment strategy preference for one video content is different for another video content and this is clear in different pairs. An important

Table 5.2 – Barnard's exact test between two pairs per content

SRC/HRCs		1	2	3	4	5	6	7	8	9	10	11	12	#⟵	#⟶	#No	Sig.	Total sig.
12	9	⟵	-	-	⟶	⟵	⟵	⟵	⟵	-	-	⟵	⟵	7	1	4	×	⟵
3	9	⟵	-	⟵	-	⟵	-	-	⟵	⟵	⟵	-	⟵	7	0	5	-	⟵
3	12	-	⟶	⟵	⟵	⟶	-	⟶	⟵	-	⟵	-	-	4	3	5	×	⟵
1	9	⟵	⟶	⟵	-	-	⟵	-	⟵	-	-	⟵	-	5	1	6	×	⟵
1	12	-	⟶	⟵	-	-	⟶	⟶	⟶	-	⟶	⟶	-	1	6	5	×	⟶
1	3	⟶	⟶	⟶	⟶	⟶	⟶	-	⟶	⟶	⟶	-	⟶	0	10	2	⟶	⟶
13	1	⟵	⟵	⟶	⟵	⟵	-	⟵	-	⟵	-	⟵	-	7	1	4	×	⟵
0	1	⟵	⟵	-	⟵	⟵	-	⟵	⟵	⟵	⟵	⟵	-	9	0	3	⟵	⟵
0	13	-	-	⟵	-	⟵	⟵	-	⟵	-	⟵	⟵	-	6	0	6	-	⟵
2	3	⟶	-	-	⟶	-	-	-	⟶	-	-	-	-	0	3	9	-	⟶
2	0	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶	⟶	0	12	0	⟶	⟶
6	12	⟵	⟵	-	-	-	-	-	⟵	⟵	-	-	⟵	5	0	7	-	⟵
6	0	-	-	⟶	⟶	-	⟶	-	-	-	⟶	-	-	0	4	8	-	⟶
6	2	⟵	⟵	⟶	⟶	-	-	-	⟵	⟵	-	-	⟵	5	2	5	×	⟵
16	9	⟵	⟵	-	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	11	0	1	⟵	⟵
16	0	⟵	⟵	-	⟶	⟵	-	⟵	-	⟵	-	⟵	-	6	1	5	×	⟵
16	2	⟵	⟵	⟶	-	⟵	⟵	⟵	-	⟵	-	⟵	⟵	8	1	3	×	⟵
16	6	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	12	0	0	⟵	⟵
7	9	⟵	⟵	⟵	⟵	⟵	⟵	⟵	-	⟵	⟶	⟵	⟵	10	1	1	×	⟵
7	13	⟵	⟵	-	-	⟵	⟵	-	⟶	⟵	⟶	⟵	⟵	7	2	3	×	⟵
7	16	⟶	-	⟵	-	⟵	-	-	⟶	-	-	⟶	-	2	3	7	×	-
17	12	⟵	⟵	⟵	-	⟵	⟵	⟵	⟵	⟵	⟵	⟵	⟵	11	0	1	⟵	⟵
17	3	⟵	⟵	-	⟵	⟵	⟵	⟵	-	⟵	⟵	⟵	⟵	10	0	2	⟵	⟵
17	13	-	⟵	⟵	⟵	-	⟵	⟵	⟵	⟵	⟵	⟵	⟵	10	0	2	⟵	⟵
17	2	⟵	⟵	-	⟵	-	⟵	⟵	⟵	⟵	⟵	-	⟵	9	0	3	⟵	⟵
17	6	⟵	⟵	⟵	-	⟵	⟵	⟵	-	⟵	⟵	⟵	⟵	10	0	2	⟵	⟵
17	7	⟵	⟵	⟵	⟵	-	-	⟵	⟵	⟵	⟵	⟵	-	9	0	3	⟵	⟵

question here is raised "What is the impact of involving video properties to select the appropriate EC strategy to better enhance the QoE?" One of the Barnard's test intuitive assumption is that if HRC17 is significantly different from HRC07 and HRC07 is significantly different from HRC09, we can say that the HRC17 is significantly superior compared to HRC09. Using this property, we can conclude that the proposed algorithm is significantly different from other HRCs except HRC 16 since there is no evidence of preference.

5.5 Conclusion

In this chapter, two main contributions are introduced and listed in Box 5.3. The proposed scheme is significantly preferred to the other temporal MDC schemes, but the number of tested sequences is too small for generalization. In addition, this chapter also highlights the fact that the preference of the MDC scheme depends on the video content itself. Hence, more investigations are required to identify these content features.

Box 5.3 – Contributions

— A new temporal MDC scheme which is characterized by standard compatibility, redundancy tuning, lightweight complexity, and suitability for n-MDC schemes. This scheme includes the process of generating descriptions and the process of reconstructing video sequences when primary data is lost. Coding unit splitting and the temporal distance properties are used to train a weighting coefficient to reconstruct the lost primary frame from the redundant frames.

— The subjective experiment that shows the preference of the proposed scheme against other MDC schemes is introduced.

6

Content-aware adaptive multiple description scheme

6.1 Introduction

Sending video streams over error-prone channels may yield, depending on the error loss rate, unsatisfying video quality. Moreover, the need for providing error resilience tools is increased due to the prediction complexity of video coding. For instance, the latest video coding standard, high efficiency video coding (HEVC) [14], achieves 50% bitrate reduction at the same quality relative to H.264/AVC. This achievement came from new/improved coding tools, especially in the motion compensation. As a result, the error resilience in HEVC is decreased compared to H.264/AVC due to the increase of temporal dependency [24]. The ways to provide error-resilient streams are investigated in [26–28]. Multiple description coding (MDC) is introduced to be one of the promising tools to maximize the quality of experience (QoE) in the presence of network errors. A comprehensive review of multiple description coding can be found in [107–109]. MDC simplifies the network design since a feedback channel is not required therefore it is suitable for real-time applications. In addition, it performs better than other error resilience approaches at high error loss rates [110, 111]. In MDC, the video sequence is encoded into two or more different bit streams called descriptions. One of the most important design principles of MDC is that each description has to deliver videos with acceptable quality even if it is the only description received by the decoder and the highest quality will be achieved if all descriptions are received. In this chapter, a temporal domain multiple description coding is studied.
Different temporal MDC schemes are introduced in the literature. Apostolopoulos in [121] reviewed different schemes. All these schemes share the same encoding and decoding processes but different error concealment strategies. Suppose that an even frame is lost. Copying the previous even frame from the distorted description to replace the lost even frame in the buffer, copying the previous odd frame from the undistorted description, averaging the previous and the next odd frames from the undistorted description, scaling the MVs of the next odd frame from the odd description by 1/2 and use them to do the motion compensation process using the previous odd frame of the undistorted description, namely inplaceMC, and generating the MVs using the available previous and next odd frames, namely MCinterp, are the error concealment strategies that are reviewed in [121]. Different approaches [131, 138, 257] introduce side information. This side information can be a duplicate of MVs of each frame in the description or a duplicate of I-frames. In [138], a different scheme is proposed in which each description contains alternating even/odd frames and odd frames in even description are containing the motion information only predicted from the previous even frame. In [257], Chapter 5, we have introduced a new MDC scheme which is suitable for high order MDC. The lost frame is concealed using weighted averaging of correctly received redundant frames. While Radulovic et al. [131], suggested that each description alternatively contains a fine quantization frame (even) followed by coarse quantization frame (odd). Table 5.1 shows the list of the above-mentioned MDC schemes and the corresponding hypothetical reference circuits (HRCs) that are used in the subjective evaluation in [257].
Moreover, content types and their underlying characteristics have a high impact in video coding. In [5], the authors show the impact of content types in setting up a subjective experiment. In [7], content characteristics are used in improving objective measures of video quality. While in [9], it was shown that the video content influences the video encoding. The following scenario will illustrate the motivation to study the content influence in MDC. Let us consider the quality of a slide show video sequence that is encoded with SDC, 2-MDC, and 4-MDC and sent through an error-prone channel. If the frame loss hits a frame that has the same content as the previous frame, copying previous frame

error concealment strategy of SDC is perfect. Now, consider that the frame loss hits a scene cut (new slide content), the copying strategy error concealment will yield unsatisfying output video. In this situation a side information from the other descriptions will be useful. In another example, let us consider the quality of a high motion-intensity video sequence. If the frame loss hits non-key frames, error concealment other than copying strategy is needed to conceal the frame and it depends on spatio-temporal variation of the content. This chapter address the following question, "is it better to use SDC, 2-MDC, or 4-MDC?".
The subjective experiment [257], Chapter 5, investigated the above mentioned MDC schemes in the presence of frame losses. In this chapter, the results of this subjective experiment are more investigated and analysed to study the influence of considering content features to help recognize the preferred scheme to be used for the transmission.
To sum up, this chapter raises the research questions listed in Box 6.1 and the chapter structure is illustrated in Box 6.2.

Box 6.1 – Research Questions

This chapter aims to answer the following research questions:
- Which content features would help to build an adaptive MDC scheme?
  In addition to that, the following secondary research question is investigated too:
  + Can we trade-off between quality and bitrate in MDC schemes by not always using a specific MDC scheme? In other words "Is it better to use SDC, 2-MDC, or 4-MDC for a specific content?"

Box 6.2 – Chapter structure

This chapter is structured as described in Figure 6.1. Data preparation part will be demonstrated in Section 6.2 and in Section 6.4. The part of building the adaptive model is illustrated in Section 6.3 and in Section 6.5. The training results will be discussed in Section 6.5.2.

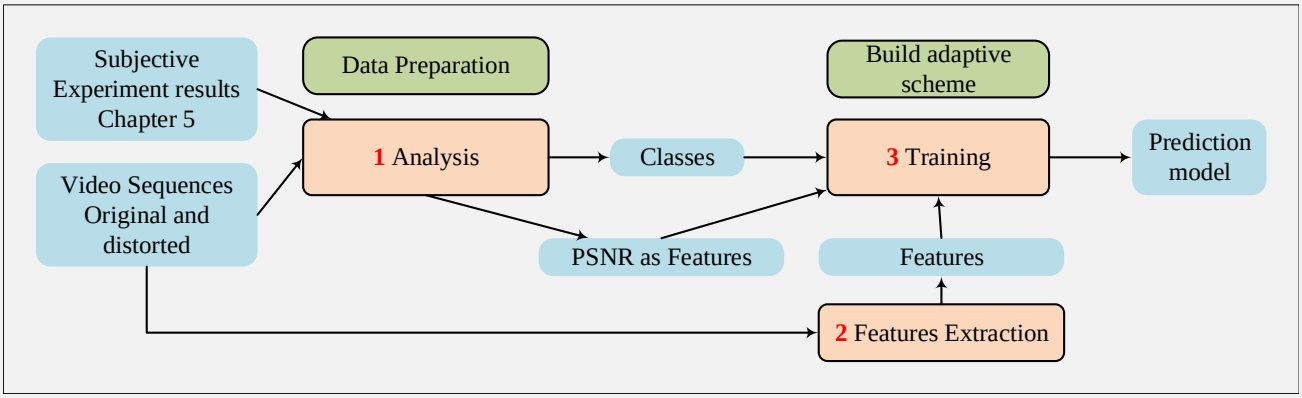


Figure 6.1 – Chapter 6 Structure

6.2 Framework overview

In this section, an overview of the proposed framework is introduced. Figure 6.2 demonstrates the framework stages. Firstly, content features are extracted from 37 video sequences as described in Section 6.3. In order to conduct a subjective experiment, a limited number of sequences needs to be selected. Therefore, a sequence selection process as illustrated in Section 6.3 is followed to reduce the number of video sequences. 12 video sequences are chosen at the end. After that, the 12 sequences are encoded using different SDC and MDC schemes, loss-impaired, decoded, and subjectively evaluated as described in [257] and are summarized in Section 6.4.1. The subjective data is analysed to know the preference scheme of each content. As illustrated in Section 6.4.2, additional analysis to introduce labels/classes and to calculate the quality of service of each scheme, i.e. the PSNR values of affected frames. In this chapter, 5% percentile of PSNR value of affected frames is selected as QoS measure. Finally, the prediction model,

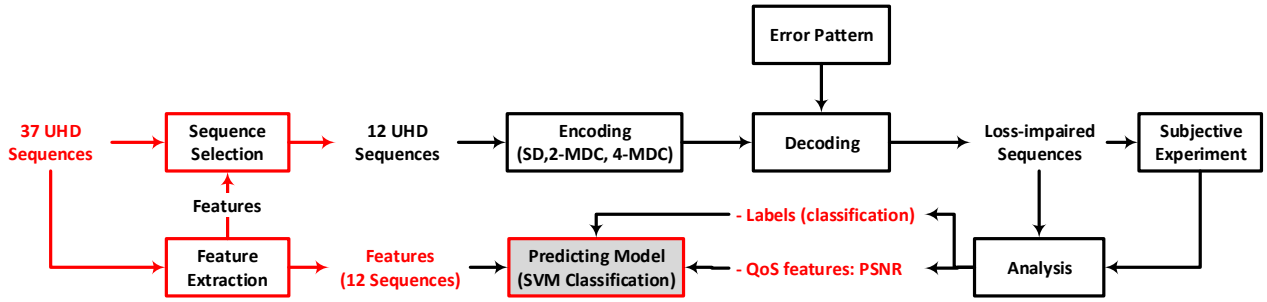


Figure 6.2 – Framework overview. The Black boxes show the work of [257] while the red ones show the contribution of this Chapter.

that is fed with the feature set and the labels, is trained using SVM classification as demonstrated in Section 6.5. The black boxes indicate that they are introduced in Chapter 5, while red boxes indicate the contributions of this chapter.

6.3 Content Features and content selection

As discussed in Section 5.4.1.1, there are different types of content features. Spatial and temporal features are extracted from the luminance frame (Y), chrominance frames (Cb and Cr), from the spatial and the frequency domain. A complete list of the extracted features is listed in Chapter 3. Other features are also extracted to help select the sequences for the experiment. Motion intensity maps [203], encoding bitrate, and camera motion descriptors [204]. The above mentioned features are extracted from 37 UHD video sequences and are normalized linearly between [0,1]. Then, each feature is categorized into 3 classes according to their normalized values. For instance, labels of 1, 2, and 3 will be assigned to the feature's value that lies in the range [0,0.33], (0.33,0.67], and (0.67,1] respectively. Features that are represented with histograms, tree classification using Jensen-Shannon divergence as distance metric [210] is used to cluster the video samples to 3 classes. Finally, the contents that cover all the features classes are selected. The thumbnails of these sequences are shown in Figure 3.24.

6.4 Observations and Problem statement

6.4.1 Overview of the subjective experiment

In this subsection, an overview of the subjective experiment will be introduced. The full description can be found in [257]. 12 out of 37 UHD video sources are selected as described in Section 6.3 and are used in the subjective experiment. They are from different content providers: 4 from Shanghai Jiao Tong University (SJTU) [205], 2 from Ultra Video Group [206], 2 from Sveriges Television AB (SVT) [207], 2 from Blender Foundation [208], and 2 from MediAVentures [209]. Figure 3.24 shows the thumbnails of the video sources. The video sequences are encoded as single description (SD), 2 descriptions MDC, and 4 descriptions MDC. All are encoded with QP=32, intra period of 32. Motion search range of 64, 128, and 256 are used for single description, 2-MDC, and 4-MDC respectively. The same error pattern is inserted to all generated videos. The 34^{th}, 50^{th}, 162^{nd}, and 178^{th} frames are dropped and concealed with different error concealment techniques as shown in Table 5.1. The pair comparison (PC) method from ITU-T Rec. P.910 [252] was selected to compare the MDC schemes subjectively.

6.4.2 Observations

6.4.2.1 Bradley-Terry model

Bradley-Terry model [258] is a linear model that analyses pair comparison preference in order to map their probabilities to scales. Given K stimuli, suppose that the pair (A_i, A_j) are two stimuli, and X_i, X_j are the number of A_i beats A_j and the number of A_j beats A_i respectively. The probability that the observers choose A_i over A_j is $P(X_i > X_j)$ and it is defined as:

$$P(X_i > X_j) \equiv \pi_{ij} = \frac{\pi_i}{\pi_i + \pi_j}, i \neq j \tag{6.1}$$

Where $\pi_i > 0$ and $\sum_{i=1}^{K} \pi_i = 1$. The value that describes a stimulus (A_i) on the scale is calculated as $V_i = log(\pi_i)$. Since the π_i value is less than one, the Bradley-Terry score V_i is a negative value.

In this experiment, the Bradley-Terry (BT) test is used to obtain the HRCs scale for each content [259]. Figure 6.3 shows the results. The scale is offset such that HRC00, which represents the single description coding, is set to zero to easily read the figures. The confidence intervals in the subplots belong to the fitting model and does not represent the observer's confidence . In [260], the author shows a method to calculate the significance between the BT scales. Since the BT scale depends on the goodness of the fitting, the significance results between the HRCs are not necessarily coherent with Barnard's test which is an unconditional test.

It was observed that 4 schemes are selected as preference for the video sequences; the single description coding

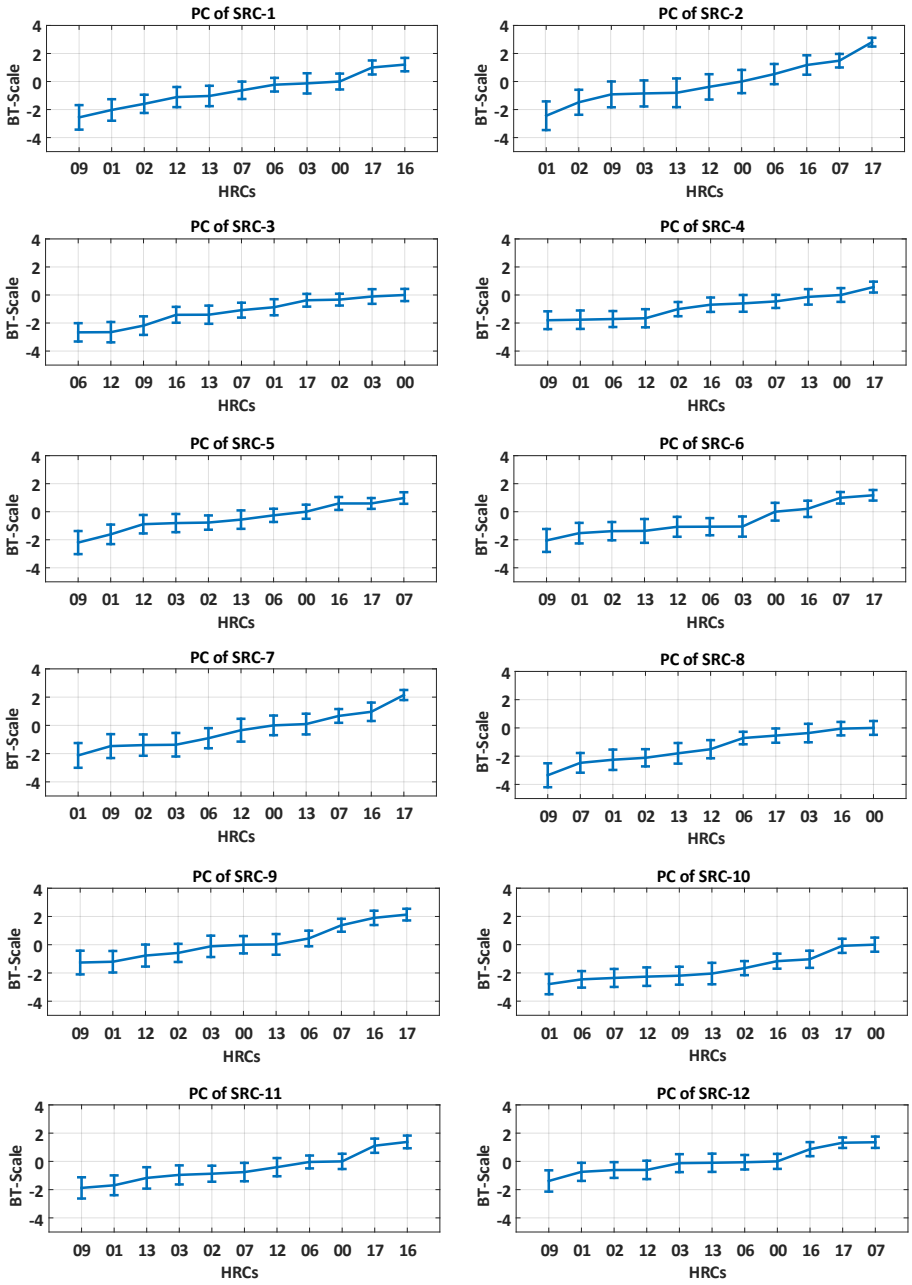


Figure 6.3 – Bradley-Terry Scale for each video content

scheme (HRC00) and three MDC schemes (HRC07/HRC15, HRC08/HRC16, and HRC17) and are labelled with A, B, C, and D respectively. Table 6.1 shows the coding scheme preference label for each content. HRC08 and HRC16 share the same error concealment technique and for the goodness of bitrate utilization, HRC08 is considered. HRC16 will be more effective in high error loss rate channels. The same observation for HRC07 and HRC15. Figure 6.4 shows the bitrate of each HRC with respect to HRC00 (Class A). Bitrate is content dependent, for instance HRC17, which is a 4-MDC scheme, consumes high bitrate and it ranges between 1.5 and 3.5. The bitrate consumption of HRC15 and HRC16 will lie in approximately the same bitrate range of HRC17. Moreover, an important observation that distinguishes the HRCs is observed. We refer to this observation as QoS of using the MDC error concealment strategy and it is measured with PSNR. Figure 6.5 shows the variations of the PSNR values of different HRCs of source 12.

Table 6.1 – Associated label/class for each video source

Source	SRC01	SRC02	SRC03	SRC04	SRC05	SRC06	SRC07	SRC08	SRC09	SRC10	SRC11	SRC12
Label/Class	C	D	A	D	B	D	D	A	D	A	C	B

These variations are content and error concealment technique dependent as can be seen in Section 6.5.

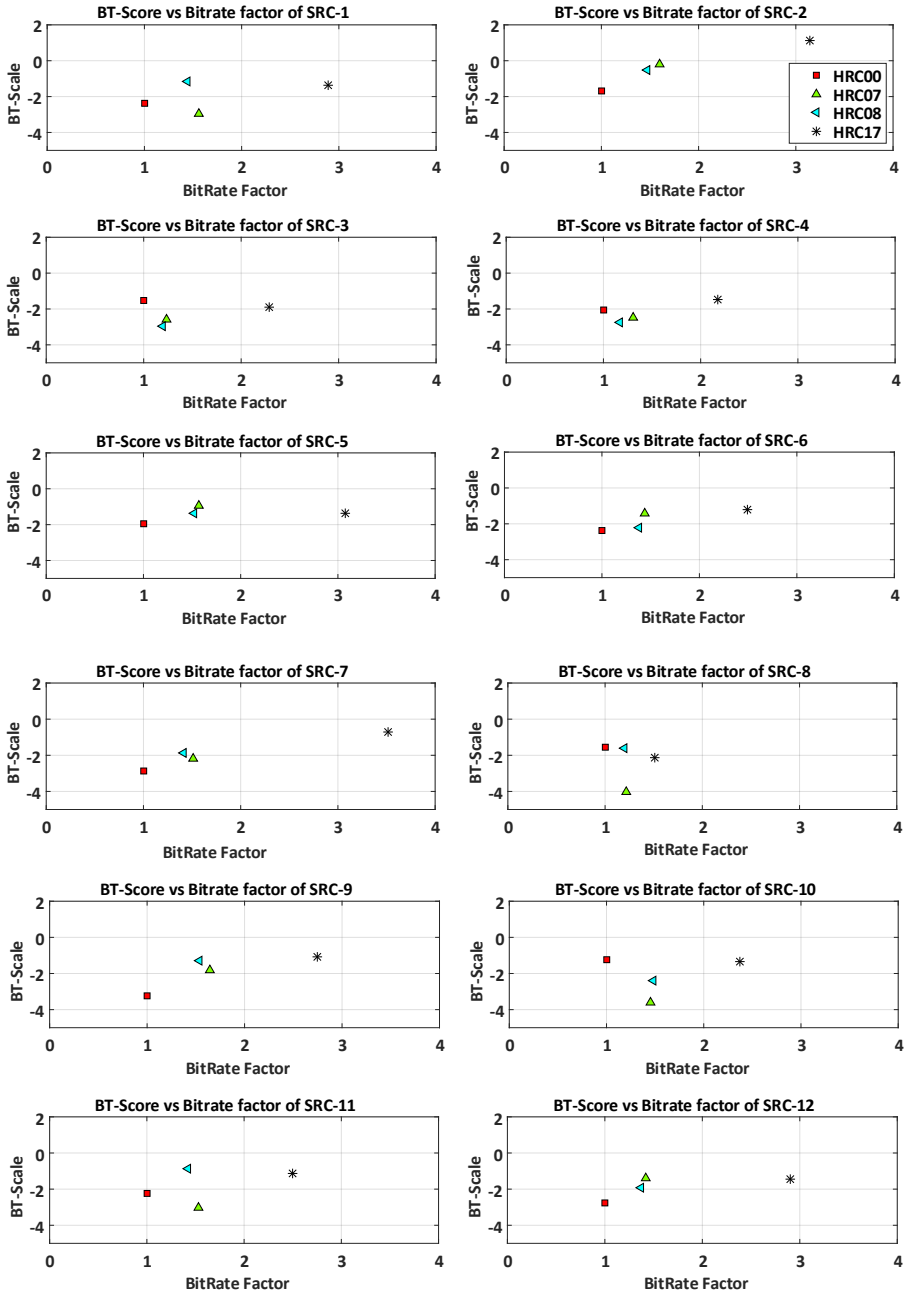


Figure 6.4 – Bradley-Terry Scale vs bitrate increase factor with respect to SDC for each video content

6.4.3 Problem statement

As noticed in the observations section, Figure 6.4, and Figure 6.5, the scheme decision depends on the content and on the error concealment technique that is used in each scheme. Therefore the content features and the effects of using different error concealment techniques, i.e. the PSNR values as QoS measure, are considered to build a prediction model to select the appropriate scheme to be used for the transmission. This experiment uses 12 video sources which may not be sufficient to build a generalized prediction model. Specifically, we analyse the influence of the content

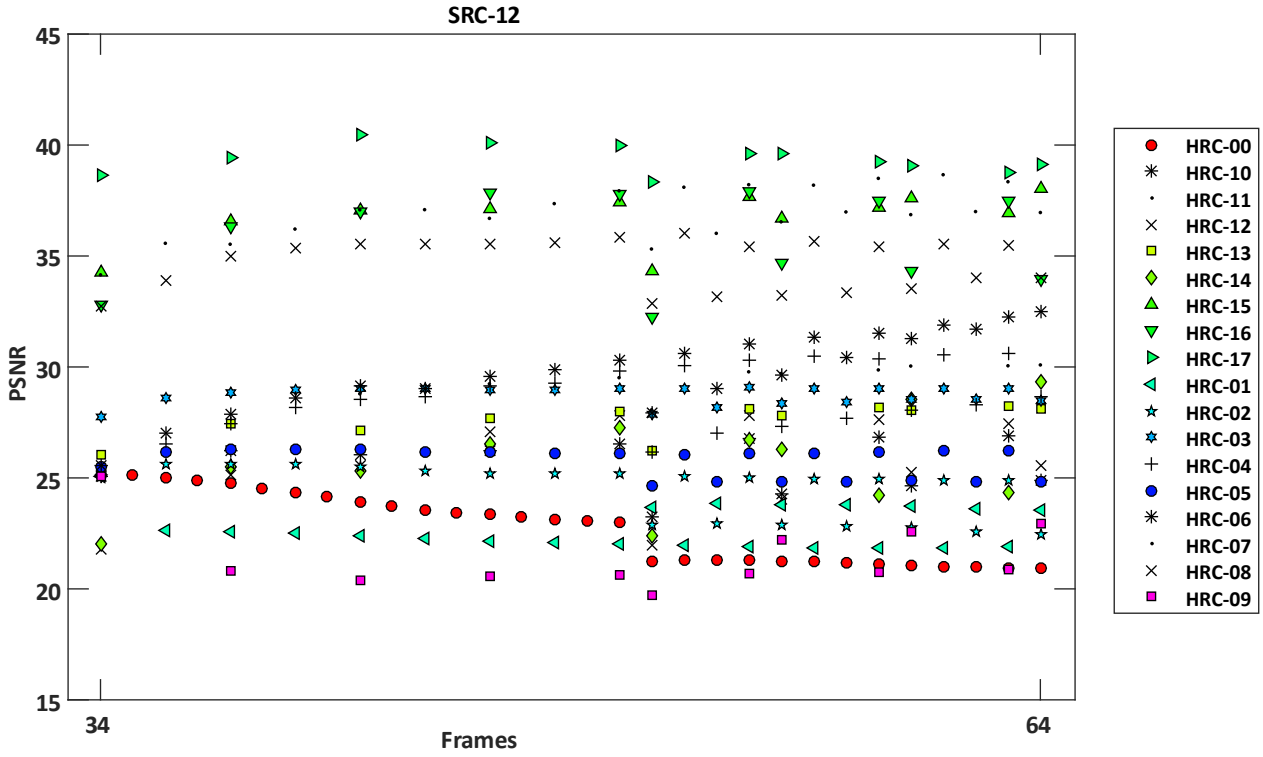


Figure 6.5 – PSNR values of affected frames of SRC 12 (From frame 34 to 64)

features in optimizing the MDC as an error resilience tool to maximize the quality of experience whatever the bitrate is. Trading-off the bitrate and the quality is a recommended future work.

6.5 The prediction model

6.5.1 Model training

The feature set in each training sample represents two types of features: the content features and the 5% percentile value of PSNR values of affected frames for each HRC. Each training sample is labelled with the corresponding HRC, Table 6.1. The main goal of this machine learning process is to highlight the features that have an impact on prediction the suitable transmission scheme. Due to the large number of features, ReliefF features ranking algorithm [261] is applied to rank the features. The top 5 ranked features are selected to train the model. These features are in order: the PSNR values of HRC11 (PSNR11), the entropy ratio between laplacian subband levels 4 and 5 (LEntropy45) and between 3 and 5 (LEntropy35), the PSNR values of HRC03 (PSNR3), and the block-based standard deviation of the contrast of GLCM (GLCMSTDCONT). We found that the HRC11 and HRC03 belong to the same error concealment technique and they are redundant, so the PSNR value of HRC11 is considered. The same observation is found in the entropy ratio of different subbands. In summary, three features are considered in the training model: PSNR11, LEntropy45, and GLCMSTDCONT. SVM training algorithm in Orange software [248] is used to train the model.

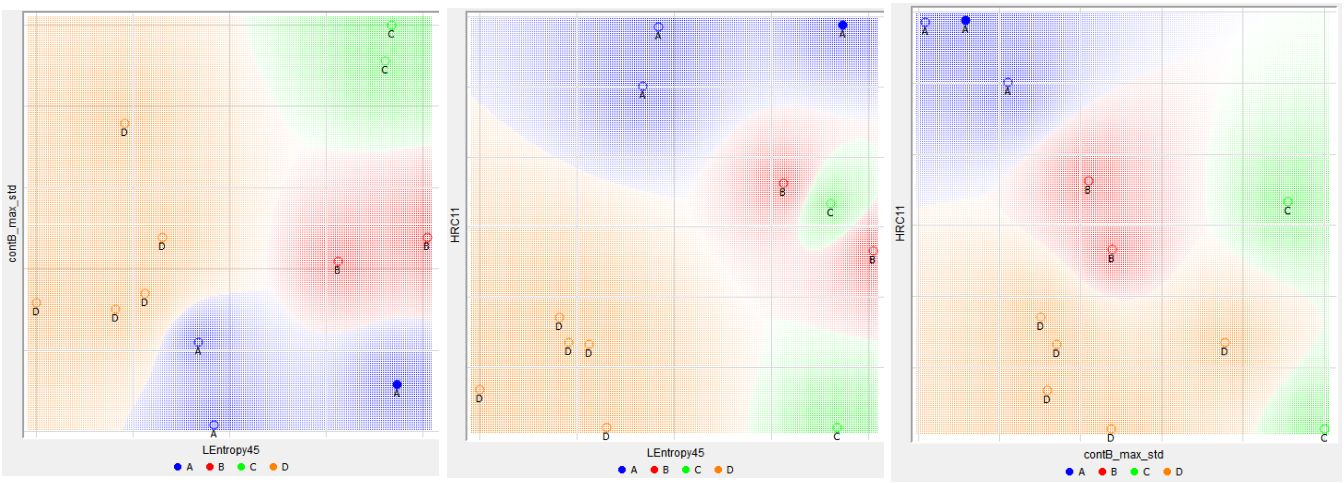
6.5.2 Results

The performance of the above-mentioned model is validated by following the following steps. First, the SVM parameters are trained. Then, the cross validation technique with 4-fold is used to validate the model. Finally, the classification accuracy (CA), area under the curve (AUC) are used to evaluate the prediction model and they are 91.7% and 0.98 respectively. Having low variations on block-based contrast of GLCM along the video means that the video may have low variations in texture and using simple copying algorithm will yield a better quality which is clear in Class A samples. Video that has high variations in contrast is preferred to be concealed with redundant motion vector (Class C). High value of PSNR11 means that the concealed frames are so close to the undistorted frame and hence the error degradation will be small, which is clear in Class A. Table 6.2 shows the confusion matrix of the prediction model. SRC-10 (Treeshade) is misclassified to B instead of A. The LEntropy45 feature's value of SRC-10 is high compared to other sequences in class A and it is so close to class B sequences, blue filled circles in Figure 6.6.

The three sub figures in Figure 6.6 show the separation between classes using any 2 features.

Table 6.2 – Confusion Matrix of the classification

Class	A	B	C	D	
A	2	1	0	0	3
B	0	2	0	0	2
C	0	0	2	0	2
D	0	0	0	5	5
	2	3	2	5	12



(a) GLCMSTDCONT vs. LEntropy45 (b) HRC11 vs. LEntropy45 (c) HRC11 vs. GLCMSTDCONT

Figure 6.6 – The separation between classes using two features

6.6 Conclusion

The main contribution of this chapter is listed in the Box 6.3. This chapter highlights and uses two types of features: the content features and the minimum at 5% percentile PSNR values when different error concealment algorithms are used. The number of samples that are used in this experiment may not be sufficient to build a generalized framework. In addition, more investigations to trade-off between the quality and the coding efficiency are recommended as a future work.

Box 6.3 – Contributions

— An adaptive content-aware framework to predict the suitable description scheme (SDC, 2-MDC, or 4-MDC) to be transmitted over an error-prone channel in order to maximize the quality of experience. The contrast of Gray Level Co-occurrence Matrix and the ratio of entropy of Laplacian levels 4 and 5 features are used to build the adaptive MDC scheme.

7

MDC-based Video Quality Evaluation Framework

7.1 Introduction

In Chapters 5 and 6, temporal-domain MDC schemes are discussed. The main shortcoming that we observed is that the better the quality you ask for the higher the amount of redundancy you need to send. One solution is proposed in Chapter 6. It introduces an adaptive content-aware MDC scheme in which a specific mode of transmission is recommended, i.e. SDC, 2-MDC, or 4-MDC. It means that for some contents, the perceived quality will not change if we use MDC. Another solution is application and network structure dependent approach. For instance, having an interactive sending/receiving application scenario may reduce the amount of redundant data.
The user datagram protocol (UDP) and the transmission control protocol (TCP) are the most used transport layer protocols. The main key elements in deciding which protocol to use are the delay and the loss. For instance, voice over IP, video conference calls, and broadcasting applications use UDP since they are lightweight applications that care about the delay and not too much about the lost data. While applications like, HTTP, emails, and FTP use TCP since they are heavyweight applications that care about lost data and not too much about the delay. When you have an application that, indeed, needs to trade-off between the delay and the lost data, you need to design your protocol. In this case, the application will be, for instance, lightweight and care about the delay and the lost data as well.
For instance, firstly, using TCP for sending MDC description doesn't make sense since it makes sure that the primary frames are received correctly (except for cases in which the packets cannot be received after a certain number of trials and/or a certain time) so, why do we need MDC in the first place? If we assume that we need to send MDC over TCP channels, making sure that the redundant data is correctly received or not is time consuming (if the primary data received) and making the network congestion-able. Secondly, using UDP for sending MDC descriptions is not the best choice as well (especially when the bitrate of MDC scheme is 4 times the bitrate of SDC and the loss rate is low) because preventing sending some redundant data is applicable if the server reports receiving of primary data before sending the corresponding redundant data. We use TCP and UDP protocols to give examples because they are the widely used protocols but there are other protocols that we don't consider.
To improve the transmission with UDP towards the direction of TCP we allow for out-of-order reception of data packets and allow for sending acknowledgment. In this chapter, an interactive way that works on the application layer to deal with the redundant data is proposed, as discussed later in Section 7.2. Sending redundant data of a current frame is delayed to be sent after the next scheduled frame if the sender doesn't receive an acknowledgment from the receiver. In this way the amount of redundant data to be sent is limited. This limitation, of course, depends on round-trip time and the amount of primary data that needs to be sent per second. Starting from the motivation inspired from the fact that the amount of sending redundant data may be reduced when using a specific application scenario, this chapter provides a possible answer to the research questions that are listed in Box 7.1. The structure of this chapter is illustrated in Box 7.2.

7.2 MDC Evaluation Framework

The reliable way to evaluate the quality of perceived videos that are transmitted over error-prone networks is to build a real-world application scenario and then subjectively evaluate the perceived video quality. Since this approach

Box 7.1 – Research Questions

This chapter aims to answer the following research questions:
- How can the quality of temporal-MDC scheme be evaluated?
  In addition to that, the following secondary research question is investigated too:
  + How to take advantage of TCP and UDP protocols to build a good networking structure that allows to reduce the amount of redundant MDC data to be sent.

Box 7.2 – Chapter structure

This chapter is structured as described in Figure 7.1. In Section 7.2, the framework to evaluate the perceived video quality of MDC schemes is proposed. The simulation results of the proposed framework are analysed and discussed in Section 7.3. The parts with (∗) are planned as future work.

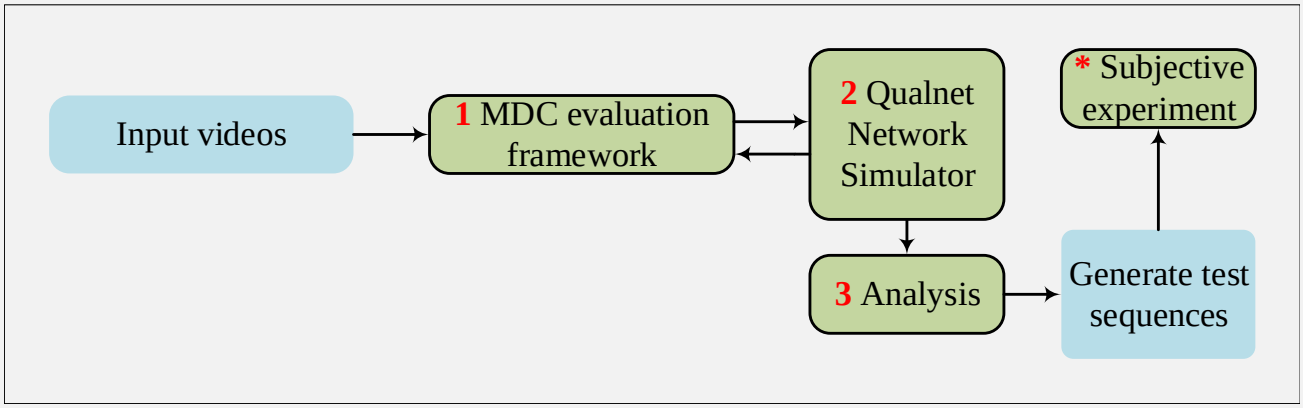


Figure 7.1 – Chapter 7 Structure

is time consuming and expensive, network simulators, like Qualnet and NS2, are the best alternative. In this work, we follow the idea that is introduced in [262, 263]. In [262], an evaluation framework for MPEG and H.264/AVC is presented to evaluate the perceived video quality using NS2 simulator. While in [263], Yi et al. presented a framework to evaluate the perceived quality of AVC/SVC. In this Section, we present a framework, as shown in Figure 7.2, to evaluate the perceived video quality of MDC schemes. We will use video traces to characterize an encoded video for network simulation. Video traces only contain metadata information about the actual video stream that is required for the analysis such as the NALUnit size and the creation time in the form of discrete events usually written in clear text files. We characterize the video bitstream (i.e. the video traffic) of each MDC scheme using HEVCESBrowser [264], a tool for analysing HEVC(h265) bitstreams, and a script, that is developed for this work. The proposed framework can be adapted to work with different networks and in this work, an Ad-hoc network is used to test the proposed framework.

7.2.1 Bitstream Extractor at encoding side

This process is meant to let the MDC and SDC encoders generate the bitstream files. This bitstream contains information about each NAL unit. Once the YUV file is ready to encode, the SDC encoder generates one stream only. While in n-MDC, n streams/descriptions are generated. Then, the HEVC stream browser is used to extract the basic information from each stream/descriptor. Figure 7.3 shows an example. The NAL unit information is extracted; the length (size), type, and other information like frame type or parameter set type.

7.2.2 Traffic generator

After the bitstream information of each description is ready, the trace information, that is required for the simulator to do the job, will be generated. This trace file represents the traffic information. A traffic generator program is implemented. The program takes as input the output of the bitstream extractor. Other supplemental inputs like the type of the MDC are provided as well to distinguish the primary and the secondary/redundant data. Each line of the traffic trace file represents a discrete event for the simulator and it provides the following information:

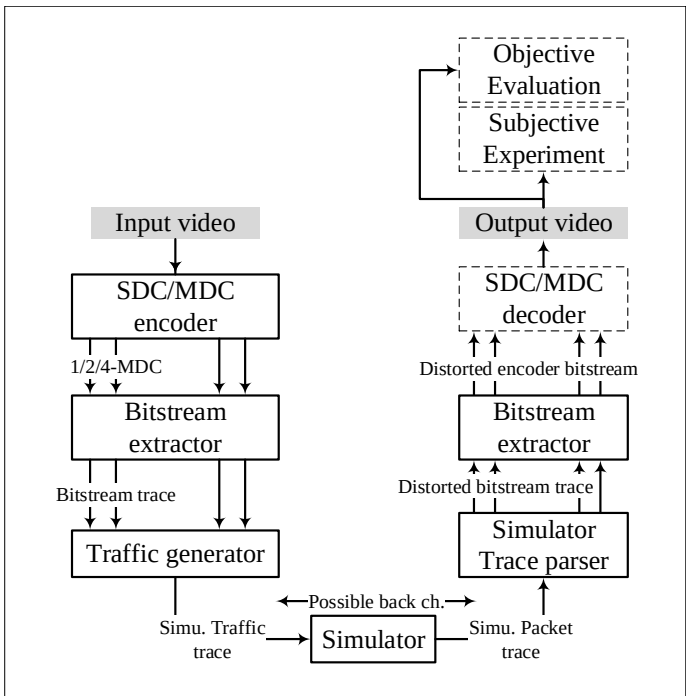


Figure 7.2 – MDC schemes Evaluation Framework

	Offset	Length	Nal Unit Type	Info
1	0x0 (0)	28	NAL_VPS	Video parameter set
2	0x1c (28)	39	NAL_SPS	Sequence parameter set
3	0x43 (67)	10	NAL_PPS	Picture parameter set
4	0x4d (77)	1193642	NAL_IDR_W_RADL	IDR Slice
5	0x1236f7 (1193719)	1564	NAL_TRAIL_N	P Slice
6	0x123d13 (1195283)	12322	NAL_TRAIL_N	P Slice
7	0x126d35 (1207605)	21746	NAL_TRAIL_N	P Slice
8	0x12c227 (1229351)	123022	NAL_TRAIL_R	P Slice
9	0x14a2b5 (1352373)	14703	NAL_TRAIL_N	P Slice
10	0x14dc24 (1367076)	35161	NAL_TRAIL_N	P Slice
11	0x15657d (1402237)	28113	NAL_TRAIL_N	P Slice
12	0x15d34e (1430350)	214155	NAL_TRAIL_R	P Slice

Figure 7.3 – Bitstream information extraction using HEVCESBrowser

— **Time:** The relative time of when the simulator will start to execute one event, i.e. when to send the data.
— **Length:** The amount of bytes that need to be transmitted for each event.
— **Frame type:** This is a flag indicating whether the event transmits primary data (flag=1) or transmits secondary/redundant data (flag=2).
— **Description number:** This represents which part of the MDC bitstream is transmitted, in an n-MDC bitstream, the number ranges from 0 to $n-1$.
— **Reference number:** This number to: 1) represent the NALUnit (here is one frame) order of the primary data in the bitstream 2) links each secondary/redundant data with the corresponding primary data.

Please note that the order of the traffic trace is important. Our strategy is to give the server the opportunity to acknowledge the receiving of the primary data (via feedback channel) and accordingly the client will not send the secondary/redundant data if the acknowledgment is received on time. In order to reach that goal, we propose, as shown in Figure 7.4, to send the primary data of current frame with the secondary/redundant data of previous frame.

Time	Length(bytes)	Primary/Secondary	#Description	#Reference
33.33MS	140737	1	1	1
33.33MS	141116	1	2	2
33.33MS	141680	1	3	3
0.00MS	3064	2	1	2
33.33MS	142726	1	4	4
0.00MS	5611	2	1	3
0.00MS	3030	2	2	3

Figure 7.4 – Traffic trace for MDC scheme (HRC17) for the first primary and secondary frames

7.2.3 The Simulator: the changes to the network structure

The traffic trace application that is part of the simulator is changed in order to fulfil the new requirements. The new requirements are shown in red texts and arrows of Figure 7.5. According to The Qualnet simulator definitions, the client is responsible for sending the data while the server is responsible for receiving the data. On the client side the following changes are applied: the first one is the changing of the header's packet information. It is changed to be able to parse each line of the traffic trace events. The second change is to update the packets sending process to prevent sending secondary/redundant data if the primary data is acknowledged by the server. The third change is to create a list to maintain the list of the acknowledged primary data. On the server side, the following changes are applied. The first change is to send an acknowledgment packet through a feedback channel to the client if the primary data is correctly received. The second change is to create a file to keep information about receiving each primary or secondary data. In addition to this log file, the simulator creates statistics file that keeps information for each network layer.

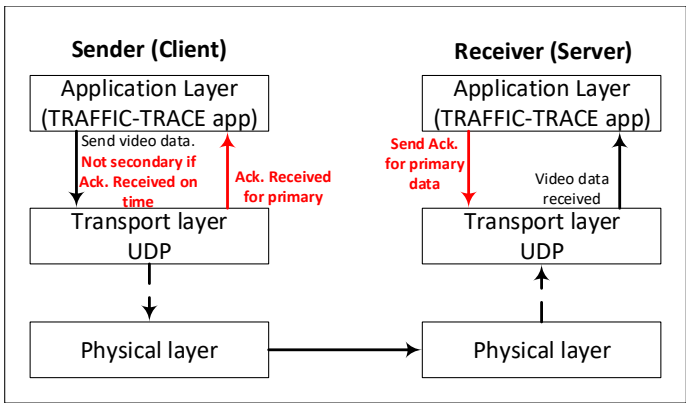


Figure 7.5 – The proposed Qualnet simulator network layers structure. Red texts and arrows represent the updates to the existing structure.

7.2.4 Simulator Trace Parser

This process is meant for the analysis of the simulator log and statistics files. It provides information for each description, specifically, the order of received primary and secondary data. Further information about the bandwidth, sent/received bytes, jitter,delay, etc. is provided as well.

7.2.5 Bitstream Extractor at decoding side

After having all the information about each description, we set a flag for each frame (flag=0 or flag=1 if the primary/redundant data is lost or not respectively) and these flags are saved in a text file for each description. This file represents the error pattern that the decoder needs to decode the distorted stream. Zero means that the NAL unit is received correctly while one means the NAL unit is not received correctly.

7.2.6 Decoding

Once the error pattern is generated for each description, a robust decoder, for instance [251], is used to decode the distorted video sequences. Here, each MDC scheme has error recovery process that uses some/all available descriptions to reconstruct the lost NAL unit.

7.2.7 Quality Evaluation

In order to run the quality estimation/comparison process, a set of HRCs has to be selected. These HRCs should have similar bitrate but actually all of the bitrates that we have to use should be comparable against each other. Therefore, video contents have to be encoded with different bitrate budgets. Then, we use the simulator, such as Qualnet, to run a simulation over a specified network in order to know the practical bitrate consumption for each MDC scheme. Finally, the HRCs that have bitrates that are close to the SDC scheme will be selected to run the quality estimation process either objectively or subjectively.

7.3 Performance Analysis using Qualnet

7.3.1 The Test Conditions

In this experiment, only two sources out of twelves will be used to illustrate the steps of the analysis. The two sources are: Source #2 (CampfireParty) and Source #11 (Wood). These two sources are selected due to their behaviour to the bitrate increase factor as shown in Figure 3.3 (Page 34). CampfireParty and Wood video sequences have a low and high bitrate increase factor, respectively. The video sequences are encoded with the different MDC schemes MDC encoders (adapted from HM12.1): HRC00 (SDC), HRC07 (2-MDC), HRC08 (2-MDC), HRC15 (4-MDC), HRC16 (4-MDC), and HRC17 (4-MDC). These HRCs are chosen as labels as explained in Chapter 6. The video sequences are encoded with different rate control options: 1.5^nMbps, where $n \in \{1, 2, \ldots, 8\}$.

7.3.2 The Network Scenario

The simulation is performed using Qualnet (version 5) with 20 nodes in a 150×150 area as shown in Figure 7.6. The TRAFFIC-TRACE application is used for sending packets to the server side. In this example, as shown in Figure 7.6, node #1 sends video data to node #16. The scenario has the following main parameters:

— Mobility: Random Way Point Model (RWP), speed: 10 m/s.
— Transport protocol: UDP protocol, Packet size 2048 bytes.
— Radio Type: 802.11a, Data rate: 54 Mbps.
— Routing protocol: OLSRv2 NIIGATA
— Network Protocol: IPv4 .

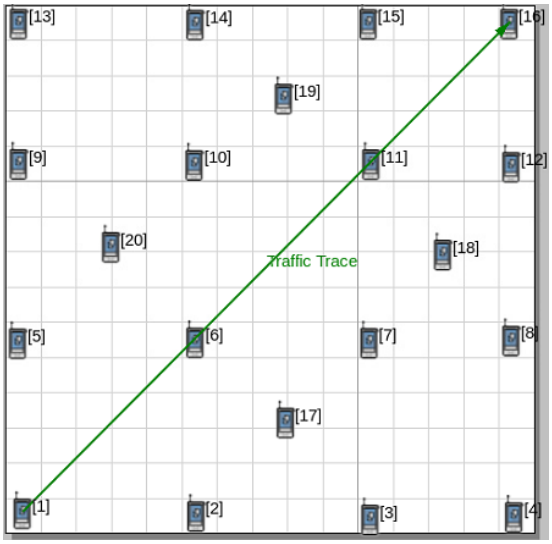


Figure 7.6 – The network scenario for the Ad-hoc network

7.3.3 Simulation Results

7.3.3.1 Bitrate analysis experiment of Chapters 5 and 6 conditions

The bitrate increase factors of the MDC schemes, that are listed in Table 7.1, are shown for two cases; the first one is the maximum case as discussed in Chapter 5. The second one is the case under the proposed transmission scenario. A letter N will be added to the HRC name to reflect it is conducted under the above-mentioned specific network condition. For instance, HRC07 will become HRC07_N01. The results in Table 7.1 are for video sequences that are encoded using $QP = 32$ as explained in Chapters 5 and 6.

Table 7.1 – Comparison of bitrate increase factors of different MDC schemes relative to the SDC. The coloured cells refer to the selected class of each source as shown in Table 6.1 (if source name is coloured, it means that it refers to class A (SDC)).

Src		HRC07_N01			HRC08_N01			HRC15_N01			HRC16_N01			HRC17_N01		
		Max.	Effect.	Diff.	Max.	Effect.	Diff.	Max.	Effect.	Diff.	Max.	Effect.	Diff.	Max.	Effect.	Diff.
1	BBB_seq1	1.57	1.34	0.23	1.45	1.35	0.10	2.84	1.89	0.95	2.31	1.93	0.38	3.05	1.91	1.14
2	BBB_seq2	1.65	1.37	0.28	1.52	1.36	0.16	3.10	1.96	1.14	2.57	2.02	0.56	3.52	2.02	1.50
3	Beauty	1.23	1.10	0.14	1.19	1.10	0.09	1.68	1.20	0.48	1.50	1.20	0.30	2.27	1.40	0.87
4	CampfireParty	1.30	1.11	0.19	1.16	1.09	0.07	1.88	1.29	0.58	1.45	1.45	0.00	2.16	1.40	0.76
5	CrowdRun	1.56	1.56	0.01	1.51	1.51	0.01	2.60	2.57	0.03	2.35	2.34	0.01	2.91	2.89	0.02
6	PROMO_WIENER_2	1.50	1.23	0.26	1.40	1.24	0.16	2.36	1.44	0.92	1.99	1.46	0.53	3.39	1.47	1.93
7	ParkJoy	1.44	1.43	0.01	1.38	1.37	0.01	2.20	2.18	0.01	1.92	1.92	0.00	2.47	2.46	0.00
8	TOS_3	1.64	1.41	0.23	1.54	1.43	0.11	2.83	1.98	0.85	2.61	2.10	0.51	2.80	1.98	0.82
9	TallBuildings	1.22	1.11	0.11	1.20	1.12	0.07	1.72	1.35	0.38	1.61	1.38	0.22	1.61	1.34	0.28
10	TreeShade	1.45	1.34	0.11	1.48	1.38	0.10	2.10	1.70	0.40	2.31	2.11	0.20	2.43	1.81	0.62
11	Wood	1.53	1.38	0.15	1.42	1.37	0.05	2.72	2.15	0.57	2.26	2.10	0.16	2.47	2.08	0.40
12	YachtRide	1.42	1.30	0.12	1.37	1.28	0.09	2.13	1.73	0.39	1.89	1.64	0.24	2.88	2.14	0.74
Average		1.46	**1.30**	**0.15**	1.39	**1.30**	**0.09**	2.35	**1.79**	**0.56**	2.06	**1.81**	**0.26**	2.66	**1.91**	**0.76**

As it can be observed, the amount of redundant data to be sent is reduced by factors of 1.93 in source #6 but not even slightly reduced in source #7. The coloured cells refer to the selected class of each source as shown in Table 6.1 on page 79 (if source name is coloured, it means that it refers to class A (SDC)). The content-aware scenario that is proposed in Chapter 6, together with good/smart networking implementation, provides a promising solution for using temporal-MDC scheme as one way to maximize the quality of experience.

7.3.3.2 Choosing HRCs for testing

A major criticism to the conducted subjective experiment in Chapter 5 is that the comparison is done with sequences encoded at the same quality level. Hence, the bitrate budget of each HRC differs from others. That was done because we do not know the effective bitrate of each MDC scheme. Now, after we are capable of knowing the effective bitrate budget of each MDC under a specified networking scenario, we can compare the perceived video sequences at the same bitrate budget. As can be noticed in Table 7.2, to test the quality of the CampireParty sequence at a bitrate budget of 7.76Mbps±10%, the shaded HRCs are going to be selected for the quality testing.
Regarding the end-to-end delay, Table 7.3, it is clear that increasing the bitrate budget increases the delay. Moreover, the delay is also increasing if the amount of redundant data increased too, i.e. the amount of data per second is increasing. This leads to increase the network congestion. It can be noticed, as well, that the end-to-end delays of the selected MDC schemes for testing are sometimes smaller and sometimes larger than the SDC (@7.59Mbps) scheme by 0.73 and 1.7 seconds, respectively. While at budget of 25.63Mbps, the end-to-end difference is neglectable. The effect of end-to-end delay is application dependent, and it might be applicable for Ad-hoc networks to have such delay especially when UHD videos are transmitted. Further experiments have to be done with different resolutions to study this parameter/characteristic.
Jitter is often defined as variations in packet delay. It is an important issue for the interactive real-time application, such as voice over IP. To avoid the jitter issues, a proper buffer size in the server (receiver) should be allocated. As noticed in Table 7.4, the jitter of the selected MDC schemes are smaller compared to the SDC scheme but it seems an issue when the amount of redundant data is much bigger than SDC scheme in high bitrate budget.

Table 7.2 – Maximum and effective bitrate consumption (in Mbps) for different bitrate budgets and different MDC schemes

SRC	Rate	SDC_N01	HRC07_N01		HRC08_N01		HRC15_N01		HRC16_N01		HRC17_N01	
			Max.	Effect.	Max.	Effect.	Max.	Effect.	Max.	Effect.	Max.	Effect.
CampfireParty	2.25	2.31	3.91	2.47	4.48	4.01	7.95	3.23	8.79	6.95	8.86	5.98
	3.38	3.45	5.04	3.60	6.70	6.07	9.13	4.59	13.20	10.72	13.27	9.37
	5.06	5.15	7.53	5.43	10.05	9.22	13.11	7.20	19.78	16.96	19.88	14.27
	7.59	7.76	11.27	8.65	15.19	14.17	19.37	11.63	29.64	28.41	29.81	22.34
	11.39	11.78	17.05	13.16	22.70	21.46	29.03	19.32	44.14	44.09	44.91	34.64
	17.09	17.59	25.61	20.44	33.94	33.81	43.56	30.11	65.71	65.67	67.53	66.90
	25.63	26.21	38.41	31.37	50.70	50.51	65.55	63.67	98.21	98.13	101.05	100.11
Wood	2.25	2.54	4.46	2.92	4.64	4.41	9.98	4.76	9.07	7.69	9.15	7.17
	3.38	3.75	5.69	4.38	6.89	6.54	11.37	6.62	13.46	11.40	13.62	10.47
	5.06	5.50	8.36	6.46	10.25	9.69	15.94	9.93	20.02	17.30	20.31	15.29
	7.59	8.22	12.37	9.54	15.35	14.64	23.03	14.86	30.00	29.97	30.40	23.87
	11.39	12.13	18.23	14.51	22.88	21.98	33.71	22.83	44.60	44.56	45.39	36.76
	17.09	17.94	26.90	21.77	34.13	34.01	48.80	39.73	66.15	66.10	67.83	67.24
	25.63	26.64	39.64	33.59	50.93	50.75	71.37	69.24	95.59	94.28	101.32	100.42

7.4 Conclusion

The main contribution of this chapter is listed in the Box 7.3. It is observed that the content-aware scenario that is proposed in Chapter 6, together with feedback-capable networking implementation, proposed in this Chapter, provide a promising solution to use the temporal-MDC scheme as one way to maximize the quality of experience.

Box 7.3 – Contributions

— Quality evaluation framework for temporal-MDC schemes is proposed. The framework introduces an interactive networking structure that helps reducing the amount of redundant data to be sent. During this work, all the steps of the proposed framework except the quality evaluation (subjectively and objectively) stage are conducted. This is due to time and computing power limitations.

Table 7.3 – End-to-End delay (in seconds) for different MDC schemes transmitted over ad-hoc network with different bitrate budgets

SRC	Rate	SDC_N01	HRC07_N01	HRC08_N01	HRC15_N01	HRC16_N01	HRC17_N01
CampfireParty	2.25	0.09	0.21	0.21	0.45	0.32	0.34
	3.38	0.14	0.30	0.33	0.60	0.72	0.59
	5.06	0.24	0.52	0.66	0.91	1.47	1.25
	7.59	0.65	1.18	1.57	2.67	4.76	4.27
	11.39	1.33	2.45	4.29	5.81	9.21	8.94
	17.09	4.28	4.54	7.85	9.99	17.06	21.56
	25.63	7.61	9.80	13.86	20.88	27.72	31.12
Wood	2.25	0.25	0.55	0.53	0.85	0.82	0.67
	3.38	0.38	0.76	0.90	1.25	1.40	1.28
	5.06	0.63	1.15	1.37	2.44	2.78	3.92
	7.59	1.40	2.80	3.68	6.64	6.82	6.69
	11.39	3.89	6.28	6.84	9.14	12.43	11.46
	17.09	6.43	9.32	10.39	13.00	19.36	22.67
	25.63	7.91	11.33	15.98	23.16	28.28	32.50

Table 7.4 – Jitter (in milliseconds) for different MDC schemes transmitted over ad-hoc network with different bitrate budgets

SRC	Rate	SDC_N01	HRC07_N01	HRC08_N01	HRC15_N01	HRC16_N01	HRC17_N01
CampfireParty	2.25	2.6	1.7	2.0	1.2	2.2	1.5
	3.38	3.2	2.0	3.1	1.5	3.7	1.8
	5.06	3.8	2.5	5.2	1.6	7.7	2.6
	7.59	5.4	3.5	11	3.1	15	4.2
	11.39	9.4	7.3	20	5.3	24	4.4
	17.09	15	12	42	4.9	36	4.8
	25.63	19	11	63	4.7	53	8.2
Wood	2.25	7.6	4.1	4.5	2.0	4.4	2.9
	3.38	9.6	5.1	5.8	2.6	5.9	3.6
	5.06	11	6.1	7.9	4.2	9.3	4.6
	7.59	16	11	16	6.3	17	7.3
	11.39	23	12	24	5.3	25	5.9
	17.09	22	13	43	5.9	36	6.1
	25.63	26	15	65	6.2	59	9.6

Inpainting-based error concealment (EC) technique in video communication

8

Inpainting-based error concealment for low-delay video communication

8.1 Introduction

Ensuring error resilience has become more complex in recent video coding standards due to the increased complexity of their prediction processes [265]. Conventional error resilience techniques [26] can be categorized into three main classes depending on where the processing is performed: forward-error-concealment (encoding side), post-processing error concealment (decoder side), and interactive error concealment (encoder and decoder sides). In this work we focus on error concealment by post-processing. In post-processing techniques, the decoder utilizes the spatial and/or temporal redundancies to reconstruct the damaged/lost area in a video frame. Spatial techniques [29, 30]utilize available surrounding pixels to reconstruct the missing pixels. They are not efficient for large areas, non-constant areas, and in terms of complexity. They usually reconstruct the texture but not the structure. The work in [163] is an extension of [30] in which a spatio-temporal selective extrapolation strategy is used to reconstruct the missing area. Temporal techniques use available motion information to predict the missing motion vectors (MVs), for instance, by interpolating [31] or by selecting the MV that minimizes the side match distortion [32]. Despite providing information about whether the current area is moving or not, this technique is efficient only for low-motion and smooth sequences and for small areas since the precision of predicted MVs is not guaranteed. Thus, the structure (of copied data) is reconstructed but not the texture.
The target of any error concealment algorithms is twofold: reconstructing a satisfying reconstruction of a lost area and reducing the miss match between the encoded and the reconstructed blocks which yields reducing the error propagation effect. To achieve that we need to reconstruct the texture and the structure of a missing area and that can be done using inpainting techniques. A review of inpainting techniques can be found in [164]. In this work we focus on exemplar-based inpainting in which each lost patch is reconstructed by copying the best match from the known area. Specifically, inpainting algorithms have many target applications and in this work, we are interested in error concealment as a target application.
Inpainting-based error concealment algorithms are introduced in [34, 165]. The algorithms have three main steps: inpainting the moving foreground object, inpainting the stationary background temporally and spatially as in [166]. In the first step, inpainting the moving foreground object, the moving pixels are identified as in [34] using Bilinear Motion Field Interpolation (BMFI) [31]. Then, the best match of a moving patch is reconstructed from the neighbouring frames. In the second step, inpainting the stationary background temporally, the best match of a moving patch is reconstructed from the co-allocated patches of the neighbouring frames. The remaining pixels are reconstructed in the third step, inpainting the stationary background spatially. There is some room of improvements. In this chapter, we try to improve one of the state-of-the-art inpainting-based error concealment algorithm [34] by raising the research questions that are listed in Box 8.1. The structure of this chapter is illustrated in Box 8.2

8.2 Inpainting-based error concealment strategy

Inpainting-based error concealment algorithms are introduced in [34, 165]. The algorithms have three main steps: inpainting the moving foreground object, inpainting the stationary background temporally and spatially as in [166]. In this work, a modified version of [34] is introduced with the following contributions:

Box 8.1 – Research Questions

This chapter aims to answer the following research questions:

— What is the information, content indicators, that needs to be considered as inputs of the inpainting-based EC algorithm?

In addition to that, the following secondary research question is investigated too:

+ How to adapt the state-of-the-art inpainting-based EC algorithms to be suitable for low delay communication?

Box 8.2 – Chapter structure

This chapter is structured as described in Figure 8.1. Generation of the motion map is shown in Section 8.2.1, and the inpainting process is demonstrated in Section 8.2.2. The experimental results are shown in Section 8.3. Finally, we sum up with the conclusions in Section 8.4.

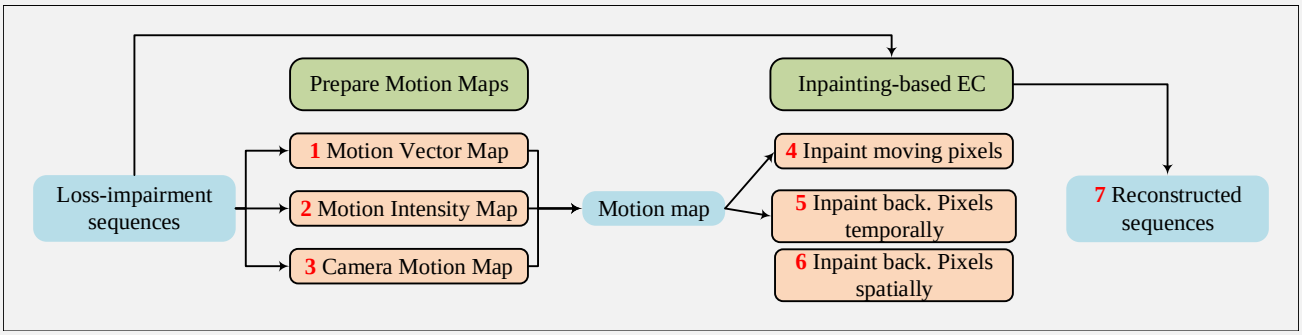


Figure 8.1 – Chapter 8 Structure

— The quality of the results depend on the input M_c which indicates whether a pixel p is moving $M_c(p) = 1$ or not $M_c(p) = 0$. The quality also depends on the strategy that replaces the simple copy strategy of the best patch match by other strategies like LLE [266] and NMF [267]. The latter factor is investigated in [268] and it is shown that the performance of the inpainting algorithm is improved. In this paper we will investigate the former factor by introducing a concept of motion map M_c that includes the predicted motion vectors M_{mv}, the pixel-based motion intensity M_{pi} and the motion vector of interests (MVI) that relate to camera motion M_{cm}. It will be shown that the proposed motion map will significantly improve the performance of the inpainting.

— The algorithm in [34] works on the sequence level. The process does not start once the error occurs, but it waits until more frames are available. Then, it first searches for the highest priority frame to start with. That means that there may occur more than one error and it also means that the concealment might be performed out of temporal order. This strategy is not practical for some video applications, since, in video communication, once the error is detected in a frame, especially those that are used as reference for coming frames, it must be concealed before the decoder continues. In this paper, the error concealment strategy is optimized for low-delay configuration.

— Using a full search strategy or fixed window size is not efficient in terms of complexity and quality respectively. In this paper an adaptive search window size for temporal and spatial inpainting is introduced.

— Trying to reduce the spatio-temporal artifacts, a simple blending strategy is employed using Poisson blending [269] with the proposed mask strategy.

8.2.1 Motion Map

In this subsection, the concept of the motion map is illustrated. The motion map M_c is computed as: $M_c = M_{mv} \vee M_{pi} \vee M_{cm}$, where $(\vee)$ is the logical OR operator, and the M_{mv}, M_{pi}, and M_{cm} will be described on the following subsections.

8.2.1.1 Motion Vector Map

As in [34], the lost MVs are predicted using Bilinear Motion Field Interpolation (BMFI) [31]. MV components V_x and V_y are threshold to determine whether the pixel p belongs to a moving object $M_{mv}(p) = 1$ or not $M_{mv}(p) = 0$. In this work, a threshold of 1 is used such that $M_{mv}(p) = 1$ if $V_x(p)$ or $V_y(p) > 1$. Figure 8.2

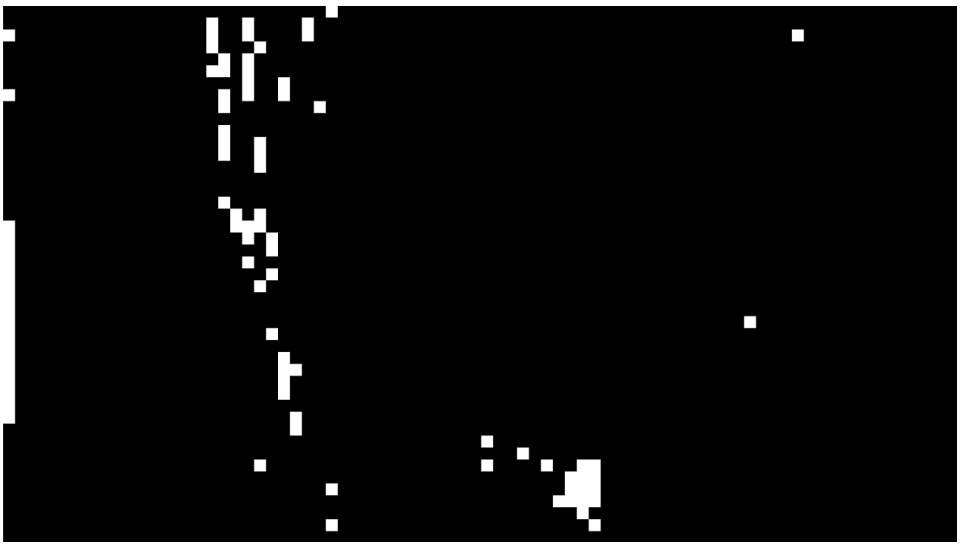


Figure 8.2 – Motion vector map for source 1

8.2.1.2 Motion Intensity Map

In order to measure the pixel-based motion intensity, the pixel change ratio map (PCRM) [270] strategy is used. This algorithm assumes that that a high intensity of motion yields a large change in pixel intensities over a video shot. In this work the shot is represented by up to 8 previous frames. $M_{pi}(p) = 1$ if $PCRM(p) > th_i$ and $M_{pi}(p) = 0$ if $PCRM(p) \leq th_i$, where th_i motion intensity threshold. In this work it is set to 0.25 to exclude the pixels that have low intensity changes over the video shot. Figure 8.3

Figure 8.3 – Motion Intensity Map for source 1

8.2.1.3 Camera Motion Map

In [271], the MVs of up to 8 previous frames are analysed to obtain motion vectors of interest (MVI). MVIs identify the spatial region where the motion information has a direct relationship with the camera movements [271]. In this work, the MVI of each frame is computed and assigned to M_{cm}. Figure 8.4

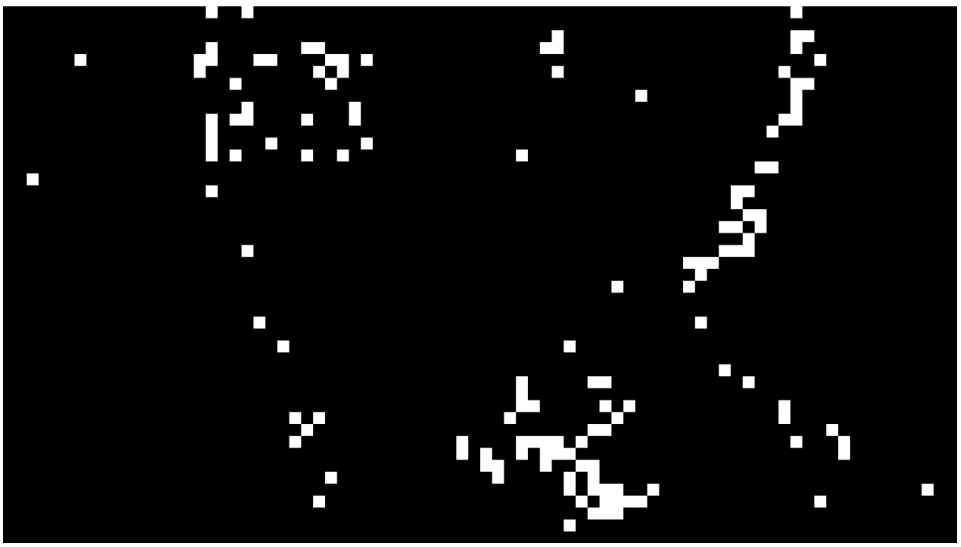


Figure 8.4 – Camera motion map for source 1

8.2.2 Inpainting Process

In this section, the three main steps of inpainting-based error concealment method will be illustrated. Let some blocks be lost in frame F at time t (F_t). The frame F_t has known/source area Φ and lost/target area Ω to be filled. The fill front $\delta\Omega$ is defined as a contour that separates known and lost areas. The key elements of exemplar-based inpainting algorithms are the filling order (or patch priority) of lost area and the texture synthesis, i.e. finding the best match of the current processed patch. These two key elements will be illustrated in the following steps.

8.2.2.1 Inpainting Moving Objects

Once the error occurs, the error concealment (EC) process starts filling the lost area patch by patch using the following steps: for each pixel p of the fill front $\delta\Omega$, compute the patch Ψ_p priority, Eq. 8.1, as in [34,165], where $C(p)$ is the confidence term and $D(p)$ is the data term. The confidence term, Eq. 8.2 represents the ratio of known data to the patch area. While the data term, Eq. 8.3, gives more priority to the patches that have orthogonal motion direction ($\nabla M_c^{\perp}$) to the fill front $\delta\Omega$. n_p is the normal to the fill front $\delta\Omega$ at p, and α is a normalizing constant ($\alpha = 255$).

$$P(p) = C(p)D(p) \tag{8.1}$$

$$C(p) = \frac{\sum_{q\in\Psi_p\cap(F-\Omega)} C(q)}{|\Psi_p|} \tag{8.2}$$

$$D(p) = \frac{|\nabla M_c^{\perp}.n_p|}{\alpha} \tag{8.3}$$

The next step now is to synthesize the patch that has the highest priority $\Psi_{\hat{p}}$, where $\hat{p} = \arg\max_{p\in\delta\Omega} P(p)$. The block matching algorithm is used to find the best matching patch Ψ_q of the known part of $\Psi_{\hat{p}}$ within a search window w in the previous/reference frame using the sum of squared differences (SSD) of colour and MV components (R, G, B, V_x, V_y) of known pixels of $\Psi_{\hat{p}}$ and all candidates in the search process. Where w is equal to the double of the largest value of MV components of the surrounding area. The pixels values of Ψ_q are copied to the co-located unknown pixels of $\Psi_{\hat{p}}$.
The aforementioned steps are repeated until all moving and damaged pixels are concealed, the confidence term of the copied pixels $\Psi_{\hat{p}}$ is updated and the motion map M_c is also updated by copying the M_c of Ψ_q to the $\Psi_{\hat{p}}$.

8.2.2.2 Inpainting The Stationary Background Temporally

In the previous step, all the moving pixels are concealed. In this section, the steps for inpainting the stationary background temporally are demonstrated.
Following almost the same process of filling-in the moving pixels, the priority term, Eq. 8.1, for each patch centred at p, where $p \in \delta\Omega$ needs to be computed. First, the confidence term of each pixel that is either damaged or moving is set to $C(p) = 0$ and $C(p) = 1$ otherwise. The data term is defined to measure the amount of available temporal information ($M_t(p)$) in the up to 8 previous frames. Hence, the data term is defined [165] as Eq. 8.4, where $M_t(p)$ is 0 if p is either a moving or a damaged pixel, else $M_t(p)$ is 1. The time index t indicates the relative position of up to 8 previous frames from $t = 0$ (the current frame) and β is a normalizing factor that represents the number of previous

frames used to compute the data term.

$$D(p) = \frac{\sum_{p \in \delta\Omega, t=-\delta n \ldots 0} M_t(p)}{\beta} \tag{8.4}$$

The next step is to copy the patch Ψ_q from the nearest frame to the unknown part of the patch $\Psi_{\hat{p}}$ that has the highest priority. Then, the confidence terms of previously damaged pixels are updated. The process iterates until no more temporal information needs copying, i.e. $D(p) = 0, \forall p \in \delta\Omega$. That means that the remaining pixels of the stationary background have to be inpainted spatially.

8.2.2.3 Inpainting the Stationary Background Spatially

In this section, the steps for spatially inpainting the remaining pixels of the stationary background will be demonstrated. This process follows the algorithm that is described in [166] exactly except for the search window size w_p. The search window size is adaptively changed for each process patch as follows. First, the minimum and maximum allowed window size is computed as Eq. 8.5, and Eq. 8.6, where $d(p)$ is the nearest distance between unknown pixels to the known pixels. Second, for each patch the search window size is set to Eq. 8.7. This adaptive procedure is to trade-off the quality and the complexity of the spatial inpainting.

$$min_w = 2 * patchSize \tag{8.5}$$

$$max_w = 2 * max(d) \tag{8.6}$$

$$w_p = max(min_w, d(p) * \frac{max_w}{min_w}) \tag{8.7}$$

8.2.3 Blending Step

In [272], the Poisson blending [269] is used to reduce the artifacts of the inpainting process and it was shown that it improves the performance since the inpainting algorithm is based on frames registration. In this work, we first demonstrate the blending of the inpainted frame temporally using the motion-compensated frame of the lost frame and using the lost area as a mask. It is observed that this process will not improve the quality if not making it worse since the structure of the lost area is not respected. This is because of the predicted motion vectors. Therefore, the blending mask M_{blend} is changed to blend only the pixels that are far enough from the edges d_{edge} and have a low motion magnitude MV_{mag}. Hence, the pixel will be blended if $M_{blend}(p) = \frac{d_{edge}}{MV_{mag}} > th_{blend}$. In this work, the th_{blend} is set to 2. It was observed that this blending mask gives better results than the former method. Unfortunately, in general, this blending process is not improving the inpainted frame as assumed since the proposed motion map maintains the structure and the texture of the inpainted frame.

8.3 Experimental results

The proposed algorithms and other state-of-the-art algorithms [31, 34] are implemented using MATLAB. Spatial-only method [166] is compared in [34] and for the sake of complexity, it is not tested in this work. Eight 1280×720 video sequences are used in the experiment, Figure 3.25. In each frame, 5%, 10%, and 20% of the 64×64 blocks are randomly lost and inpainted using different error concealment methods. For the sake of fair comparison, each source share the same error pattern. The patch size should be greater than the thickest structure (e.g., edges) in the source region [166]. In this work, it is set to 9 for all sequences. Figures 8.5, 8.6, and 8.7 show the results of the recovered areas from sequence 1. It can be noticed that the proposed method improves the visual quality of the recovered areas. Table 8.1 shows the performance of the different methods, in terms of difference of quality (PSNR). Method of [34] is used as reference of comparison. The results are the average of the first 45 frames in the video shot. It can be noticed that the proposed method achieves 1 to 6 dB of quality improvements depending on the video shot characteristics. In terms of complexity, the proposed algorithm is faster than the algorithm in [34] by a factor of two on average.

8.4 Conclusion

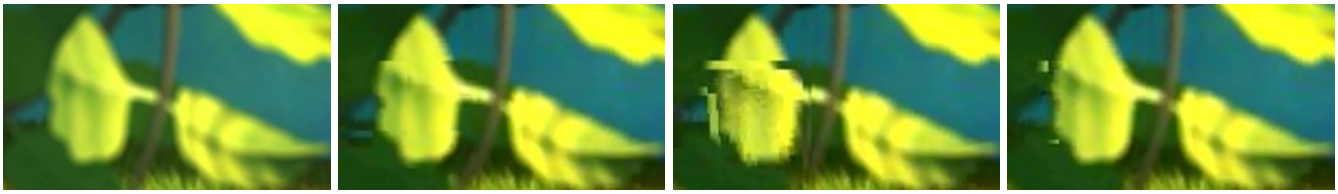
This work proposes a modified version of the inpainting-based error concealment [34] by introducing several enhancements that are listed in Box 8.3. The experimental results show that the proposed methods improve the visual quality and hence, reduce the error propagation. More investigations are required to know when the blending technique might be used. Moreover, running the proposed method in a real network environment and real coding environment is planned as future work.

Table 8.1 – Quality performance of the different EC methods.

Sequences	%lost	Δ PSNR (dB) compared to [34]		
		[31]	Proposed	Proposed with blending
Seq. 1	5	-0.2	2.0	2.0
	10	-1.1	2	1.9
	20	-2.4	1.5	1.4
Seq. 2	5	-0.6	6.0	5.6
	10	-0.7	6.0	5.7
	20	-0.7	6.3	6.2
Seq. 3	5	-8.2	1.0	0.3
	10	-9.7	0.6	-0.1
	20	-9.5	0.9	0.2
Seq. 4	5	-13.3	1.4	0.8
	10	-13.5	1.2	0.9
	20	-13.7	1.1	0.7
Seq. 5	5	-3.7	6.3	6.0
	10	-4.4	6.3	6.0
	20	-5.1	5.3	4.9
Seq. 6	5	-8.7	5.1	5.1
	10	-8.3	6.2	6.1
	20	-8.7	5.9	5.9
Seq. 7	5	-6.9	15.6	12.2
	10	-6.8	15.4	11.7
	20	-6.9	14.3	10.3
Seq. 8	5	-6.8	4.7	4.4
	10	-7.2	4.9	4.3
	20	-7.1	4.5	4.1
average	5	-6.0	5.3	4.5
	10	-6.5	5.3	4.6
	20	-6.8	5.0	4.2

(a) original (b) motion comp. [31] (c) inpaint [34] (d) inpaint (proposed)

Figure 8.5 – Example 1: Comparison of different error concealment methods for 10% of lost of sequence 1 (pink rectangle).



(a) original (b) motion comp. [31] (c) inpaint [34] (d) inpaint (proposed)

Figure 8.6 – Example 2: Comparison of different error concealment methods for 10% of lost of sequence 1 (yellow rectangle).

(a) original (b) motion comp. [31] (c) inpaint [34] (d) inpaint (proposed)

Figure 8.7 – Example 3: Comparison of different error concealment methods for 10% of lost of sequence 1 (white rectangle).

Box 8.3 – Contributions

A modified version of the inpainting-based error concealment [34] is proposed. The following improvements are achieved in the proposed algorithm:

- The concept of motion map M_c is introduced. It includes the predicted motion vectors M_{mv}, the pixel-based motion intensity M_{pi} and the motion vector of interests (MVI) that relate to camera motion M_{cm}. It was shown that the proposed motion map improves the performance of the inpainting.
- The algorithm is adapted to be practical for low-delay video communications.
- An adaptive search window size for temporal and spatial inpainting is introduced.
- Reduce the spatio-temporal artifacts using simple Poisson blending strategy with the proposed mask strategy.

9

Content-aware observer's disruption analysis of inpainting-based EC technique

9.1 Introduction

Visual scenes often contain enormous amounts of information: many orders of magnitude greater than the processing capacity of the brain. For a simplified representation of this huge data, the visual system limits the high resolution sensitivity to less than two degrees of viewing angle around the central viewpoint known as the *Foveola*. Second, from the temporal perspective, the visual system needs a finite duration of time before it semantically understands and grasps the temporal activity in a scene as well. This temporal and spatial localization of real-world information in the visual system is what we refer to as a *Spatio-Temporal Short-Term*.
The visual periphery in general refers to one of the several regions outside the central foveal area: the Para-Fovea, Peri-Fovea or Extra Peri-Fovea, the exact definition based on the photoreceptor mosaic topology which determines the visual sensitivity and acuity of an area [273–275]. Irrespective of the exact terminology, vision studies have highlighted the drop in spatial texture [211], colour [219], motion [212] and flicker [235] sensitivity across the periphery of the retina. Aspects regarding how these reduced sensitivities translate to drop in perceived quality have been less explored. In an earlier instance, it was observed that humans were sensitive to both: spatial and High Efficiency Video Coding (HEVC) based flicker distortions in the visual periphery [276], although the exact sensitivity was found to be content dependent. There has also been a significant amount of work in peripheral perception studies, through the study of visual equivalents known as *Peripheral Metamers* : stimuli that differ physically but look the same [277]. In addition, the recent work from Wallis et al [278] indicates that humans have an impressive sensitivity towards deviations from natural appearance, in viewing eccentricities as much as ten degrees.
In this chapter we aim to analyse the perceived subjective quality of videos that are subject to losses. The lost areas are reconstructed using two error concealment algorithms: the first on is the motion vector prediction and motion compensation using Bilinear Motion Field Interpolation (BMFI) [31]. The second EC algorithm is the inpainting-based EC technique that is illustrated in Chapter 8. The main goal is to study the subject's disturbance by following the hypothesis "if the user changes gaze, he is giving a lower MOS". Hence, in this chapter, we raise the research questions that are listed in Box 9.1. The structure of this chapter is shown in Box 9.2.

Box 9.1 – Research Questions

This chapter aims to answer the following research questions:
- Does the observer get disturbed with the proposed inpainting-based EC algorithm? Does that correlate with DMOS?
- Which content indicators may help in predicting this correlation?

Box 9.2 – Chapter structure

This chapter is structured as descried in Figure 9.1.

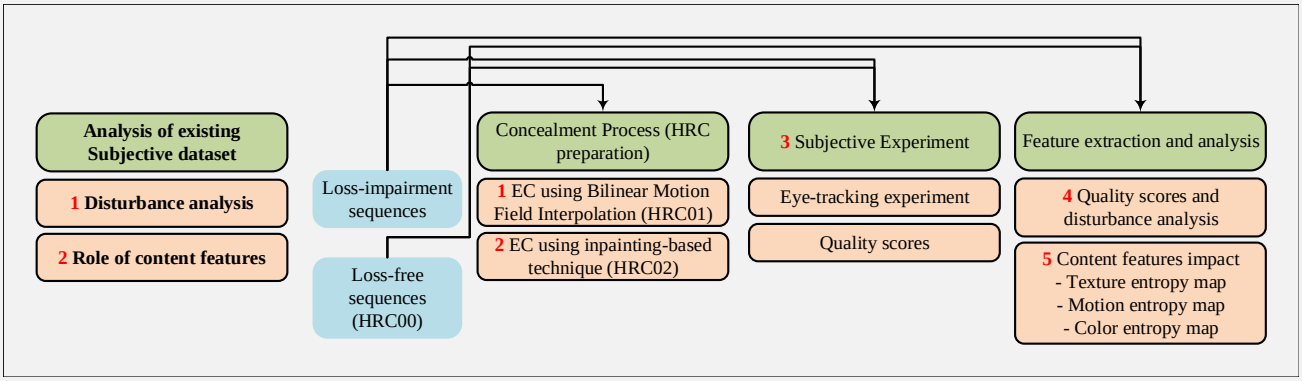


Figure 9.1 – Chapter 9 Structure

9.2 Observations from existing loss-impairment video dataset

In this Section, we particularly analyse the perceived subjective quality of videos containing H.264/AVC transmission impairments [279], incident at various degrees of retinal eccentricities of subjects. We relate the perceived drop in quality, to five basic types of features that are important from a perceptive standpoint: texture, colour, flicker, motion trajectory distortions and also the semantic importance of the underlying regions.

9.2.1 Subjective Experiment

The subjective experiment consists of two parts: The first involved the measurement of quality scores in addition to gaze recording as explained in [279]. The second experiment was performed much later in order to assess the importance of each object in the presented scenes.

9.2.1.1 Experimental Setup and Test Subjects

The experimental setup is described in [279] and we restrict ourselves only to the relevant details here. 30 naive human observers were each presented with 20 different videos from the VQEG dataset, under 5 different conditions - only 3 of them being relevant here : (a) Control condition where no transmission impairment is embedded. (b) Transmission impairment in a salient area for 400ms and (c) Transmission impairment in a non-salient area for 400ms. The Joint Video Team (JVT) loss simulator was used to introduce packet loss into the H.264/AVC bit stream to produce a transmission impairment that lasted exactly for 0.4 secs corresponding to our *short-term*. To have a better control regarding the location and extent of the loss patterns, a fixed number of 45 macro blocks (MB) per slice was chosen, and the error was restricted to this single slice only.
The experiment was designed according to ITU Rec. BT.500 and the videos were presented on a LVM-401W full HD screen by TVlogic with a size of 40" and a native resolution of 1920 × 1080 pixels and frame rate 25fps.

9.2.1.2 Measuring Subjective opinion

The 5-point impairment scale was used to assess the annoyance of the distortions in the sequences. Here, the subjects assigned one of the following adjectival ratings to each of the sequences: 'Imperceptible (5)', 'Perceptible, but not annoying (4)', 'Slightly annoying (3)', 'Annoying (2)', and 'Very annoying (1)'. Scores obtained for the pristine undistorted sequences were then subtracted from the scores obtained for test cases with transmission errors for each individual subject and video, in order to obtain the 1200 difference scores (30 observers × 20 videos × 2 impairments) for the analysis.

9.2.1.3 The Eye-Tracking experiment

The eye-movement patterns of the subjects were recorded throughout the test, with the scoring in Section 9.2.1.2 also performed using the eye-tracker. The SMI Hi-Speed eye-tracker was used to obtain 500 gaze data samples per second. Calibration was performed before displaying the actual video to minimize errors due to bad calibration.

9.2.1.4 Measuring Object Importance

It is important to understand as to, which objects human subjects regard to be the most important in a video, so that the semantic importance feature can be calculated. As explained in 3.3.6, the object semantic importance is calculated. The importance score for every individual object in every video were then averaged among the subjects to obtain an average importance that is less affected by individual variations and has a better precision.

9.2.2 Feature Analysis

This section lists the content features/indicators that will be used in this work. Each item in the list refers to the section where more details can be found.

— Viewing Eccentricity 3.3.1
— Distortion in Texture 3.3.2
— Distortion in Colour 3.3.3
— Role of Semantic Importance 3.3.6
— Distortion in Motion Trajectories 3.3.4
— Distortion in Temporal Harmonics(Flicker) 3.3.5

9.2.3 Observations

Each of the 1200 difference opinion scores (DOS) corresponding to that of each subject in each viewing (30 observers x 20 videos x 2 impairments) are examined in 12 separate groups in accordance to the impairment viewing eccentricities derived from the corresponding Eye Tracking data. In addition, content feature analysis is performed in order to relate the perceived quality drop at each individual eccentricity to the underlying video content.

9.2.3.1 Effect of viewing eccentricity on perceived quality

Investigating the relation between the viewing eccentricity and the drop in the quality score (as compared to the pristine reference), shows us that viewing eccentricity is a major factor that determines perceived quality. It is seen from Figure 9.2, that the drop in quality score of subjects is dependent on the eccentricity at which they observed the distortion. The drop in *Cortical Magnification* factor in the V1 area follows a similar characteristic as well [280]. The eccentricity at which the distortion was incident upon the retina of the subject is therefore a very important determinant of perceived quality. This is mainly due to the nature of photoreceptor arrangement in the early vision stages of the human visual system, which in turn affects its resolving power.

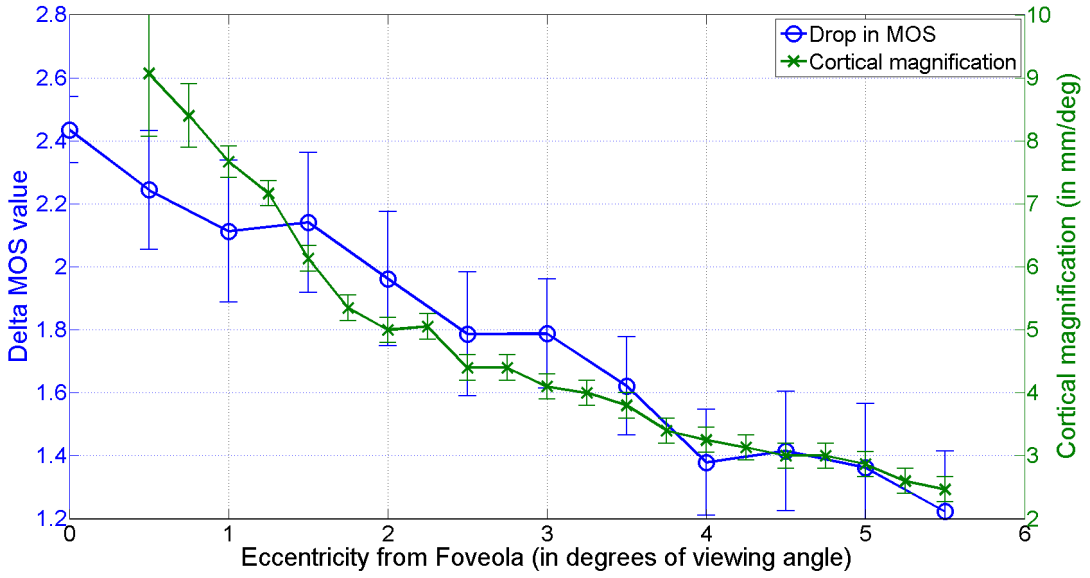


Figure 9.2 – Variation of difference scores of subjects and cortical magnification factor [280] with viewing eccentricity along with the 95 percent confidence intervals

9.2.3.2 Overall effect of the content features

For analysing the overall effect of the content features (at all eccentricities), as shown in Figure 9.3, we use two different methods: The correlation score obtained by correlating the normalized feature responses with the drop in

MOS and second, the weights obtained by fitting a linear model (weights constrained between -1 and 1) to predict the drop in MOS (also normalized 1). Both indicators show a strong influence of the flicker distortion phenomenon (Low and Middle frequencies) and both the colour opposition channel distortions (R-G and B-Y) on the overall quality. In case of temporal harmonics (or flicker) however, the increase in the energy of temporal harmonics in the distorted case as compared to the reference, reverses the sign of the feature value.

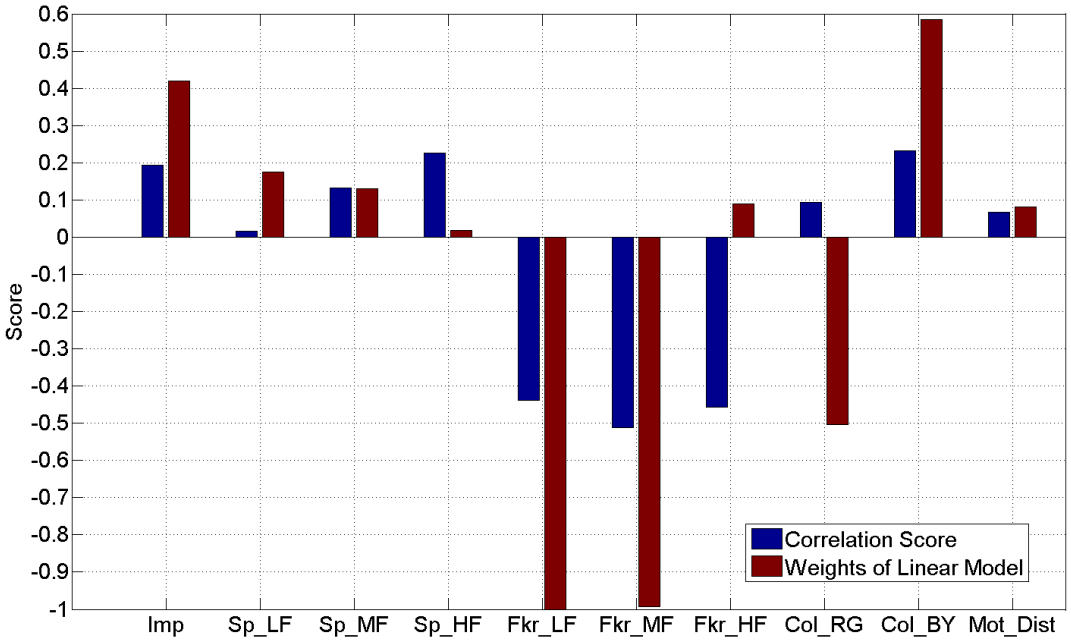


Figure 9.3 – Pearson correlation coefficients and the Linear model weights, of the normalized feature responses versus the normalized difference opinion score. Negative correlation indicates an improving difference opinion score with reducing feature response. Please refer to Section 3.3 to have more detail about the feature labels.

9.2.3.3 Effect of content features in the periphery

Perhaps the most important part of the analysis is to obtain a relation between the various features and the perceived drop in quality at *every individual eccentricity*. Fitting a separate linear model at every individual eccentricity shows us the importance of each feature as in Figure 9.4. While semantic importance, motion distortions and spatial low frequency distortions are important features that determine the drop in perceived quality in the fovea, flicker and colour opposition channel distortions are perhaps the most important quality indicators at six degrees of visual periphery. The colour opposition channel distortions in particular, maintain their importance at all eccentricities.

9.2.4 Discussion

We are able to observe that the perceived drop in quality across the visual periphery is closely related to the Cortical Magnification fall-off characteristics of the V1 cortical region. Additionally, we see that while object importance and low frequency spatial distortions are important indicators of quality in the central foveal region, the low-medium temporal distortions ($< 9.3Hz$) and colour distortions are the most important determinants of quality in the periphery. We therefore conclude that, although human observers are more forgiving of distortions they viewed peripherally, they are nevertheless not totally blind towards it: the effects of flicker and colour distortions being particularly important.
It is noteworthy that any single feature fails to produce a dramatically high correlation as seen in earlier studies [281]. The current work, however, attempts to study this multi-dimensional feature dependency, by projecting the quality scores separately into each individual feature dimension.
Because more than 90 percent of the information in a visual scene is incident on the peripheral areas, these results serve as important indicators to produce video content whose quality is perceptually optimum.

9.3 Subjective Evaluation of inpainting-based EC

9.3.1 Source video contents

In Chapter 3.4, the contents that are used in this section are shown. The 14 source sequences are in high definition (HD) with a resolution of 1280×720 pixels. The frame rate of the video sequences varies from 25 frames per second

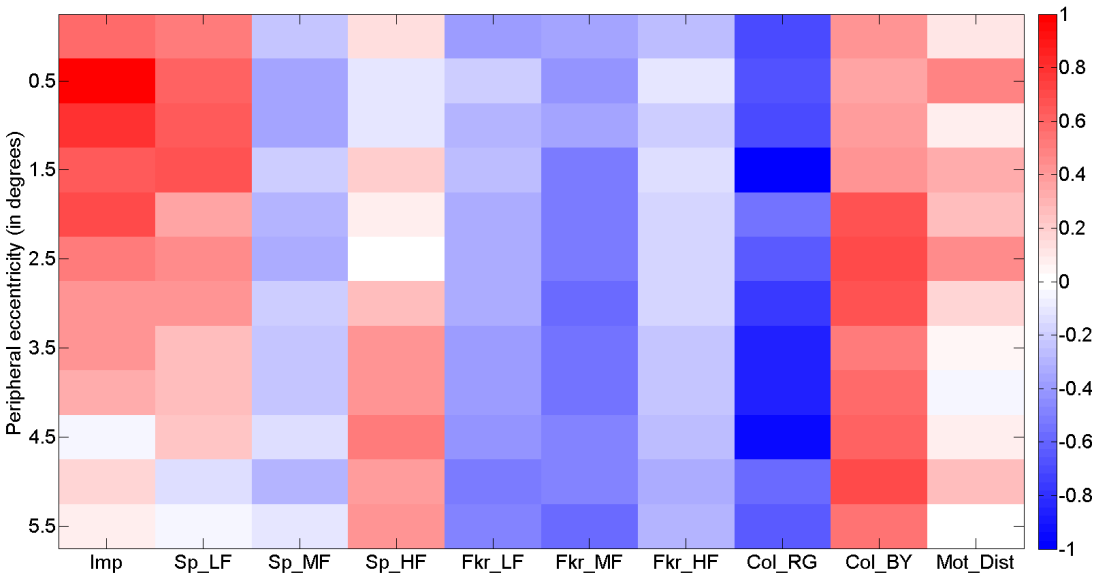


Figure 9.4 – Weights of the various features (constraint: $-1 \leq weight_i \leq 1$) in the Linear Model, that is used to predict the drop in normalized quality score. The vertical axis indicates the eccentricity at which the impairment was observed, and the horizontal axis indicates the different features that were used to predict the difference opinion score. Please refer to Section 3.3 to have more detail about the feature labels.

(fps) to 60 fps. Each sequence is 10 seconds long. Video sequences cover different video properties: motion intensity, camera motion type, spatial complexity, and colours.

9.3.2 Hypothetical reference circuit (HRC)

The 14 video contents are subject to loss-impairment and are concealed using two error concealment algorithms; the first one (HRC01) is the motion vector prediction and motion compensation using Bilinear Motion Field Interpolation (BMFI) [31]. The second EC algorithm (HRC02) is the inpainting-based EC technique that is illustrated in Chapter 8. Hence, we have 2 HRCs and one original pristine (HRC00). To sum up, 14×3=42 processed video sequences (PVSs) are generated. Each content has the same error pattern. The video sequences are divided spatially and temporally. Regarding the spatial division, each frame is divided to regions; top-left, top-right, bottom-left, bottom-right, and centre. The centre region left without distortion. Regarding the temporal division, the video sequence is divided into 5 time slots each is 2 seconds. The other slots are subject to loss-impairments in one of the spatial regions. The loss is 256×128 or 128×256. The order of appearing the loss is random and different from content to content.

9.3.3 Testing conditions

The experiment was designed according to ITU Rec. BT.500. [255]. The 5-point impairment scale was used to judge the annoyance of the distortions in the sequences. After each sequence viewing the observers has to choose one of the 5-point impairment scale: 'Imperceptible (5)', 'Perceptible, but not annoying (4)', 'Slightly annoying (3)', 'Annoying (2)', and 'Very annoying (1)'. Since all processed videos were affected by error insertions, the impairment scale is selected over the quality scale.

9.3.4 Subjective assessment

For each content, the stimulus is viewed one after another. The gaze information, eye-tracking, is recorded in the viewing using the SMI Hi-Speed eye-tracker operated in binocular viewing mode, providing 500 gaze samples per second. After the viewing of each PVS, and, of course, collecting the eye-tracking data, the observer is asked for his score for each PVS. A playlist for each observer is prepared taking into consideration that the sequences that belong to the same content are not viewed consecutively, and orders of the sequences are random. The viewing distance was 4.5H times the height of the screen. The experiment was explained to the observers using a training session prior to the test session. The test duration is about 50 minutes including training and breaks. The screen brand is Grundig FINEARTS 55 FLX 9490 SL with a 55-inch diagonal. The content is displayed on the centre of the gray screen. The ITU Recommendations BT.709-5 [254] and BT.500-13 [255] are followed to adjust the screen colour and brightness and to set up the testing room respectively. 24 non-expert observers participated in the experiment, 13 males and 11 females and the age average is 24 (19 to 30). A vision check is performed before the experiment using far and colour vision tests. Any observers with normal or corrected to normal visual acuity are allowed to do the experiment.

9.4 Analysis

9.4.1 Quality scores

9.4.1.1 Subjective scores

Figure 9.5 shows the mean opinion scores of the observers and the corresponding confidence interval (CI). The MOS of the undistorted sequences have a better quality than HRC01 and HRC02 except for Sources 4 (CrowdRun), 5 (ParkJoy), and 11 (ReadySteadyGo). Source 4 has a lot of faces and although there are distortions in the texture, the distortion does not disturb the subjects. Source 5 has different textures; trees, water, and grass. The distortion caused by HRC02 is not perceptible. The visible distortions in Source 11 do not annoying the subjects since it lies on the not of interest region.
Regarding the comparison between the two distortions; HRC01 and HRC02. HRC02 shows significance in quality with no/ignored overlap in CI in Sources 2 (Beauty), 3 (CampFireParty), 4 (CrowdRun), 5 (ParkJoy), 8 (YachtRide), 11 (ReadySteadyGo), and 12 (ResidentialBuilding). Figure 9.6 shows examples. It is observed that HRC02 shows significant quality improvements. On the other hand, HRC02 gives a higher MOS but without ignored CI with Sources 1 (BigBugBunny), 6 (TallBuilding), 10 (Library), and 13 (Marathon). Figure 9.7 shows examples. It can also be observed that HRC02 has better quality frame-wise but sequences-wise there are some overlaps in quality with HRC01. An even performance for the two HRCs is noticed on Source 14 (TrafficandBuilding), Figure 9.8. HRC01 shows better MOS than HRC02 without ignored CI on Sources 7 (Treeshade) and 9 (ConstructionField). Figure 9.9 shows that HRC01 performs better due to the lack of motion activity in the scene. In such situation, the main step that gives the inpainting-based EC technique an important value is the inpainting the background temporally. Without this step the spatial inpainting step will limit the performance of the technique.

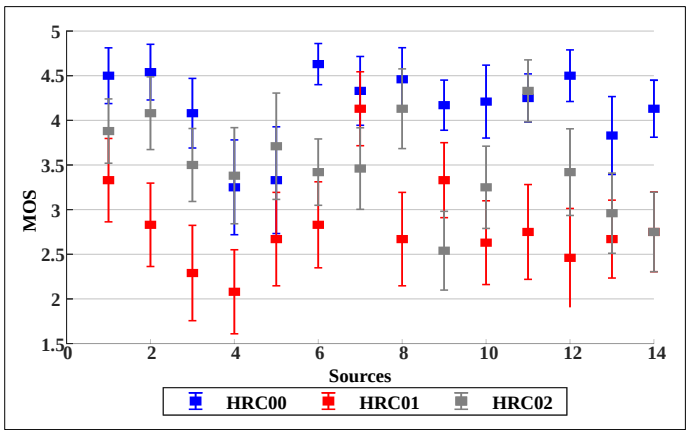


Figure 9.5 – MOS and CI for each content

9.4.1.2 Objective scores

The full-reference (FR) quality assessment tool helps gives a quality score for the distorted sequence if the reference (original) exists. In this subsection, PSNR, SSIM [61], MS-SSIM [62], VIF [245], and VMAF [282] are used to give the quality score of each distorted sequence despite the fact that these measures are not implemented for the loss-impairment sequences. VMAF (Video Multimethod Assessment Fusion) is a FR VQA tool provided by NetFlix. Figure 9.10 shows the correlation of each objective measure and the DMOS of the MOS scores for each HRC. All measures give a bad correlation with DMOS although the correlation in HRC01 is much better than the correlation of HRC02. As discussed in the previous subsection, the perceptual quality of HRC02 is much better than the perceptual quality of HRC01. It means that the error concealment using HRC01 makes kind of global distortion in the sequence which makes it easy to notice. While the error concealment with HRC02 makes a kind of local distortion that not always visible to observers and this kind of distortion make the job of FR VQA much harder. That explains the low correlation of all measure except for the VMAF since it employs the detail loss metric (DLM) [283] which measures the loss of useful visual information that affects the content visibility. It is clear that the traditional objective video quality measures are not sufficient to measure the quality of a sequence with local distortion. Following these observations, this chapter introduces in the next section the content-aware disturbance analysis that may be useful in order to model the subjects behaviour for each HRC.

Figure 9.6 – Examples when HRC02 outperforms HRC01

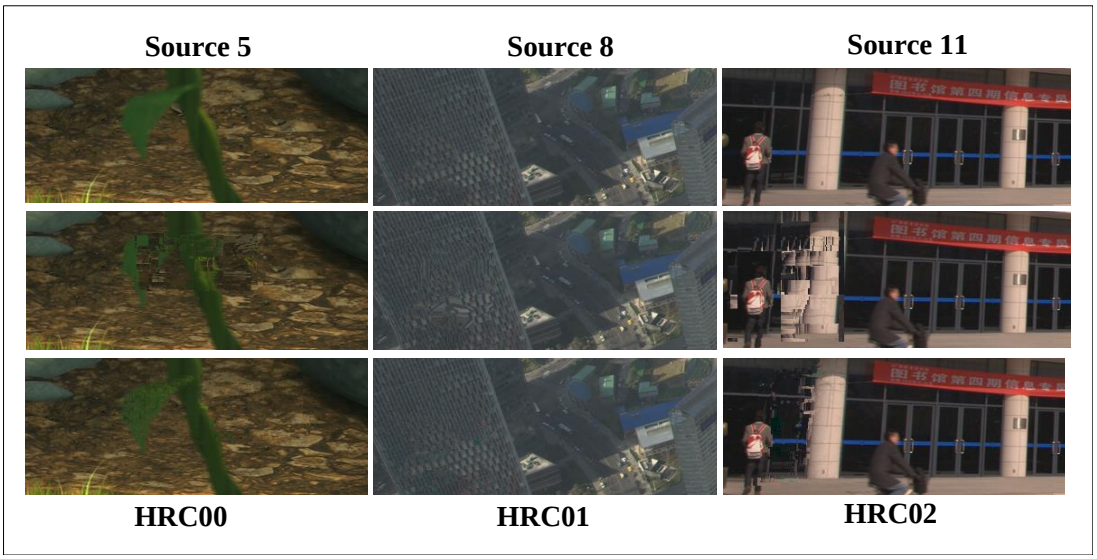


Figure 9.7 – Examples when HRC02 and HRC01 performances are questionable, but HRC02 is better in general

9.4.2 Disruption analysis

9.4.2.1 Defining Disruption

Although *gaze disruption* was previously defined as a sudden change in visual attention due to a certain unexpected event/characteristic of the video, this definition is mathematically insufficient to appropriately quantify and identify

Figure 9.8 – Examples when HRC02 has the same quality of HRC01

Figure 9.9 – Examples when HRC02 and HRC01 performances are questionable, but HRC01 is better in general

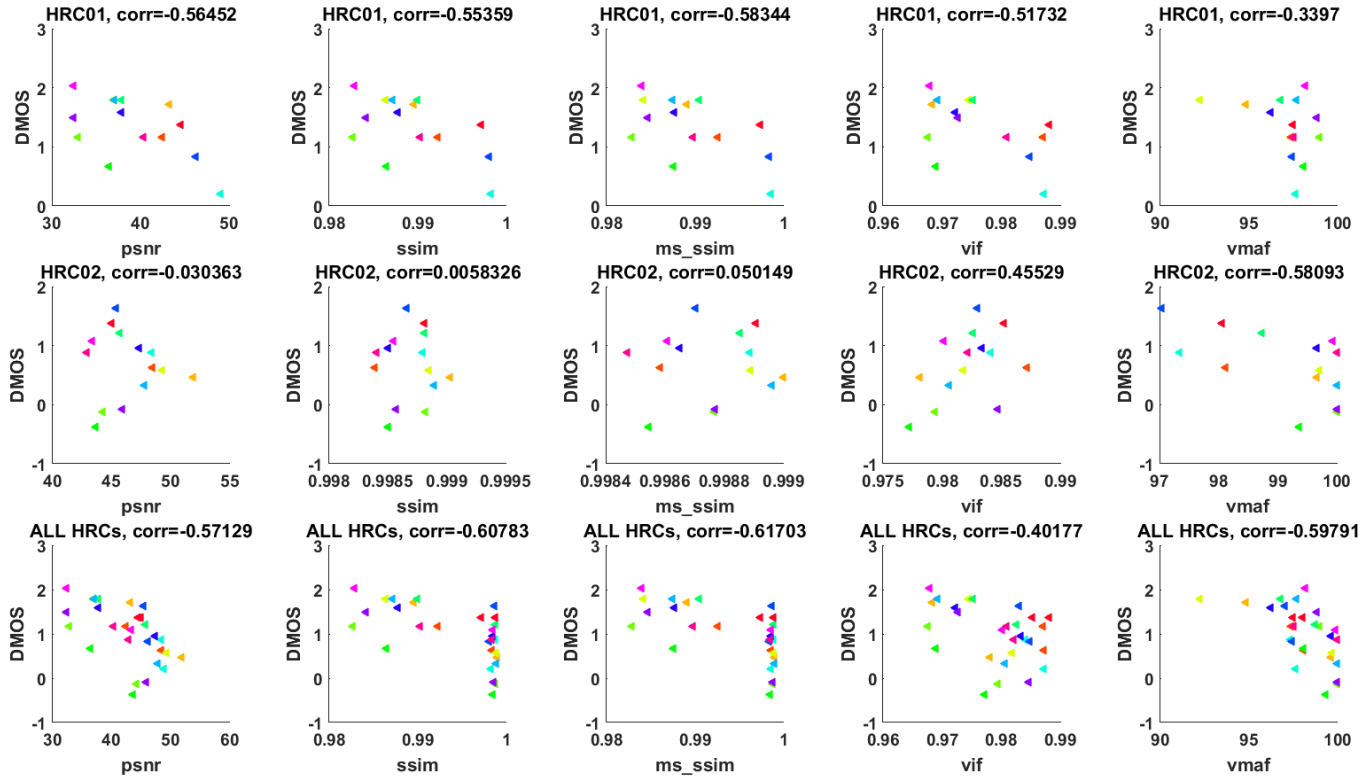


Figure 9.10 – Objective scores for each HRC and per metric type; PSNR, SSIM,MS-SSIM,VIF, VMAF

disruption. Hence this subsection aims to provide a basic definition of disruption. For more details, reader is advised to read [284]. Before defining disruption, it is essential to quantify and define two important measurements extracted from the eye-tracking data: the eccentricity of viewing just before the distortion is presented (known as the *Relative position(RP) of initial gaze*) and the eccentricity of viewing after the presentation of the transmission impairments (known as *Relative position(RP) of saccadic target*). This is shown in Figure 9.11 and serve as an important indicator of disruption. Eccentricity in this context is defined as the shortest distance (in degrees) between the point of gaze and the impaired region in the video. As disruption is defined as the change in the attended location due to the presence of a special signal, it is especially important to consider cases with a large RP of initial target and a small RP of

saccadic targets. Such a scenario indicates that the annoyance due to the impairment possibly caused the observer to shift his attention from some other point in the video towards the area the impairment has occurred. It is, however, possible that the observer executed this saccade with a natural exploratory intention and hence statistical tests are required to ascertain the purpose of such a saccade.

Impairments often disturb the natural viewing behaviour of an observer, in turn strongly affecting the RP of the

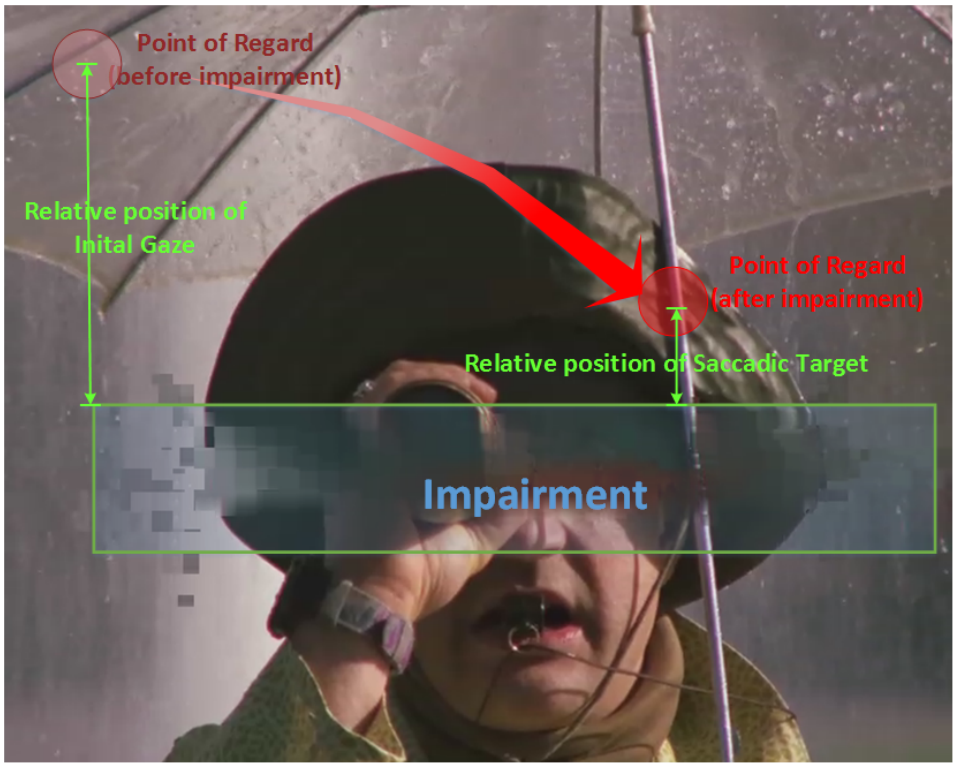


Figure 9.11 – Point of initial gaze refers to the region that the subject was initially looking at before the distortion appeared. On the other hand, saccadic targets refer to the region where the user shifted his gaze, as soon as the distortion was presented

.

saccadic target. Examining the probability of an observer being drawn towards the impairment therefore serves as a measure of disruption. Assuming that a viewer was in a RP of initial gaze X and that the impairment makes him saccade towards a RP of saccadic target say Y, we define disruption D as the probability that the saccadic amplitude $X-Y$ caused by the impairment is greater than a finite threshold δ, when examined at every possible X ranging from δ to the maximum possible viewing angle A_{max}, as shown in equation 9.1

$$D = \sum_{x_i=\delta}^{A_{max}} p((X-Y) > \delta | X = x_i) p(X = x_i) \tag{9.1}$$

A_{max} can be normally derived from the maximum display dimensions that are under consideration considering that the user cannot execute a saccade that is greater than this amplitude. Note that the summation starts from a finite δ because, it is impossible to measure disruption, if the observer was already in the impaired location even before the impairment was introduced. Hence we neglect such cases and start the summation only from a certain offset : δ. It can also be deduced that disruption can be completely and sufficiently deduced by just examining X and $X-Y$ at every possible $X = x_i$.

9.4.2.2 Disruption correlation with DMOS

The disruption D is defined in the previous subsection. Hence, this disruption is measured for each content in HRC1 and HRC2 once an error occurs.

Figure 9.12 shows the subjective scores (DMOS) against disruption (D). The measured disruption correlates well with the subjective scores. The Pearson Linear Correlation Coefficient is calculated and the value of 0.899 is recorded. In the Figure, the blue and red points represent HRC1 and HRC2 respectively. It can be noticed that the sequences, that are concealed using HRC1, have higher entropy value than HRC2. This high entropy value is an indicator of insufficiency of the EC technique to recover the lost texture and structure especially for sequences of high motion

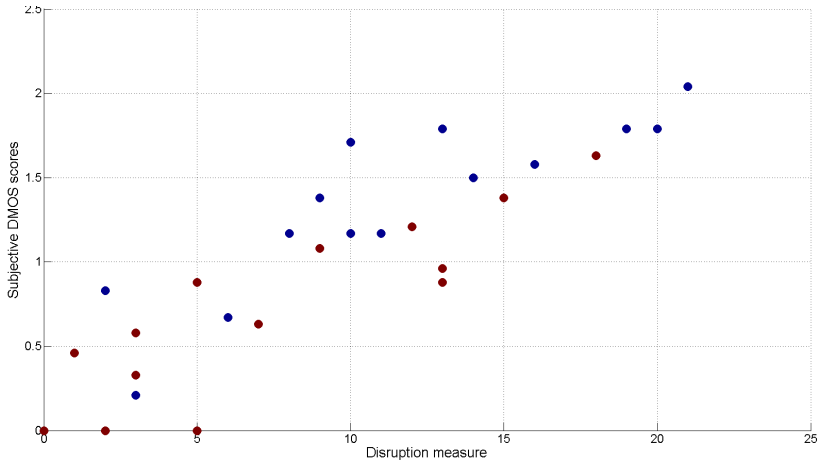


Figure 9.12 – Subjective score (DMOS) against disruption (D). The correlation is 0.899. The blue points for HRC1 and red points for HRC2.

intensity and for sequences of high spatial details.
Moreover, it can be observed that there are sequences (3, 5, and 10) have $DMOS = 0$ and a small amount of disruption. It means that there are few subjects who noticed the distortion.

9.5 Content-aware disruption analysis

The first among the many features to consider in the disruption model is the effect of the content. It is plausible that the saccades, or more specifically saccadic disruptions in the gaze-data can arise due to the nature of the content itself and hence, it would be naive to disregard the effects of the content when assessing the effect of disruptions. The following content features are used, in this experiment, to analyse the subject gaze patterns and scores: Texture entropy map, Section 3.3.2, colour entropy map, Section 3.3.3, and motion entropy map, Section 3.3.4.

9.5.1 Role of Entropy

Morandi et al [285]., have highlighted that while unusual details and unpredictable contours in the picture results in shift of attention towards them and a high concentration of fixations, boring textures and predictable regions often seem uninteresting. They conclude that informativeness of a region, as determined in terms of its *recognizance* had a huge role to play in the final density of the fixated regions.
This principle has been used in various forms by many other studies like [286, 287]. In [286] for example where entropy gain along a feature dimension is measured by the *Incremental Coding length*. Their basic principle was to ensure that their system must respond placidly to common stimuli and be alert to anomalous ones, in line with the above principles. Further, evidence for such an approach has also been found in the *Sparse coding theory* in the visual cortex region. Measurement of large coding length increments helps the computational system achieve attention selectivity in any space: texture, colour or temporal, based on the feature dimension, because such a system responds very aggressively to frequently activated features. A very similar strategy is followed in the models of [287], where saliency is defined as *given the surrounding area, it is the minimum uncertainty of the local region (namely the minimum conditional entropy)*. They especially define conditional entropy in their work, as the lossy coding length of Multivariate-Gaussian data.

9.5.1.1 Calculation of entropy

Entropy in the present work is calculated within the *short-term spatio-temporal* construct. At every pixel of every frame, we consider a short term filtering window of 2 degrees of visual angle (120 pixels at standard viewing distance) and 500ms temporal extent centred at that pixel as shown in Figure 9.13. In the Figure, we see an example of a cylindrical spatio-temporal window centred at the considered pixel. In case we intend to keep the calculations causal, we might additionally consider the past temporal neighbours alone, as opposed to considering the candidates from the future, as well. All the feature values (like motion, colour for example) within this short term tube are collected together and used (jointly) for the calculation of entropy.
It is believed that such an entropy calculation within every small window will point out the anomalies and disturbances in the flows (possibly due to a distortion) along a certain feature dimension.

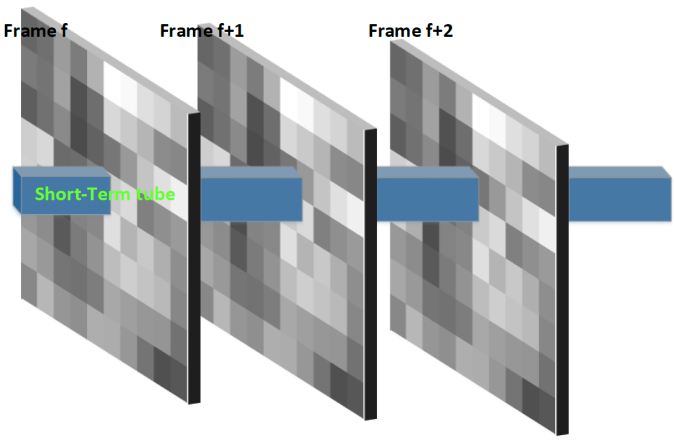


Figure 9.13 – Calculation of entropy occurs over a tube that encompasses a spatial neighbourhood and several temporal neighbours

9.5.1.2 Calculation of Differential cumulative sum of entropy (CSE)

The following steps are followed to calculate an objective value that represents the difference in the entropy maps of reference and distorted sequences.

- once an error occurs, a set of n presumed scan-paths are generated as explained in [284].
- for each scan-path pair (reference, distorted),
  - calculate the cumulative sum of entropy (CSE).
  - calculate the cumulative difference (CDE) of entropy as in 9.2
- calculate the average per sequence for the sequence.

$$CDE(n) = CSE_{ref}(n) - CSE_{dist}(n) \tag{9.2}$$

9.5.2 Texture entropy map

Section 3.3.2 describes how we calculate the entropy for each pixel in the video shot. Then, the CDE is calculated for each sequence. Figure 9.14 shows the subjective score (DMOS) against logarithmic scale of texture CDE. The Pearson Linear Correlation Coefficient is calculated and the value of 0.63 is recorded. Although this correlation is not high, the texture entropy map can highlight the importance of texture features, i.e. this entropy map can be improved. Besides, this indicator correlates better than the objective measures.
It is observed that the texture CDE of sequences that are concealed using HRC1 is higher than those that are used HRC2 except for sequence #7 (TreeShade) and sequence #8 (YachtRide). This observation is consistent with the perceived MOS scores for sequence #7. This sequence has no camera motion and very small motion which make it so perfect for HRC1. For sequence #8, the inpainting EC (HRC2) caused a high change in the texture properties especially the dynamic texture. This change is captured by the objective CDE but not the subjects.

9.5.3 Colour entropy map

Section 3.3.3 describes how we calculate the entropy for each pixel in the video shot. Then, the CDE is calculated for each sequence. Figure 9.15 shows the subjective score (DMOS) against logarithmic scale of colour CDE. The Pearson Linear Correlation Coefficient is calculated and the value of 0.7 is recorded. Although this correlation is not high, the texture entropy map can highlight the importance of texture features, i.e. this entropy map can be improved. Besides, this indicator correlates better than the objective measures and the texture entropy map.
It is observed that the colour CDE of sequences that are concealed using HRC1 is higher than those that are used HRC2 except for sequence #7 (TreeShade) and sequence #8 (YachtRide). This observation is similar to the CDE of texture. What is different is that the rank/order of sequences texture and colour CDEs is different. This could be one step towards introducing a hybrid entropy measure.

9.6 Conclusion

The main contributions of this chapter are listed in Box 9.3. In this chapter, the subject disruption for an existing database and for the inpainting-based EC algorithm are analysed in order to know if such disruptions are related to the perceived quality. The visual disruption and MOS scores are calculated. As expected, traditional objective metrics struggled to even achieve a strong correlation, while visual disruption achieved a correlation of 0.899 with the recorded

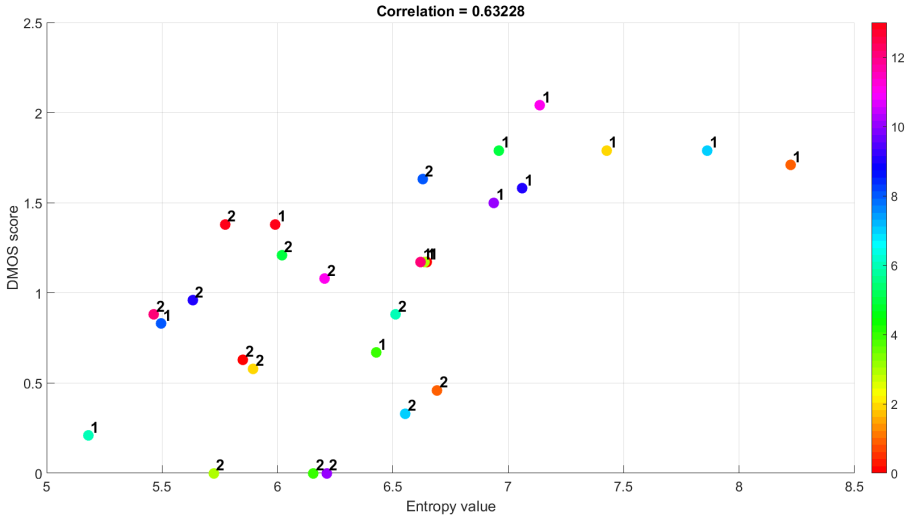


Figure 9.14 – Subjective score (DMOS) against logarithmic scale of texture CDE. The correlation is 0.63. The colour bar refers to the content and the number on each point refers to the HRC number.

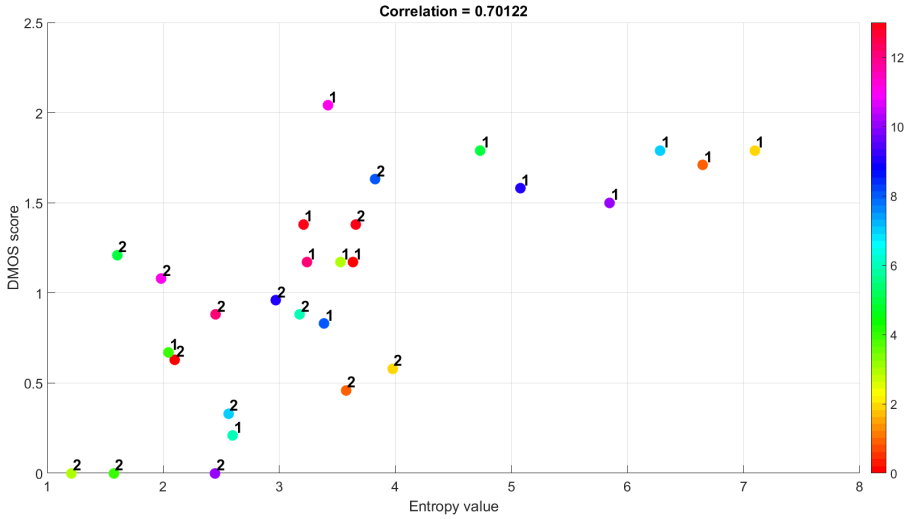


Figure 9.15 – Subjective score (DMOS) against logarithmic scale of colour CDE. The correlation is 0.7. The colour bar refers to the content and the number on each point refers to the HRC number.

opinion scores. These findings support the claim that the disruption is a good indicator of the perceived local quality. The proposed content features (texture, colour, and motion entropy maps) are good steps to be further analysed by the VQA researchers to implement and introduce objective video quality measurement for loss-impaired sequence.

Box 9.3 – Contributions

- It is observed that the disruption measure has a high correlation with the perceived DMOS. In addition to that, it is shown that the inpainting-based EC technique achieves a better perceived quality with respect to one of the-state-of-the-art EC techniques.
- Three content features are introduced to study the subject disruption as one step forward to help measure the quality of the perceived degraded videos. The features are: texture, colour, and motion entropy maps.

Role of measured content characteristics in quality assessment

10

Influence of content and coding conditions on different full-reference video quality measures

10.1 Introduction

Typical industrial video distribution chains may continuously monitor the video quality at several processing steps, at the camera capture, on the contribution channel to the studio, for the distribution to the customer, and finally at the customer side. In this work, the application focus would be on those parts where a reference video is available for comparison to a degraded video using Full-Reference (FR) video quality measures and measurement needs to be performed in real-time, potentially on low-performance network equipment. The reference video may either be available explicitly, for example as input to an encoder step, or implicitly, for example using a (camouflaged) test video during a regular operation. A huge number of FR algorithms have been developed and are still in development by researchers in industry and academia ranging from very low to very high computational demands. The evaluation of these methods is usually performed by comparing their prediction performance to ground truth data obtained in subjective experiments, a typical example being the validation experiments by the Video Quality Experts Group (VQEG) [288] that led to several Recommendations of the International Telecommunication Union (ITU-T J.144, J.247, J.341).
Performance evaluation by subjective experiments may be seen as mandatory and thus necessary but not sufficient due to the limited number of test cases with respect to the above-mentioned application scenario. It is usually considered as ground truth for the training, verification, and validation of objective video quality measures. The number of test cases that may be obtained in subjective assessment is however limited. Less than 200 video sequences of about 10 seconds in one session may be evaluated when using one of the most efficient methods, Absolute Category Rating [194]. In addition, with a reasonable number of observers, only about 75% of the test cases are pairwise distinguishable with confidence intervals around 0.3 on a five-point scale [289]. Even with recent collections of available databases, notably the Qualinet Database [290], the choice of available annotated databases for a particular usage scenario stays limited. Automatic performance analysis only using objective measurements provides a complementary approach. Two alternatives shall be mentioned here. First, the creation of dedicated test sequences in which the performance is expected to be known a priori such as increasing strength of a distortion [291]. The second possibility is to create and evaluate successively a large-scale database [240, 292]. A review of the large-scale database can be found in 3.4.4. It is complementary, as it may be expected that performance anomalies, such as outliers, may be detected that may be missing in the limited selection performed for subjective assessment. In [293], Ciaramello and Reibman explained a similar approach for image quality predictors by creating a large amount of test images with specific degradations.
This chapter is aiming to use the large-scale database in order to conduct analysis and see observations that cannot be obtained with the subjective experiments. Hence, in this Chapter, we follow the research questions listed in Box 10.1. The structure of this Chapter is outlined in Box 10.2

Box 10.1 – Goals

This chapter aims to answer the following research questions:

— How do different full-reference video quality measures behave in terms of ranking for the error-free and loss-impairment sequences?

— Characterize the behaviour of FR video quality measures at frame and sequence levels with respect to video content and coding parameters.

Box 10.2 – Chapter structure

This chapter will be organized and structured as shown in Figure 10.1. In Section 10.2, the behaviour of the quality measures in terms of ranking/agreement will be studied. In Sections 10.3.1 and 10.3.2 characterize the impact of video content and coding parameters on the agreement of FR quality measures at frame and sequence levels, respectively.

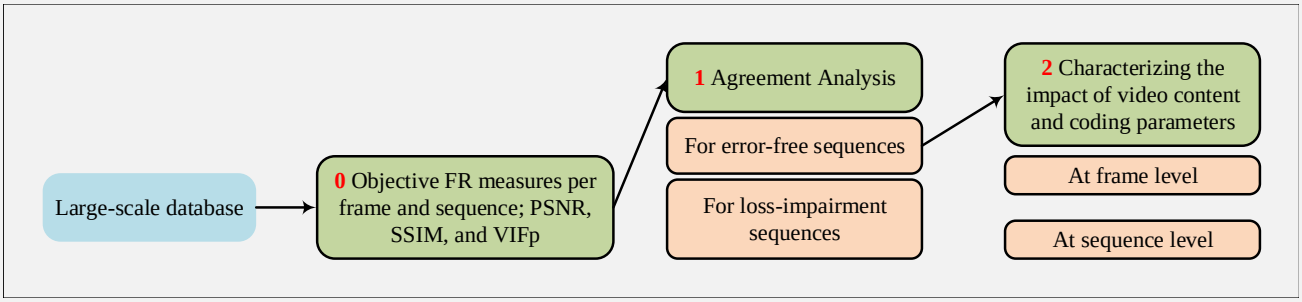


Figure 10.1 – Chapter 10 Structure

10.2 FR measures agreement for loss-impaired sequences

The focus of this section is on the characterization of three well-known objective video quality predictors in terms of ranking agreement and distance with realistic coding and lossy network transmission conditions. On other words, we study the stability of the measures with the ultimate goal of better understanding the intrinsic limits of the measures themselves. To this aim, we present an extension to the database [292] which adds a large number of objective quality evaluations when compressed video streams are subject to data loss. It shall be noted that none of the three measures was specifically designed for measuring degradations due to packet loss, notably concealment artifacts and time varying quality. However, they have been used repeatedly in the literature in order to measure such scenarios regardless of such considerations. Transmission degradations have therefore been considered as being at least in the extended scope of application: prediction with a limited accuracy was expected using these measurement algorithms. Having several measures for the same video sequence naturally yields to the question if such measures are consistent in ranking. In other words, given two processed video sequences (PVS), do all the measures agree about which is the one with the highest quality score? This agreement can be expressed mathematically as in Eq. 10.1. The underlying idea is that if one or more of these measures do not agree, this condition should deserve further investigation. In [240], a similar approach is used when dealing with PVSs that do not contain data loss impairments.

$$Agreement = \begin{cases} 1 & |\sum_{Q\in\{PSNR,SSIM,VIFP\}} sign(Q(A)-Q(B))| = 3 \\ 0 & else \end{cases} \tag{10.1}$$

Table 10.1 reports results for the nine sources considered in this work. The first data column shows the percentage of comparisons with disagreement among all comparisons within the same source (i.e., 1,230,055,200 pairs). The next three columns show how many cases (as a percentage out of all disagreement cases) can be ascribed to each measure (PSNR, SSIM, VIFP). These results can be directly compared with the ones in [240] showing that, in the considered scenario, the loss impairment tends to increase the amount of disagreement. Moreover, for most sources the share of disagreement attributed to PSNR increases compared to the case in [240] (i.e., no losses), whereas there is a decrease for the VIFP measure. This behaviour may suggest that some measures are more influenced (and perhaps are more sensitive) to loss impairment than others.

However, the previous results should be taken carefully since very limited variations around the equivalence between

Table 10.1 – Reasons of disagreement among quality measurements for each sequence. *src09* is not included due to the PSNR issue: infinite values for some encoded frames are present.

Source	% of disagreement	% due to PSNR	% due to SSIM	% due to VIFP
src01	12.74	38.81	41.60	19.59
src02	4.29	61.37	23.97	14.66
src03	12.07	45.47	26.42	28.11
src04	10.41	57.51	22.55	19.94
src05	4.11	47.26	32.27	20.47
src06	9.98	71.81	12.43	15.76
src07	5.64	65.27	11.89	22.84
src08	5.46	59.19	19.73	21.07
src10	12.44	46.67	32.12	21.21

the pairs could yield disagreement that may be due to potentially tiny modifications of the characteristics of the PVS. Therefore, for each algorithm we introduce a normalized difference by linearly rescaling the results in the interval [0..1]. Normalized values are denoted by the hat symbol (e.g., $\widehat{PSNR}$). Then, the individual differences of all the measurements for a sequence pair are combined in a single normalized difference $\hat{d}$ by using the Euclidean distance:

$$\hat{d} = \sqrt{\Delta\widehat{PSNR}^2 + \Delta\widehat{SSIM}^2 + \Delta\widehat{VIF}^2} \tag{10.2}$$

so that the results can then be plotted in one dimension using a histogram for each source, as suggested in [240]. Figure 10.2 presents a sample histogram showing the reason of disagreement as a function of the normalized difference for *src03*. As expected, the amount of disagreement decreases as the distance increases. Moreover, the figure shows a smooth reduction trend for all the three considered measures, though the share is generally higher than the one in [240]. This result is in part different from [240], in particular for *src03*, where the share strongly fluctuates as the distance increases. This effect was described as being potentially attributed to the characteristic of the sequence content.
Since the share of agreement increases more slowly compared to [240], we plotted the histograms for a larger distance, as shown in Fig. 10.3 for *src01* and *src06*. For convenience, a black vertical line shows the point up to which Fig. 10.2 and all histograms in [240] have been plotted. It can be noticed that the disagreement spans over a larger interval of normalized distances: this is expected since loss impairments typically have stronger influence on the measurements. However, strong differences in the reasons of disagreement can be observed: *src01* is dominated by SSIM disagreement, whereas PSNR is the main reason of disagreement for *src06*. Therefore, content characteristics appear to play a significant role in this regard. The analysis presented in Figure 10.3 shows that our approach can help investigate the

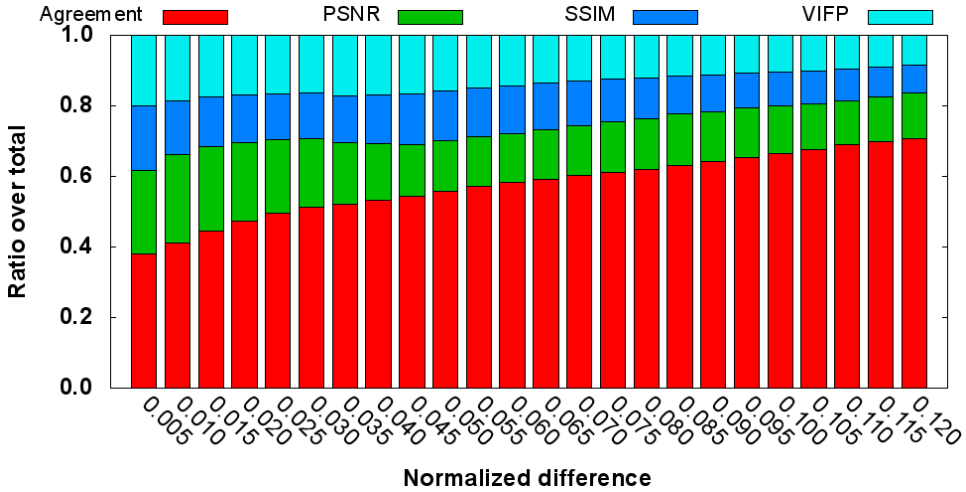


Figure 10.2 – Reason of disagreement (expressed as a ratio over the total pairs) between the various algorithms as a function of the normalized difference for *src03*.

stability of video quality measures even when a large number of test sequences are involved, which makes subjective assessments impossible. However, we remark that our aim is not substituting subjective assessment. On the contrary,

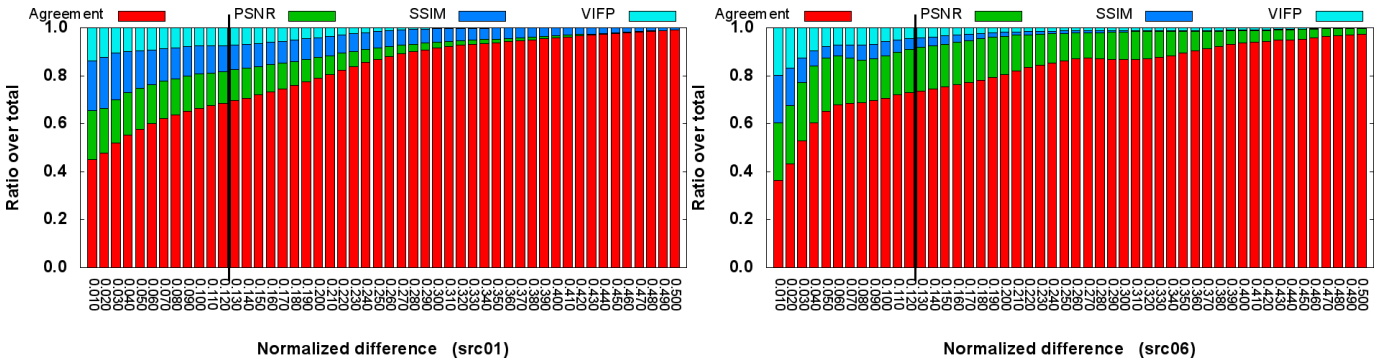


Figure 10.3 – Reason of disagreement (expressed as a ratio over the total pairs) between the various algorithms as a function of the normalized difference for *src01* and *src06*. To simplify comparisons, the black vertical line shows the point up to which Fig. 10.2 and all histograms in [240] have been plotted.

we aim at identifying potential shortcomings in terms of, e.g., stability and agreement, of existing video quality metrics that could not be investigated without resorting to large scale assessment. Interesting conditions and outlier situations can be identified, further studied, and analysed. We remark that such results can only be achieved by means of using a large scale database. In particular, we noted how some peculiar observations on the behaviour of metrics such as PSNR are due to the use of a large number of combinations of coding parameters.

10.3 Impact of content and coding condition in FR agreement consistency

Analysis methods and preliminary conclusions using such agreement analysis were proposed by the authors for coding and packet-loss scenarios [240, 294] and discussed in previous sections. The analysis used either pairwise comparisons or additional indicators that were fitted either to improve coherence or to analyse the behaviour of the measures. Using the same type of analysis, this section proposes an evaluation of objective measurements that is difficult to achieve in subjective assessment: characterization of single frame prediction performance in the context of a video sequence. By frame-wise analysis, important insight may be gained concerning the scope of application of a measure, for example regarding suitable temporal pooling strategies, i.e. required smoothing for outliers or rate-distortion applications. In the latter case, different distortion measures may be considered in order to improve the smoothness of the perceived video quality optimization. Due to the still limited size of the current large-scale database and the selection of the objective measures dictated by the available processing power, this study focuses on presenting innovative analysis methods rather than generalizable results. Two types of analysis are shown as depicted in Fig. 10.4. The first, pairwise ranking comparison of consecutive frames as a measure of coherence is presented in Section 10.3.1. The second is introduced in Section 10.3.2 in which different source videos and coding parameters are frame-wise compared providing insight into influence of content and coding structure decisions on coherence.

10.3.1 Consistency measure on consecutive frames

In this section, the continuity of agreements and disagreements is analysed within one HRC, i.e. within one coding condition, for each source content. The continuity of agreements is measured in a sequential manner as formerly described in Eq. 10.1 and as illustrated in Figure, 10.4 with solid arrows. Once the disagreement is happening between frames (A and B), there is high probability that frame B disagrees too with frames before A. For the given large-scale database, there are 5952 HRCs for each 250-frame source. Hence, a 5952x249 agreement/disagreement matrix is calculated for each source. Then, for each of these sources, the columns are summed and divided by the number of HRCs. This type of analysis shows the temporal behaviour of different objective video quality metrics, namely PSNR, SSIM, and VIFP. Fig. 10.5 shows the variations of the number of disagreements over time for two source contents: source number 6 and 10. The darker the bar for a particular frame, the higher the fraction of disagreement between the frame represented on the X-axis and its previous frame.
It is difficult to make general interpretations from such an overall analysis. A better strategy is to consider agreement with respect to the different sources or coding conditions as described in the following subsections as well as in Section 10.3.2. On the other hand, what can already be observed in this high-level analysis is that the highest number of agreements (white peaks) are happening when agreement between video quality measures is calculated between Intra-frames and their successive/preceding Inter-frame. Further analysis of the data reveals that this is because the Intra-frame has a notable higher quality compared to next/previous Inter-frame from the quality measure point of

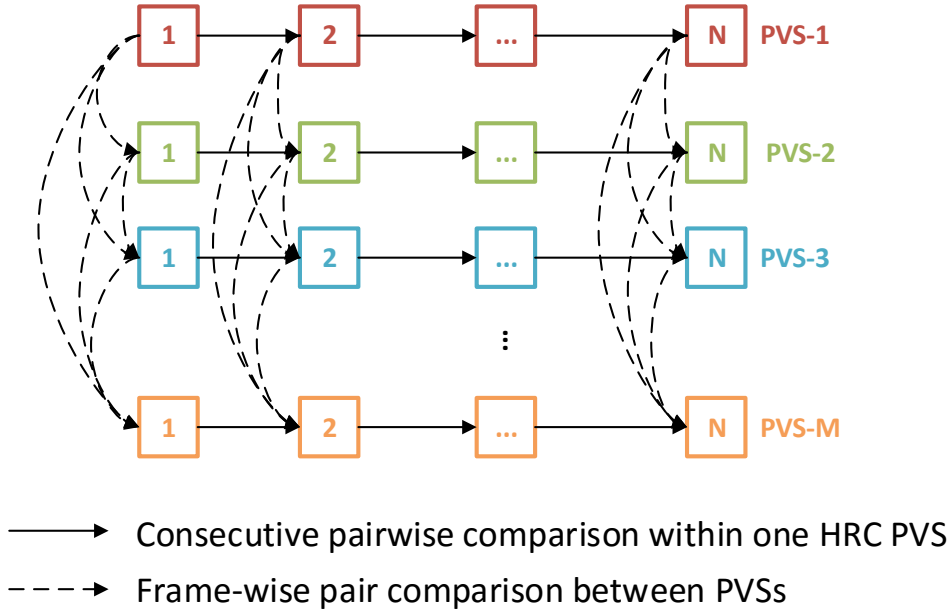


Figure 10.4 – Illustrations for the two types of analysis that are demonstrated in Section 10.3.

view. The used encoder configurations imply a higher quality to the Intra-frame compared to Inter-frames such that all measures easily agree on which frame is highest in quality. Thus, when measuring improvements for upcoming algorithms, it is advantageous to compare all available objective measures with respect to the content in order to provide a thorough analysis of the proposal.

10.3.1.1 The impact of content

When analysing the data in more detail, the influence of the content types and characteristics clearly appears. From the data, it can be observed that the number of disagreements varies from one content type to the other. In Fig. 10.6, the fraction of disagreement for each quality measure is displayed. It can be observed that the contribution of each quality measure to the overall disagreement is very clear. The majority of disagreement in SRC3 is due to PSNR, while the majority of disagreements in SRC10 is due to SSIM, the figure is not presented here due to space limitations. From these observations, it can be concluded that depending on the type of the source content, PSNR, SSIM, and VIF can act differently. Thus, when measuring improvements of algorithms, it is advantageous to compare all available objective measures with respect to the content in order to provide a thorough analysis.

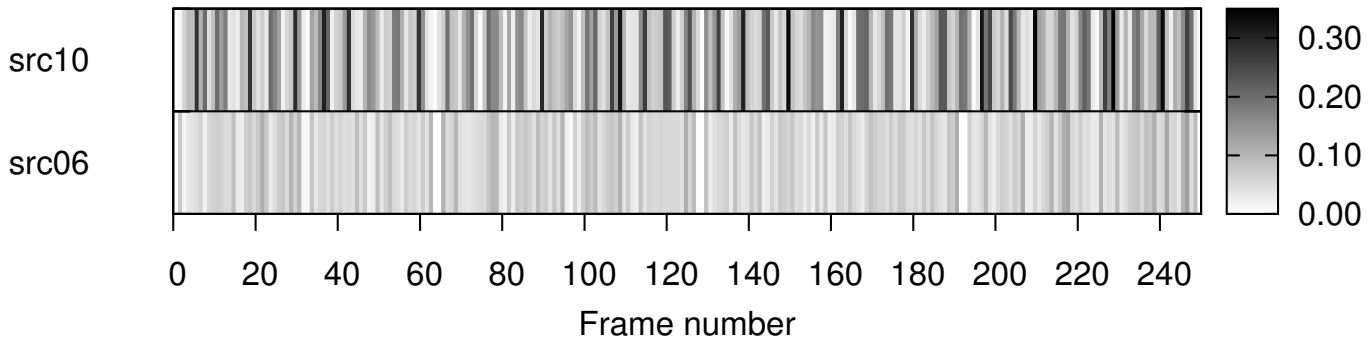


Figure 10.5 – The variations of the number of disagreements over time for two source contents: source number 6 and 10.

10.3.1.2 The impact of Intraperiod

As mentioned in the high-level analysis, the Intraperiod is a very important factor in understanding the temporal behaviour of the quality measures. Fig. 10.7 shows this effect. It demonstrates the variation of disagreements of HRCs of SRC6. It is obvious that the disagreement fractions between Intra-frames and next/previous frames are very low

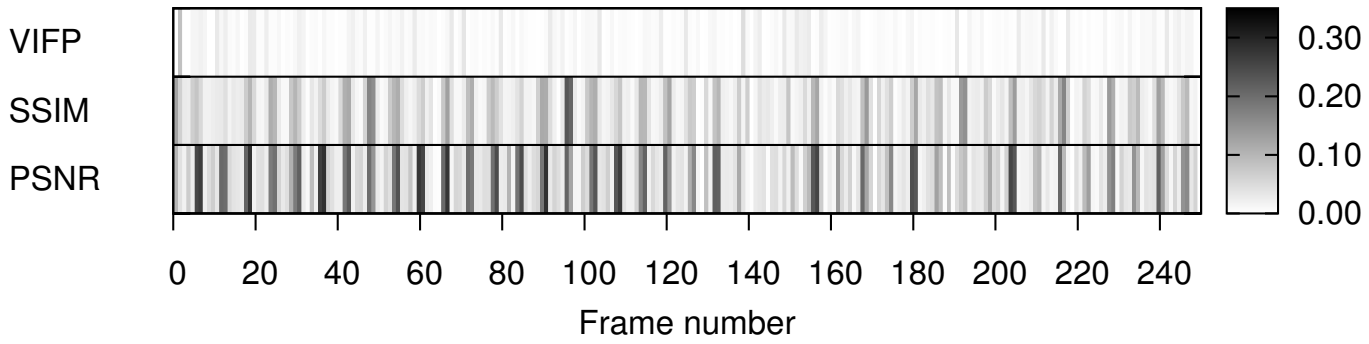


Figure 10.6 – The cause of disagreements in SRC3.

compared to other frames. Similar observations can be made for all Intraperiods (8, 16, 32, and 64) and also for the other source contents. The capability of the quality measures to agree when comparing two frames of notable difference in quality is the main reason for this phenomenon. Hence, when a source is encoded with coding conditions that only differ in the Intraperiod, temporal pooling strategies for calculating the video quality score may be examined and this effect may be taken further into account.

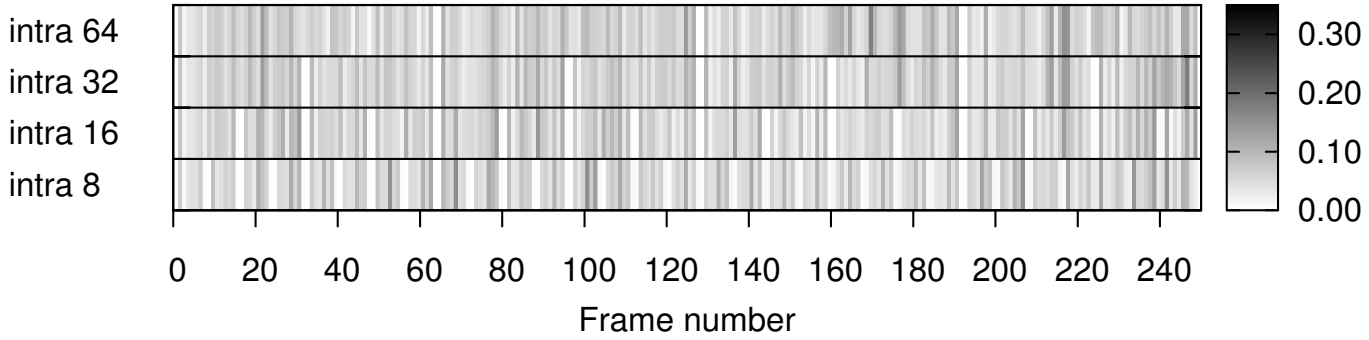


Figure 10.7 – The variation of disagreement fractions of HRCs with Intraperiod of 8, 16, 32, and 64 of SRC6.

10.3.1.3 The impact of GOP structure

Low-delay and the hierarchical structure of GOP configurations are widely used in different application scenarios. In this work, the consistency of quality measures is categorized to show the role of hierarchical GOP structure with different sizes and a low-delay configuration of size four. Fig. 10.8 shows this role for SRC6. The number of disagreement in the low-delay configuration is higher than the number in the hierarchical coding structures. This observation stands for all source contents except for SRC3. This behaviour of low-delay might be due to its configuration of using not only the previous frame but also -5,-9, and -13 frames relative to the first frame of the GOP. Moreover, in low-delay there is only one layer and the quality of the inter-frames are very similar, which yields a high inconsistency between the quality measurements.

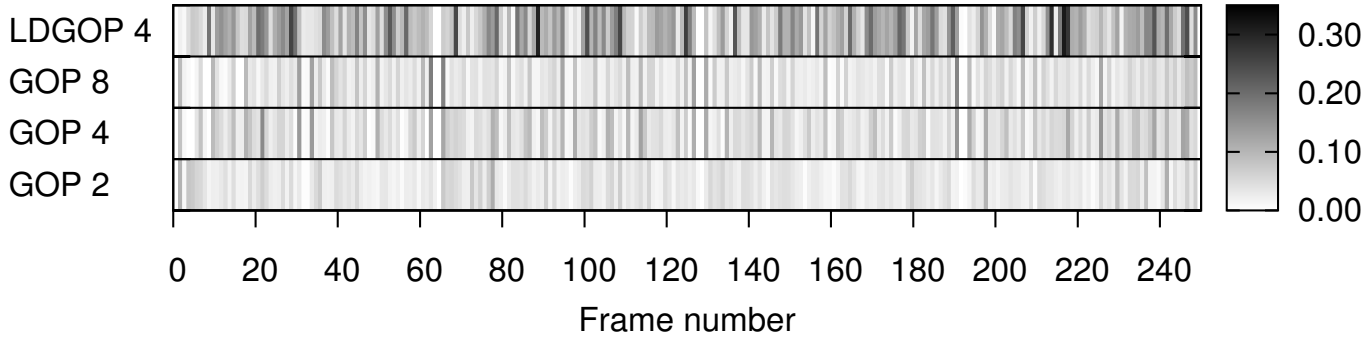


Figure 10.8 – The variation of disagreement fractions of HRCs with hierarchical GOP of 2, 4, and 8 and the low delay of 4.

10.3.1.4 The impact of QP and rate control

An interesting observation can also be made for the impact of using a constant quantization parameter or a rate control configuration. Very low and very high disagreement fractions periodically alternate at the beginning and the middle of the GOP while this is not observed when rate control is used. Fig. 10.9 shows this observation for SRC7. In this source, the fraction of disagreements for some frames is higher than 50% when constant QP is used while it is not the case for the rate control option.

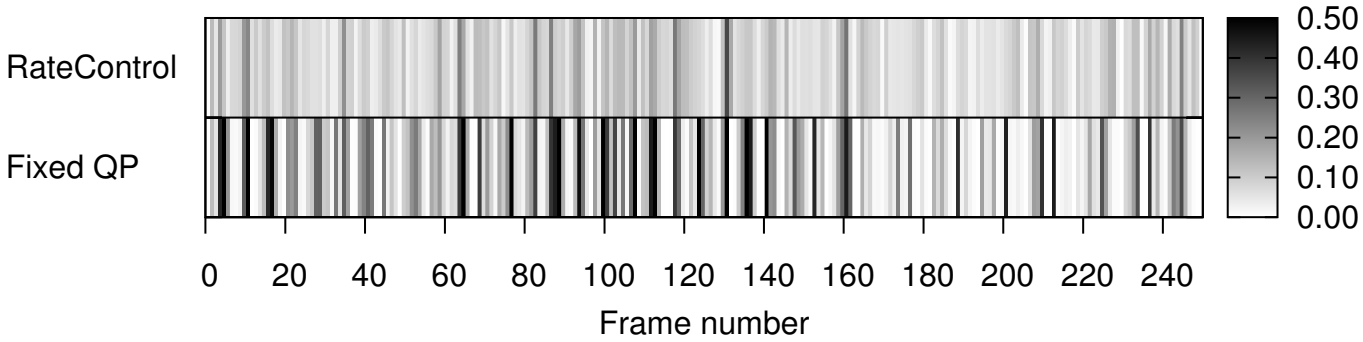


Figure 10.9 – The variation of disagreement fractions of HRCs with constant QP and rate control options.

10.3.2 Consistency with respect to source content and coding parameters

In this section the agreement of the measures is analysed across PVS, i.e. by considering two PVS at a time and comparing their quality measure values for each single frame (see the dashed arrows in Fig. 10.4). Consider, for instance, the sequence-level values of each of those metrics for two different PVSs. Two cases are possible: either all the measures agree (Case *Agree*) on which PVS provides the best quality, or they do not agree (Case *Disagree*). From this point, we only consider Case *Agree*, and we investigate if such an agreement at the overall sequence level corresponds to agreement for single frames as well.

First, we observed that, for sequences for which the quality is strongly different, typically there is agreement at the frame level, i.e., comparing the measures for frames in the same position in the two sequences yields to agreement among the measures. However, when the quality difference is less pronounced, even in Case *Agree*, for some frames in the sequence there is no agreement for frames in the same position. For the purpose of this work we consider only sequences for which the agreement holds for more than 90% of the frames (Case *Agree90*). The rationale behind this choice is that when a new coding and/or processing technique is proposed, typically quality values for the overall sequence are presented to show that the new technique is better than some reference. In absence of further information, such form of presentation typically creates the expectation that the improvement holds for the large majority of the frames in the sequence. If this is not the case, it might be a symptom of some temporal irregularities that should be better investigated directly by the proponents.

In the rest of the section, we will focus on Case *Agree90* by investigating how the disagreement between corresponding frames in different sequences is influenced by the coding parameters. By fixing the value of most of the coding parameters, we obtain a set of sequences from which we choose the Case *Agree90* ones. The latter ones are compared one against each other, yielding to $N(N-1)/2$ comparisons when N sequences are considered.

As a first example, we consider the number of slices per frame. Fig. 10.10 shows, for each frame position, the fraction of frames in that position that disagree among all the performed comparisons, and for which the reason of disagreement is the PSNR. This operation is repeated for similar sets in which only the number of slices per picture changes. It can be observed that the number of frames and their temporal position is very similar, therefore it seems that the number of slices does not significantly affect the number of disagreement. This method allows to intuitively see the difference and their position for a few different conditions. However it is impractical to perform large scale analysis. Therefore, instead of visually comparing the behaviour over time of the fraction of disagreement, we propose to compute a similarity index, i.e., the absolute value of the correlation coefficient. Such an approach also allows to provide a quantitative measurement of the similarity. The previous figure 10.10 can be compactly represented by the data in

	1 sl/frame	2 sl/frame	4 sl/frame	1500 B/slice
1 sl/frame	1,000	0.988	0.976	0.969
2 sl/frame	0.988	1,000	0.974	0.973
4 sl/frame	0.976	0.974	1,000	0.966
1500 B/slice	0.969	0.973	0.966	1,000

Table 10.2 – Correlation coefficient among the results of Fig. 10.10.

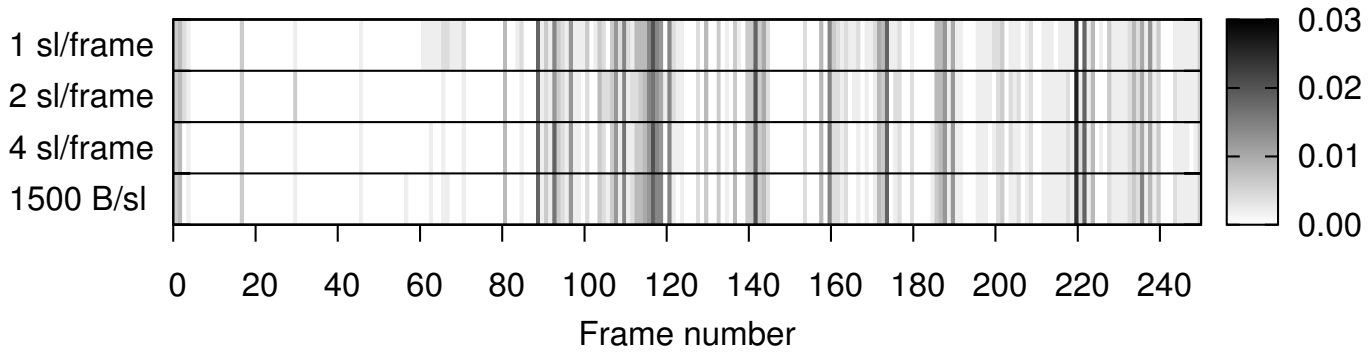


Figure 10.10 – Fraction of frames in disagreement for different number of slices per frame. Fixed QP, GOP size 8, intra refresh 16, open GOP.

Table 10.2. To further improve the scaleability of the method, we represent such data using matrices with different gray values, where the darker is the gray level, the higher is the absolute correlation. Fig. 10.11 shows the same data of the previous table in this form. The image is obviously symmetric along the diagonal as the values in the table.

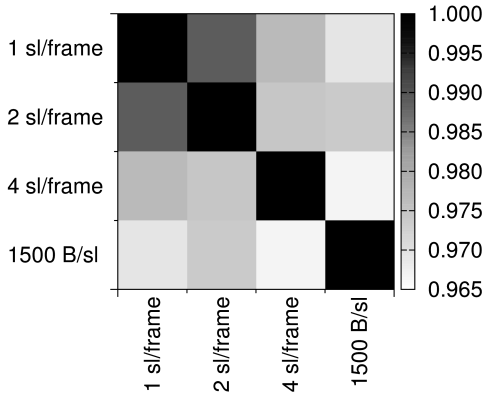


Figure 10.11 – Graphical representation of correlation coefficients shown in Table 10.2.

We adopt this technique to analyse the influence of the major coding parameters. When the correlation is close to one, the parameter has almost no impact, whereas lower values show much higher influence. First, we consider the fixed quantization parameter (QP) case, as done in most of the video coding works [295], and we vary only one parameter at a time. When all combinations of all the other coding parameters, including the source sequence, are considered, instead of only a subset as in Fig. 10.11, results are similar, as shown in the left part of Fig. 10.12.

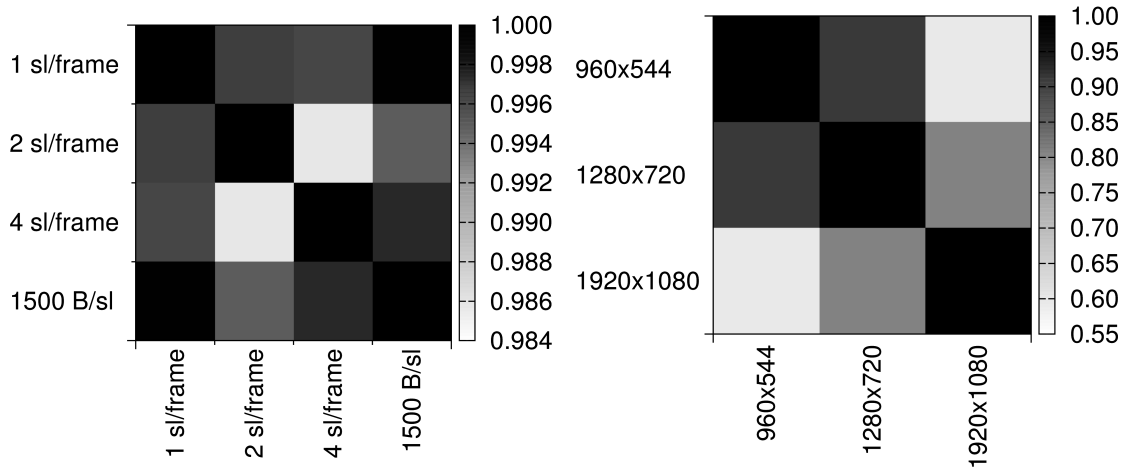


Figure 10.12 – Correlation coefficients between the cases (HRCs); left only slice size parameter is changed and right only resolution is varied.

The same behaviour happens for the open or closed GOP parameter (not shown in figures, correlation equal to 0.906), and partly for the resolution as in the right side of Fig. 10.12. The more interesting parameters are the Intraperiod and the GOP size. Significant variations can be observed, especially when they are considered jointly as

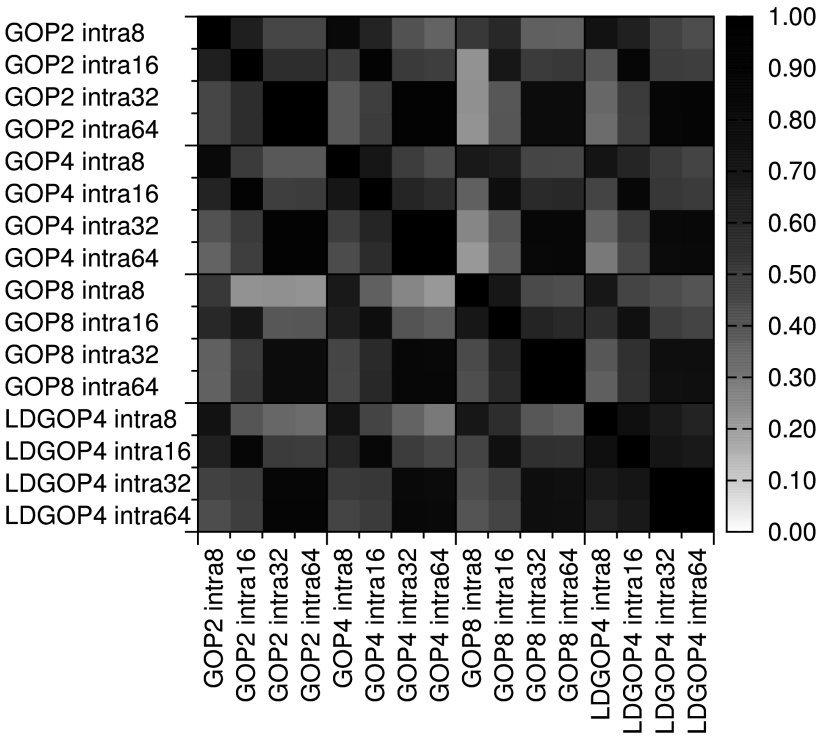


Figure 10.13 – Correlation coefficients between the cases in which all but the GOP size and intra refresh parameters are varied.

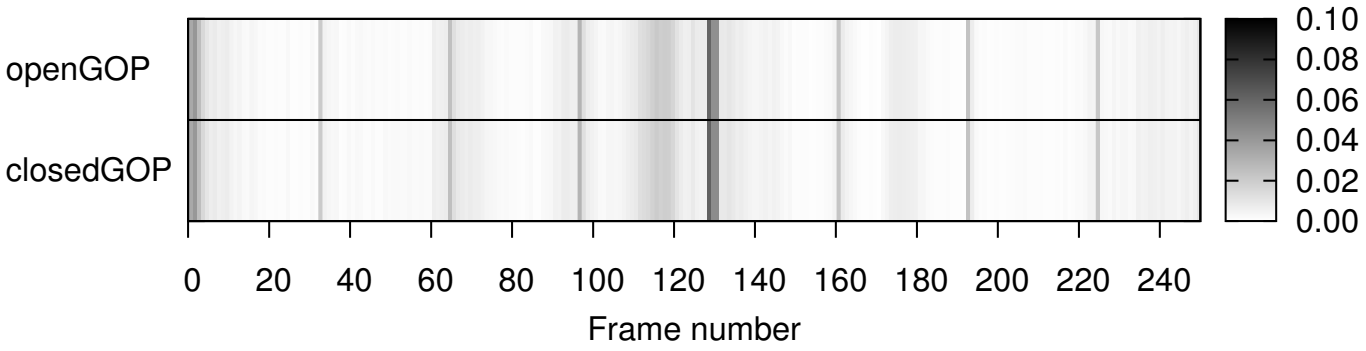


Figure 10.14 – Fraction of frames in disagreement for different number of slices per frame. HM rate control, LDGOP size 4, intra refresh 32. Note the peaks (darker vertical lines) at multiples of 32 frames.

in Fig. 10.13. In particular, it seems that when the GOP size is small and the intra period is large, there might be a strong impact on the position of disagreements, whereas with the largest GOP size the effect is reduced. With the low-delay GOP configuration (LDGOP) correlation is very high, meaning that the influence of the intra refresh rate is much more reduced. When the rate control algorithm of the HM test model software [296] is used instead of the fixed QP parameter, interesting observations can be made in the data, in particular when they are represented as a function of the frame position in the sequence. Fig. 10.14 is an example of such condition. The two rows are almost equal since they only differ for the open or closed GOP parameter which, as previously stated, has very little influence. For instance, in the first part a high fraction of disagreement is visible. This can be ascribed to the fact that an initial, fixed, QP is used by the HM rate control algorithm, which then quickly adapts to the requested bitrate. Moreover, note the peaks which appear in correspondence of the periodicity dictated by the intra refresh rate, i.e., when frames with I-type blocks only are inserted. By further experiments we determined that this behaviour is probably due to the inclusion of some source sequences which seems particularly difficult for the HM rate control when a frame with I-type blocks is inserted. This observation underlines the importance of performing such types of analysis on a large database with multiple coding parameters and several different content types. Although our database is somehow limited in the latter aspect, nevertheless such effects can already be observed.

Finally, we consider the fraction of disagreement by considering the same sets of comparisons but computing the fraction of disagreement for all the three measures. Figure 10.15 shows an example of the typical situation. While some behaviours are common for all metrics, e.g., the initial frames and the periodicity of the peaks, others seem to be peculiar of the measures. However, the latter often have a lower intensity.

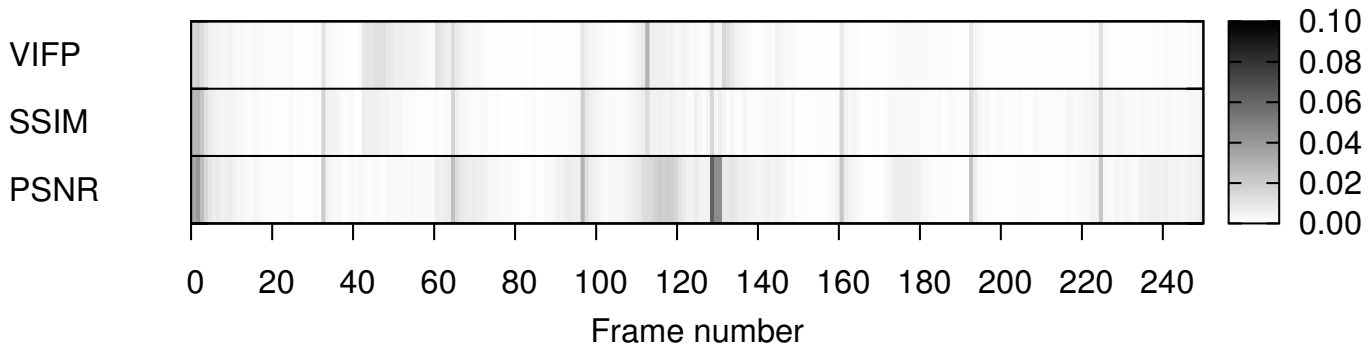


Figure 10.15 – Fraction of frames in disagreement separated for each measure. HM rate control, LDGOP size 4, intra refresh 32, open GOP.

10.4 Conclusions

This work showed how statistical analysis of a large-scale database including about half a million video sequences distorted by data loss can provide insights on the behaviour and particular limits of widespread simple objective video quality measures that are used partly out of scope. The main observations of the analysis that has been conducted in this Chapter are listed in Box 10.3. While these results concern future developments of coding and quality measurement algorithms, further work on the large-scale database approach requires a significant extension of the samples, both sequences and algorithm results, which is currently limited by the computational resources and the availability of implementations of objective measurement algorithms. Methodical work on analysis methods using statistical methods will continue towards the identification of particular cases that require inclusion in subjective experiments and the characterization of objective measures.

Box 10.3 – Main observations

— The agreement between the three tested measures PSNR, SSIM, and VIFP showed that the results of their predictions are similar, notably in the high and low quality range, less so in the middle range. It was further noted that the disagreement of the measurements is more pronounced in case of packet loss than for coding-only conditions which may be seen as a first step towards an automatic identification of the scope of application for objective measures. Thanks to the large size of the analysed dataset, some important effects on the characterization of the performance were highlighted that are not evident when a limited set of contents and parameters is considered.

— It may have been expected that disagreement between several objective measures exists on a frame-level even if the measures agree on a sequence level. However, the particular patterns of this disagreement point to two important conclusions. The first conclusion is that the usage of one single measure may not be sufficient. In particular, it may be beneficial to analyse the usage of several complementary algorithms within the coding loop, i.e. for rate-distortion optimization. In addition, it should be noted that performance bias may occur when improvements are measured only objectively and only using one single method, thus weakening such proposals. The second conclusion is that the pronounced correlation between content characteristics and encoder parameter selection encourages further analysis, for example with respect to the efficiency of rate-control algorithms. Some coding factors are almost not influential, whereas others have a strong impact, suggesting that quality comparisons among sequences without considering the detailed behaviour of the quality over the frames in the sequence itself could be strongly misleading.

11

Content and Machine Learning Based No-Reference (NR) VQA

11.0.1 Introduction

New image or video coding standards introduce new or improved coding tools in order to improve the rate-distortion performance. Each standard may be characterized by the type and amount of degradation that is added to the encoded image or sequence [186]. A lot of efforts have been done in identifying these coding degradations for different standards [187], notably for H.264/AVC. In addition to their importance in guiding the improvements in coding standard, understanding such degradations is also important for objective quality measures especially when there is no information about the original source. This type of quality assessment is called No-Reference (NR) measurement. A classification of no-reference quality estimation models has been reviewed in [188] and a variety of algorithms has been discussed. Although H.264/AVC NR measures can be adapted to the High Efficiency Video Coding (HEVC) use case, some publications have specifically addressed HEVC. In [189], a no-Reference Pixel (NR-P) based method is proposed in which the quality estimation for loss-impaired sequences is measured by calculating the temporal variations of the power spectrum across the decoded frames. As stated in [189], the model has correlation scores between 0.7 and 0.8 and works well for low-to-medium temporal activity sequences. This calls for integrating further content characteristics, either pixel-based or bitstream features, in objective video quality measurement models. In [167, 190–192], No-Reference Bitstream (NR-B) based models are introduced. In [167], Kanumuri et al, modeled the visibility of packet-loss in MPEG-2 video using pixel and bitstream based features. In [190], the authors train a neural network using subjective scores as well as packet loss rate, frame type, GOP structure, Intraperiod, percentage of damaged frames, and percentage of frames at different temporal levels. In [192], the authors use the QP and the spatial information (SI) to introduce a two-parameter NR-B method to estimate the perceptual quality (DMOS) of encoded HEVC sequences. The SI, as in [192], is calculated as the weighted sum of the DC difference values of inconsistent transform units (TU) and their respective neighbouring TUs based on the ratio of the TU edge length. In [191], the authors rely on the bitstream features to predict the perceptual video quality. Please, refer to Section 2.6 for more related work. While these measurement tools provide promising results, the question of generalization in the real application remains and may thus be seen as the main limitation of the proposed measures in the literature. In this chapter, the content dependent and machine learning based NR VQA models are introduced for error-free and loss-impairment sequences, respectively. We try to answer the research questions that are listed in Box11.1. The structure of this Chapter is illustrated in Box 11.2.

Box 11.1 – Goals

This chapter aims to answer the following research questions:

— What is the impact of using pixel-based content features in building machine learning based NR VQA for error-free and loss-impairment sequences along with channel and coding parameters?

Box 11.2 – Chapter structure

This chapter will be organized and structured as shown in Figure 11.1.

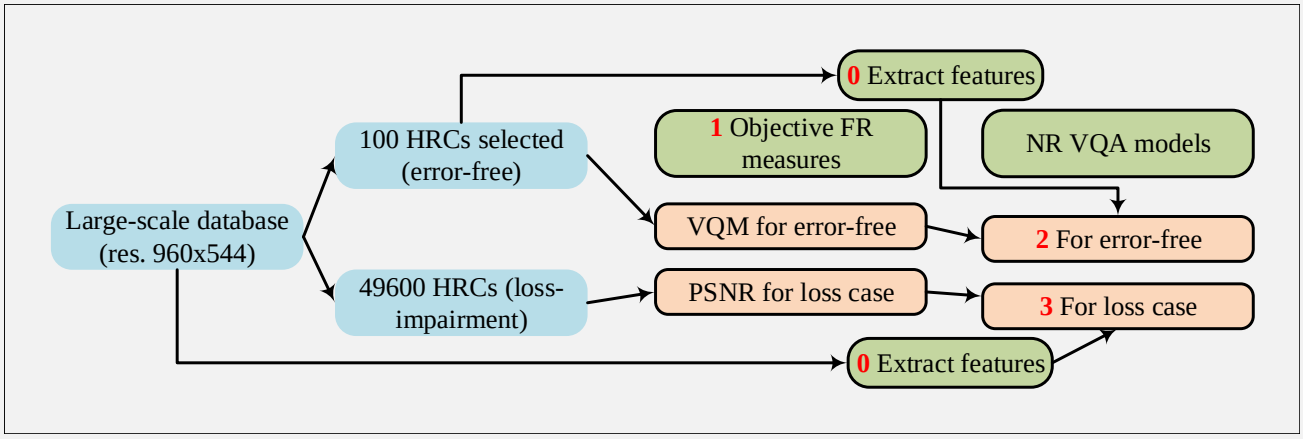


Figure 11.1 – Chapter 11 Structure

11.1 The pixel-based content features

The pixel-based content features, that are used in this paper, have been listed in Chapter 3 and used in [297,298]. A brief overview of these features will be listed here. The features cover spatial and temporal characteristics that are extracted from the luminance frame (Y), and the chrominance frames (Cb and Cr), in the spatial domain or in the frequency domain. The features are extracted on both block or frame levels. For the features that are extracted at the block level, the Minkowski sum with different power is applied to obtain a scalar value of each frame, then several statistical measures (e.g., mean, maximum, standard deviation, etc.) are applied to get a scalar value that represents the video sequence. In addition to those features, standard deviation, the variance, the skewness, and the kurtosis of the motion intensity histogram that is computed using pixel change ratio map (PCRM) [203] are calculated. In total, 284 features are extracted.

11.2 Feature selection process

Firstly, all HRCs are encoded and then an objective full-reference measure is used to estimate the quality, the VQM in our case. Then, the pixel-based features is extracted from the decoded output and finally the support vector regression (SVR) is used to train the model. The feature selection Algorithm 1 as in [297] is used to get the features that are required for the regression process (SVR). It is an exhaustive process of adding each feature one by one until no improvement is introduced. Epsilon-SVR (LIBSVM tool [299]) with radial basis function is used to train the model with n-fold cross validation. The SVR parameters are trained before applying the training algorithm. Table11.1 shows a summary of inputs and outputs of the proposed NR VQA models.

11.3 Content-dependent NR VQA model for Error-free sequences

In this Section, a link will be established between Full Reference and future Hybrid No-Reference measures by showing how to predict the results of one of the measures, in this case VQM, to information extracted from the transmitted video sequences, notably content features characteristics.
Figure 11.2 shows the training model that is used in the experiments. In the training phase, the following inputs are prepared as explained in Table 11.1. The samples, i.e the processed video sequences (PVSs), are 1000 (100 HRCs×10 SRCs). The variables are the content features, listed in Section 3.2, that are extracted from the decoded sequences. Finally, the responses are the quality score of VQM [300], the objective video quality metric. Then, the feature selection process, Algorithm 1, is applied with cross-validation approach to select the content features that are going to be used in the training model. The 11 selected features are listed in Table 11.1. These features contain different types of features, i.e. motion and texture features. The model is validated using the training data and two different random subsets from the large-scale database, 100 HRCs each. Figure 11.3 shows the results of the testing process. Pearson Linear Correlation Coefficient (PLCC) and Root Mean Squared Error (RMSE) are used to compute the prediction

Table 11.1 – Summary of inputs and outputs of the proposed NR VQA models

	Error-free NR model	Loss-impaired NR model
Number of PVSs (samples)	100 HRCs (in Random) x 10 SRC	1984 HRCs x 25 Error x 10 SRC
Features (variables)	Pixel-based features from encoded seq.	Pixel-based of original, encoding param., and channel param.
Model response	VQM quality score	Q, here is PSNR
Performance measure (PLCC)	0.9830	0.932
Number of selected features	11	18
List of features	— Histogram dissimilarity of low and medium frequency DCT maps. — Mean of (V) chrominance component. — Mean of SI std. — Contrast of GLCM (2). — Std of Motion intensity Hist. — Skewness of cross-correlation — Kurtosis of TI. — Kurtosis of SI13, — Skewness of (U) chrominance component. — Entropy ratio extracted from Laplacian based features (Laplacian pyramid level 5 over original)	— Encoding Parameters (8) — Channel Parameters (2) — Number affected frames (1) — Content features (7) — Temporal information — Correlation of,GLCM (2) — Energy of GLCM — Entropy of GLCM — DCT based smoothness — MPEG-7 short length of zero of spatial distribution of the objects

model performance, see the title of each sub-figure in Figure 11.3. The model maintains its stability across other validation sets.
Hence, the possibility to correlate the values of video quality measures with content features has interesting implications. For the case analysed here (VQM) it seems that the video quality metric value can be predicted quite reliably by a subset of content selected by using the algorithm proposed in this work. This could be the first step towards designing a hybrid No-Reference quality metric that can show good agreement with full reference metrics such as VQM.

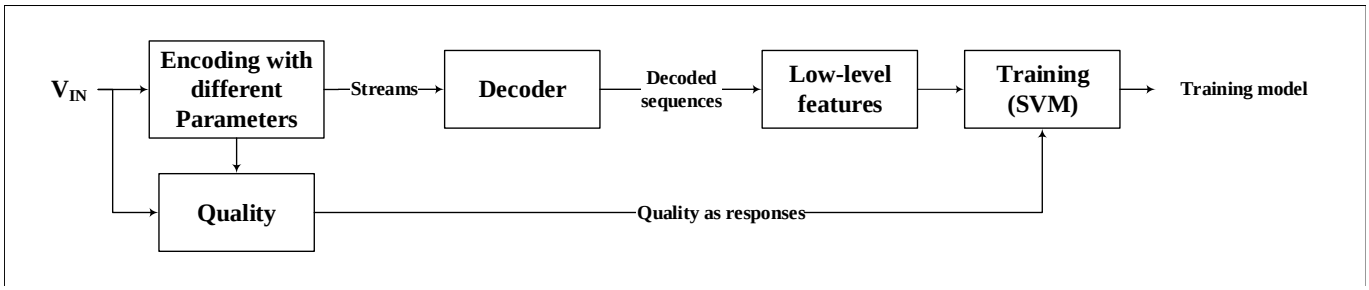


Figure 11.2 – NR-Reference VQA model for error-free sequences

11.4 Content-dependent NR VQA model for loss-impaired sequences

In this Section, a link will be established between Full Reference and future Hybrid No-Reference measures by showing how to predict the results of one of the measures, in this case PSNR, to information extracted from the transmitted video sequences, notably content features, coding parameters, and network characteristics.

We focus on a method to predict the behaviour of one of the measures, namely PSNR, from a set of features by employing a machine learning approach, so that difficult-to-predict situations (i.e., outliers) can be identified. The main goal of this machine learning process is to highlight the content features, besides the encoding and channel

Algorithm 1 Training algorithm

Input: (data,target) {data is $m \times n$, where m is the samples and n is number of features, and target is the response of each data sample (ΔPSNR)}
Output: F {set of selected features}

$F \Leftarrow data$(channel and encoding parameters)
Train SVM model with 5-fold cross validation
$pBest$ = Save the performance results of the training as best performance
for $i = 1 : n$ **do**
  for $i = 1 : n$ **do**
    $F' \leftarrow F + data(i)$
    Train SVM model with 5-fold cross validation using F'
    $p(i) = performance$ {Save the training results}
  end for
  $[p'\ indx] \leftarrow max(p)$
  if $p' > pBest$ **then**
    $F \leftarrow F + data(indx)$
    $pBest \leftarrow p'$
  else
    break
  end if
  $p \leftarrow 0$
end for

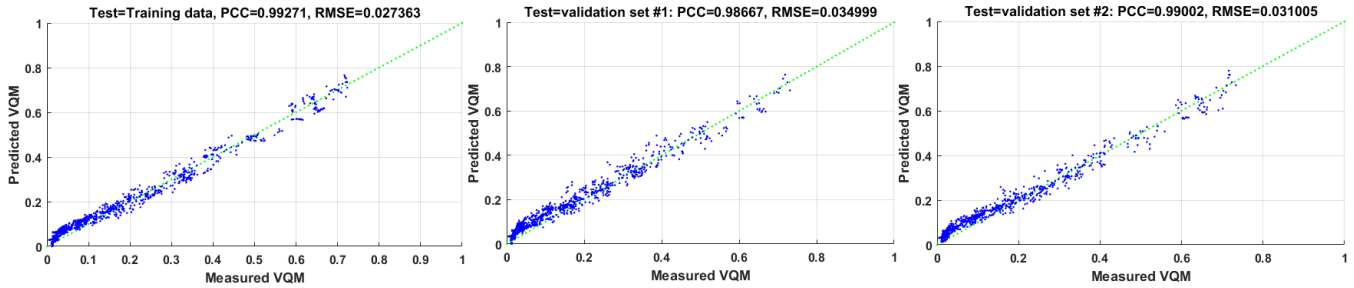


Figure 11.3 – NR-Reference VQA test results for error-free sequences

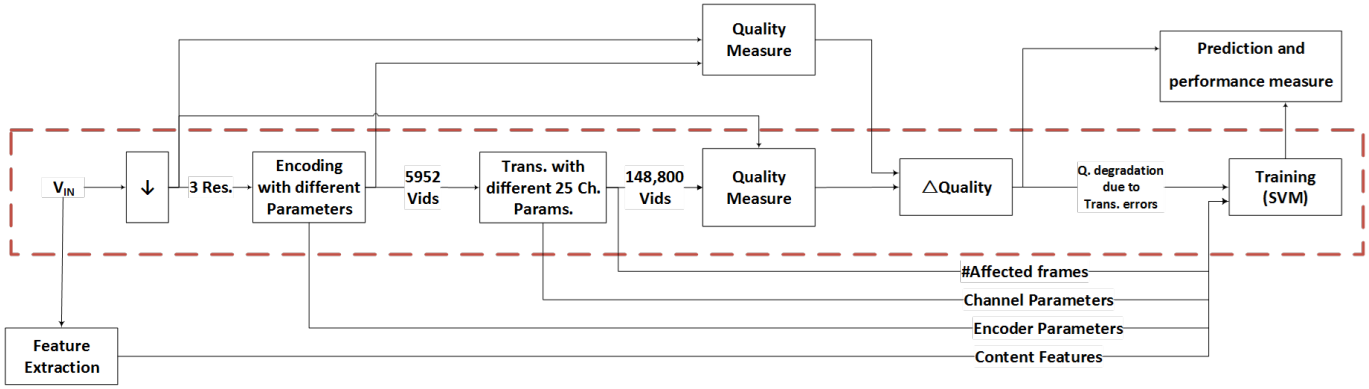


Figure 11.4 – NR-Reference VQA model for loss-impairment sequences

parameters, that have an impact on improving the prediction of the difference in quality between sequences with and without packet loss, i.e. ΔPSNR, that is calculated as the difference between the PSNR of each condition with coding-only degradation and the PSNR of the same condition with applied packet loss pattern. Figure 11.4 shows the training model that is used in the experiments. In the training phase, the following inputs are prepared as explained in Table 11.1. The samples, i.e the processed video sequences (PVSs), are 496000 (1984 HRCs of 960×544 resolution ×10 SRCs ×25 error patterns). The variables are the content features listed in Section 3.2, coding parameters, and network characteristics, that are extracted from the original sequences due to the complexity issues. We assume that the features that are extracted from the original signal are good approximation for the features that will be extracted from the decoded sequences. Finally, the responses are the ΔPSNR. Then, the feature selection process, Algorithm 1, is applied with cross-validation approach to select the features that are going to be used in the training model. The 18 selected features are listed in Table 11.1. The channel parameters and the encoding parameters are fixed to be part

of the selected features. After the end of the training, 7 content features only are selected.
Table 11.2 shows the performance of the prediction model using the test data set. Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Squared Error (RMSE) are used to compute the prediction model performance. In addition, Figure 11.5 shows the correlation between the predicted data against the original data. It can be observed that when ΔPSNR decreases (i.e. the encoded and degraded sequences are approximately of the same quality), the prediction error increases.

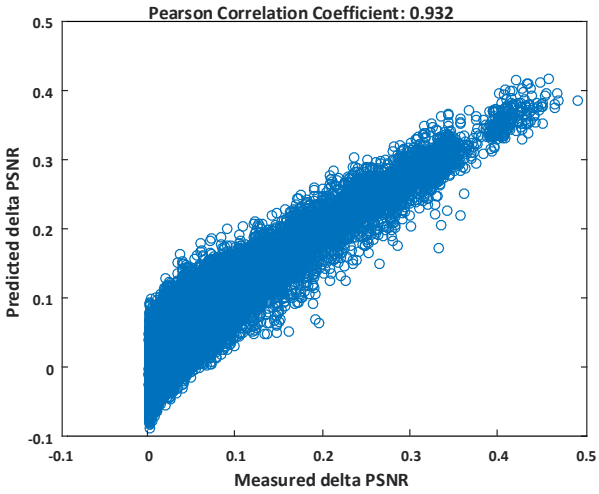


Figure 11.5 – Performance of predicting ΔPSNR from content features only. Please note the lack of significant outliers.

Table 11.2 – Performance of the predicting model

Performance	PCC	SROCC	RMSE
Test data (reduced feature set)	0.9320	0.833	0.0305
Train data (reduced feature set)	0.9310	0.832	0.0305
Test data (All features)	0.9144	0.770	0.0368
Train data (All features)	0.9135	0.769	0.0368

11.4.1 Analysis based on ΔPSNR prediction

By modelling on all the features, the importance of every feature can be derived. Such analysis indicated that counting the number of frames that get affected by packet loss (frames_affected) is one of the appropriate features that correlate well with simple objective measures. Aided by this observation, questions arise about how the amount of affected frames influences the agreement between different quality measures. In Fig. 11.6, this analysis has been performed on all the sequences of src5. Other sources show a similar behaviour. On the Y-axis, the number of affected frames of all src5 sequences has been displayed with respect to the PSNR of these sequences on the X-axis. Every dot represents a sequence of src5, each having a different number of affected frames, caused by the packet loss scenario and the video stream structure, and each having a different PSNR. First of all, when looking at the distribution of the points, it can be observed that the number of affected frames does not behave linearly with respect to the PSNR that the model needs to learn. Using simple reasoning, one would expect that when many frames are affected by a slice loss, a strong PSNR reduction would result from this loss. In contrast, it can be observed that a large part of the range of the objective measure can be obtained from any different amount of frames affected. Obviously, even the smallest impact or change of a frame is considered as an additional affected frame. Because there is not a simple linear relation between the feature and the measure, it could be useful to design other features from which further linear correlation can be derived. Such feature would be able to provide more insight than the ones that are available in this set. So, although the precision of the designed model is high, the features should not yet be regarded as a comprehensive generic set.
When analysing the disagreement properties of all these sequences, the intensities of the plotted sequences need to be investigated. Light dots indicate sequences which have a low maximum disagreement with other sequences in the set. Dark points, on the other hand, depict regions where the objective measures can reach a high disagreement when comparing all sequences of src5. This high disagreement means that when one sequence is compared with all the others

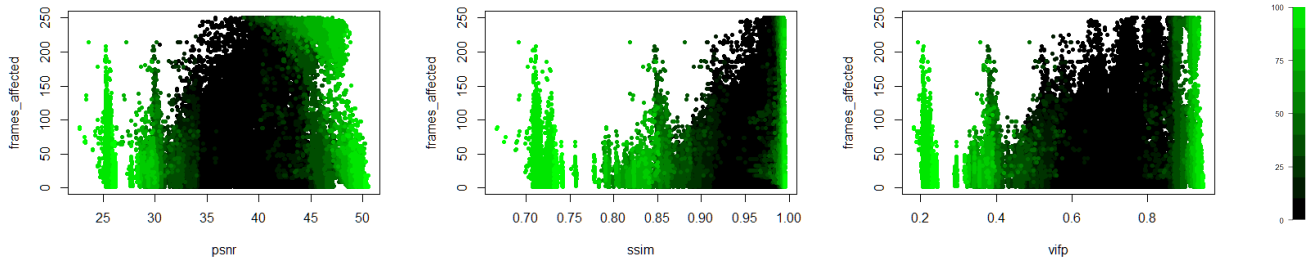


Figure 11.6 – The analysed objective measures, namely PSNR, SSIM, and VIFP plotted together with the transmission feature frames_affected of all src5 sequences. Darker dots indicate high disagreement, lighter dots indicate small disagreement.

from the same source, disagreement is measured as the normalized amount of sequences on which PSNR disagrees. When taking the maximum value resulting from all comparisons with other sequences, a high value is coloured darker in the plot. Fig. 11.6 also provides these plots with respect to SSIM and VIFP. It may be observed that agreement mainly occurs in the high and low PSNR range. So, when objective quality measures indicate a very high or a very low quality, it is more likely for the measures to agree. On the other hand, when operating at more average values, PSNR, SSIM, and VIFP tend not to agree. In both the plots of SSIM and VIFP, it can be observed that the frames_affected indicator influences the rate of agreement in an insignificant way. For PSNR, the frames_affected indicator influences the amount of agreement in the high PSNR quality range. This agreement persists longer at lower PSNR values when frames_affected is high.

11.5 Conclusion

In this chapter, we conduct experiments to find a possible correlation between the values of video quality measures and the content, coding, and channel characteristics. This possibility has interesting implications. For the cases analysed here (VQM and PSNR) it seems that the video quality metric value can be predicted quite reliably by a subset of content and channel features selected by using the algorithm proposed in this work. This could be the first step towards designing a hybrid No-Reference quality metric that can show good agreement with traditional full reference metrics such as PSNR. The list of contributions for this Chapter is listed in Box 11.3. It was shown that the results of the Full-Reference measures may be predicted from content features, coding, and packet-loss parameters with a high correlation, even if a reduced set of only seven parameters are used. This indicates that the complete image data may not be required in order to achieve the typical prediction performance of the evaluated measures. This is important for Reduced-Reference and No-Reference measures. Furthermore, although the number of affected frames provides the highest importance in predicting the PSNR within the developed model, there is no easily interpretable correlation between this feature and PSNR. Therefore, it would be beneficial to look further for features that can provide this easy to understand knowledge. Additionally, this analysis provides insight in the ranges in which the objective quality measures PSNR, SSIM, and VIFP agree. Especially in the average quality regions, comparing different sequences results in higher disagreement of the measures. This region is certainly of higher interest when performing subjective evaluation in order to further improve the large database approach of quality metric investigation.

Box 11.3 – Contributions

- A content-based NR VQA is built for error-free and loss-impairment video sequences along with coding, and channel characteristics. It predicts the behaviour of the full-reference VQA. The following features are found useful for the prediction model:
  - Channel parameters: loss rate, average and burst length.
  - Coding parameters: GOP size/type, intra-period, number of slices, open/close GOP, QP.
  - Channel and coding parameters: number of frames hit, number of slices hit, and number of affected frames.
  - Content-based features:
    + Gray-level co-occurrence matrix properties,
    + Chrominance information,
    + Spatial and temporal information,
    + Cross-correlation,
    + DCT and Laplacian based properties,
    + Motion intensity, and
    + MPEG-7 motion activity descriptor.

12

HRC Selection Algorithms and Improved Performance Measures for Learning-based Video Quality Assessment Algorithms

12.1 Introduction

As discussed in the Chapter 11, while the NR VQA tools provide promising results, the question of generalization in the real application remains and may thus be seen as the main limitation of the proposed measures in the literature. In order to tackle the problem of general applicability, we proposed in [294], also discussed in Section 11.4, to use a large-scale database to predict the behaviour of the objective measures with full-reference (FR) video quality metrics for loss-impaired sequences using encoding, channel, and content features. Following the conclusions from [294], we extend this work in Section 11.3 by including pixel-based features to predict the video quality.
In [292], the limitations of subjective experiments and future goals beyond the large-scale database are discussed. In this Chapter, we address one of the main limitations of the previous work, notably the fact that dealing with 1,984 HRCs for one resolution of a single content is often a large effort. While this may be justified in performance verification, it may be prohibitive when iteratively developing a new measurement method. In [301], we therefore propose two HRC subset selection algorithms for this purpose that aim at reducing the size of the large database with limited loss of generality. In Figure 12.2 the pseudo-code of the two subset selection algorithms is presented; each one elaborates the algorithm for selecting a subset of HRCs for a specific target. More details will be found in Section 12.2. The first one shows the selection that is optimized for HRCs that cover different ranges of (PSNR, Bitrate) values. The second shows the selection that is optimized for the HRCs in terms of content, i.e. those that behave differently depending on the sources. In order to verify whether these targets are reached, we compare these HRC subsets with several randomly selected subsets. The comparison approach is based on the typical development cycle for an objective measure: First, a training dataset (one of the subsets) is selected; second, machine learning is applied to a number of content and quality indicators in order to optimize the prediction of the ground truth data for this training dataset; third, the trained model is applied to a validation dataset (another one of the aforementioned subsets). Measuring the suitability of the different subsets for the training stage and for the verification stage is not straightforward. In the video quality assessment community often the Pearson Linear Correlation Coefficient (PLCC), Spearman's Rank Order Correlation Coefficient (SROCC), and the Root Mean Squared Error (RMSE) are used. We discuss advantages and shortcomings of these evaluation measures and propose further measures based on the analysis of the results obtained with the performed subset analysis. The subset selection for the training stage introduces a bias in the quality measurement model which may be chosen such that particular degradations or content characteristics are predicted with a better performance at the expense of worse performance for others as it is demonstrated in this contribution with a subset aimed at content variety and a subset that is aiming at predicting rate-distortion scenarios. In general, the approach of informed subset selection from a large-scale database may be a solution to the bias often introduced when training objective models on available or newly created subjective datasets. Instead of selecting the degradations a priori, the experimenter would create a large set of candidate sequences, run the automatic subset

selection tool with a specific target, and only then evaluate the subset in a subjective experiment.
In this Chapter, we raise the research questions listed in Box 12.1 in order to tackle the above-mentioned issues. This Chapter is structured as detailed in Box 12.2.

Box 12.1 – Goals

This chapter aims to answer the following research questions:
- Can a representative subset be selected from a large-scale database such that this small-scale database can be further analysed and the conclusions drawn on the small-scale database also apply to the large scale database?
- In case that the PLCC and RMSE cannot report the goodness of a model, what other performance measures that we need to report this goodness?
  In addition to that, the following secondary research question is investigated too:
  + How does the goodness analysis work when the content sources are different as well as the HRCs?

Box 12.2 – Chapter structure

This chapter will be organized and structured as shown in Figure 12.1. The HRC subsets preparation will be demonstrated in Section 12.2. The extracted features for video sequences and the images will be discussed in Sections 12.3.1, 12.3.2, and 12.5. The training phase of the trained models are illustrated in Section 12.3.4. The testing results are shown in Section 12.3.5. Finally, the proposed improved performance measures are discussed and demonstrated in Section 12.4.

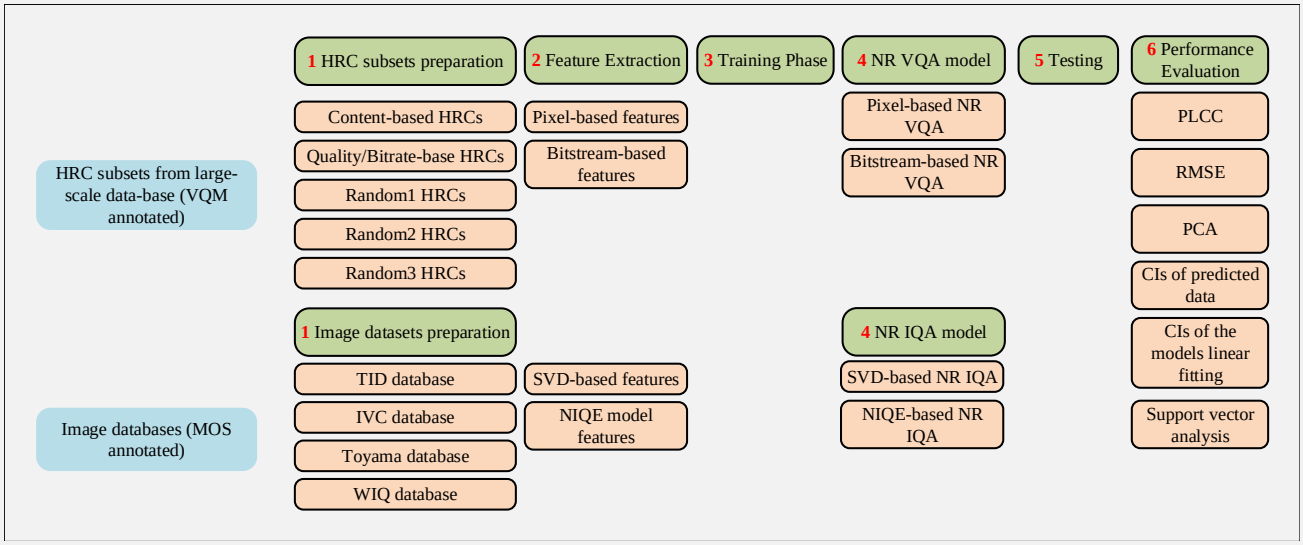


Figure 12.1 – Chapter 12 Structure

12.2 Goal-driven Large-scale Database Subset Generation

In Section 8.1, we discussed the limitations of the subjective experiments and the goals beyond the large-scale database. In this section, one goal beyond the generation of the large-scale database is discussed. Identifying target HRCs for a subjective experiment or for training a no-reference (NR) quality measure is challenging. Different correlation scores may be obtained if one tests an objective video quality (VQ) measurement using two different databases [302]. The reason could be the lack of content variety in the databases or the use of different HRCs in the experiments. Generally speaking, neither choosing different quality levels, i.e. different QPs or different bitrate budgets, nor selecting different content types is the optimal way to generate the database. What we need is to choose the HRCs that cover a wide range of the targets. If the target is a quality measure, e.g. the PSNR, we need to select

HRCs that cover all ranges of bitrate and quality. If the target is the content, we need to select HRCs that behaves differently with the contents. Dealing with the full set of 1984 HRCs for one resolution of a content is often computationally expensive. Therefore, in this section, a demonstration of two algorithms, Figure 12.2, to select a subset of the HRCs is discussed. Figure 12.2 shows two flowcharts. Each elaborates the algorithm of selecting a subset of HRCs for a specific target. The left flowchart shows the selection that is optimized for HRCs that cover different ranges of (PSNR, Bitrate). The right flowchart shows the selection that is optimized for the HRCs in terms of contents, i.e. those that behave differently with sources. The following subsections demonstrate the two algorithms.

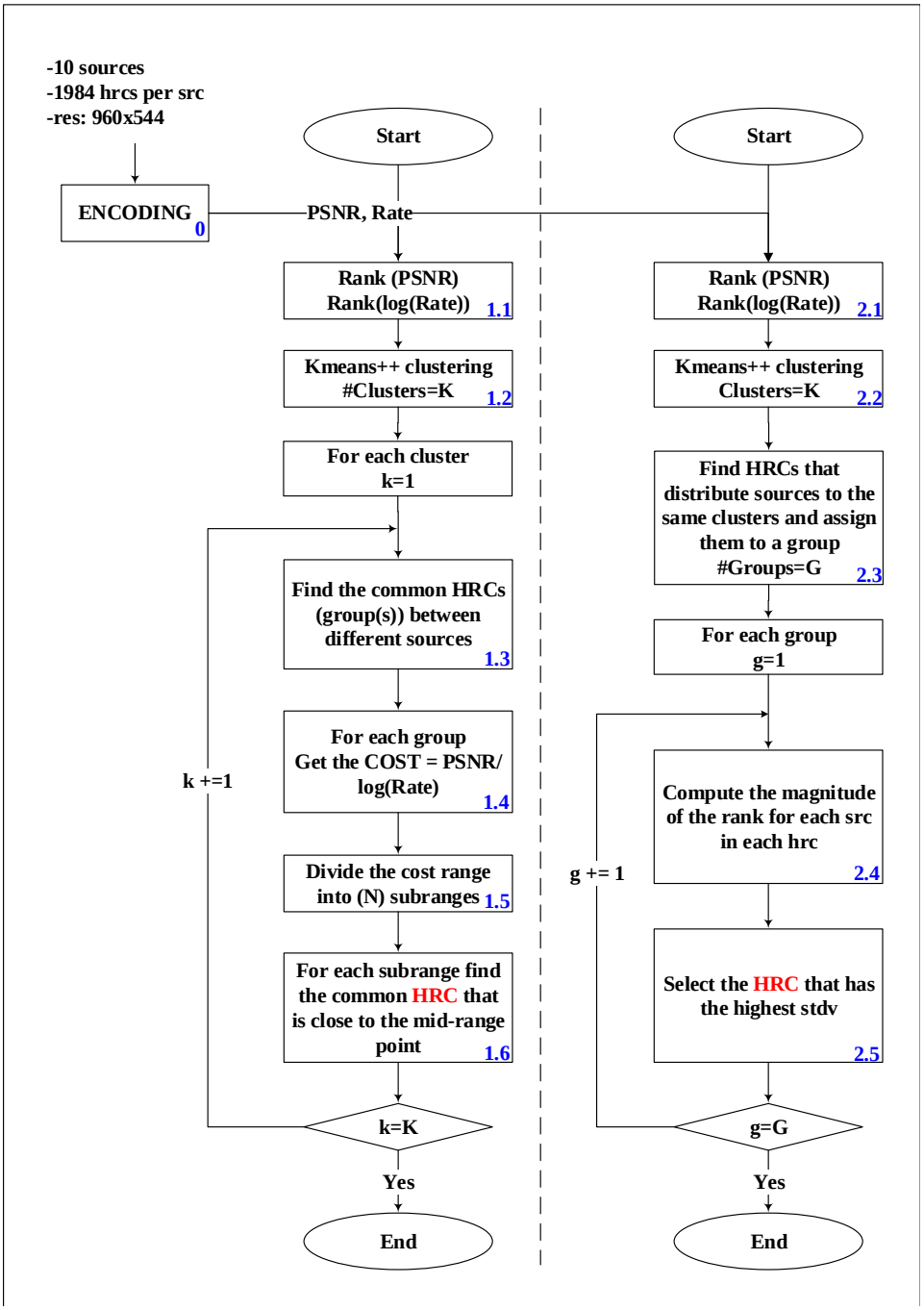


Figure 12.2 – Two algorithms for selecting large-scale database subsets for different targets. Left) Selection is optimized on HRCs that cover different ranges of (PSNR, Bitrate). Right) Selection is optimized on the HRCs in terms of contents (i.e. those that assign sources to different clusters)

.

12.2.1 Quality/Bitrate-driven HRCs Subset

In this subsection, the algorithm for selecting HRCs that cover a wide range of PSNR and bitrate values is demonstrated. Please refer to the flowchart in the left part of Fig. 12.2. At a specific quality level or in a specific

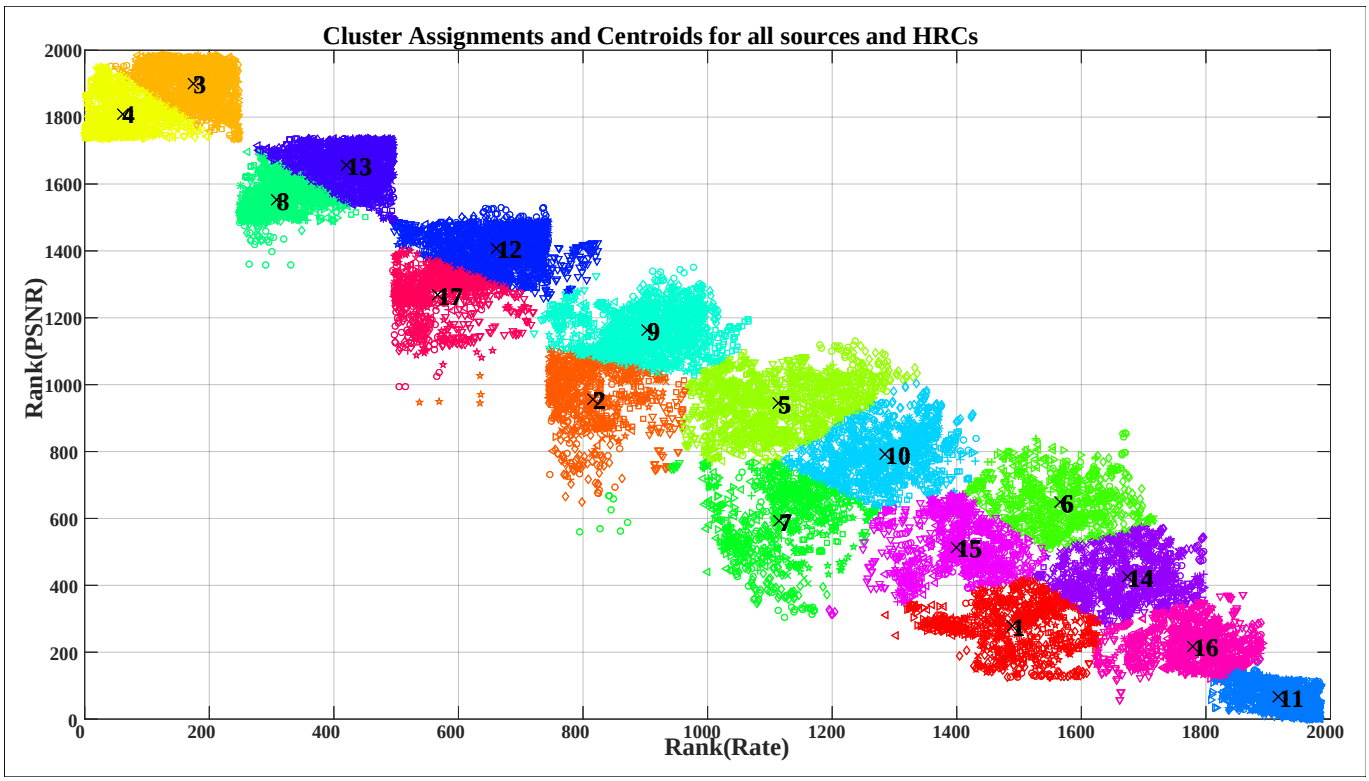


Figure 12.3 – Rank(PSNR) against Rank(Rate) of all HRCs and contents. Numbers and colours indicate the cluster number.

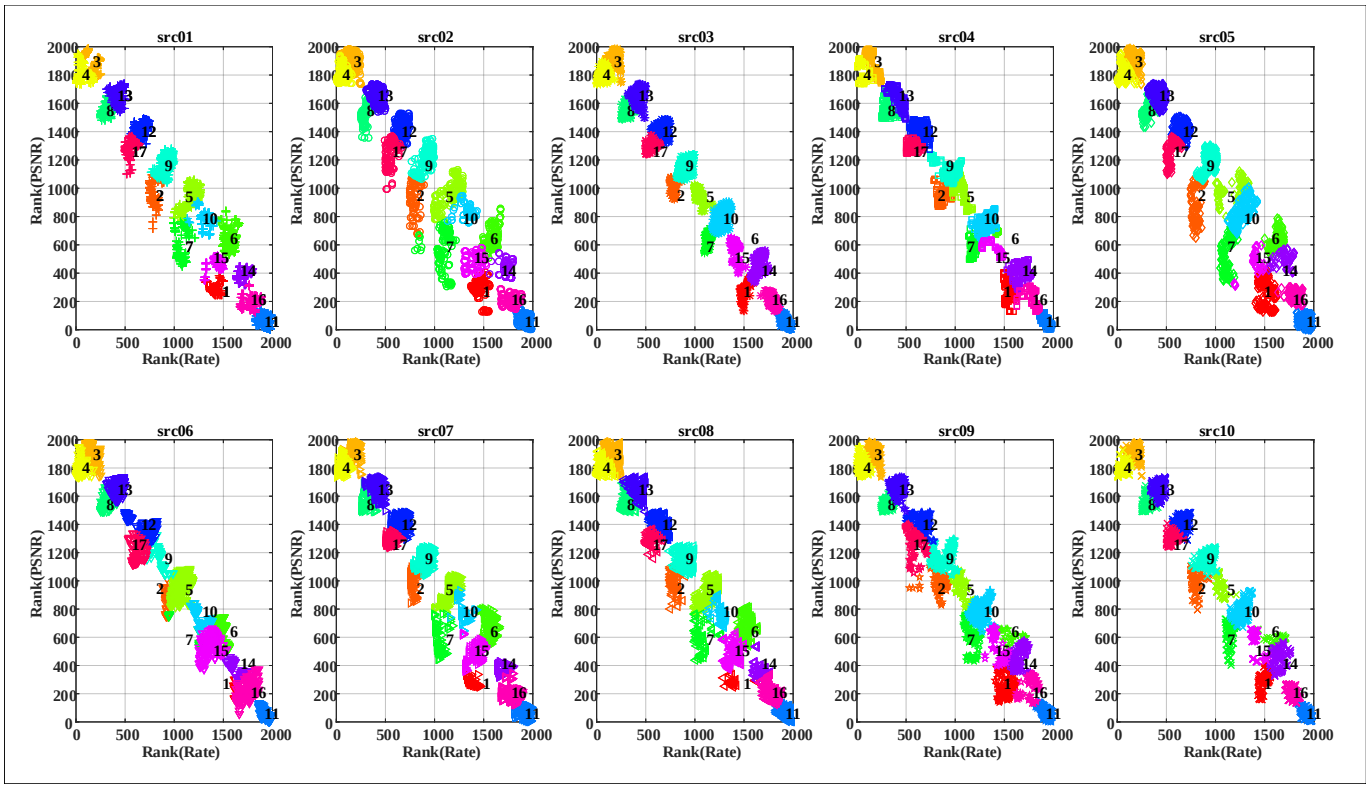


Figure 12.4 – Rank(PSNR) against Rank(Rate) per content of all HRCs. Numbers and colours indicate the cluster number.

quality range, the higher the quality the higher the bitrate. This intuitive assumption is followed as the main idea of the selection process. On the other hand, this assumption might be deviated from this assumption when a specific encoding parameter is changed, such as slice parameters. This deviation is exploited to identify the behaviour of each HRC in terms of quality and bitrate. The following steps are followed.

- Step 0: all sources are encoded using all HRCs, then the quality measure and the bitrate are calculated for each HRC.
- Step 1.1: rank the HRCs according to the quality measure and the bitrate in ascending order. Fig. 12.3 shows all pairs of rank(PSNR, Rate) of all sources while Figure 12.4 shows the pairs per content.
- Step 1.2: kmeans++ [303] clustering algorithm is used to cluster the HRCs according to their ranks in the quality measure. Different number of clusters are tested to select the optimal number of clusters. Figure 12.3 shows the 17 coloured clusters and their centroids for all rank pairs for all HRCs while Fig. 12.4 shows the cluster assignments and their centroids per content. From these two Figures 12.3 and 12.4, one can observe the following. The intuitive assumption is stable in the very low quality and very high quality in all contents although there are changes in other encoding parameters. The deviation of this assumption in the middle range of quality is obvious and it points to the impact of other encoding parameters and to the content.
- Step 1.3: as it can be been observed from Fig. 12.4, each cluster has a different number of HRCs for different sources. For instance, SRC-03 does not have an HRC that belongs to cluster number 6 and has many of them in cluster 14. Therefore, in order to get all HRCs that cover a wide range of qualities and bitrates, each cluster is divided into groups. Each group represents HRCs that are common between content sources. For instance, the first group contains HRCs that are common between $1^{st}, 2^{nd}$, and 10^{th} content sources. The second group contains the HRCs of the 8^{th} source since there are no common HRCs with other sources. The third group contains the common HRCs of the rest of the sources.
- Step 1.4: for each group, the quality per rate ($Cost = PSNR/log(Rate)$) is calculated to characterize each rank pair.
- Step 1.5: for each group, the $Cost$ values are ordered and divided into N subranges. The value of N affects the number of HRCs to be selected for each group. The total number of HRCs is 32, 61, 83, and 109 if N equals to 1,2,3, and 4 respectively.
- Step 1.6: for each subrange in each group, compute the mid-subrange point and then select the closest HRC to this point. Therefore, all ranges of quality and bitrate values are covered.

12.2.2 Content-driven HRCs Subset

In this subsection, the algorithm for selecting HRCs that behave differently with the contents is discussed, please refer to the flowchart in the right part of Fig. 12.2. The intuitive assumption that has already been discussed in the previous subsection, Section 12.2.1, is followed and exploited to identify the behaviour of each HRC with different content sources. The following steps are followed.

- Steps 0, 2.1, and 2.2 are similar to steps 0, 1.1, and 1.2 of the quality/bitrate-driven HRCs algorithm respectively.
- Step 2.3: in this algorithm, we care about the behaviour of each HRC with different contents. The HRCs that distribute source contents to same clusters are grouped. For instance, if one HRC distributes 3 contents out of 10 to clusters 2 and 5 respectively and another HRC distributes 4 contents out of 10 to clusters 2 and 5 respectively, then, the two HRCs belong to the same group. This decision is made because it is observed that this can happen between neighbouring clusters due to clustering error. In total, there are 97 groups for this dataset.
- Steps 2.4 and 2.5: for each group, in order to characterize each rank pair, the magnitude of rank of each content per HRC is computed and then the HRC that has the highest standard deviation is selected to represent the behaviour of this group. Thereby, we reduce the effect of clustering error and ensure that redundant HRCs are avoided.

12.2.3 Selected HRCs for each subset

In this Section, the selected HRCs' qualities and bitrate(s) values are shown to confirm the output of each algorithm of the subset generation. Figures 12.5, 12.6, and 12.7 show the quality measure (PSNR) against the logarithmic bitrate of all HRCs, quality/bitrate-driven HRCs, and content-driven HRCs per content source respectively. It can be observed that the quality/bitrate-driven HRCs cover the whole range of quality and bitrate values for each source content, while, on the other hand, the content-driven HRCs do not present the same behaviour. Moreover, as it can be seen in Fig. 12.8 and 12.8, the distribution of quality and bitrate rank points are regularly distributed in quality/bitrate-driven subset over all source contents while, in content-driven subset, it can be noticed that the quality and bitrate rank points are not regularly distributed over all the contents and are distributed roughly in the area of middle qualities and middle bitrate(s). The standard deviation of the ranks' magnitudes of each HRC is another indicator that shows that the quality/bitrate-driven HRCs is not content representative. HRCs that have low standard deviation values in content-driven subset are not selected, which means that there are similar-behaviour HRCs of higher standard deviation that strongly distinguish the HRCs from others in terms of content.

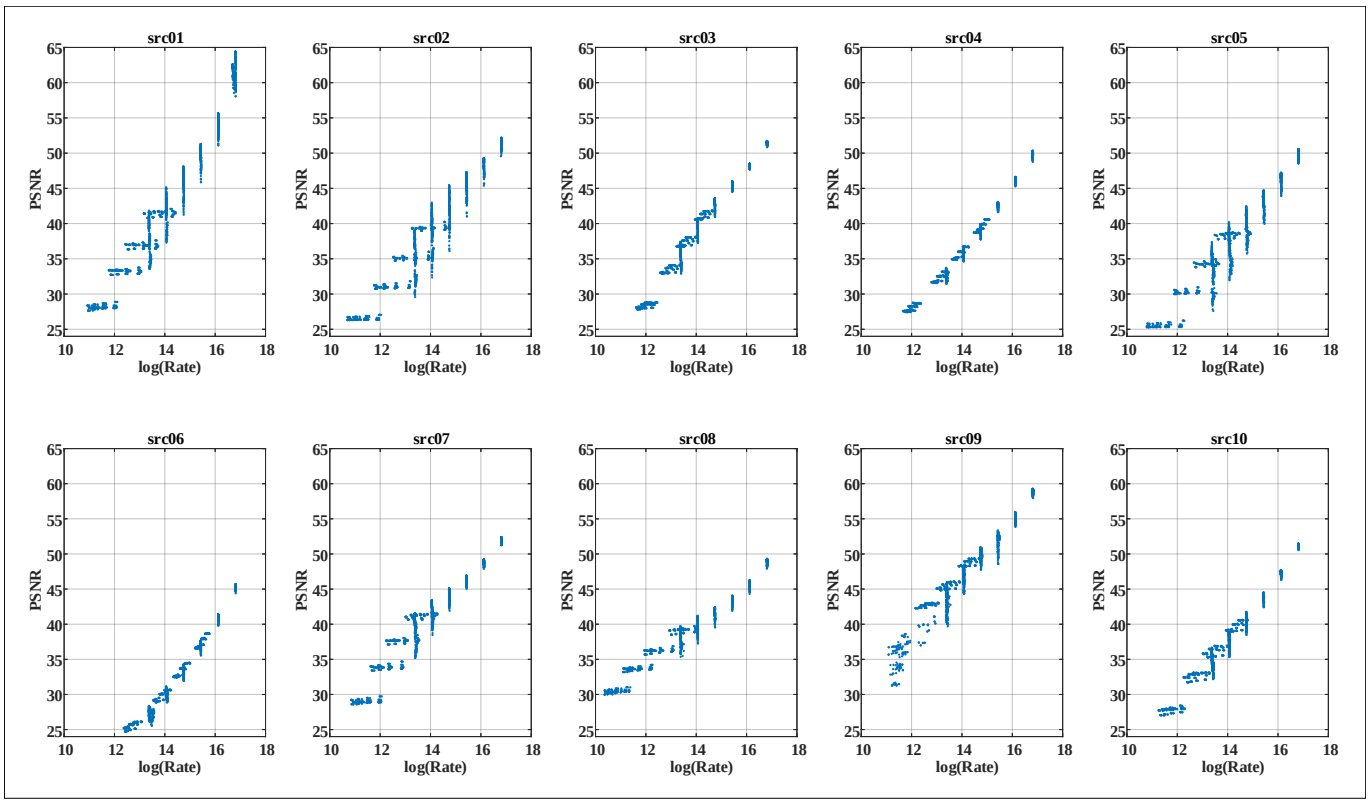


Figure 12.5 – PSNR against log(Rate) of all HRCs per contents.

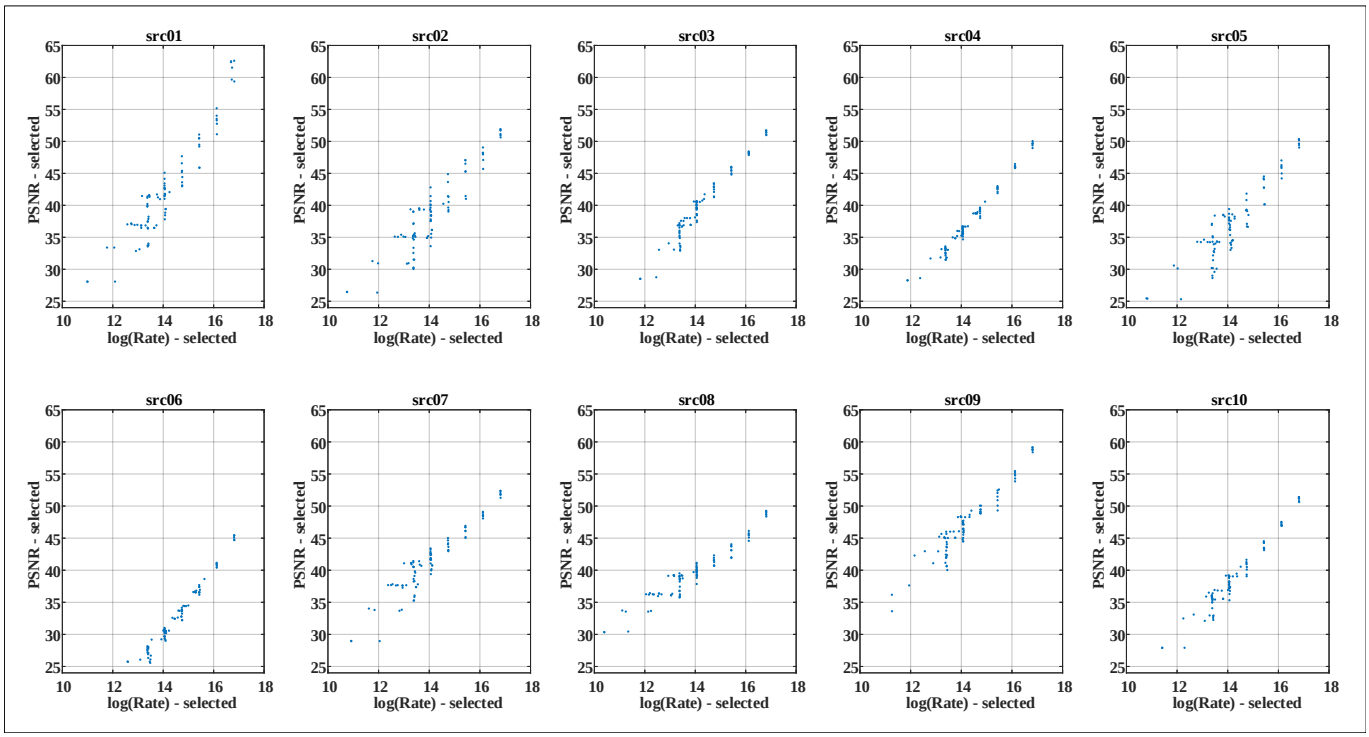


Figure 12.6 – PSNR against log(Rate) of all HRCs per contents of selected HRCs for the quality/bitrate-driven subset.

12.3 No-reference video quality measure

12.3.1 The pixel-based content features

The pixel-based content features used in this paper have been listed in [297] and used in [297, 298]. The features cover spatial and temporal characteristics that are extracted from the luminance frame (Y), and the chrominance frames (Cb and Cr), in the spatial or frequency domain. The features are extracted on both block and frame levels.

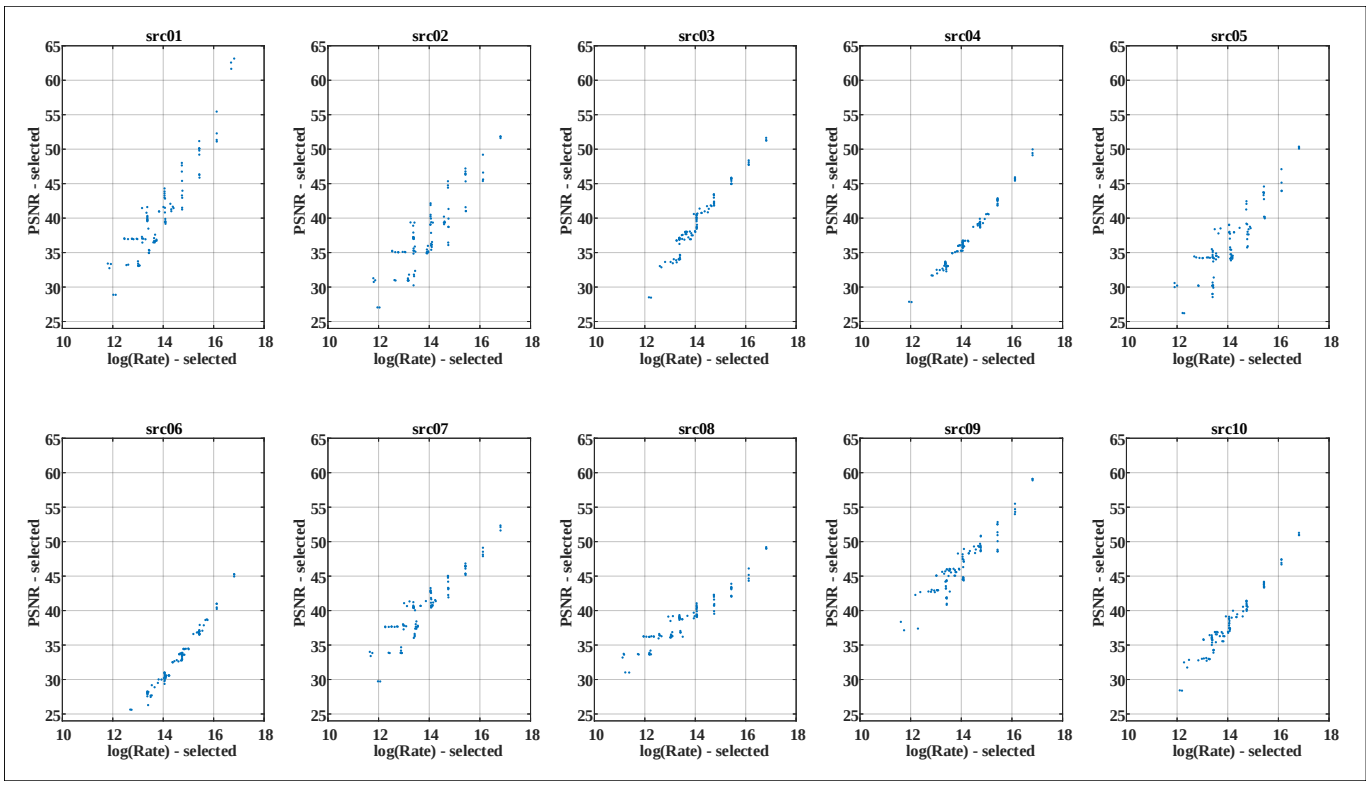


Figure 12.7 – PSNR against log(Rate) of all HRCs per contents of selected HRCs for the content-driven subset.

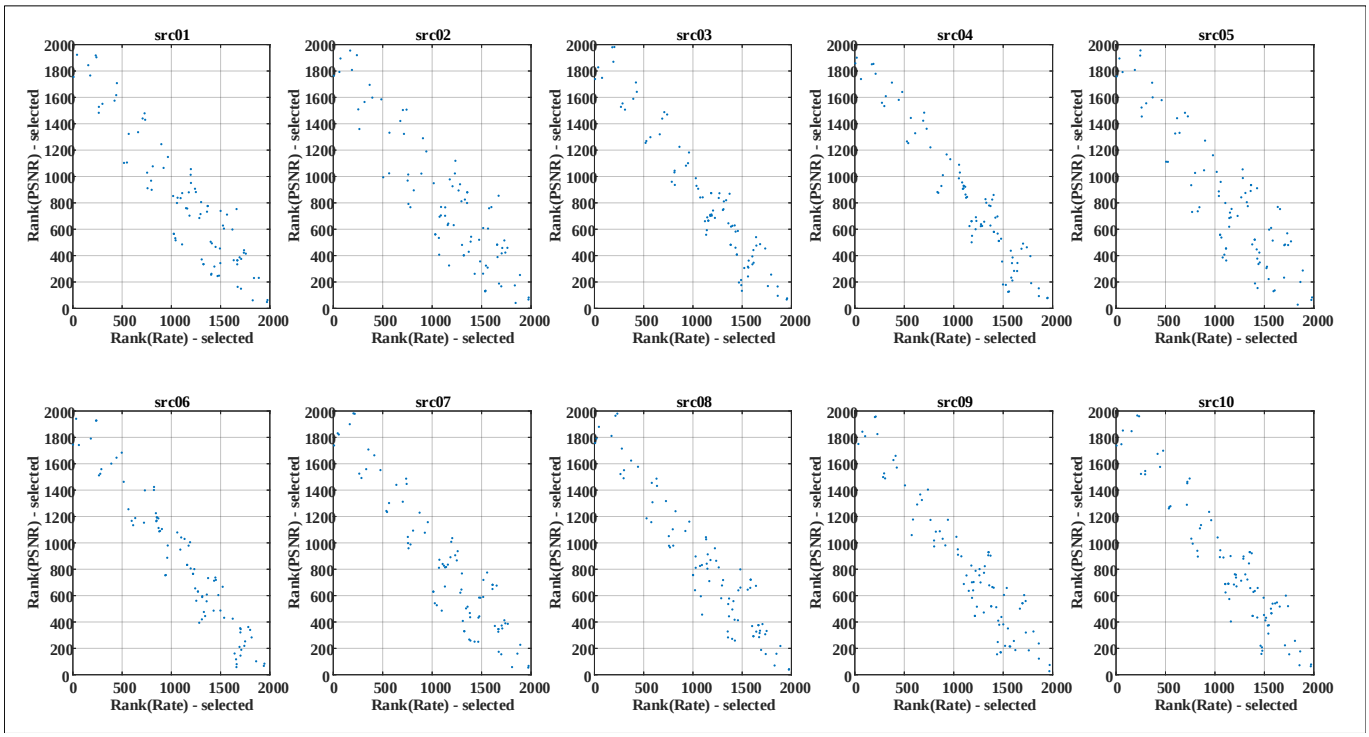


Figure 12.8 – Rank of PSNR against Rank of log(Rate) of all HRCs per contents of selected HRCs for the quality/bitrate-driven subset.

For the features that are extracted at the block level, the Minkowski sum with different power is applied to obtain a scalar value of each frame, then several statistical measures (e.g., mean, maximum, standard deviation, etc.) are applied to get a scalar value that represents the video sequence. In addition to those features, standard deviation, the variance, the skewness, and the kurtosis of the motion intensity histogram that is computed using a pixel change ratio map (PCRM) [203] are calculated. In total, 284 features are extracted from a subset of the encoded sequences in the large-scale database [292].

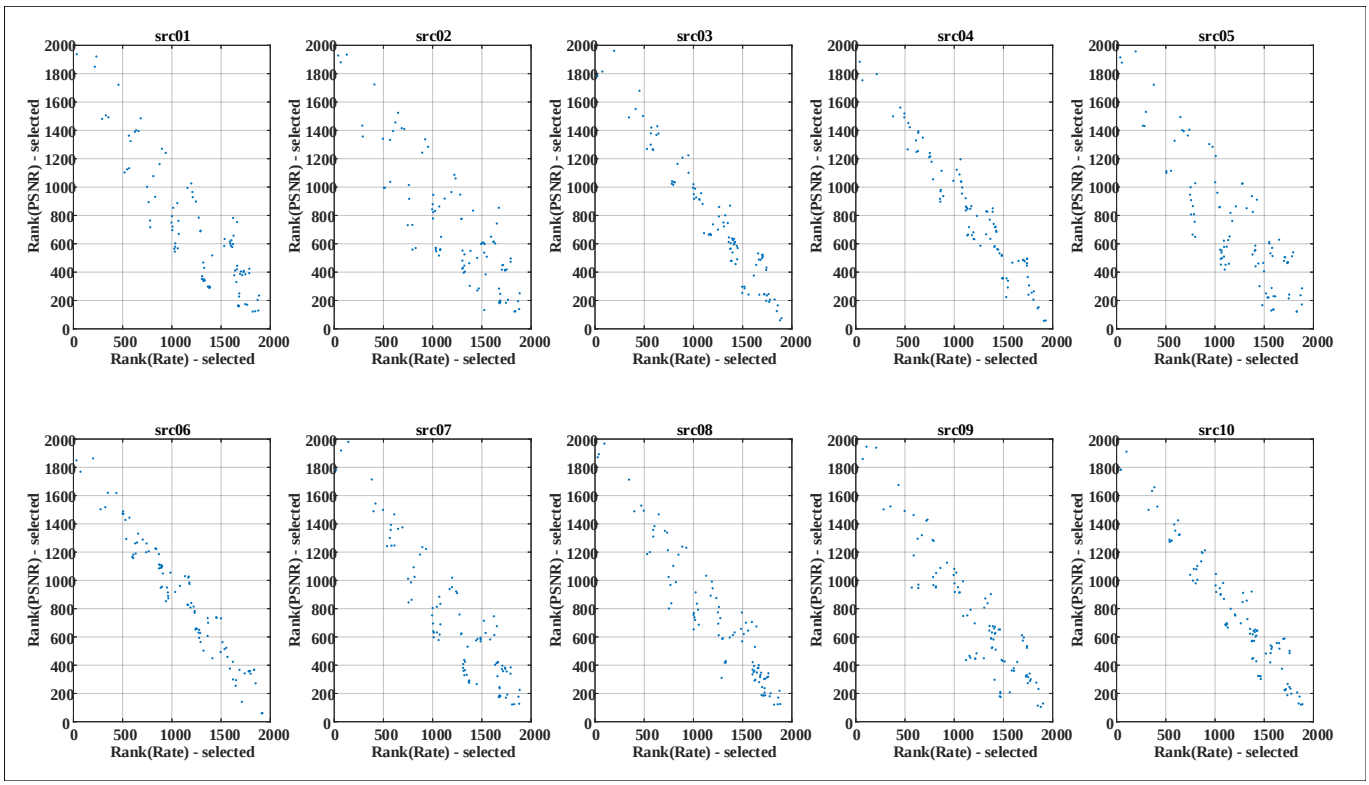


Figure 12.9 – Rank of PSNR against Rank of log(Rate) of all HRCs per contents of selected HRCs for the content-driven subset.

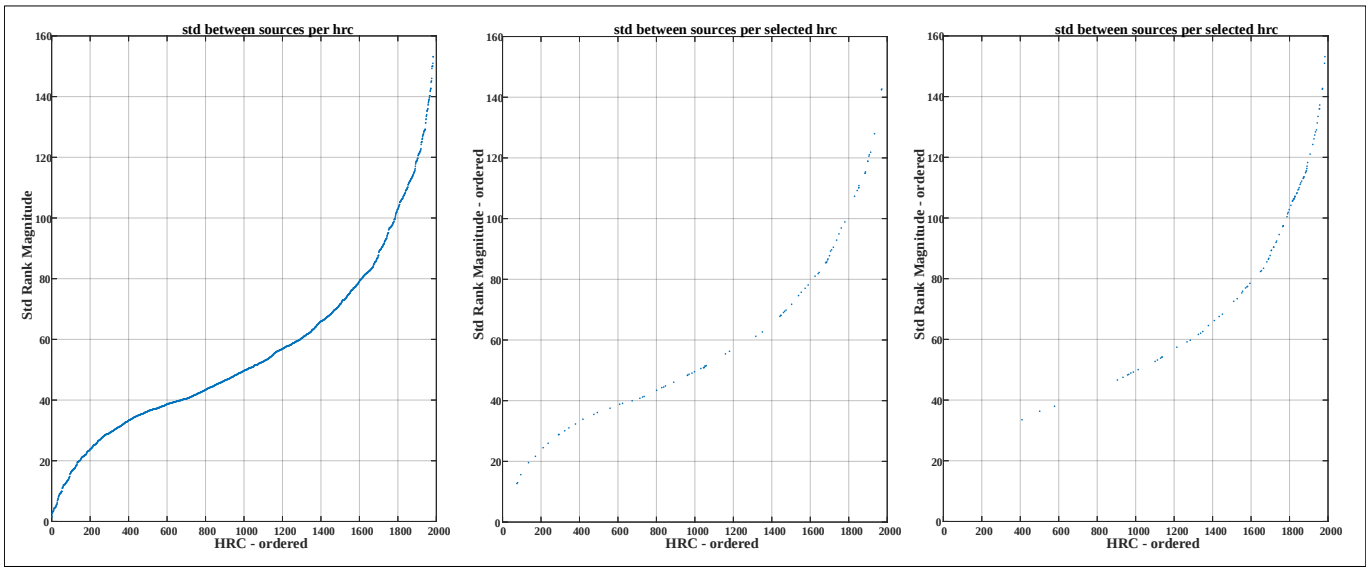


Figure 12.10 – Standard deviation of rank magnitudes for each HRCs. left) all HRCs. centre) Selected HRCs of quality/bitrate-driven subset. right) Selected HRCs of content-driven subset.

12.3.2 Bitstream features

In [191], Shahid et. al. use 52 bitstream features in order to perform perceptual quality estimation of HEVC coded videos. Ratios of various used CU sizes and of various prediction modes of intra and inter frames, and statistics of different levels of quantization parameters and motion vectors are considered in these features. The features are extracted as follows:

- The bitstream information extractor (HMIX) [292] is used to generate the '.xml' file from the encoded stream file.
- HMIXParser, developed for this work, is used to extract the bitstream features. Firstly, the frame-level features are extracted and then sequence-level features are calculated using a pooling strategy based on the average value.

12.3.3 Subset description

Five HRC subsets are used in this work. Two HRC subsets are selected using HRC generation algorithms [301], see Figure 12.2. The first one shows the selection that is optimized for HRCs that cover different ranges of (PSNR, Bitrate) values. The second shows the selection that is optimized for the HRCs in terms of contents, i.e. those that behave differently depending on the sources. The other three datasets use a random selection. Figure 12.11 shows the histograms of the quality scores (PSNR) for the five subsets. This histogram will be useful, for instance, when testing HRCs that are under represented in the subset (i.e. PSNR > 50). These subsets will be named as follows: HRC_1, HRC_2, HRC_3, HRC_4, and HRC_5 and correspond to Content-driven subset, quality/bitrate-driven subset, and the three random-based subsets respectively. Note that the number of HRCs in each subset are 97, 83, 100, 100, and 100, respectively. The number of HRCs is not identical due to the selection algorithms but sufficiently close for comparison.

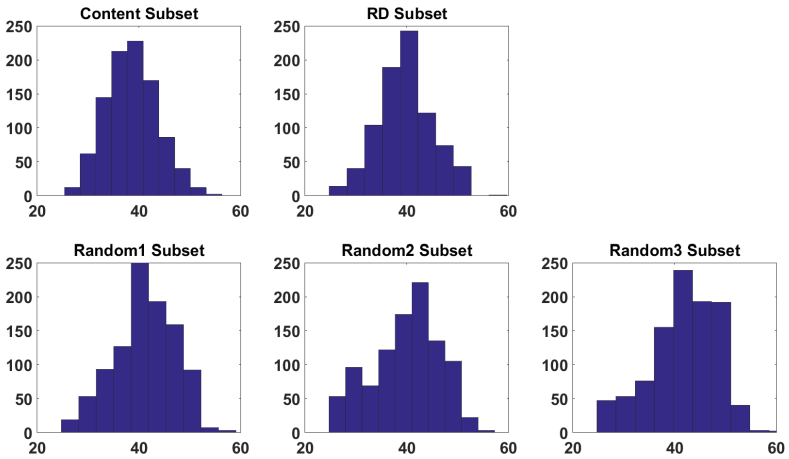


Figure 12.11 – Histograms of the quality scores for the five subsets

12.3.4 Feature selection process

Figure 11.2 shows the model that is used in the experiments. Firstly, all HRCs are encoded and then an objective full-reference measure is used to estimate the quality. We relied on VQM. Then, the pixel-based features are extracted from the decoded output and finally the support vector regression (SVR) is used to train the model. The feature selection algorithm in [297] is used to get the features that are required for the regression process (SVR). Epsilon-SVR (LIBSVM tool [299]) with radial basis function is used to train the model with 10-fold cross validation. Before the training is starting, the parameters of the SVR (C, G, and epsilon) are optimized by selecting one combination of different C, G, and epsilon values. Five features selection processes (SP) have been carried out: the first one $SP1$ for content-driven subset HRC_1, the second one $SP2$ for quality/bitrate-driven subset HRC_2, and the $SP3$, $SP4$, and $SP5$ for the three random subsets HRC_3, HRC_4, and HRC_5, respectively. These processes have been carried out for the pixel-based NR VQA and other five selection processes have been carried out for bitstream-based NR VQA. In the training phase, an exhaustive process of adding each feature one by one is applied. In the training process of $SP1$, 16 features are selected to be used for the SVR training. LIBSVM reports the squared correlation coefficient (SCC) as performance criterion. The SCC when using the 16 features is 0.9728. On the other hand, 14 features are selected in $SP2$ with an SCC of 0.9735. Figure 12.12 and Figure 12.13 show two features for the selected HRCs. The first feature (DCTHis13) is the histogram dissimilarity of DCT based feature maps using low and high frequency maps. The second feature (entrB_p4_mean) is the mean of entropy of 64x64 gray level co-occurrence matrix using Minkowski pooling (p=4). It can be observed that the features cover different ranges of values which make them useful for the training model. In $SP3/4/5$, 45, 11, and 8 features are selected, respectively, with SCC of 0.9883, 0.9830, and 0.9828. It can be observed that the number of selected features largely depends on the training data. Because of the over-fitting problem, the model that has the highest correlation is not necessarily the best one. This can be tested when the trained model is further validated with other datasets.

12.3.5 Training and testing results: the impact of content features

After the features have been selected for the training model for the five HRC subsets, each training model is trained and tested using all other HRC subsets including the subset that is used in the training phase. The experiments are divided into three categories: the first category will show the overall impact of the pixel-based content features in the

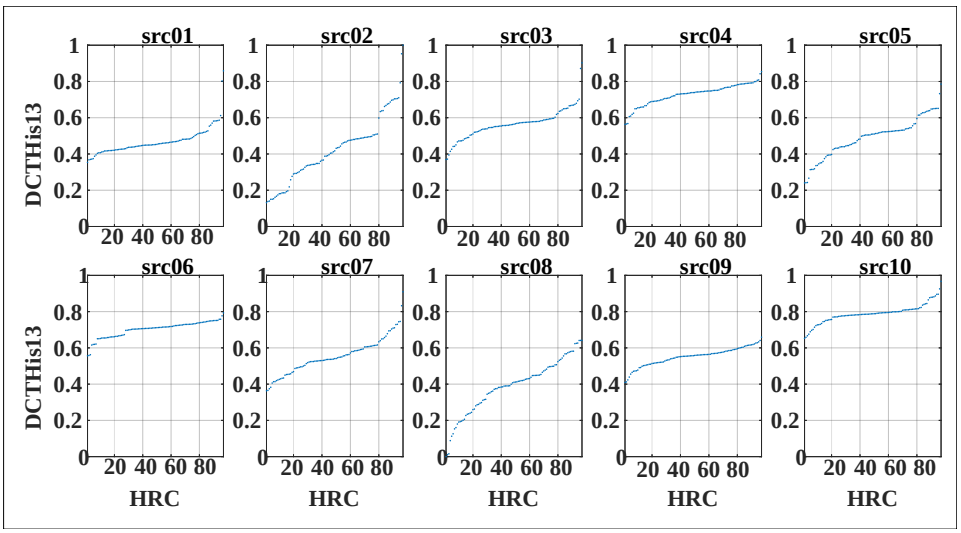


Figure 12.12 – DCT-based histogram dissimilarity feature of low and high frequency maps for content-driven subset.

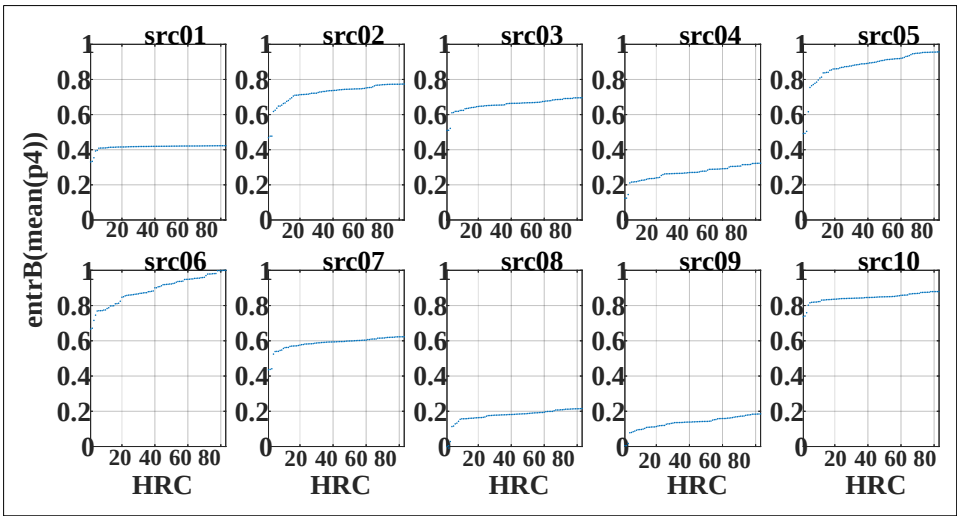


Figure 12.13 – The mean of entropy feature of 64x64 gray level co-occurrence matrix using Minkowski pooling (p=4) of all sources for quality/bitrate-driven subset.

different datasets. The second category will study the impact of pixel-based features per content. The third category will show which HRC group behaves differently when using the features. The performance of all experiments are measured using Pearson Linear Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE).

In the first category, 25 experiments (referenced to as $X(row, column)$) are conducted as shown in Figure 12.14. The rows of the figure represent the different training models that are trained, from top to bottom, using HRC_1, HRC_2, HRC_3, HRC_4, and HRC_5, while the columns represents the test data for each model and they are, from left to right, HRC_1, HRC_2, HRC_3, HRC_4, and HRC_5. Hence, the diagonal represents the evaluation of the model using the training data as the input. The green line represents $y = x$, while the red line represents the fitting line of each experiment. The first observation is the stability of content-driven, quality/bitrate-driven, and random 2 prediction models. It can be noticed when looking at the performance row by row, there is a stable high performance (PCC higher than 0.95) for rows HRC_1, HRC_2 and HRC_4. Although the prediction model using HRC_3 is stable, it is still a random process and, as it can be noticed in the other random-based prediction models, the correlation scores are not stable when using HRC_5 and the fitting line deviates by a notable offset. Further discussion on the performance measure PCC will be presented in Section 12.4. Moreover, the predicted VQM in $X(2,1)$ is better correlated than $X(1,2)$. This can be explained as follows: both experiments try to predict VQM but the HRCs in $X(2,1)$ cover a wide range of quality/Bitrate values while this is not the case for $X(1,2)$. Therefore, the training model has a better ability to predict the VQM value. Hence, this is an indication that the selection algorithm for quality/bitrate-driven works well.

Another observation can be made by looking column-wise at the correlation of the experiments. This can suggest which HRCs are challenging to a certain model. The two challenging sets are, in order, quality/bitrate-based HRCs and content-based HRCs. Table 12.1 shows the analysis of PCC by calculating the absolute mean difference of the

Table 12.1 – Correlation analysis, expressed as a percentage, for the NR VQA models using SVR

		PCC						Difference to train PCC					Absolute
		HRC_1	HRC_2	HRC_3	HRC_4	HRC_5	Average	HRC_1	HRC_2	HRC_3	HRC_4	HRC_5	mean difference
Pixel-based	HRC_1	98.8	97.1	96.3	96.6	96.5	96.6	0	1.7	2.53	2.21	2.29	2.19
	HRC_2	98.3	98.8	98.8	99.1	98.9	**98.8**	0.53	0	0.01	-0.32	-0.13	**0.02**
	HRC_3	97.8	96.1	99.5	97.2	90.5	95.4	1.71	3.44	0	2.30	9.03	4.12
	HRC_4	95.1	98.2	98.7	99.3	99.0	97.7	4.13	1.12	0.60	0	0.27	1.53
	HRC_5	62.7	60.0	68.6	91.9	99.2	70.8	36.52	39.22	30.66	7.33	0	28.43
Bitstream-based	HRC_1	98.0	97.3	97.4	97.5	97.8	97.5	0	0.73	0.59	0.49	0.26	0.52
	HRC_2	97.2	98.2	97.9	98.1	98.2	**97.8**	1.07	0	0.35	0.10	0.02	**0.38**
	HRC_3	96.5	97.4	98.5	98.4	98.0	97.6	1.96	1.04	0	0.04	0.44	0.87
	HRC_4	95.7	97.1	98.1	99.0	98.4	97.3	3.34	1.94	0.99	0	0.69	1.74
	HRC_5	96.8	97.6	97.7	98.4	98.9	97.6	2.09	1.25	1.16	0.46	0	1.24

correlation coefficients. It shows that the quality/bitrate-based and the HRC_3 subsets perform better in the pixel-based models. While the quality/bitrate-based and the content-based subsets perform well than others.
In the second category of the experiments, the influence of the content will be shown. One content is left out during the training and then the model is tested on the content that is left out. Figure 12.15 shows a typical example where source 5 is left out of the training set and used for evaluation. Comparing this figure with the results of models that include all contents in the training, it can be observed that the correlation is reduced and also the residual error is increased. That is an indication of content importance and how the absence of some content HRCs would affect the training model. Finally, in order to show that the HRCs subset selection algorithms work well, we compare the results on the diagonal. In general, on average, Random 1 HRCs set has a lower correlation: this suggests that each content in the subset is valuable. The content and the quality/bitrate HRC subsets come next. When leaving one sample out from a subset that has many samples, a negligible drop in correlation means that this sample is redundant, whereas a huge drop in correlation means that the subset is not good enough. By considering this assumption, the content and the quality/bitrate HRC subsets are the ones that perform well.
In the third category of the experiments, the influence of individual HRCs cannot be seen by removing one HRC from the training phase and then tested with this HRC since, in this experiment, only 10 sources are used and there are HRCs that share the same encoder conditions. Therefore, one HRC group, i.e., coding condition, is removed from the training phase and the model is tested with this group. It is observed, as shown in Table 12.2, that the main HRC groups that have the highest impact are the quality groups whereas other groups such as 'Open/Closed GOP', 'Intraperiod', and 'Slice Arg.' have stable results and higher PCC compared to quality groups. In general, removing one of these HRCs groups will highly impact the training model. For instance, including HRCs of low quality (QP=46) and high quality (8 Mbps, 16 Mbps, and QP=26) will help the model in better predicting the quality of new sample videos.

12.3.6 Training and testing results for bit-stream based no-reference model

Figure 12.16 shows the same training/testing experiments done for the pixel-based features, but this time for bit-stream-based features NR VQA. Here, the samples (HRCs) are common between the pixel-based model and the pixel-based model. It can be observed that all HRC subsets have a high correlation which makes it quite difficult to distinguish between them. Since samples and features are important inputs for the training, the following conclusions can be drawn: first, the bit-stream features are optimal for the prediction, so all subsets have a high correlation. Second, the performance measures (PCC and RMSE) are not indicative of the significance of the HRCs. Further discussion will be elaborated in Section 12.4.

12.3.7 Results from different machine learning algorithms

The two NR VQA models are trained using Stochastic Gradient Boosted Regression Trees algorithm, which recently has been shown objectively to be the state-of-the-art approach on structured data [304]. Furthermore, XGBoost is used, which is the state-of-the-art variation of the Stochastic Gradient Boosted Regression Trees algorithm [305]. In recent years, the popularity of this algorithm has risen dramatically due to its performance results in many machine learning competitions. For example, on the Kaggle platform , 17 out of 29 challenge winning solutions in 2015 used XGBoost. These XGBoost-based approaches outperformed both neural network and support vector machine-based solutions [306]. Apart from its success in machine learning competitions, XGBoost has also been proven to work well for practical applications such as train occupancy prediction [307], offshore wind turbine power prediction [308] and ads click-through prediction [309]. The success of XGBoost is often attributed to several aspects such as the fact that it is an ensemble model, requires little hyper parameter tuning, can deal with sparse data, requires no feature scaling and is very scalable due to its out-of-core learning ability [306].
The aim of this step is to observe some similarities and some dissimilarities when using different machine learning

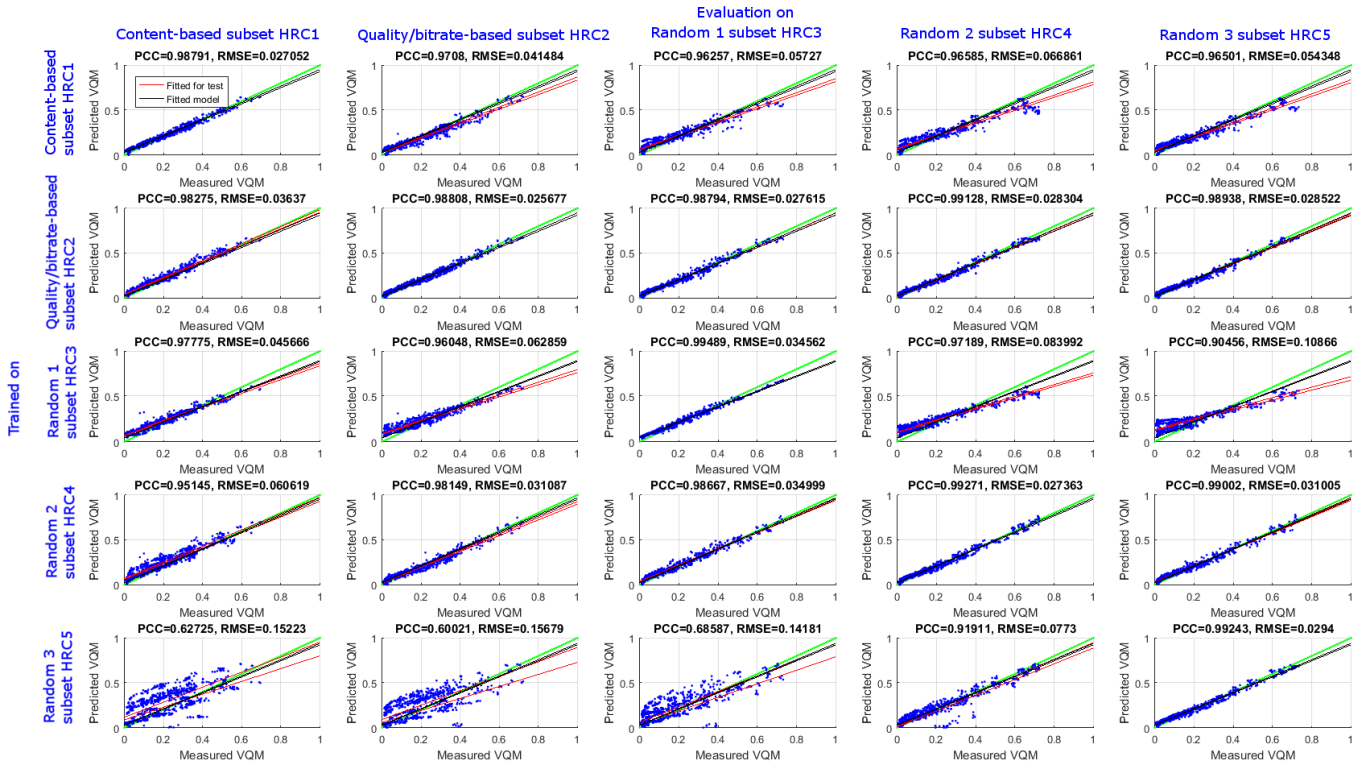


Figure 12.14 – The PCC and the RMSE for the 25 experiments of the pixel-based model. Rows: the different training models that are trained using $HRC_1, HRC_2, HRC_3, HRC_4$, and HRC_5,. Columns: the test data for each model, from the left, $HRC_1, HRC_2, HRC_3, HRC_4$, and HRC_5. The green line is the reference ($y = x$).

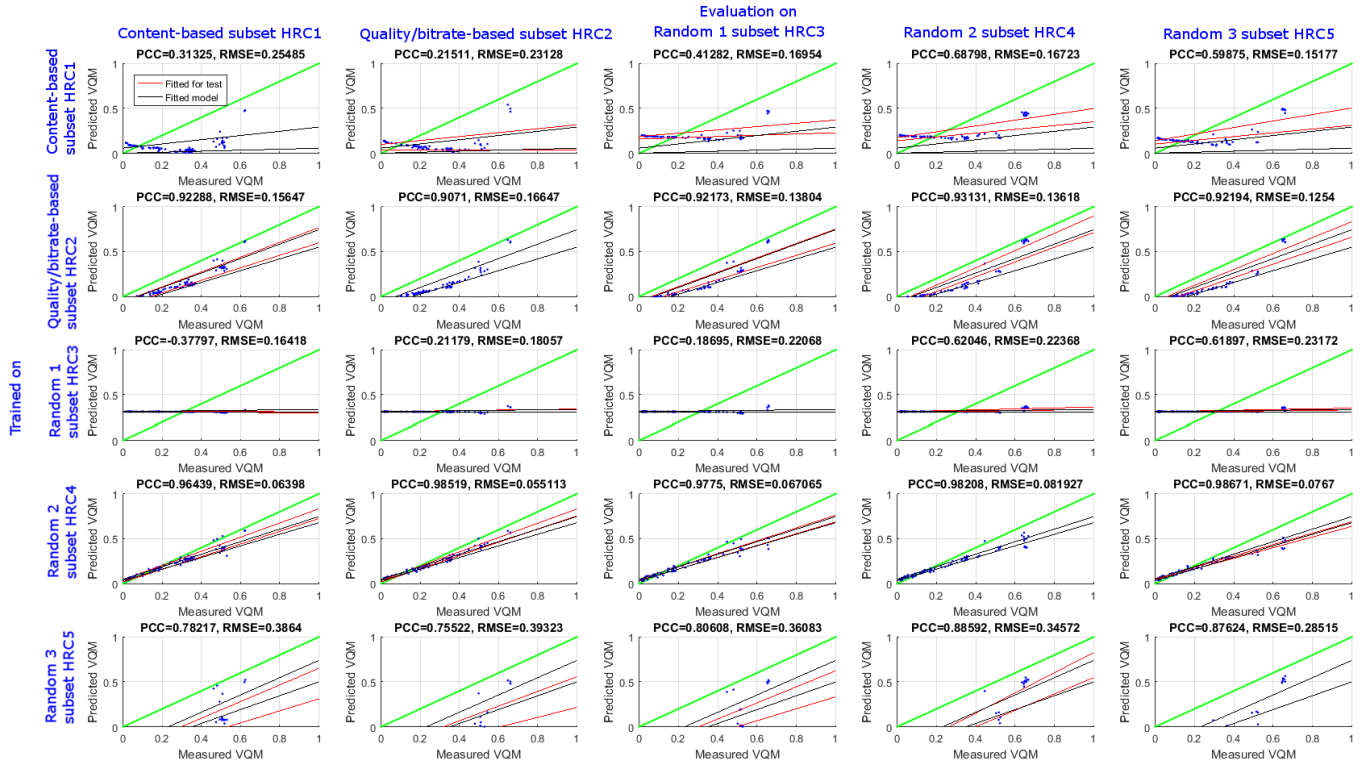


Figure 12.15 – The PCC and the RMSE for the 25 experiments that are trained without source 5 and tested with source 5 HRCs. Rows: the different training models that are trained using $HRC_1, HRC_2, HRC_3, HRC_4$, and HRC_5. Columns: the test data for each model, from the left, $HRC_1, HRC_2, HRC_3, HRC_4$, and HRC_5. The green line is the reference ($y = x$).

algorithms. The five models are trained using the same selected features for the pixel-based and bitstream-based VQA

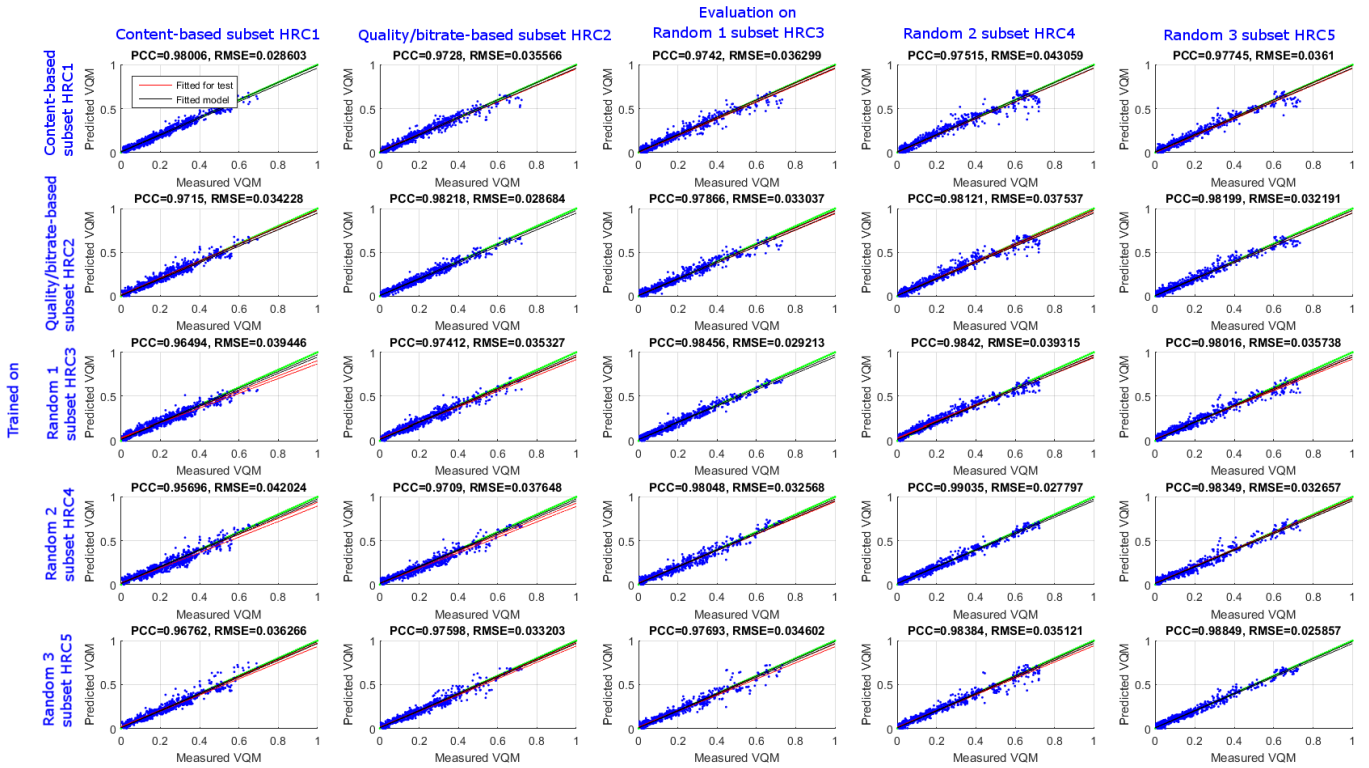


Figure 12.16 – The PCC and the RMSE for the 25 experiments of bitstream-based model. Rows: the different training models that are trained using $HRC_1, HRC_2, HRC_3, HRC_4$, and HRC_5,. Columns: the test data for each model, from the left, $HRC_1, HRC_2, HRC_3, HRC_4$, and HRC_5. The green line is the reference ($y = x$).

models. In comparisons with SVR technique and using PCC performance measure, the XGBoost results confirm that HRC_5 is the worst subset and it disagrees in the performance order of other subsets. The absolute mean difference of PCC is considered as the stability measure between two different machine learning algorithms. Therefore, because they almost agree when using the absolute mean difference of the PCC performance measure, see Tables 12.1 and 12.3, we will complete the rest of this paper with using SVR technique.
After comparing the two machine-learning algorithms, the two NR VQA models using XGBoost are subsequently compared. From Table 12.3, it can be concluded that content-based and quality/bitrate-based HRC subsets provide the most optimal results with respect to the correlation analysis; the absolute mean difference performance measure, Regarding the pixel-based model, HRC_3 shows better performance on average, then HRC_1 HRC_2 HRC_4 HRC_5 come next in order. It should be noted that it is difficult to distinguish the difference between the $HRC_{1,2}$. On the other hand, in bitstream-based models, the average correlation shows that the content-based subset is the worst and it is very hard to distinguish the difference between the $HRC_{2,3,4}$. But when analysing the absolute mean difference of PCC, the content-based and the quality/bitrate-based subsets perform better than others for the both NR VQA models.

12.4 Performance measures for models and (sub)sets

As explained in the introduction, one of the main goals of this work is to have an HRCs subset that can represent the large scale database. Hence, a set of analyses for the predicted values should be identified in order to judge the datasets. In the previous section, it is observed that the usual PCC and RMSE are not enough to judge a dataset. In this section, other analyses are proposed for performance evaluation.
Please note that a prerequisite of all the following measures is that the input data is restricted to the unit interval, zero to one. This can be achieved by linear rescaling in most cases.

12.4.1 Analysis of the residuals using PCA

12.4.1.1 Redundancy in the training data $P_{\text{RPCA\_T}}$

Purpose: measure the goodness of the training data in the training process by Analysis of the residuals using PCA.
Idea: find the systematic redundancies in the training data that should be avoided such as redundant HRCs or

Table 12.2 – PCC of the prediction using leave-one-out strategy, i.e. leave one HRC group out.

HRC groups			Data sets				
			HRC_1	HRC_2	HRC_3	HRC_4	HRC_5
GOP		GOP2	0.99	0.98	0.99	0.99	0.99
		GOP4	0.98	0.99	1.00	0.99	0.99
		GOP8	0.96	0.98	0.99	0.99	0.99
		LDGOP4	0.98	0.99	0.99	0.98	0.99
Quality Control	Bitrate	500000	0.96	0.97	0.96	0.97	0.97
		500001	0.95	0.96	0.94	0.94	0.95
		1000000	0.98	0.98	0.98	0.96	0.97
		1000001	0.98	0.98	0.97	0.97	0.97
		2000000	0.97	0.97	0.98	0.95	0.94
		2000001	0.99	0.93	0.98	0.96	0.95
		4000000	0.83	0.80	0.91	0.86	0.88
		4000001	0.88	0.87	0.93	0.94	0.91
		8000000	0.58	0.34	0.72	0.67	0.63
		8000001	0.55	0.25	0.77	0.75	0.62
		16000000	-	0.13	0.50	0.22	0.04
		16000001	0.23	0.13	0.54	0.20	-0.02
	QP	26	0.92	0.93	0.93	0.95	0.91
		32	0.92	0.97	0.94	0.92	0.95
		38	0.94	0.96	0.90	0.93	0.91
		46	0.08	0.42	0.29	0.54	0.22
Open/ close GOP		1	0.96	0.98	0.99	0.99	0.99
		2	0.99	0.99	0.99	0.99	0.99
Intra-period		8	0.97	0.98	0.98	0.99	1.00
		16	0.99	0.98	1.00	0.99	0.99
		32	0.97	0.98	0.99	0.99	0.98
		64	0.97	0.99	0.99	0.99	0.99
Slice Arg.		0	0.98	0.98	1.00	0.99	0.99
		2	0.99	0.98	1.00	0.99	1.00
		4	0.99	0.98	0.99	0.99	0.98
		1500	0.98	0.99	0.99	0.99	0.99

redundant contents. By identifying similar behaviour of the RMSE for two contents over all HRCs, redundancies can be identified. Optimality is reached if the HRCs behave differently for any two contents of the subset. The same applies to the HRC analysis: Optimality is reached if the contents behave differently for any two HRCs of the subset.
Process: train the model and evaluate it on the training data. Calculate the residual errors of the prediction by the model. For the content analysis (dimension=SRC), first create a vector per content that contains the residual errors for each HRC. Perform a PCA on these m vectors. Calculate the sum of the Eigenvalues of the first n components of the total m components. The default value should be $n = 0.2m$. Perform the same operations by creating a vector per HRC (dimension=HRC).
Reporting: Use $P^{\text{dimension}}_{\text{RPCA\_T}}(\frac{n}{m}, m) = x$, i.e. $P^{\text{SRC}}_{\text{RPCA\_T}}(0.2, 10) = 0.9$.
Interpretation: the lower the value, the better because the explained variance is low in the first n components, i.e. the remaining components have significant information.
Example and further explanations: The titles of the subplots in Figure 12.17 show the sum of the first two principal components, i.e $P^{\text{SRC}}_{\text{RPCA\_T}}(0.2, 10)$. The higher the value, the higher the possibility of existing systematic redundancy. As shown in the diagonal of Figure 12.17, it is observed that the quality/bitrate-based subset has the lowest explained variance in the first two components. That is an indication that the HRCs are valuable in the subset. Using the RMSE would not provide the same information, as can be seen from Fig. 12.14 because the best subset with respect to different HRCs is not easy to identify.

12.4.1.2 Redundancy in the validation data $P_{\text{RPCA\_V}}$

Purpose: measure the goodness of the validation data in the subset and model comparison process by analysing the residuals using PCA. In other words, characterizing which subset is challenging for the trained models.
Idea: similar to $P_{\text{RPCA\_T}}$, find the systematic redundancies in the validation data. Redundancy should be avoided,

Table 12.3 – Correlation analysis, expressed as a percentage, for the NR VQA models using XGBoost

		PCC						Difference to train PCC					Absolute
		HRC_1	HRC_2	HRC_3	HRC_4	HRC_5	Average	HRC_1	HRC_2	HRC_3	HRC_4	HRC_5	mean difference
Pixel-based	HRC_1	96.5	95.8	96.9	97.2	97.0	96.7	0	0.69	-0.40	-0.78	-0.58	0.27
	HRC_2	94.6	96.4	96.3	97.0	96.8	96.2	1.76	0	0.08	-0.59	-0.47	**0.20**
	HRC_3	98.2	98.5	99.9	99.4	99.4	**98.9**	1.72	1.38	0	0.46	0.46	1.00
	HRC_4	87.1	89.5	92.0	95.9	95.1	90.9	8.76	6.35	3.89	0	0.77	4.94
	HRC_5	82.9	85.5	90.3	93.7	94.3	88.1	11.35	8.74	4.01	0.59	0	6.17
Bitstream-based	HRC_1	98.0	95.9	97.0	96.5	97.0	96.6	0	2.13	1.02	1.46	1.03	**1.41**
	HRC_2	97.9	100.0	98.8	98.5	98.9	**98.5**	2.10	0	1.21	1.47	1.06	1.46
	HRC_3	97.2	97.6	99.6	98.7	98.4	98.0	2.47	1.98	0	0.94	1.18	1.64
	HRC_4	97.1	97.8	98.8	99.9	99.0	98.2	2.80	2.14	1.12	0	0.97	1.76
	HRC_5	94.5	96.9	97.4	98.2	99.7	96.7	5.25	2.89	2.37	1.58	0	3.02

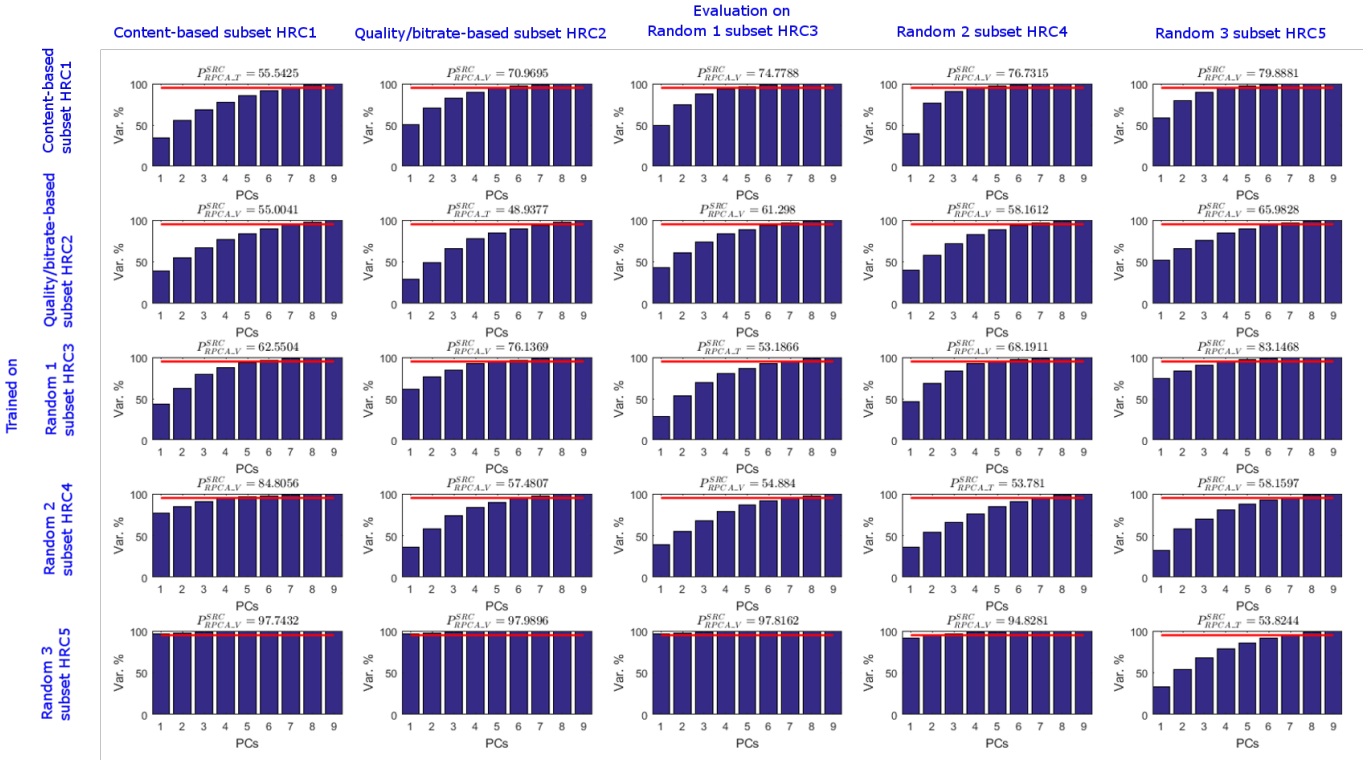


Figure 12.17 – The cumulative sum of explained variances of the principal components. The red lines indicate when the model reach a 95% of cumulative variances.

both as redundant HRC or redundant content.

Process: Evaluate the already trained model on the validation data without retraining. Then, follow the process of Section 12.4.1.1 in order to obtain $P_{\text{RPCA\_V}}$.

Reporting: Use $P^{\text{dimension}}_{\text{RPCA\_V}}(\frac{n}{m}, m) = x$, i.e. $P^{\text{SRC}}_{\text{RPCA\_V}}(0.2, 10) = 0.4$.

Interpretation: The lower the value of $P_{\text{RPCA\_V}}$, the better the performance as less redundancy is found in the considered dimension in the validation. There are two sources of redundancy in the validation: The first one is the same as with training, i.e. the used set contains systematic redundancies, the second source is that the model may behave alike for different conditions, such as not considering a certain degradation at all. When using several subsets for training and validation, further analysis on these two can be obtained by comparing the graphs cross-wisely, i.e. $X(n,:)$ and $X(:,n)$. Good models should provide low values of $P_{\text{RPCA\_V}}$ row-wise in $X(n,:)$ and good subsets for verification are characterized by high values of $P_{\text{RPCA\_V}}$ column-wise, i.e. $X(:,m)$ because a high value indicates that a model is challenged by the subset m, i.e. the model cannot reliably predict this subset.

Example: following the above interpretation in Figure 12.17, models that were trained on the specific subsets are performing in the following rank order: quality/bitrate-based, random 2, random 1, content-based, random 3. The subsets in decreasing order of goodness are quality/bitrate-based, content-based, random 2, random 1 and finally random 3.

12.4.2 Analysis of confidence intervals (CIs) of the different models fittings

In this section, further performance measures for the models are explained. These analyses are based on two different notions of confidence intervals. When fitting a model, the parameters of the model are determined based on training data. The more training data is available and the better the model fits, the smaller the confidence intervals for each calculated model parameter. In this text, this is called the model confidence, *model-C*. When the model is used for prediction, a certain percentage (usually 95%) of the predicted data lies in a corridor bounded by the upper and lower confidence intervals. This is called the data confidence, *data-C*.

12.4.2.1 Model's prediction performance on particular validation dataset

Purpose: measure the goodness of the trained model with respect to its reliability of predicting validation data.
Idea: Determine the confidence interval corridor for the model predicting its own training data. Then, count the number of validation data points that fall into this corridor.
Process: Train the model on the training data. Evaluate it on the validation dataset. The 95% confidence interval boundaries for *data-C*are obtained by using a function such as MATLAB's [1] *polycon* function. This function is applied on the training data in order to get the two boundary lines that are parallel to the fitting line, i.e. $y \pm \delta$. The validation data is then predicted by the same model and for each data point it is determined whether it is inside the previously determined confidence interval boundary. The ratio of inliers i and outliers o of the total number of data points in the validation set n is reported. This is similar to the well-known outlier ratio with respect to the standard error but takes into consideration training and validation. This analysis is taken further in Section 12.4.3.
Reporting: Use $P_{\text{DCI\_V}}(\delta, n)=\frac{i}{o}$, i.e. $P_{\text{DCI\_V}}(0.12, 100)=0.3$.
Interpretation: The higher the ratio, the better the model predicts the validation data with respect to its own training data.
Example: as shown in Figure 12.18, the black lines are the boundaries of *data-C*of the trained model and the black points are the predicted data points of the trained data. The red points are the predicted data points of the validation data. In addition, the red lines show the boundaries of *data-C*using the validation subset (further exploited in 12.4.3). Each sub figure reports the $P_{\text{DCI\_V}}$. For instance, the fifth row $X(5,:)$ showing the validation of the model trained on Random3 shows that the spread of content and quality/bitrate based subset is largest compared to the other subsets which is reflected in the value of $P_{\text{DCI\_V}}$. Since the content-based model is not designed to have a wide range of quality and bitrate, the predicted VQM values of content-based HRCs lie mostly outside the area of the *data-C*CIs for other models. Therefore, its HRCs are challenging for other models, especially random-based models.

12.4.2.2 Model determined by its training data

Purpose: measure the goodness of the model by analysing the area of the confidence interval spread by the *model-C*.
Idea: the size of the area of the model parameter's confidence spread provides information about the exactness with which the model parameters can be determined by the training data.
Process: determine the confidence interval values for each of the trained model parameter on the training data of size n. For a linear model, gradient and offset have a confidence interval that is provided by the fitting function, e.g. the MATLAB[1] function $fitlm$ in conjunction with $coefCI$. Determine the maximum confidence boundaries that are spread by the uncertainty, e.g. for a linear model the lower bound is determined by the line $y = (a - CI_a)x + (b - CI)$. Calculate the area of uncertainty x, e.g. for a linear model between the lower and upper bound $y = (a+CI_a)x+(b+CI)$. This analysis is taken further in Section 12.4.3.
Reporting: Use $P_{\text{MCI\_T}}(n)=x$, i.e. $P_{\text{MCI\_T}}(120)=0.4$.
Interpretation: the lower the value, the better the model is able to predict its training data. Please note that a low value may also indicate overtraining or irrelevant training data. Hence, we will consider this value when considering the interaction between the training data, and the validation data in Section 12.4.3.
Example: Figure 12.14 and Figure 12.15 show in each subplot 5 lines. The green line represents $y = x$. The two black ones are the CIs of a fitted model, therefore each row in the figures shows the same two black lines. The value of $P_{\text{MCI\_T}}$(n) for the five models are, respectively, 0.014, 0.014, 0.008, 0.011, and 0.010. The sign of the over-fitting is obvious for random-based subsets. Going further than $P_{\text{MCI\_T}}$, the two red lines represent the *model-C*CIs when trained on the validation set. This leads to ideas to observe the amount of overlap between training a model on one or the other subset. A good model is characterized by red lines located between black lines. As shown in Figure 12.14, the amount of overlap in quality/bitrate-based model is the largest.

1. MATLAB functions are given here for exact reproducibility, other software such as Octave or R has similar functionality

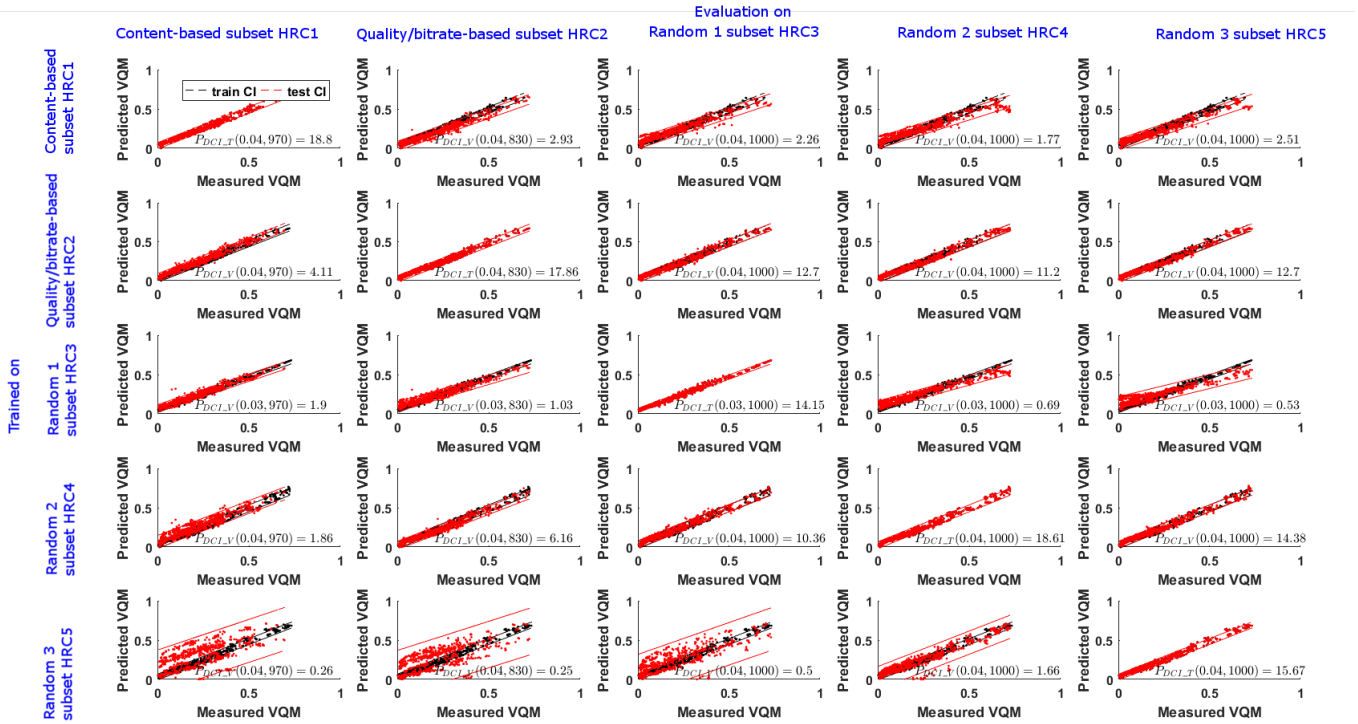


Figure 12.18 – How much the predicted VQM values lie in area of confidence interval of the fitted data.

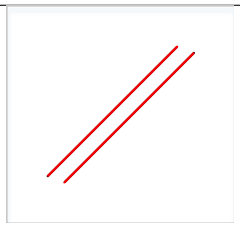
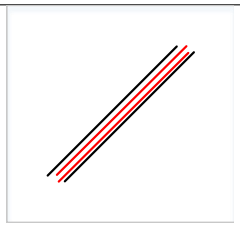
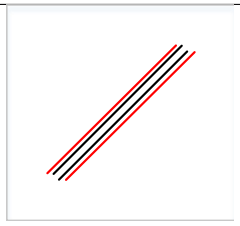
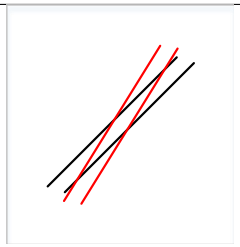
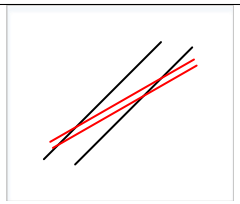
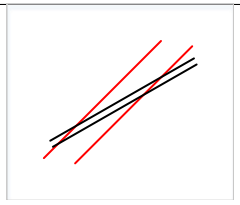
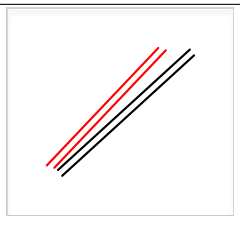
12.4.3 Interaction between the model training, the training data, and the validation data

In this subsection, the goodness of the trained model on its own training data and on a specific validation subset is studied. The analysis can be applied to the model fit (*model-C*CI) or to the data fitting ability (*data-C*CI). The following three attributes of the CI analysis are used. The area between the CI boundaries of the training (black lines, denoted as b), the area between the CI boundaries of the validation (red lines, denoted as r), and finally the area of the intersection between the two areas (denoted as i). The main conditions with respect to line intersections are explained in Table 12.4.

12.4.3.1 Goodness of data prediction using a trained model on validation data

Purpose: provide an absolute number for the prediction performance of a trained model on a validation dataset taking into consideration the training dataset.
Idea: The model prediction performance can be characterized by the *data-C*CI of the validation data. The smaller the CI, the better the model. Taking into consideration the training process, the smaller the CI on the training data, the better the model. Finally, taking into consideration the interaction between the training and the validation, the larger the intersection between the CI, the better the fitting.
Process: Train the model on the training data. Calculate the area of the *data-C*CI corridor similar to Subsection 12.4.2.1 in order to obtain the area b. Perform the same operation without retraining on the validation data in order to obtain r. Calculate the intersection between the two corridors in order to obtain i.
Reporting: The goodness value is reported as $P_{\text{GData}}^{(\text{b,r,i})} = \frac{i}{\max(b,r)^2}$, e.g. $P_{\text{GData}}^{(0.5,0.4,0.3)} = 1.2$
Interpretation: The higher the value, the better the model's performance and the data that the model was trained on. The calculation is divided into two terms, the first one being $\frac{i}{\max(b,r)}$ which gets to its maximum 1 if the intersection covers exactly the larger area and is equal to zero for no overlap. The second term is $\frac{1}{\max(b,r)}$ which gets larger, the smaller the CI areas get. The measure was designed to provide a reasonable tradeoff between these goals. The behaviour of this measure can be seen in Figure 12.19.
Example: Figure 12.20 shows 4 sub-figures, the first column is related for $P_{\text{GData}}^{(\text{b,r,i})}$ for the features-based NR VQA and bitstream-based NR VQA. In both NR VQA models, the quality/bitrate-based dataset has the largest $P_{\text{GData}}^{(\text{b,r,i})}$ value. While the content-based subsets is ranked the third and the fifth in both models respectively.

Table 12.4 – List of interesting cases for analysis of the *data-C*CI, the cases for *model-C*CI are similar. Black lines indicate the CI on the training data. For simplicity it is assumed that these are fixed which is true in most practical cases. Red lines indicate the CI on the validation data.

Case	Icon	Condition	Note
1		$b = r = i$	Typical case for validating on the training data, this is considered the perfect fitting, i.e. all three areas are identical. Refer for example to the main diagonal $X(n,n)$ in Fig. 12.14. In this case, $G = \frac{1}{\max(b,r)}$. To compare between different models or data, the lower the $\max(b,r)$, i.e. the smaller the larger CI, the better.
2		$r = i$	The validation data is better predicted than the training data and the CI lie completely within the boundaries of the trained model. This is likely to be a default of the validation data and thus reduces the goodness as compared to Case 1. In this case, $G = \frac{r}{b^2}$.
3		$b = i$	The validation data is less well predicted than the training data but the validation CI covers completely the training CI. This is considered a case of overfitting of the model and should thus be penalized compared to case 1. In this case, $G = \frac{b}{r^2}$.
4		$b \approx r$	This is the typical case of slight deviation between training and validation. The goodness depends mainly on the intersection area. In this case, $G = \frac{i}{\max(r,b)^2}$.
5		$b >> r$	These cases indicate a larger misalignment either of the training CI or the validation CI with respect to the model fit and thus are a combination of case 4 with the cases 2 and 3 respectively. In these cases, the smaller intersection penalizes the goodness compared to the case 4 as the value of i is smaller in $G = \frac{i}{\max(r,b)^2}$
6		$b << r$	
7		$i = 0$	This is the worst case, the validation data does succeed in being predicted by the model, thus $G = 0$. Please note that this may also be an indication of a missing alignment between the training and validation data. An additional alignment step may be required in particular for models that were trained on different conditions (e.g. different video encoder).

12.4.3.2 Goodness of two datasets for determining linear model parameters

Purpose: evaluate the model's stability when using different datasets.
Idea: A good model should provide a stable linear relation to any given dataset. This is similar to Subsection 12.4.3.1

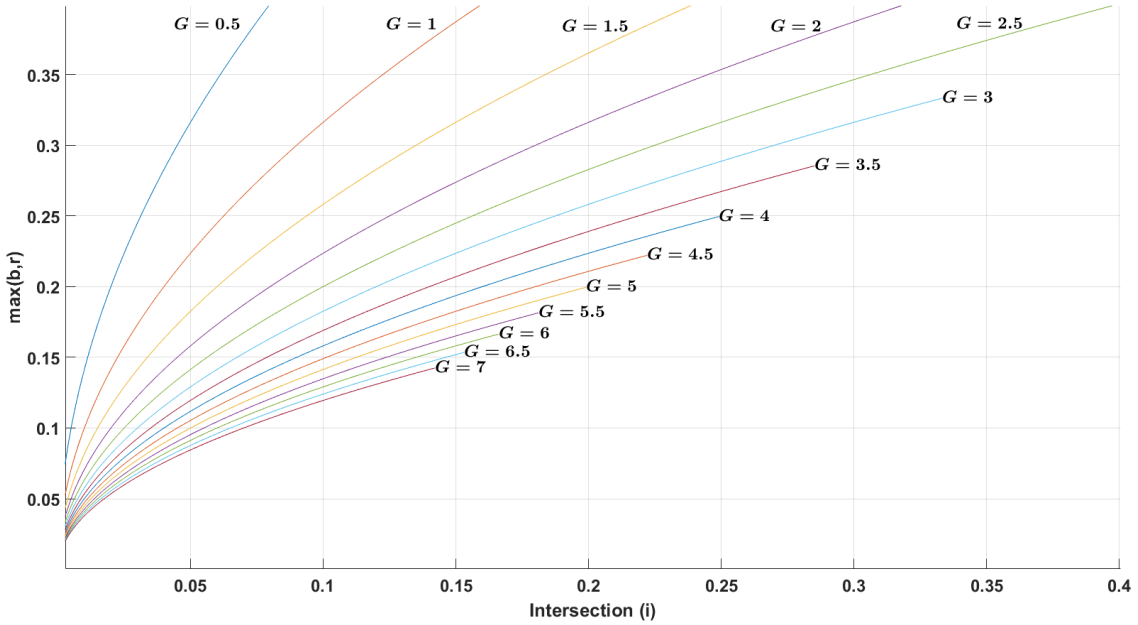


Figure 12.19 – The behaviour of G with different $max(b,r)$ and i values.

with exchanging the *data-C*with the *model-C*
Process: Train the model on the training data. Perform a linear fitting on the training data and calculate the area b=$P_{\text{MCI_T}}$as explained in Subsection 12.4.2.2. Perform a linear fitting on the validation data and calculate the area r similarly. Calculate the intersection between the two areas i.
Reporting: The goodness value is reported as $P^{(\text{b,r,i})}_{\text{GModel}} = \frac{i}{\max(b,r)^2}$, e.g. $P^{(0.5,0.4,0.3)}_{\text{GModel}} = 1.2$

Interpretation: The higher the value the better. The same explanation as in Subsection 12.4.3.1 holds but in this case, the stability of the model to predict different datasets is analysed.
Example: The second column of Figure 12.20 is related for $P^{(\text{b,r,i})}_{\text{GModel}}$ for the features-based NR VQA and bitstream-based NR VQA. In both NR VQA models, the quality/bitrate-based dataset has the largest $P^{(\text{b,r,i})}_{\text{GModel}}$(b,r,i) value. While the content-based subsets is ranked the third and the second in both models respectively.

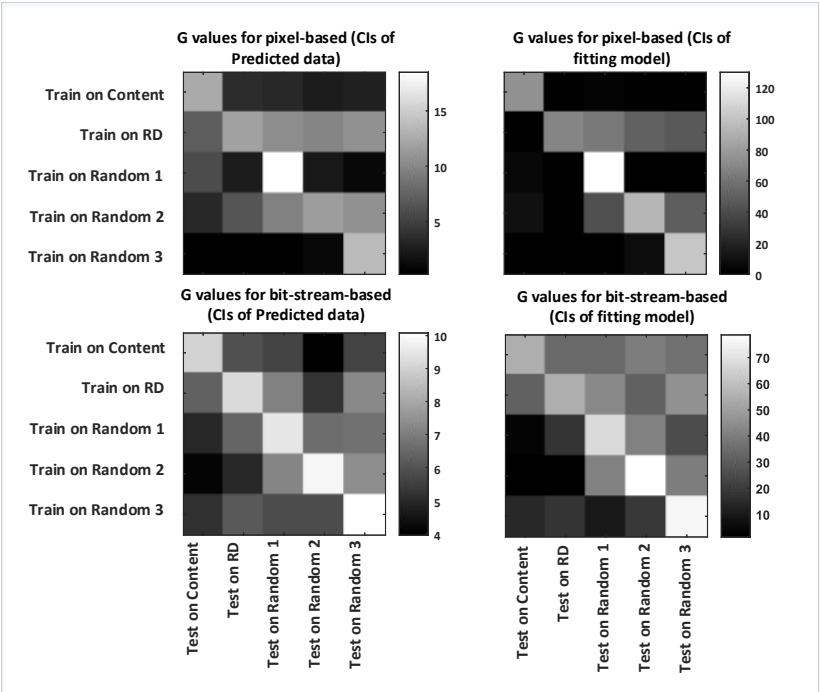


Figure 12.20 – The G values for the CIs analysis for the pixel-based and bit-stream-based NR VQA models

Table 12.5 – All HRC subsets ranks for each performance measure and for the pixel-based and the bit-stream-based NR VQA measures.

Performance measure	Pixel-based NR VQA (Proposed)					Bit-stream-based NR VQA				
	Content	RD	Rand 1	Rand 2	Rand 3	Content	RD	Rand 1	Rand 2	Rand 3
PCC Cross-dataset	3	1	4	2	5	4	1	3	5	2
PCC Leave-one-out	2	1	5	3	4	2	4	3	5	1
PCC Challenging HRCs	2	1	3	4	5	1	2	3	5	4
RMSE Cross-dataset	3	1	4	2	5	5	1	4	3	2
RMSE Leave-one-out	2	1	4	3	5	3	5	4	1	2
RMSE Challenging HRCs	1	2	3	4	5	1	3	5	2	4
$P^{\mathrm{SRC}}_{\mathrm{RPCA\_T}}(\frac{n}{m},m)$, $P^{\mathrm{SRC}}_{\mathrm{RPCA\_V}}(\frac{n}{m},m)$	3	1	4	2	5	1	2	1	1	3
$P_{\mathrm{DCI\_V}}(\delta,n)=\frac{i}{o}$	3	1	4	2	5	2	1	3	5	4
$P^{(\mathrm{b,r,i})}_{\mathrm{GModel}}$	3	1	4	2	5	2	1	3	4	5
$P^{(\mathrm{b,r,i})}_{\mathrm{GData}}$	3	1	4	2	5	5	1	2	3	4
Average	**2.5**	**1.1**	3.9	2.6	4.9	**2.6**	**2.1**	3.10	3.4	3.10

12.4.4 Comparing the performance of HRC subsets

As discussed and observed in the previous sections, all the aforementioned performance measures yield different results for different HRC subsets. Therefore, in this subsection, all the results are put together in order to judge the HRC subsets. A rank-order technique is applied in order to get a final score for each HRCs set. Since we have 5 HRC subsets, each HRCs set will have an order number for each performance measure discussed in this paper and then a comparison between the two NR VQA measures is shown. Table 12.5 shows all HRC subsets ranks for each performance measure and for the pixel-based and the bit-stream-based NR VQA measures. From the table, it can be surmised that the systematic way of selecting the HRC set to be used on the experiments performs better than the random selection that covers different ranges of bitrate and quality. The key element of using such technique in evaluation is that when a given performance measure cannot give clear indications about which model is better than others, another performance measure can.

12.4.5 Detailed Analysis of Support Vectors

Support vector (SV) based machine learning is one of the widespread methodologies for regression fitting. Important insight can be gained from support vectors because they are actual data points from the training dataset. In most cases this means that some of the created conditions (resulting in PVS) are deemed of foremost importance for representing the whole training dataset.

Purpose: Evaluate the efficiency of the distribution and the weighting of the selected support vectors with respect to the ground-truth quality.

Idea: Because the SVs are training data points, each support vector is assigned one ground truth quality score. The machine learning should choose SV that equally spread over the predicted quality range. In other words, if the training chooses SV in a small quality subrange, the prediction may get unstable if confronted with conditions outside this particular quality range and the chosen SV may be redundant. The weighting of the SV needs to be taken into consideration.

Process: Train the algorithm on the training data and extract the SVs and their weights. Identify which training data point corresponds to each SV. Retrieve the ground-truth quality score (i.e. on which the algorithm was trained).

Reporting: Visualize the data in one or several scatterplots: on the x-axis the ground-truth quality score and on the y-axis the main parameter(s) of the condition for the SV (ex. bitrate). The size of the dots indicates the weight.

Interpretation: The more widespread the data points over the quality range, the better and the more stable the training result. Higher density of data points and/or higher weights in certain quality ranges should be analysed. They may either indicate redundant or overrepresented training data points or shortcomings of the algorithm in distinguishing between these closely related conditions. In order to distinguish further between such conditions, additional factors (ex. quality indicators) may need to be added to the prediction algorithm.

Example: Figure 12.21 is a typical result of this analysis. It shows the same prediction algorithm trained on three different training datasets (from left to right: $\mathrm{HRC}_{1,2,3}$). Here, VQM is used as ground-truth quality score and the "main parameter of the condition for the SV" is the quality control parameter being either QP (if<52) or bitrate (if>52). A major problem can be observed in the case on the right side. It is evident that the density of SV is higher on the low subrange and on the high subrange of the VQM scores. This is an indication of redundant HRCs in the set. On the other hand, the density of support vectors in the content-based (left) and the quality/bitrate-based (middle) training subsets is mostly uniform over the VQM scores.

The example showed the results for the pixel-based NR-VQA model. For the bit-stream-based NR VQA model, this strong difference is not observed, i.e. the SVs for each HRC subset cover different ranges of VQM score levels. This may be due to the usage of different factors, i.e. indicators, in the prediction algorithm. Notably the QP value is included as one of the quality indicators for the prediction algorithm. This leads to the effect that the output of the

prediction algorithm is mostly determined by QP or at least more stable when the QP value is similar. Therefore, the SVM may avoid selecting redundant SVs.

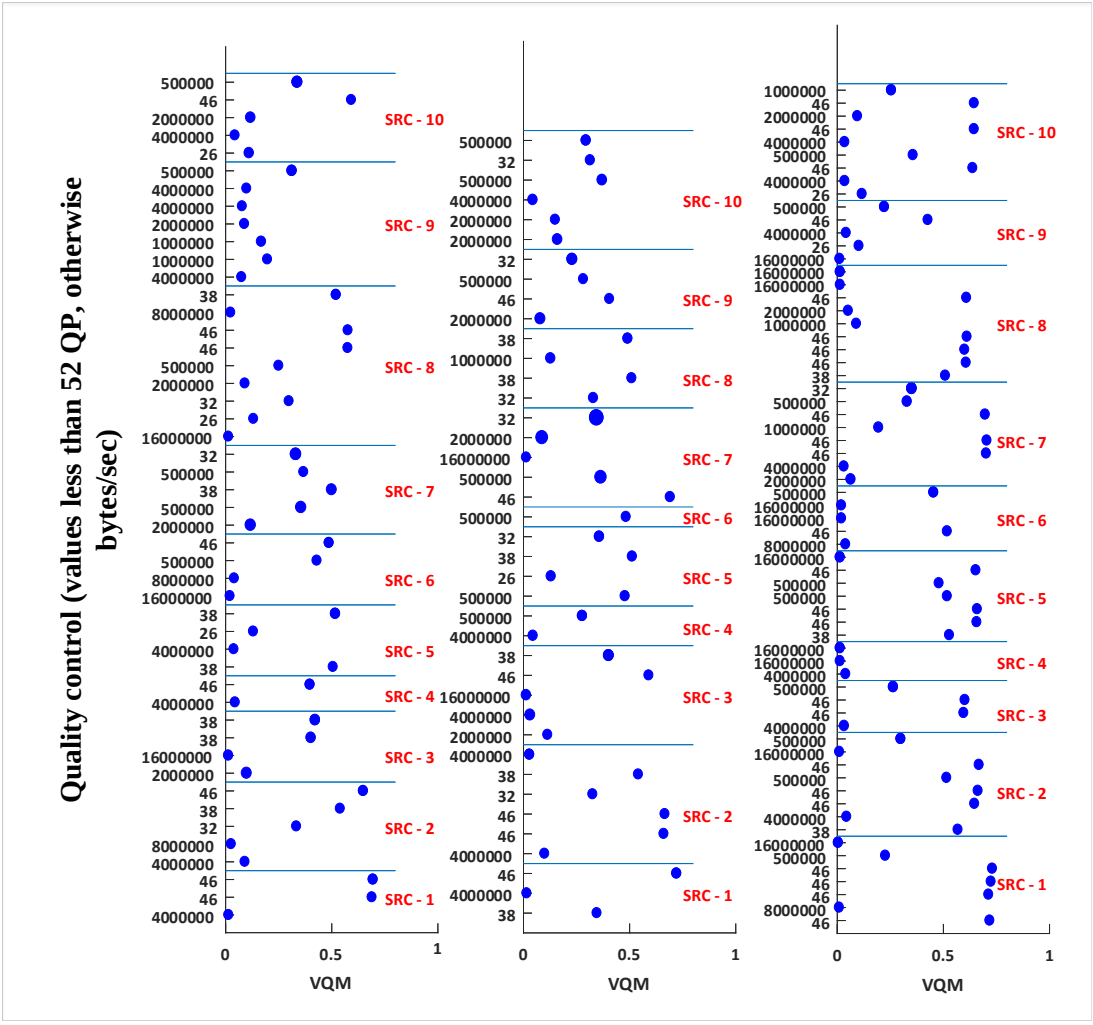


Figure 12.21 – The VQM quality score and the quality control parameter that are assigned to each SV of the following models: (left) HRC_1, (middle) HRC_2, and (right) HRC_3. The size of the dots indicates the weight of each SV.

12.5 No-reference image quality measure

We first present the methodology on a large-scale database that is evaluated objectively using the VQM algorithm. This ensures that sufficiently large subsets of the database can be extracted without database alignment issue. In this Section, we then present a sample application on a typical quality assessment case: the performance evaluation on several subjective datasets that can be seen as subsets of possible images/videos.

As discussed in the introduction, in this section the aim is not to compare two NR IQA measures, but to see how they work with different image datasets that are different in size, content, and image distortion types and levels. In [310], a machine-learning-based NR IQA measure is introduced. It uses Single Value Decomposition (SVD) based features as input for the machine learning algorithm. Here, the machine learning can be seen as a feature pooling technique; 256 features are extracted from the distorted images. In [311], natural scene statistic (NSS) based features are extracted from patches that correspond to original images. The resulting Natural Image Quality Evaluator (NIQE) is an opinion and distortion unaware model. These NR IQA measures are trained with different datasets that differ in content, size, and number of distortions. These datasets are: the TID database [312] (68 reference and 1700 distorted images), the IVC database [313] (10 reference and 185 distorted images), the Toyama database [314] (14 reference and 168 distorted images), and the WIQ database [315] (7 reference and 80 distorted images). TID images are distorted using 17 distortion types. IVC database uses four distortion types while the Toyama database uses two distortion types. Finally, WIQ uses four distortion types.

12.5.1 Evaluation method

The predicted quality is fitted using a 5-parameter logistic function as recommended in [316], and the correlation and the RMSE measures are size-weighted to calculate the average correlation and RMSE.

12.5.1.1 PCC

Figure 12.23 and Figure 12.24 show 16 experiments and the corresponding PCC and RMSE. In the SVD-based and NIQE models, the TID dataset performs better than the other datasets, i.e., the WIQ, IVC, and Toyama datasets, respectively. It seems that the WIQ dataset is not large enough to be used to build a NIQE model since it shows a very low correlation when it is tested on the training data. Hence, this dataset will be excluded in the final ranking order in the overall ranking Table 12.6. Regarding the distortion types that are challenging for the models, TID and the WIQ datasets show that there are, indeed, distortion types and levels that are challenging for other models. This result is expected since TID has very different distortion types and WIQ has different distortions than the ones in the IVC and Toyama databases.

12.5.1.2 RMSE

With the RMSE performance measure in SVD-based IQA, WIQ performs better than others. Then TID, IVC and Toyama come next in order. For the case of the NIQE model, the WIQ comes first, and then IVC, Toyama, and TID come next in order. Regarding the distortion types that are challenging for the models, IVC contains distortion types that are often challenging for other models.

12.5.1.3 $P^{\mathbf{SRC}}_{\mathbf{RPCA\_T}}$ and $P^{\mathbf{SRC}}_{\mathbf{RPCA\_V}}$

This measure is not applied here since the distortions are not similar in terms of distortion type or distortion level.

12.5.1.4 $P^{(\mathbf{b,r,i})}_{\mathbf{GData}}$

As discussed in Section 12.4.3.1, this measure tries to provide an absolute number for the prediction performance of a trained model on a validation dataset taking into consideration the training dataset. The first column of Figure 12.22 shows the $P^{(\mathrm{b,r,i})}_{\mathrm{GData}}$ of both image quality assessment. In the SVD-model and the NIQE-model, IVC and TID datasets have the ability to predict the quality within the confidence intervals corridors respectively.

12.5.1.5 $P^{(\mathbf{b,r,i})}_{\mathbf{GModel}}$

This measure reports the model's stability when using different datasets, Section 12.4.3.2. This is done by measuring the overlap between the two models (G), Table 12.4. As it can be seen in the second column of Figure 12.22, Toyama and IVC datasets have a higher stability in SVD and NIQE models respectively. In the SVD-based, the G value of the WIQ model is very small compared to the others: this is due to the bad fitting model for the training data, i.e. the black area is very large. Therefore, this dataset is not suitable for training for both models.

12.5.1.6 Performance comparisons

As discussed and observed in the previous sections, all the aforementioned performance measures give different results for different image datasets. In this subsection, all of them are put together in order to judge the image datasets. A rank-order technique is applied in order to get a final score for each image dataset. Since we have 3 image datasets (WIQ is excluded), each set will have an order number for each performance measure discussed in this paper and then a comparison between the two NR IQA measures is shown. Table 12.6 shows all the ranks for each performance measure and for the SVD-based and the NIQE model NR IQA measures. As can be seen from the table, there is no clear indication about which dataset can be used as a generalized dataset. In the SVD-based model, the TID dataset is the winner, while in the NIQE model the IVC dataset is the winner. This observation already considers the exclusion of the WIQ dataset due to its limitation in size and in types of distortions that are not in common with the other datasets. In conclusion, we recommend that different datasets should be tested with different performance measures when a new objective NR IQA tool is introduced.

12.6 Conclusion

This Chapter introduces the contributions that are listed in Box 12.3. In this Chapter, we discussed the effects of different training and validation datasets on the performance of objective quality measurement algorithms. As an example study, we used five subsets for training and validation; two were targeted towards different goals, three were

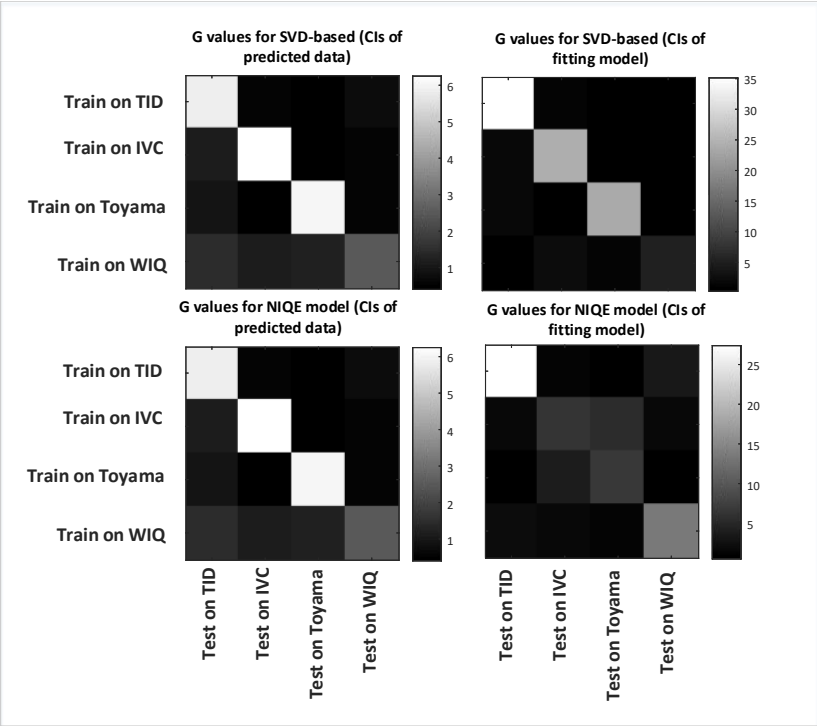


Figure 12.22 – The G values for the CIs analysis for the SVD-based and NIQE-model NR IQA models.

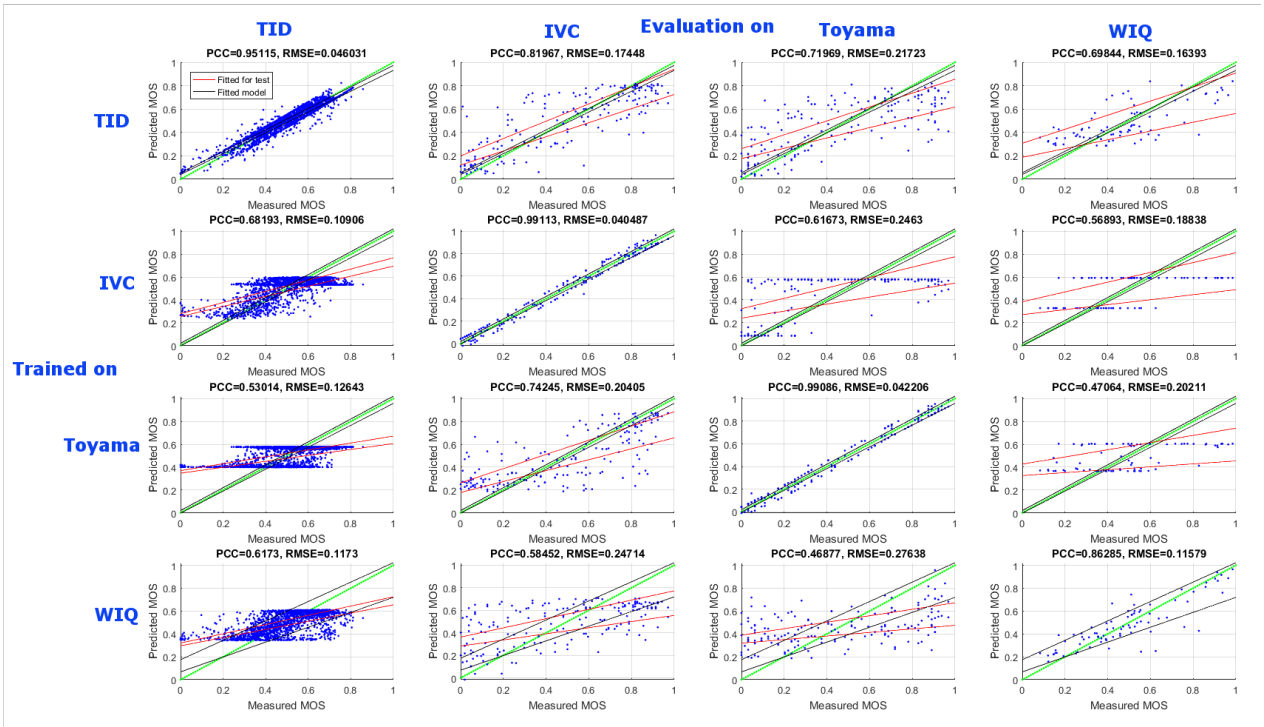


Figure 12.23 – The PCC and the RMSE for the 16 experiments that are trained and tested with source 4 image datasets of SVD-based NR IQA measure. Rows: the different training models that are trained using $TID, IVC, Toyama$, and WIQ. Columns: the test data for each model, from the left, $TID, IVC, Toyama$, and WIQ.

random. In the study, two NR VQA algorithms with typical quality indicators were trained by SVR.
We analysed the outcome of this widespread approach with state-of-the-art performance measures and identified important shortcomings. We therefore proposed several novel performance measures in three categories: The first category analyses the residual errors to find the systematic redundancies in the training and evaluation subsets. The second category provides insight on the training by using the confidence intervals of models fitting and the confidence interval of the predicted data. The third category is specific to SVR and the analysis of the density of SV over the

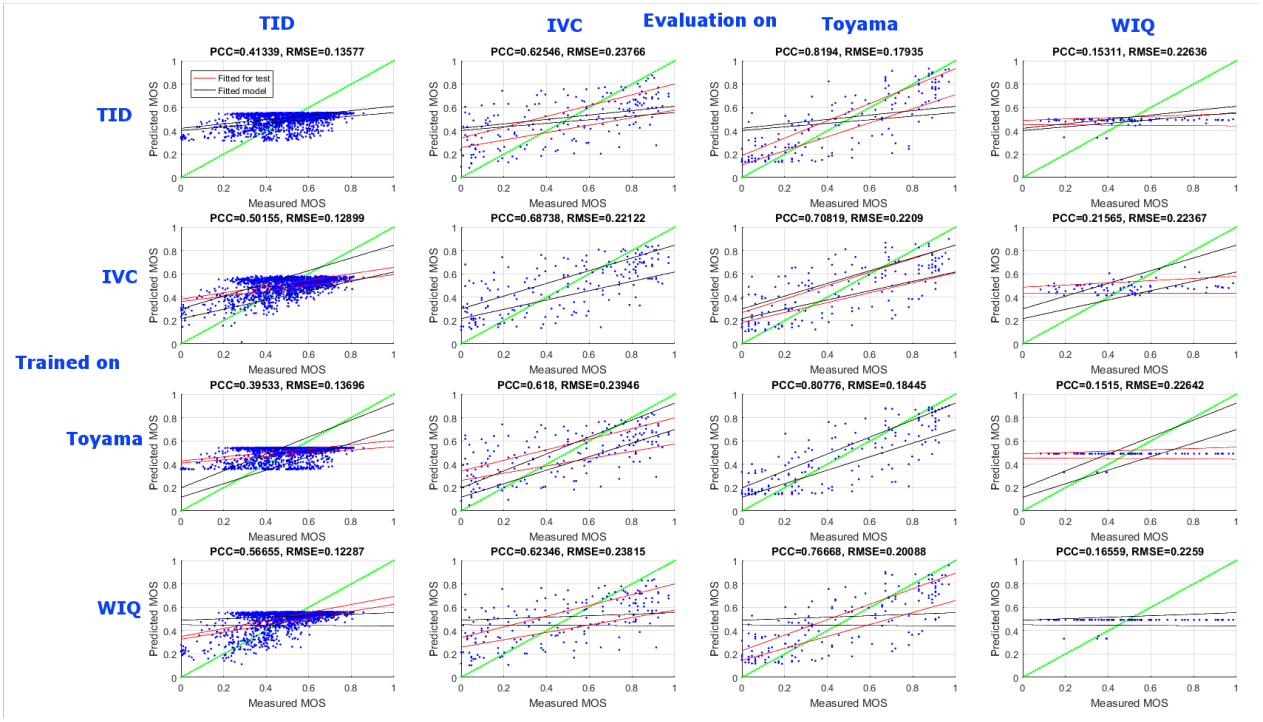


Figure 12.24 – The PCC and the RMSE for the 16 experiments that are trained and tested with source 4 image datasets of NSS-based NR IQA measure. Rows: the different training models that are trained using $TID, IVC, Toyama$, and WIQ. Columns: the test data for each model, from the left, $TID, IVC, Toyama$, and WIQ.

Table 12.6 – All image datasets ranks for each performance measure and for the SVD-based and NIQE NR VQA measures.

Evaluation	SVD-based NR IQA				NIQE NR IQA			
	TID	IVC	Toyama	WIQ	TID	IVC	Toyama	WIQ
PCC Cross-dataset	1	2	3	-	1	2	3	-
PCC Challenging HRCs	1	3	2	-	1	2	3	-
RMSE Cross-Dataset	1	2	3	-	3	1	2	-
RMSE Challenging HRCs	1	3	2	-	3	1	2	-
$P^{(\mathrm{b,r,i})}_{\mathrm{GModel}}$	3	2	1	-	2	1	3	-
$P^{(\mathrm{b,r,i})}_{\mathrm{GData}}$	3	1	2	-	1	2	3	-
Average	**2.17**	3.00	3.00	-	2.33	**2.00**	3.5	-

quality range.
An example study on image quality databases with subjective scores illustrates the usefulness of the performance measures.
The newly proposed performance measures are presented such that they can easily be reproduced. It would be very beneficial to report such measures in future proposals of video quality assessment algorithms in order to enable an in-depth analysis and a comparison across the proposals of different authors in the domain who often use varying datasets for training and validation.
Further performance measures may be required, in particular when training with other machine learning algorithms such as deep-learning.

Box 12.3 – Contributions

- Two subset selection algorithms are proposed. They are targeting a wide range of a specific target; quality/bitrate or content targets. Specifically, a small-scale set is selected from a large-scale database such that this small-scale database can be further analysed and the conclusions drawn on the small-scale database also apply to the large scale database.
- The following new performance measures are proposed for learning-based video quality assessment algorithms:
  - Measures depend on analysing the residual error using PCA,
  - Measures depend on analysing the confidence intervals of the predicted data, and
  - Measures depend on analysing the confidence intervals of the linear coefficients of the trained and tested models.

VI

Conclusion and future perspectives

13

Conclusions and future perspectives

13.1 Conclusions

In this dissertation, a research effort has been conducted to show how the underlying *content features/indicators* can be used for enhancing the video delivery chain performance. The following conclusions have been drawn regarding each delivery chain component.

Pre-encoding process (Part II): the *pixel-based features* have an impact in predicting the encoder behaviour of a specific content with respect to bitrate, distortion, and complexity. The proposed framework of predicting the encoder parameter values is able to link the content features with the parameter value that trades-off the bitrate, distortion, and complexity. The visualization tool, that shows the impact of the complexity in the rate-distortion optimization, shows that the proposed model can be applied in two different scenarios: 1) when the complexity matters: Chapter 4 follows this assumption. In Chapter 4, a model for predicting coding parameter values using the underlying content features is proposed. The model trades-off rate (R), distortion (D), and complexity (C). 2) when complexity doesn't matter: two dimensions will be used (bitrate and distortion) in order to select the configuration that gives better results. It was observed that using the complete set of encoder tools doesn't guarantee better results.

Source coding (multiple description coding) (Part III): in this part, beside showing the impact of pixel-based features, the impact of good/smart networking structure in reducing the amount of redundant data to be sent is shown as well. Regarding the impact of *pixel-based features*, it was observed that the CTU split decision and the distance between the secondary frame (used in the recovery process) and the reference frame of this secondary frame have an influence in generating the weights that are going to be used in the recovery process. Moreover, the pixel-based features have the ability to be used in building an adaptive content-aware multiple description scheme. Regarding the impact of good/smart networking structure, it was observed that the amount of redundant data to be sent may be reduced if the proposed network structure is used. This approach allows different types of applications that use error-prone channels to deliver video content with satisfying perceived quality.

Inpainting-based error concealment (Part IV): the impact of using *pixel-based features* is twofold in this part. The first impact is observed on using motion information derived from the motion vectors, motion intensity, and camera motion maps. These motion maps are combined with the logical OR operator to generate one map. Since having good motion information participates in better structure reconstruction of the lost area, this motion map is used as input for inpainting-based error concealment. After finishing from reconstructing moving pixels, the job, then, is easy for spatial inpainting to reconstruct the texture. The second impact is observed when the user disruption, when looking at impaired sequences, is analysed using with proposed entropy maps, that are computed from texture, colour, and motion information. These entropy maps are used as indicators to predict the observer's disruption in impaired sequences.

Quality assessment (Part V): in this part, the large-scale database is used as a good alternative to analyze the VQA measures and to evaluate their performance for scenarios that subjective experiments cannot evaluate due to its limitations. Hence, we can draw the following conclusion for different different quality assessment aspects. 1) As the whole PhD topic aims to, in this part, the impact of *pixel-based features* are used to build a prediction model that predicts the behaviour of the full-reference quality measures for error-free and loss-impaired sequences. 2) The encoder behaviour of different video sequences of the large-scale database is analysed and led to propose an algorithm to select a

set of encoding conditions that cover a variety of targets. The targets might be content-based or quality/bitrate-based. The goal behind introducing the selection algorithms is to generate a representative small-scale database that is able to reflect the behavior of the large-scale database. 3) It is observed that depending on the correlation measures and the mean square error as performance evaluation measure doesn't, in fact, report the right conclusion about the compared models. Hence, improved measures that help evaluate the performance of the objective measurement with different datasets are introduced. The measures depend on analysing the residual error using PCA, depend on analysing the confidence intervals of the predicted data, and depend on analysing the confidence intervals of the linear coefficients of the trained and tested models.

13.2 Future perspectives

Through the thesis, utilizing the content features have been studied in different aspects. The final goal of this dissertation is to highlight the relationship between content characteristics and optimizing the video delivery chain. As a result, these studies are attracting researches due to variety of topics and aspects that can be addressed. Hence, this work can be extended as follows:

— In this work, we bring complexity awareness to the prediction of the encoding parameter values with the help of content features. Basically it is targeting the industry with limited computation power. On the other hand, some enhancements can be done in different aspects: 1) The content features might be linked with the encoder parameter values without any awareness of complexity as well. This could fit industry that has a superior computing power. 2) The proposed optimization process used PSNR as a quality measure which can be replaced with other metrics that reflect some aspects of human visual system.

— The developed entropy maps that are used in the disruption analysis can be used to optimize the multiple description coding. Instead of sending all the redundant data, we can employ the entropy maps to select regions that are highly important to maintain the consistency of the recovered lost primary frames. On the other hand, the subjective experiment has to be conducted to compare the temporal MDC. Moreover, a simulation with different datarates capabilities has to be done as well.

— Regarding the large-scale database, further analysis is required to study the impact of content features in the disagreement of FR VQA within a short-term. For instance, if the quality measures agree for 10 consecutive frames and disagree for frame 11. Studying such cases let us know if this disagreement is due to a notable change in content features or due to the measure inconsistency.

— In this work, two subset selection algorithms are proposed to generate a representative subset that is able to reflect the behaviour of the large-scale database. Combining these two algorithms in one algorithm may generate a robust subset. Further analysis has to be done in this aspect.

— Since integrating one property of TCP helps building a robust networking structure of best-effort networks for temporal-MDC schemes, further investigations are required to 1) target other application scenarios like internet of things application, for instance, 2) utilize the multi-path property of the routing protocols.

— The pixel-based features that are used in this dissertation are limited, i.e. they are collected from the state-of-the-art except for the features that are used in the disruption analysis. Hence, more analysis are required to include more features that might be extracted from the signal based statistics.

Bibliography

[1] Cisco, *The Zettabyte Era: Trends and Analysis*, 2016. 9

[2] ——, *Cisco Visual Networking Index: Forecast and Methodology, 2015–2020*, 2016. 9

[3] M. H. Pinson, K. Sue Boyd, J. Hooker, and K. Muntean, "How to choose video sequences for video quality assessment," in *Proceedings of the seventh International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) Scottsdale, USA*, 2013. 9, 31, 46, 59

[4] VQEG, *The Validation Of Objective Models Of Multimedia Quality Assessment*, PHASE I 2008. 9, 31, 46, 59

[5] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2d and 3d subjective video quality tests," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–12, 2013. 9, 31, 75

[6] J. Korhonen and J. You, "Improving objective video quality assessment with content analysis," in *Proceedings of the fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) Scottsdale, USA*, 2010. 9, 29, 31

[7] B. Ortiz-Jaramillo, J. Niño-Castañeda, L. Platiša, and W. Philips, "Content-aware objective video quality assessment," *Journal of Electronic Imaging*, vol. 25, no. 1, pp. 013 011–013 011, 2016. 9, 29, 31, 75

[8] A. Aldahdooh, E. Masala, O. Janssens, G. Van Wallendael, and M. Barkowsky, "Comparing simple video quality measures for loss-impaired video sequences on a large-scale database," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, June 2016, pp. 1–6. 9, 31

[9] Y. Pitrey, M. Barkowsky, R. Pépion, P. Le Callet, and H. Hlavacs, "Influence of the source content and encoding configuration on the perceived quality for scalable video coding," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2012, pp. 82 911K–82 911K. 9, 29, 31, 75

[10] D. Agrafiotis, D. R. Bull, and C. N. Canagarajah, "Enhanced error concealment with mode selection," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 8, pp. 960–973, 2006. 9, 27, 28, 31

[11] S. Cen and P. C. Cosman, "Decision trees for error concealment in video decoding," *Multimedia, IEEE Transactions on*, vol. 5, no. 1, pp. 1–7, 2003. 9, 27, 31

[12] Z. Rongfu, Z. Yuanhua, and H. Xiaodong, "Content-adaptive spatial error concealment for video communication," *Consumer Electronics, IEEE Transactions on*, vol. 50, no. 1, pp. 335–341, 2004. 9, 28, 31

[13] S.-C. Huang and S.-Y. Kuo, "Optimization of hybridized error concealment for h. 264," *Broadcasting, IEEE Transactions on*, vol. 54, no. 3, pp. 499–516, 2008. 9, 28, 31

[14] ITU, "ITU-T H.265 : High efficiency video coding," *http://www.itu.int/rec/T-REC-H.265-201504-I*, 2013, http://www.itu.int/rec/T-REC-H.265-201504-I. 10, 53, 67, 75

[15] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012. 10, 53, 67, 70

[16] R. Frederick, V. Jacobson, and P. Design, "RTP: A transport protocol for real-time applications," *IETF RFC3550*, 2003. 10

[17] ISO, "Information technology - generic coding of moving pictures and associated audio information: Systems," *ISO/IEC*, 2014. 10

[18] ISO/IEC, "Information technology—coding of audio-visual objects—part 12: Iso base media file format," *ISO/IEC 14496-12 (4th ed.)*, 2012. 10

[19] ——, "Information technology - dynamic adaptive streaming over http (dash) - part 1: Media presentation description and segment formats," *ISO/IEC 23009-1 (2nd ed.)*, 2014. 10

[20] A. C. Begen, T. Akgul, and M. Baugher, "Watching video over the web: Part 1: Streaming protocols," *Internet Computing, IEEE*, vol. 15, no. 2, pp. 54–63, 2011. 10

[21] A. Begen, T. Akgul, and M. Baugher, "Watching video over the web: Part 2: Applications, standardization, and open issues," *Internet Computing, IEEE*, vol. 15, no. 3, pp. 59–63, 2011. 10

[22] I. Sodagar, "The mpeg-dash standard for multimedia streaming over the internet." *IEEE Multimedia*, no. 18, pp. 62–67, 2011. 10

[23] T. Stockhammer, "Dynamic adaptive streaming over HTTP-design principles and standards," in *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, vol. 2014. New York, USA: ACM, 2011, pp. 2–4. 10

[24] B. Oztas, M. T. Pourazad, P. Nasiopoulos, and V. C. Leung, "A study on the HEVC performance over lossy networks," in *Electronics, Circuits and Systems (ICECS), 2012 19th IEEE International Conference on.* IEEE, 2012, pp. 785–788. 10, 26, 67, 75

[25] L. R. Siruvuri, P. Salama, and D. S. Kim, "Adaptive error resilience for video streaming," *International Journal of Digital Multimedia Broadcasting*, vol. 2009, 2009. 10

[26] Y. Wang, J. Ostermann, and Y.-Q. Zhang, *Video processing and communications.* Prentice Hall Upper Saddle River, 2002, vol. 5. 10, 21, 22, 23, 24, 26, 67, 75, 93

[27] S. Wenger, "H. 264/AVC over IP," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 645–656, 2003. 10, 24, 67, 75

[28] T. Schierl, M. M. Hannuksela, Y.-K. Wang, and S. Wenger, "System layer integration of high efficiency video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1871–1884, 2012. 10, 24, 25, 67, 75

[29] Y. Wang, Q. F. Zhu, and L. Shaw, "Maximally smooth image recovery in transform coding," *IEEE Transactions on Communications*, vol. 41, no. 10, pp. 1544–1551, Oct 1993. 11, 28, 93

[30] K. Meisinger and A. Kaup, "Spatial error concealment of corrupted image data using frequency selective extrapolation," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 3, May 2004, pp. iii–209–12 vol.3. 11, 28, 93

[31] M. E. Al-Mualla, N. Canagarajah, and D. R. Bull, "Error concealment using motion field interpolation," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, Oct 1998, pp. 512–516 vol.3. 11, 28, 93, 95, 97, 98, 99, 101, 105

[32] W. M. Lam, A. R. Reibman, and B. Liu, "Recovery of lost or erroneously received motion vectors," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 5, April 1993, pp. 417–420 vol.5. 11, 28, 93

[33] P. Le Callet, S. Möller, A. Perkis *et al.*, "Qualinet white paper on definitions of quality of experience," *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003)*, 2012. 11, 21

[34] M. Ebdelli, O. L. Meur, and C. Guillemot, "Loss concealment based on video inpainting for robust video communication," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 1910–1914. 13, 28, 93, 94, 95, 96, 97, 98, 99

[35] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *Signal Processing Magazine, IEEE*, vol. 15, no. 6, pp. 74–90, 1998. 19, 20, 24

[36] B. Meng, O. C. Au, C.-W. Wong, and H.-K. Lam, "Efficient intra-prediction mode selection for 4× 4 blocks in H. 264," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 3. IEEE, 2003, pp. III–521. 20

[37] S. Saponara, M. Casula, L. Fanucci, F. Rovati, and D. Alfonso, "Dynamic control of motion estimation search parameters for low complex h. 264/avc video coding," in *Consumer Electronics, 2006. ICCE'06. 2006 Digest of Technical Papers. International Conference on.* IEEE, 2006, pp. 481–482. 20

[38] L. Zhang and W. Gao, "Reusable architecture and complexity-controllable algorithm for the integer/fractional motion estimation of h. 264," *Consumer Electronics, IEEE Transactions on*, vol. 53, no. 2, pp. 749–756, 2007. 20

[39] B. La, M. Eom, and Y. Choe, "Fast mode decision for intra prediction in H. 264/AVC encoder," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 5. IEEE, 2007, pp. V–321. 20

[40] J.-H. Kim and B.-G. Kim, "Fast block mode decision algorithm in H. 264/AVC video coding," *Journal of Visual Communication and Image Representation*, vol. 19, no. 3, pp. 175–183, 2008. 20

[41] L. Su, Y. Lu, F. Wu, S. Li, and W. Gao, "Complexity-constrained h. 264 video encoding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 4, pp. 477–490, 2009. 20

[42] M.-C. Chien, J.-Y. Huang, and P.-C. Chang, "Complexity control for h. 264 video encoding over power-scalable embedded systems," in *Proc. IEEE Symp. on Consumer Electronics*, 2009, pp. 43–46. 20

[43] M.-Y. Chiu and W.-C. Siu, "Computationally-scalable motion estimation algorithm for h. 264/avc video coding," *Consumer Electronics, IEEE Transactions on*, vol. 56, no. 2, pp. 895–903, 2010. 20

[44] J. Leng, L. Sun, T. Ikenaga, and S. Sakaida, "Content based hierarchical fast coding unit decision algorithm for HEVC," in *Multimedia and Signal Processing (CMSP), 2011 International Conference on*, vol. 1. IEEE, 2011, pp. 56–59. 20

[45] X. Li, M. Wien, and J.-R. Ohm, "Rate-complexity-distortion optimization for hybrid video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 7, pp. 957–970, 2011. 20, 63

[46] G. Corrêa, P. Assuncao, L. A. da Silva Cruz, and L. Agostini, "Adaptive coding tree for complexity control of high efficiency video encoders," in *Picture Coding Symposium (PCS), 2012.* IEEE, 2012, pp. 425–428. 20

[47] X. Shen and L. Yu, "CU splitting early termination based on weighted SVM," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, pp. 1–11, 2013. 20

[48] L. Shen, Z. Zhang, and Z. Liu, "Effective CU size decision for HEVC intracoding," *Image Processing, IEEE Transactions on*, vol. 23, no. 10, pp. 4232–4241, 2014. 20

[49] ——, "Adaptive inter-mode decision for hevc jointly utilizing inter-level and spatiotemporal correlations," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 10, pp. 1709–1722, 2014. 20

[50] Y.-F. Cen, W.-L. Wang, and X.-W. Yao, "A fast cu depth decision mechanism for hevc," *Information Processing Letters*, vol. 115, no. 9, pp. 719–724, 2015. 20

[51] J. Wang, L. Dong, and Y. Xu, "A fast inter prediction algorithm based on rate-distortion cost in hevc," 2015. 20

[52] V. A. Nguyen and M. N. Do, "Efficient coding unit size selection for hevc downsizing transcoding," in *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on.* IEEE, 2015, pp. 1286–1289. 20

[53] H. Chen, R. Xie, and L. Zhang, "Gradient based fast mode and depth decision for high efficiency intra frame video coding," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2015 IEEE International Symposium on.* IEEE, 2015, pp. 1–6. 20

[54] D. Zhao, S. Zhu, and S. Gao, "A novel fast intra-prediction algorithm for high-efficiency video coding based on structural similarity," *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 23, pp. 4212–4218, 2015. 20

[55] J. He, W. Yang, and J. Wang, "Fast hevc coding unit decision based on bp-neural network," *International Journal of Grid and Distributed Computing*, vol. 8, no. 4, pp. 289–300, 2015. 20

[56] G. Correa, P. A. Assuncao, L. Volcan Agostini, and L. A. da Silva Cruz, "Fast hevc encoding decisions using data mining," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 25, no. 4, pp. 660–673, 2015. 20

[57] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, 1996. 21

[58] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," in *ITU-T RECOMMENDATION, P*, 1999. 21, 22

[59] I. Recommendation, "500-11,"methodology for the subjective assessment of the quality of television pictures," recommendation itu-r bt. 500-11," *ITU Telecom. Standardization Sector of ITU*, 2002. 21, 22

[60] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 165–182, 2011. 22

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004. 22, 57, 106

[62] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2. IEEE, 2003, pp. 1398–1402. 22, 57, 106

[63] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *Broadcasting, IEEE Transactions on*, vol. 50, no. 3, pp. 312–322, 2004. 22, 57

[64] A. B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *Journal of Electronic imaging*, vol. 10, no. 1, pp. 20–29, 2001. 22

[65] A. P. Hekstra, J. G. Beerends, D. Ledermann, F. De Caluwe, S. Kohler, R. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss, "Pvqm–a perceptual video quality measure," *Signal processing: Image communication*, vol. 17, no. 10, pp. 781–798, 2002. 22

[66] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: A review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, 1998. 22, 23, 27

[67] Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, "Error resilient video coding techniques," *Signal Processing Magazine, IEEE*, vol. 17, no. 4, pp. 61–82, 2000. 22, 23, 26

[68] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial & Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960. 22

[69] S. Lin and D. Costello, "Error control coding: Fundamentals and applications," 1983. 22

[70] A. Albanese, J. Blomer, J. Edmonds, M. Luby, and M. Sudan, "Priority encoding transmission," *Information Theory, IEEE Transactions on*, vol. 42, no. 6, pp. 1737–1744, 1996. 22

[71] ITU-T, "Video coding for low bit rate communication," *International Telecommunication Union*, version 1, 1996; version 2, 1998; version 3, 2000. 22, 23

[72] E. Steinbach, N. Farber, and B. Girod, "Standard compatible extension of h. 263 for robust video transmission in mobile environments," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 7, no. 6, pp. 872–881, 1997. 22, 24

[73] T.-H. Lee and P.-C. Chang, "Error-robust h. 263 video coding system," in *Photonics Asia 2002*. International Society for Optics and Photonics, 2002, pp. 209–218. 22

[74] T. Stockhammer, M. M. Hannuksela, and T. Wiegand, "H. 264/avc in wireless environments," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 657–673, 2003. 22, 24

[75] S. Wenger, "Video redundancy coding in h. 263," in *Proc. AVSPN*, vol. 97, 1997. 23

[76] ISO, "Information technology – coding of audio-visual objects – part 2: Visual," *ISO/IEC*, 3rd ed. 2004. 23

[77] R. Koenen, "Overview of the mpeg-4 standard," *ISO/IEC JTC1/SC29/WG11 N*, vol. 1730, pp. 11–13, 2002. 23

[78] M. Wien, *High Efficiency Video Coding - Coding tools and Specification*. Springer, 2015. 23

[79] V. Sze, M. Budagavi, and G. J. Sullivan, *High Efficiency Video Coding (HEVC) - Algorithms and Architectures*. Springer, 2014. 23

[80] S. Wenger, G. Knorr, J. Ott, and F. Kossentini, "Error resilience support in h. 263," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 7, pp. 867–877, 1998. 23

[81] A. Katsaggelos, F. Ishtiaq, L. Kondi, M. Hong, M. Banham, and J. Brailean, "Error resilience and concealment in video coding," in *Proc. European Signal Processing Conf., Rhodes, Greece*, 1998, p. 221. 23

[82] ITU-T RECOMMENDATION H.264, "Advanced video coding for generic audiovisual services," *ISO/IEC*, 2003. 23

[83] S. Kumar, L. Xu, M. K. Mandal, and S. Panchanathan, "Error resiliency schemes in h. 264/avc standard," *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 425–450, 2006. 23

[84] S. Wenger and T. Stockhammer, "H. 26l over ip and h. 324 framework," *ITU-T SG16 Doc. VCEGN52*, 2001. 23

[85] J. C. Ye and Y. Chen, "Flexible data partitioning mode for streaming video," *JVT-D136*, 2002. 24

[86] T. Stockhammer, "Independent data partitions a and b," *JVT-C132*, 2002. 24

[87] ITU-T RECOMMENDATION H.265, "High efficiency video coding," *ISO/IEC*, 2013. 24

[88] P. Salama, N. B. Shroff, and E. J. Delp, "Error concealment in encoded video streams," in *Signal recovery techniques for image and video compression and transmission*. Springer, 1998, pp. 199–233. 24, 27

[89] M. Karczewicz and R. Kurceren, "The sp-and si-frames design for h. 264/avc," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 637–644, 2003. 24

[90] E. Setton and B. Girod, "Video streaming with sp and si frames," in *Visual Communications and Image Processing 2005*. International Society for Optics and Photonics, 2005, pp. 59 606F–59 606F. 24

[91] X. Zhou, W.-y. Kung, and C.-C. J. Kuo, "Error resilient h. 264 video with sp/si coded macroblocks," in *The 14th International Packet Video Workshop*, 2004. 24

[92] X. Zhou, W.-y. Kung, and C.-C. Kuo, "A robust h. 264 video streaming scheme for portable devices," in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*. IEEE, 2005, pp. 3263–3266. 24

[93] X. Zhou, W.-Y. Kung, and C.-C. J. Kuo, "Improved error resilient h. 264 coding scheme using sp/si macroblocks," in *Electronic Imaging 2005*. International Society for Optics and Photonics, 2005, pp. 1031–1042. 24

[94] T. Stockhammer, D. Kontopodis, and T. Wiegand, "Rate-distortion optimization for jvt/h. 26l video coding in packet loss environment," in *Int. Packet Video Workshop*, 2002. 24

[95] M. H. Willebeek-LeMair, Z.-Y. Shae, and Y.-C. Chang, "Robust h. 263 video coding for transmission over the internet," in *INFOCOM'98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1. IEEE, 1998, pp. 225–232. 24

[96] J. Y. Liao and J. D. Villasenor, "Adaptive intra update for video coding over noisy channels," in *Image Processing, 1996. Proceedings., International Conference on*, vol. 3. IEEE, 1996, pp. 763–766. 24

[97] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging h. 263 standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 6, no. 2, pp. 182–190, 1996. 24

[98] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *Selected Areas in Communications, IEEE Journal on*, vol. 18, no. 6, pp. 966–976, 2000. 24

[99] P. Haskell and D. Messerschmitt, "Resynchronization of motion compensated video affected by atm cell loss," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 3. IEEE, 1992, pp. 545–548. 24

[100] G. Côté and F. Kossentini, "Optimal intra coding of blocks for robust video communication over the internet," *Signal Processing: Image Communication*, vol. 15, no. 1, pp. 25–34, 1999. 24

[101] G. Côté, S. Shirani, and F. Kossentini, "Optimal mode selection and synchronization for robust video communications over error-prone networks," *Selected Areas in Communications, IEEE Journal on*, vol. 18, no. 6, pp. 952–965, 2000. 24

[102] T. Wiegand, N. Farber, K. Stuhlmuller, and B. Girod, "Error-resilient video transmission using long-term memory motion-compensated prediction," *Selected Areas in Communications, IEEE Journal on*, vol. 18, no. 6, pp. 1050–1062, 2000. 24

[103] N. Färber, E. Steinbach, and B. Girod, "Robust h. 263 compatible video transmission over wireless channels," in *in Proceedings of the Picture Coding Symposium*. Citeseer, 1996. 24

[104] J. Y. Liao and J. Villasenor, "Adaptive intra block update for robust transmission of h. 263," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 10, no. 1, pp. 30–35, 2000. 24

[105] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h. 264/avc standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103–1120, 2007. 24

[106] P. Helle, H. Lakshman, M. Siekmann, J. Stegemann, T. Hinz, H. Schwarz, D. Marpe, and T. Wiegand, "A scalable video coding extension of hevc," in *Data Compression Conference (DCC), 2013*. IEEE, 2013, pp. 201–210. 24

[107] V. K. Goyal, "Multiple description coding: Compression meets the network," *Signal Processing Magazine, IEEE*, vol. 18, no. 5, pp. 74–93, 2001. 25, 67, 75

[108] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 57–70, 2005. 25, 67, 75

[109] M. Kazemi, S. Shirmohammadi, and K. H. Sadeghi, "A review of multiple description coding techniques for error-resilient video delivery," *Multimedia Systems*, vol. 20, no. 3, pp. 283–309, 2014. 25, 67, 75

[110] J. Chakareski, S. Han, and B. Girod, "Layered coding vs. multiple descriptions for video streaming over multiple paths," *Multimedia Systems*, vol. 10, no. 4, pp. 275–285, 2005. 25, 67, 75

[111] Y.-C. Lee, J. Kim, Y. Altunbasak, and R. M. Mersereau, "Layered coded vs. multiple description coded video over error-prone networks," *Signal Processing: Image Communication*, vol. 18, no. 5, pp. 337–356, 2003. 25, 67, 75

[112] R. Bernardini, M. Durigon, R. Rinaldo, L. Celetto, and A. Vitali, "Polyphase spatial subsampling multiple description coding of video streams with h264," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 5. IEEE, 2004, pp. 3213–3216. 25

[113] S.-K. A. Yeung and B. Zeng, "Multiple description coding in the quincunx sub-sampling lattice with diamond-shape dct," in *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*. IEEE, 2009, pp. 3182–3185. 25

[114] Z. Wei, K.-K. Ma, and C. Cai, "Prediction-compensated polyphase multiple description image coding with adaptive redundancy control," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 3, pp. 465–478, 2012. 25

[115] S. Shirani, M. Gallant, and F. Kossentini, "Multiple description image coding using pre-and post-processing," in *Information Technology: Coding and Computing, 2001. Proceedings. International Conference on*. IEEE, 2001, pp. 35–39. 25

[116] D. Wang, N. Canagarajah, D. Redmill, and D. Bull, "Multiple description video coding based on zero padding," in *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on*, vol. 2. IEEE, 2004, pp. II–205. 25

[117] T. Tillo and G. Olmo, "Data-dependent pre-and postprocessing multiple description coding of images," *Image Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 1269–1280, 2007. 25

[118] Y. Yapıcı, B. Demir, S. Ertürk, and O. Urhan, "Downsampling-based multiple description image coding using optimal filtering," *Journal of Electronic Imaging*, vol. 17, no. 3, pp. 033 018–033 018, 2008. 25

[119] S. Shirani, "Content-based multiple description image coding," *Multimedia, IEEE Transactions on*, vol. 8, no. 2, pp. 411–419, 2006. 25

[120] J. G. Apostolopoulos, "Error-resilient video compression through the use of multiple states," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3. IEEE, 2000, pp. 352–355. 25

[121] ——, "Reliable video communication over lossy packet networks using multiple state encoding and path diversity," in *Photonics West 2001-Electronic Imaging*. International Society for Optics and Photonics, 2000, pp. 392–409. 25, 26, 75

[122] T. Tillo and G. Olmo, "Low complexity pre postprocessing multiple description coding for video streaming," in *Information and Communication Technologies: From Theory to Applications, 2004. Proceedings. 2004 International Conference on*. IEEE, 2004, pp. 519–520. 25

[123] O. Campana and R. Contiero, "An h. 264/avc video coder based on multiple description scalar quantizer," in *Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on*. IEEE, 2006, pp. 1049–1053. 25

[124] V. Parameswaran, A. Kannur, and B. Li, "Adapting quantization offset in multiple description coding for error resilient video transmission," *Journal of Visual Communication and Image Representation*, vol. 20, no. 7, pp. 491–503, 2009. 25

[125] V. A. Vaishampayan, "Design of multiple description scalar quantizers," *Information Theory, IEEE Transactions on*, vol. 39, no. 3, pp. 821–834, 1993. 25

[126] K. R. Matty and L. P. Kondi, "Balanced multiple description video coding using optimal partitioning of the dct coefficients," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 7, pp. 928–934, 2005. 25

[127] D. Comas, R. Singh, A. Ortega, and F. Marqués, "Unbalanced multiple-description video coding with rate-distortion optimization," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 81–90, 2003. 25

[128] D.-M. Chung and Y. Wang, "Multiple description image coding using signal decomposition and reconstruction based on lapped orthogonal transforms," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9, no. 6, pp. 895–908, 1999. 25

[129] A. R. Reibman, H. Jafarkhani, Y. Wang, M. T. Orchard, and R. Puri, "Multiple-description video coding using motion-compensated temporal prediction," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 3, pp. 193–204, 2002. 25

[130] C.-W. Hsiao and W.-J. Tsai, "Hybrid multiple description coding based on h. 264," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 1, pp. 76–87, 2010. 25

[131] I. Radulovic, P. Frossard, Y.-K. Wang, M. M. Hannuksela, and A. Hallapuro, "Multiple description video coding with H. 264/AVC redundant pictures," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 20, no. 1, pp. 144–148, 2010. 25, 26, 70, 75

[132] C. Lin, Y. Zhao, H. Bai, and K. H. Goh, "Multiple description coding for scalable video coding with redundant slice," in *Networks (ICON), 2012 18th IEEE International Conference on*. IEEE, 2012, pp. 344–348. 25

[133] R. Choupani, S. Wong, and M. Tolun, "Optimized multiple description coding for temporal video scalability," in *Advances in Computational Science, Engineering and Information Technology*. Springer, 2013, pp. 167–176. 25

[134] B. Li, J. Xu, F. Wu, and H. Li, "Constrained temporal motion vector prediction for error resilience," *JCTVC-D139*, 2011. 26

[135] J.-L. Lin, Y.-W. Huang, C.-M. Fu, C.-Y. Chen, Y.-P. Tsai, and S. Lei, "Syntax for amvp parsing error control," *JCTVC-D126*, 2011. 26

[136] B. Li, J. Xu, and H. Li, "Parsing robustness in high efficiency video coding-analysis and improvement," in *Visual Communications and Image Processing (VCIP), 2011 IEEE*. IEEE, 2011, pp. 1–4. 26

[137] J. Carreira, V. D. Silva, E. Ekmekcioglu, A. Kondoz, P. Assuncao, and S. Faria, "Dynamic motion vector refreshing for enhanced error resilience in hevc," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 22nd European*. IEEE, 2014, pp. 281–285. 26

[138] R. Kibria and J. Kim, "H. 264/avc-based multiple description coding for wireless video transmission," in *International Conference on Communications*, 2008. 26, 75

[139] S. Kumar and L. Xu, "Rvlc decoding scheme for improved data recovery in mpeg-4 video coding standard," *Real-Time Imaging*, vol. 10, no. 5, pp. 315–323, 2004. 26

[140] L. Superiori, O. Nemethova, and M. Rupp, "Performance of a h. 264/avc error detection algorithm based on syntax analysis." in *MoMM*, 2006, pp. 49–58. 26

[141] S. GARY, "Joint model reference encoding methods and decoding concealment methods," *JVT-I049*, 2003. 27

[142] S. Shirani, F. Kossentini, and R. Ward, "Error concealment methods, a comparative study," in *Electrical and Computer Engineering, 1999 IEEE Canadian Conference on*, vol. 2. IEEE, 1999, pp. 835–840. 27

[143] Y. Wang, Q.-F. Zhu, and L. Shaw, "Maximally smooth image recovery in transform coding," *Communications, IEEE Transactions on*, vol. 41, no. 10, pp. 1544–1551, 1993. 27

[144] J. W. Park, J. Kim, and S. U. Lee, "Dct coefficients recovery-based error concealment technique and its application to the mpeg-2 bit stream error," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 7, no. 6, pp. 845–854, 1997. 27

[145] S. Shirani, F. Kossentini, and R. Ward, "Reconstruction of baseline jpeg coded images in error prone environments," *Image Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 1292–1299, 2000. 27

[146] Z. Alkachouh and M. G. Bellanger, "Fast dct-based spatial domain interpolation of blocks in images," *Image Processing, IEEE Transactions on*, vol. 9, no. 4, pp. 729–732, 2000. 27

[147] S. Aign and K. Fazel, "Temporal and spatial error concealment techniques for hierarchical mpeg-2 video codec," in *Communications, 1995. ICC'95 Seattle,'Gateway to Globalization', 1995 IEEE International Conference on*, vol. 3. IEEE, 1995, pp. 1778–1783. 27

[148] P. Salama, N. B. Shroff, E. J. Coyle, and E. J. Delp, "Error concealment techniques for encoded video streams," in *Image Processing, International Conference on*, vol. 1. IEEE Computer Society, 1995, pp. 9–9. 27

[149] Y. Xu and Y. Zhou, "H. 264 video communication based refined error concealment schemes," *Consumer Electronics, IEEE Transactions on*, vol. 50, no. 4, pp. 1135–1141, 2004. 27

[150] K. Meisinger and A. Kaup, "Spatial error concealment of corrupted image data using frequency selective extrapolation," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3. IEEE, 2004, pp. iii–209. 27

[151] ——, "Minimizing a weighted error criterion for spatial error concealment of missing image data," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 2. IEEE, 2004, pp. 813–816. 27

[152] A. Kaup, K. Meisinger, and T. Aach, "Frequency selective signal extrapolation with applications to error concealment in image communication," *AEU-International Journal of Electronics and Communications*, vol. 59, no. 3, pp. 147–156, 2005. 27

[153] W. Kwok and H. Sun, "Multi-directional interpolation for spatial error concealment," *Consumer Electronics, IEEE Transactions on*, vol. 39, no. 3, pp. 455–460, 1993. 27

[154] W. Kim, J. Koo, and J. Jeong, "Fine directional interpolation for spatial error concealment," *Consumer Electronics, IEEE Transactions on*, vol. 52, no. 3, pp. 1050–1056, 2006. 27

[155] D. Agrafiotis, D. R. Bull, and N. Canagarajah, "Spatial error concealment with edge related perceptual considerations," *Signal Processing: Image Communication*, vol. 21, no. 2, pp. 130–142, 2006. 27, 28

[156] H. Sun and W. Kwok, "Concealment of damaged block transform coded images using projections onto convex sets," *Image Processing, IEEE Transactions on*, vol. 4, no. 4, pp. 470–477, 1995. 27

[157] X. Li and M. T. Orchard, "Novel sequential error-concealment techniques using orientation adaptive interpolation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 10, pp. 857–864, 2002. 27

[158] S.-C. Hsia, "An edge-oriented spatial interpolation for consecutive block error concealment," *Signal Processing Letters, IEEE*, vol. 11, no. 6, pp. 577–580, 2004. 27

[159] J. He, S. He, Z. Chen, and Y. Huang, "An improved h. 264 adaptive intra error concealment algorithm," in *Image and Signal Processing (CISP), 2012 5th International Congress on*. IEEE, 2012, pp. 183–187. 27

[160] J.-W. Suh and Y.-S. Ho, "Error concealment based on directional interpolation," *Consumer Electronics, IEEE Transactions on*, vol. 43, no. 3, pp. 295–302, 1997. 27

[161] ——, "Error concealment techniques for digital tv," *Broadcasting, IEEE Transactions on*, vol. 48, no. 4, pp. 299–306, 2002. 27

[162] D. Agrafiotis, D. R. Bull, and N. Canagarajah, "Enhanced spatial error concealment with directional entropy based interpolation switching," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*. IEEE, 2006, pp. 4–pp. 28

[163] K. Meisinger and A. Kaup, "Spatiotemporal selective extrapolation for 3-d signals and its applications in video communications," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2348–2360, Sept 2007. 28, 93

[164] C. Guillemot and O. L. Meur, "Image inpainting : Overview and recent advances," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 127–144, Jan 2014. 28, 93

[165] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video inpainting of occluding and occluded objects," in *IEEE International Conference on Image Processing 2005*, vol. 2, Sept 2005, pp. II–69–72. 28, 93, 96

[166] A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, Sept 2004. 28, 93, 97

[167] S. Kanumuri, P. C. Cosman, A. R. Reibman, and V. A. Vaishampayan, "Modeling packet-loss visibility in mpeg-2 video," *Multimedia, IEEE Transactions on*, vol. 8, no. 2, pp. 341–355, 2006. 28, 29, 125

[168] A. R. Reibman, S. Kanumuri, V. Vaishampayan, and P. C. Cosman, "Visibility of individual packet losses in mpeg-2 video," in *Image Processing, 2004. ICIP'04. 2004 International Conference on*, vol. 1. IEEE, 2004, pp. 171–174. 28

[169] A. R. Reibman and D. Poole, "Predicting packet-loss visibility using scene characteristics," in *Packet Video 2007*. IEEE, 2007, pp. 308–317. 28

[170] Y. J. Liang, J. G. Apostolopoulos, and B. Girod, "Analysis of packet loss for compressed video: Does burst-length matter?" in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 5. IEEE, 2003, pp. V–684. 28

[171] T. Rahrer, R. Fiandra, S. Wright *et al.*, "Triple-play services quality of experience (qoe) requirements," in *DSL-Forum. Technical Report TR-126*, 2006. 28

[172] F. Boulos, B. Parrein, P. Le Callet, and D. Hands, "Perceptual effects of packet loss on h. 264/avc encoded videos," in *Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM-09*, 2009. 28

[173] V. Adzic, H. Kalva, and B. Furht, "Exploring visual temporal masking for video compression," in *Consumer Electronics (ICCE), 2013 IEEE International Conference on*. IEEE, 2013, pp. 590–591. 28

[174] R. R. Pastrana-Vidal, J.-C. Gicquel, C. Colomes, and H. Cherifi, "Frame dropping effects on user quality perception," in *5th International Workshop on Image Analysis for Multimedia Interactice Services*, 2004. 28

[175] R. R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, and H. Cherifi, "Sporadic frame dropping impact on quality perception," in *Proceedings of SPIE*, vol. 5292, 2004, pp. 182–193. 28

[176] ——, "Temporal masking effect on dropped frames at video scene cuts," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 194–201. 28

[177] Q. Dai and R. Lehnert, "Impact of packet loss on the perceived video quality," *IEEE INTERNET*, 2010. 28

[178] T.-J. Liu, Y.-C. Lin, W. Lin, and C.-C. J. Kuo, "Visual quality assessment: recent developments, coding applications and future trends," *APSIPA Transactions on Signal and Information Processing*, vol. 2, p. e4, 2013. 29

[179] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of psnr in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, June 2008. 29

[180] P. Le Callet, S. Péchard, S. Tourancheau, A. Ninassi, and D. Barba, "Towards the next generation of video and image quality metrics: Impact of display, resolution, contents and visual attention in subjective assessment," in *Second International Workshop on Image Media Quality and its Applications, IMQA2007*, 2007, p. A2. 29

[181] R. Feghali, F. Speranza, D. Wang, and A. Vincent, "Video quality metric for bit rate control via joint adjustment of quantization and frame rate," *IEEE Transactions on Broadcasting*, vol. 53, no. 1, pp. 441–446, March 2007. 29

[182] A. Khan, L. Sun, and E. Ifeachor, "Content clustering based video quality prediction model for MPEG4 video streaming over wireless networks," in *Communications, 2009. ICC'09. IEEE International Conference on*. IEEE, 2009, pp. 1–5. 29

[183] M. Garcia, R. Schleicher, and A. Raake, "Towards a content-based parametric video quality model for IPTV," in *Proceedings of the 3rd International Workshop on Perceptual Quality of Systems (PQS'10)*, 2010. 29

[184] Y. F. Ou, Y. Xue, and Y. Wang, "Q-star: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Transactions on Image Processing*, vol. 23, no. 6, pp. 2473–2486, June 2014. 29

[185] B. Ortiz-Jaramillo, J. Niño-Castañeda, L. Platiša, and W. Philips, "Content-aware video quality assessment: predicting human perception of quality using peak signal to noise ratio and spatial/temporal activity," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2015, pp. 939 917–939 917. 29

[186] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with H.264/AVC: tools, performance, and complexity," *IEEE Circuits and Systems Magazine*, vol. 4, no. 1, pp. 7–28, First 2004. 29, 125

[187] M. Yuen and H. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, no. 3, pp. 247 – 278, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168498001285 29, 125

[188] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 40, 2014. 29, 125

[189] M. A. Aabed and G. AlRegib, "No-reference quality assessment of HEVC videos in loss-prone networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2015–2019. 29, 125

[190] J. Guo, K. Zheng, G. Hu, and L. Huang, "Packet layer model of hevc wireless video quality assessment," in *2016 11th International Conference on Computer Science Education (ICCSE)*, Aug 2016, pp. 712–717. 29, 125

[191] M. Shahid, J. Panasiuk, G. V. Wallendael, M. Barkowsky, and B. Lövström, "Predicting full-reference video quality measures using HEVC bitstream-based no-reference features," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, May 2015, pp. 1–2. 29, 125, 140

[192] K. Izumi, K. Kawamura, T. Yoshino, and S. Naito, "No reference video quality assessment based on parametric analysis of HEVC bitstream," in *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, Sept 2014, pp. 49–50. 29, 125

[193] A. Aldahdooh, M. Barkowsky, and P. L. Callet, "Content-aware adaptive multiple description coding scheme," in *2016 IEEE International Conference on Multimedia Expo Workshops(ICMEW), 22nd International Packet Video workshop*, July 2016, pp. 1–6. 31

[194] ITU-T, "ITU-T P.910 Subjective video quality assessment methods for multimedia applications," *ITU-T P.910*, 1997. 32, 33, 115

[195] S. Wolf and M. Pinson, "Video quality measurement techniques," *2002.*, 2002. 32, 33

[196] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Electronic Imaging 2003*. International Society for Optics and Photonics, 2003, pp. 87–95. 32, 33

[197] G. Srinivasan and G. Shobha, "Statistical texture analysis," in *Proceedings of world academy of science, engineering and technology*, vol. 36, 2008, pp. 1264–1269. 32

[198] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973. 32, 33

[199] J. Lewis, "Fast normalized cross-correlation," in *Vision interface*, vol. 10, 1995, pp. 120–123. 32, 33

[200] K. Zhu, V. Asari, and D. Saupe, "No-reference quality assessment of H.264/AVC encoded video based on natural scene features," in *SPIE Defense, Security, and Sensing*. International Society for Optics and Photonics, 2013, pp. 875 505–875 505. 32, 33

[201] K. Zhu, K. Hirakawa, V. Asari, and D. Saupe, "A no-reference video quality assessment based on laplacian pyramids." in *ICIP*, 2013, pp. 49–53. 32, 33

[202] S. Jeannin and A. Divakaran, "Mpeg-7 visual motion descriptors," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 720–724, 2001. 32, 33, 34

[203] H. Yi, D. Rajan, and L.-T. Chia, "A new motion histogram to index motion content in video segments," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1221–1231, 2005. 32, 33, 35, 71, 77, 126, 139

[204] M. A. Hasan, M. Xu, X. He, and C. Xu, "Camhid: Camera motion histogram descriptor and its application to cinematographic shot classification," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 10, pp. 1682–1695, 2014. 32, 36, 71, 77

[205] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *QoMEX*, 2013, pp. 34–35. 32, 59, 72, 77

[206] Ultra Video Group, *Ultra Video Group 4K sequences*, accessed Septemper 15, 2015, http://ultravideo.cs.tut.fi/#testsequences. 32, 59, 72, 77

[207] *SVT sequences hosted in VQEG*, accessed Septemper 15, 2015, ftp://vqeg.its.bldrdoc.gov/HDTV/SVT_MultiFormat/. 32, 72, 77

[208] *Blender Foundation 4K sequences*, accessed Septemper 15, 2015, http://bbb3d.renderfarming.net/download.html. 32, 72, 77

[209] *MediAVentures*, 2013, http://www.mediaventures.be/. 32, 72, 77

[210] J. A. Thomas and T. Cover, *Elements of information theory*. Wiley New York, 2006, vol. 2. 32, 72, 77

[211] R. Scobey and J. Horowitz, "Detection of image displacement by phasic cells in peripheral visual fields of the monkey," *Vision Research*, vol. 16, no. 1, pp. 15 – 24, 1976. 38, 101

[212] V. Virsu, J. Rovamo, P. Laurinen, and R. Näsänen, "Temporal contrast sensitivity and cortical magnification," *Vision research*, vol. 22, no. 9, pp. 1211–1217, 1982. 38, 43, 101

[213] T. Wolff, H.-H. Ho, J. M. Foley, and S. K. Mitra, "H. 264 coding artifacts and their relation to perceived annoyance," in *Signal Processing Conference, 2006 14th European*. IEEE, 2006, pp. 1–5. 38

[214] K. Zeng, T. Zhao, A. Rehman, and Z. Wang, "Characterizing perceptual artifacts in compressed video streams," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 90 140Q–90 140Q. 38

[215] R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Visible difference predicator for high dynamic range images," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, vol. 3. IEEE, 2004, pp. 2763–2769. 38

[216] A. B. Watson, "The cortex transform: rapid computation of simulated neural images," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 311–327, 1987. 38

[217] P. Ye and D. Doermann, "No-reference image quality assessment based on visual codebook," in *Image Processing (ICIP), 2011 18th IEEE International Conference on.* IEEE, 2011, pp. 3089–3092. 38

[218] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using gabor filters," *Pattern recognition*, vol. 24, no. 12, pp. 1167–1186, 1991. 41

[219] T. Hansen, L. Pracejus, and K. R. Gegenfurtner, "Color perception in the intermediate periphery of the visual field," *Journal of Vision*, vol. 9, no. 4, pp. 26–26, 2009. 41, 43, 101

[220] P. Gouras, "Opponent-colour cells in different layers of foveal striate cortex," *The Journal of physiology*, vol. 238, no. 3, p. 583, 1974. 41

[221] S. Winkler, "Perceptual distortion metric for digital color video," in *Electronic Imaging'99.* International Society for Optics and Photonics, 1999, pp. 175–184. 41, 43

[222] P. Le Callet and D. Barba, "Robust approach for color image quality assessment," in *Visual Communications and Image Processing 2003.* International Society for Optics and Photonics, 2003, pp. 1573–1581. 41, 43

[223] J. R. Newton and R. T. ESKEW, "Chromatic detection and discrimination in the periphery: A postreceptoral loss of color sensitivity," *Visual neuroscience*, vol. 20, no. 05, pp. 511–521, 2003. 43

[224] R. S. Hunter, "Photoelectric color difference meter," *Josa*, vol. 48, no. 12, pp. 985–995, 1958. 43

[225] J. Maunsell, T. A. Nealey, and D. D. DePriest, "Magnocellular and parvocellular contributions to responses in the middle temporal visual area (mt) of the macaque monkey," *The Journal of Neuroscience*, vol. 10, no. 10, pp. 3323–3334, 1990. 43

[226] U. Tulunay-Keesey, J. N. Ver Hoeve, and C. Terkla-McGrane, "Threshold and suprathreshold spatiotemporal response throughout adulthood," *JOSA A*, vol. 5, no. 12, pp. 2191–2200, 1988. 43

[227] W. Seiple, K. Holopigian, Y. Shnayder, and J. P. Szlyk, "Duration thresholds for target detection and identification in the peripheral visual field," *Optometry & Vision Science*, vol. 78, no. 3, pp. 169–176, 2001. 43

[228] C. van den Branden Lambrecht, D. Costantini, G. Sicuranza, and M. Kunt, "Quality assessment of motion rendition in video coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9, no. 5, pp. 766–782, Aug 1999. 43

[229] Y. Rai, P. Le Callet, and G. Cheung, "Role of hevc coding artifacts on gaze prediction in interactive video streaming systems," in *Image Processing, 2016. ICIP 2016. IEEE International Conference on*, vol. 2. IEEE, 2016, pp. II–169. 43

[230] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," 1981. 44

[231] Y. Rai, G. Cheung, and P. Le Callet, "Quantifying the relation between perceived interest and visual salience during free viewing using trellis based optimization," in *Image, Video, and Multidimensional Signal Processing, 2016 International Conference on*, vol. 9394, July 2016, pp. 93 941H–93 941H–9. 44

[232] J. Xia, Y. Shi, K. Teunissen, and I. Heynderickx, "Perceivable artifacts in compressed video and their relation to video quality," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 548–556, 2009. 45

[233] T. Wolff, H.-H. Ho, J. M. Foley, and S. K. Mitra, "Modeling subjectively perceived annoyance of h. 264/avc video as a function of perceived artifact strength," *Signal Processing*, vol. 90, no. 1, pp. 80–92, 2010. 45

[234] P. Wang, Y. Zhang, H.-M. Hu, and B. Li, "Region-classification-based rate control for flicker suppression of i-frames in hevc," in *Image Processing (ICIP), 2013 20th IEEE International Conference on.* IEEE, 2013, pp. 1986–1990. 45

[235] R. Snowden and R. Hess, "Temporal frequency filters in the human peripheral visual field," *Vision research*, vol. 32, no. 1, pp. 61–72, 1992. 45, 101

[236] Y. Rai, A. Aldahdooh, S. Ling, M. Barkowsky, and P. L. Callet, "Effect of content features on short-term video quality in the visual periphery," in *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, Sept 2016, pp. 1–6. 45

[237] J. Wang, D. M. Chandler, and P. Le Callet, "Quantifying the relationship between visual salience and visual importance," in *IS&T/SPIE Electronic Imaging.* International Society for Optics and Photonics, 2010, pp. 75 270K–75 270K. 46

[238] G. Van Wallendael, N. Staelens, E. Masala, and M. Barkowsky, "Full-hd hevc-encoded video quality assessment database," in *Ninth International Workshop on Video Processing and Quality Metrics (VPQM)*, 2015. 47, 49, 183

[239] E. Elliot, "Estimates on error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, pp. 1977–1997, Sep. 1963. 48

[240] G. Van Wallendael, N. Staelens, E. Masala, and M. Barkowsky, "Full-HD HEVC-encoded video quality assessment database," in *Ninth International Workshop on Video Processing and Quality Metrics (VPQM)*, 2015. 48, 115, 116, 117, 118, 185

[241] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "HEVC complexity and implementation analysis," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1685–1696, 2012. 53

[242] A. Aldahdooh, M. Barkowsky, and P. Le Callet, "The impact of complexity in the rate-distortion optimization: A visualization tool," in *Systems, Signals and Image Processing (IWSSIP) 2015, International Conference on*. IEEE, 2015, pp. 45–48. 55, 56, 57, 58, 61, 181, 184

[243] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *CD-ROM proceedings of the second international workshop on video processing and quality metrics, Scottsdale, USA*, vol. 4, 2006. 57

[244] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *Proceedings of the Third International Workshop on Video Processing and Quality Metrics*, vol. 4, 2007. 57

[245] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *Image Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 430–444, 2006. 57, 106

[246] *Xiph.org Video Test Media 4K sequences*, accessed May 15, 2014, https://media.xiph.org/video/derf/. 59

[247] *HEVC software reference*, accessed May 2, 2014, http://hevc.hhi.fraunhofer.de/. 59

[248] *Orange software*, accessed July 2, 2014, http://orange.biolab.si/. 60, 80

[249] G. Thomas, "Television motion measurement for datv and other applications," *NASA STI/Recon Technical Report N*, vol. 88, p. 13496, 1987. 69

[250] J. Carreira, E. Ekmekcioglu, A. Kondoz, P. Assuncao, S. Faria, and V. De Silva, "Selective motion vector redundancies for improved error resilience in HEVC," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2457–2461. 70

[251] M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnstrom, N. Staelens, and P. Le Callet, "Objective video quality assessment - towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. 98, no. 1, pp. 2–11, 2015. 72, 86

[252] ITU, "ITU-T recommendation P.910: Subjective video quality assessment methods for multimedia applications," 2008. 72, 77

[253] J. Li, M. Barkowsky, and P. Le Callet, "Subjective assessment methodology for preference of experience in 3DTV," in *IVMSP Workshop, 2013 IEEE 11th*. IEEE, 2013, pp. 1–4. 72

[254] ITU, "ITU-R recommendation BT.709-5: Parameter values for the HDTV standards for production and international programme exchange," 2002. 72, 105

[255] ITU , "ITU-R recommendation BT.500-13: Methodology for the assessment of the quality of television pictures," 2012. 72, 105

[256] G. Barnard, "A new test for 2× 2 tables," *Nature*, vol. 156, p. 177, 1945. 73

[257] A. Aldahdooh, M. Barkowsky, and P. Le Callet, "Spatio-temporal error concealment technique for high order multiple description coding schemes including subjective assessment," quality of Multimedia Experience (QoMEX) 2016 submission ID 29. Supplied as additional material `qomex2016.pdf`. 75, 76, 77, 184

[258] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952. 77

[259] F. Wickelmaier and C. Schmid, "A matlab function to estimate choice model parameters from paired-comparison data," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 1, pp. 29–40, 2004. 78

[260] J. C. Handley, "Comparative analysis of Bradley-Terry and Thurstone-Mosteller paired comparison models for image quality assessment," in *PICS*, vol. 1. Citeseer, 2001, pp. 108–112. 78

[261] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997. 80

[262] J. Klaue, B. Rathke, and A. Wolisz, "Evalvid–a framework for video transmission and quality evaluation," in *International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*. Springer, 2003, pp. 255–272. 84

[263] J. Yi, B. Parrein, and D. Radu, "Multipath routing protocol for manet: Application to h.264/svc video content delivery," in *Wireless Personal Multimedia Communications (WPMC), 2011 14th International Symposium on*, 2011. 84

[264] *HEVC stream analyser*, accessed September 2, 2016, https://github.com/virinext/hevcesbrowser/. 84

[265] B. Oztas, M. T. Pourazad, P. Nasiopoulos, and V. C. M. Leung, "A study on the hevc performance over lossy networks," in *Electronics, Circuits and Systems (ICECS), 2012 19th IEEE International Conference on*, Dec 2012, pp. 785–788. 93

[266] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000. [Online]. Available: http://science.sciencemag.org/content/290/5500/2323 94

[267] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *In NIPS*. MIT Press, 2000, pp. 556–562. 94

[268] M. Ebdelli, C. Guillemot, and O. L. Meur, "Examplar-based video inpainting with motion-compensated neighbor embedding," in *2012 19th IEEE International Conference on Image Processing*, Sept 2012, pp. 1737–1740. 94

[269] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 313–318, Jul. 2003. [Online]. Available: http://doi.acm.org/10.1145/882262.882269 94, 97

[270] H. Yi, D. Rajan, and L.-T. Chia, "A new motion histogram to index motion content in video segments," *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1221 – 1231, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865504003277 95

[271] M. Hasan, M. Xu, X. He, and C. Xu, "CAMHID: Camera motion histogram descriptor and its application to cinematographic shot classification," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 10, pp. 1682–1695, Oct 2014. 95

[272] M. Ebdelli, O. L. Meur, and C. Guillemot, "Video inpainting with short-term windows: Application to object removal and error concealment," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3034–3047, Oct 2015. 97

[273] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography," *Journal of Comparative Neurology*, vol. 292, no. 4, pp. 497–523, 1990. 101

[274] M. Iwasaki and H. Inomata, "Relation between superficial capillaries and foveal structures in the human retina." *Investigative ophthalmology & visual science*, vol. 27, no. 12, pp. 1698–1705, 1986. 101

[275] Y. Rai, M. Barkowsky, and P. Le Callet, "Does H.265 based peri and para-foveal quality flicker disrupt natural viewing patterns?" in *Systems, Signals and Image Processing (IWSSIP), 2015 International Conference on*, Sept 2015, pp. 133–136. 101

[276] ——, "Role of spatio-temporal distortions in the visual periphery in disrupting natural attention deployment," in *IS&T/SPIE Electronic Imaging*, vol. 9394, 2015, pp. 93 941H–93 941H–9. [Online]. Available: http://dx.doi.org/10.1117/12.2086371 101

[277] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011. 101

[278] T. S. A. Wallis, M. Bethge, and F. A. Wichmann, "Testing models of peripheral encoding using metamerism in an oddity paradigm," *Journal of Vision*, vol. 16, no. 2, p. 4, 2016. [Online]. Available: +http://dx.doi.org/10.1167/16.2.4 101

[279] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, "Modelling saliency awareness for objective video quality assessment," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2010. 102

[280] B. M. Harvey and S. O. Dumoulin, "The relationship between cortical magnification factor and population receptive field size in human visual cortex: constancies in cortical architecture," *The Journal of Neuroscience*, vol. 31, no. 38, pp. 13 604–13 612, 2011. 103, 185

[281] M. M. Alam, K. P. Vilankar, D. J. Field, and D. M. Chandler, "Local masking in natural images: A database and analysis," *Journal of Vision*, vol. 14, no. 8, 2014. [Online]. Available: http://www.journalofvision.org/content/14/8/22.abstract 104

[282] *Video Multimethod Assessment Fusion (VMAF) tool*, accessed March. 1st, 2017, http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html. 106

[283] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, Oct 2011. 106

[284] Y. Rai, "Visual attention for quality prediction at fine spatio-temporal scales: from perceptual weighting towards visual disruption modeling," *PhD thesis*, 2017. 108, 111

[285] N. Mackworth and A. Morandi, "The gaze selects informative details within pictures," *Perception and Psychophysics*, vol. 2, no. 11, pp. 547–552, 1967. [Online]. Available: http://dx.doi.org/10.3758/BF03210264 110

[286] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in neural information processing systems*, 2009, pp. 681–688. 110

[287] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual saliency based on conditional entropy," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 246–257. 110

[288] "Video Quality Experts Group (VQEG)," Jul. 2016. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.aspx 115

[289] M. Pinson, C. Schmidmer, L. Janowski, R. Pepion, Q. Huynh-Thu, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "Subjective and objective evaluation of an audiovisual subjective dataset for research and development," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, 2013, pp. 30–31. 115

[290] K. Fliegel and C. Timmerer, "WG4 Databases White Paper v1.5: QUALINET Multimedia Database enabling QoE Evaluations and Benchmarking Version 1.5," *Qualinet*, 2013. 115

[291] H. Liu and A. R. Reibman, "Software to stress test image quality estimators," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, June 2016, pp. 1–6. 115

[292] M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnstrom, N. Staelens, and P. Le Callet, "Objective video quality assessment – towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. E98-B, no. 1, pp. 2–11, Jan. 2015. 115, 116, 133, 139, 140

[293] F. Ciaramello and A. Reibman, "Systematic stress testing of image quality estimators," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2011, pp. 3101–3104. 115

[294] A. Aldahdooh, E. Masala, O. Janssens, G. Van Wallendael, and M. Barkowsky, "Comparing simple video quality measures for loss-impaired video sequences on a large-scale database," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2016, pp. 1–6. 118, 133

[295] F. Bossen, "Common test conditions and software reference configurations," *Doc. JCTVC-J1100*, Jul. 2012. 122

[296] K. McCann, B. Bross, W.-J. Han, I.-K. Kim, K. Sugimoto, and G. J. Sullivan, "High Efficiency Video Coding (HEVC) Test Model 12 (HM 12) Encoder Description v. 12.1 Doc. JCTVC-N1002," Nov. 2013. 123

[297] A. Aldahdooh, E. Masala, O. Janssens, G. V. Wallendael, and M. Barkowsky, "Comparing simple video quality measures for loss-impaired video sequences on a large-scale database," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, June 2016, pp. 1–6. 126, 138, 141

[298] A. Aldahdooh, M. Barkowsky, and P. L. Callet, "Content-aware adaptive multiple description coding scheme," in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–6. 126, 138

[299] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011. 126, 141

[300] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transaction on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004. 126

[301] A. Aldahdooh, E. Masala, G. Van Wallendael, and M. Barkowsky, "Framework for reproducible objective video quality research with case-study on PSNR implementations," *Elsevier Digital Signal Processing*, " in press". 133, 141

[302] S. Tourancheau, F. Autrusseau, Z. M. P. Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *2008 15th IEEE International Conference on Image Processing*, Oct 2008, pp. 365–368. 134

[303] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035. 137

[304] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Systems with Applications*, vol. 82, pp. 128 – 150, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417302397 143

[305] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1. 143

[306] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794. 143

[307] G. Vandewiele, P. Colpaert, J. Van Herwegen, O. Janssens, R. Verborgh, E. Mannens, F. Ongenae, and F. De Turck, "Predicting train occupancies based on query logs and external data sources," in *Proc. of the 26th International World Wide Web Conference: 7th Interna- tional Workshop on Location and the Web*, 2017. 143

[308] O. Janssens, N. Noppe, C. Devriendt, R. Van de Walle, and S. Van Hoecke, "Data-driven multivariate power curve modeling of offshore wind turbines," *Engineering Applications of Artificial Intelligence*, vol. 55, pp. 331–338, 2016. 143

[309] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers *et al.*, "Practical lessons from predicting clicks on ads at facebook," in *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM, 2014, pp. 1–9. 143

[310] M. Narwaria and W. Lin, "Svd-based quality metric for image and video using machine learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 347–364, April 2012. 153

[311] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, March 2013. 153

[312] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "Color image database for evaluation of image quality metrics," in *2008 IEEE 10th Workshop on Multimedia Signal Processing*, Oct 2008, pp. 403–408. 153

[313] P. Le Callet and F. Autrusseau, "Subjective quality assessment irccyn/ivc database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/. 153

[314] "Mict image quality evaluation database." [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html 153

[315] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, "Reduced-reference metric design for objective perceptual quality assessment in wireless imaging," *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525 – 547, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596509000836 153

[316] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii," 2003, available: http://www.vqeg.org. 154

Contents

List of Tables

List of Figures

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Ahmed ALDAHDOOH

Transmission vidéo «contenu»-adaptée dans le contexte HEVC

Optimisation de la compression, de la tolérance aux erreurs de la transmission, et de la qualité visuelle

Content-Aware Video Transmission in HEVC Context

Optimization of compression, of error resilience and concealment, and of visual quality

Résumé

Dans cette étude, nous utilisons des caractéristiques locales/globales en vue d'améliorer la chaîne de transmission des séquences de vidéos. Ce travail est divisé en quatres parties principales qui mettent à profit les caractéristiques de contenu vidéo. La première partie introduit un modèle de prédiction de paramètres d'un encodeur basé sur la complexité du contenu. Ce modèle utilise le débit, la distorsion, ainsi que la complexité de différentes configurations de paramètres afin d'obtenir des valeurs souhaitables (recommandées) de paramètres d'encodage. Nous identifions ensuite le lien en les caractéristiques du contenu et ces valeurs recommandées afin de construire le modèle de prédiction. La deuxième partie illustre le schéma de l'encodage à description multiple (Multiple Description Coding ou MDC, en anglais) que nous proposons dans ces travaux. Celui-ci est optimisé pour des MDC d'ordre-hauts. Le décodage correspondant et la procédure de récupération de l'erreur contenu-dépendant sont également étudiés et identifiés. La qualité de la vidéo reçue a été évaluée subjectivement. En analysant les résultats des expériences subjectives, nous introduisons alors un schéma adaptatif, c'est-à-dire adapté à la connaissance du contenu vidéo. Enfin, nous avons simulé un scénario d'application afin d'évaluer un taux de débit réaliste. Dans la troisième partie, nous utilisons une carte de déplacement, calculées au travers des propriétés de mouvement du contenu vidéo, comme entrée pour l'algorithme de masquage d'erreur par recouvrement (inpainting based error concealment algorithm). Une expérience subjective a été conduite afin d'évaluer l'algorithme et d'étudier la perturbation de l'observateur au visionnage de la vidéo traitée. La quatrième partie possèdent deux sous-parties. La première se penche sur les algorithmes de sélections par HRC pour les grandes bases de données de vidéos. La deuxième partie introduit l'évaluation de la qualité vidéo utilisant la connaissance du contenu global non-référencé.

Mots clés

Caractéristiques du contenu visuel, Résilience d'erreur, Masquage d'erreur, Enodage à description multiple, Compression vidéo, Evaluation de la qualité visuelle, Transmission vidéo, HEVC standard.

Abstract

In this work, the global/local content characteristics are utilized in order to improve the delivery chain of the video sequences. The work is divided into four main parts that take advantages of video content features. The first part introduces a joint content-complexity encoder parameters prediction model. This model uses bitrate, distortion, and complexity of different parameters configurations in order to get the recommended encoder parameters value. Then, the links between content features and the recommended values are identified. Finally, the prediction model is built using these features and the recommended encoder parameter values. The second part illustrates the proposed multiple description coding (MDC) scheme that is optimized for high-order MDC. The corresponding decoding and content-dependent error recovery procedures are also identified. The quality of the received videos is evaluated subjectively. By analyzing the subjective experiment results, an adaptive, i.e. content-aware, scheme is introduced. Finally, an application scenario is simulated to study the realistic bitrate consumption. The third part uses the motion properties of a content to introduce a motion map that will be used as an input for the modified state-of-the-art inpainting based error concealment algorithm. A subjective experiment was conducted to evaluate the algorithm and also to study the content-aware observer's disturbance when perceiving the processed videos. The fourth part has two sub-parts, the first one is about HRC selection algorithms for the large-scale video database with an improved performance evaluation measures for video quality assessment algorithms using training and validation sets. The second part introduces global content aware no-reference video quality assessment.

Key Words

Visual content characteristics, Error resilience, Error Concealment, Multiple description coding, Video compression, Visual quality assessment, Video transmission, HEVC standard.