Generalized Rate-Distortion Functions of Videos

by

Wentao Liu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2019

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Gene Cheung Associate Professor Dept. of Electrical Engineering & Computer Science York University
Supervisor(s):	Zhou Wang Professor, Dept. of Electrical and Computer Engineering University of Waterloo
Internal Member:	Amir Khandani Professor, Dept. of Electrical and Computer Engineering University of Waterloo
Internal Member:	Liang-Liang Xie Professor, Dept. of Electrical and Computer Engineering University of Waterloo
Internal-External Member:	Edward R. Vrscay Professor, Dept. of Applied Mathematics University of Waterloo

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Customers are consuming enormous digital videos every day via various kinds of video services through terrestrial, cable, and satellite communication systems or over-the-top Internet connections. To offer the best possible services using the limited capacity of video distribution systems, these video services desire precise understanding of the relationship between the perceptual quality of a video and its media attributes, for which we term it the generalized rate-distortion (GRD) function. In this thesis, we focus on accurately estimating the GRD function with a minimal number of measurement queries.

We first explore the GRD behavior of compressed digital videos in a two-dimensional space of bitrate and resolution. Our analysis on real-world GRD data reveals that all GRD functions share similar regularities, but meanwhile exhibit considerable variations across different combinations of content and encoder types. Based on the analysis, we define the theoretical space of the GRD function, which not only constructs the groundwork of the form a GRD model should take, but also determines the constraints these functions must satisfy.

We propose two computational GRD models. In the first model, we assume that the quality scores are precise, and develop a robust axial-monotonic Clough-Tocher (RAMCT) interpolation method to approximate the GRD function from a moderate number of measurements. In the second model, we show that the GRD function space is a convex set residing in a Hilbert space, and that a GRD function can be estimated by solving a projection problem onto the convex set. By analyzing GRD functions that arise in practice, we approximate the infinite-dimensional theoretical space by a low-dimensional one, based on which an empirical GRD model of few parameters is proposed.

To further reduce the number of queries, we present a novel sampling scheme based on a probabilistic model and an information measure. The proposed sampling method generates a sequence of queries by minimizing the overall informativeness of the remaining samples.

To evaluate the performance of the GRD estimation methods, we collect a large-scale database consisting of more than $4,000$ real-world GRD functions, namely the Waterloo

generalized rate-distortion (Waterloo GRD) database. Extensive comparison experiments are carried out on the database. Superiority of the two proposed GRD models over state-of-the-art approaches are attested both quantitatively and visually. Meanwhile, it is also validated that the proposed sampling algorithm consistently reduces the number of queries needed by various GRD estimation algorithms.

Finally, we show the broad application scope of the proposed GRD models by exemplifying three applications: rate-distortion curve prediction, per-title encoding profile generation, and video encoder comparison.

Acknowledgements

First and foremost, I would like to express my sincere thanks to Professor Zhou Wang for taking me on as his Ph.D student. Throughout the five years of my Ph.D study, Zhou has never been tired of listening to and discussing with me on my research projects, and always provided suggestions with his incisive and profound thoughts. Through these discussions, Zhou has taught me fundamentals of good research, writing and presentation which I will benefit from for my future career. Zhou encouraged me to explore new research topics that both of us found interesting, and immerse myself into what I had a passion for. Zhou has also built a collaborative environment in the Image and Vision Computing (IVC) Lab, where I could enjoy a happy and fruitful PhD life with fellow colleagues. I feel tremendously lucky and privileged to have had the opportunity to work with Zhou. I could not have asked for a finer mentor. I am honored to have Dr. Gene Cheung, Dr. Amir Khandani, Dr. Liang-Liang Xie and Dr. Edward R. Vrscay on my thesis committee.

I would like to thank all the amazing group members in the IVC Lab for their selfless support and help on my research and life. These include Kede Ma, Zhengfang Duanmu, Shahrukh Athar, Rasoul Mohammadi Nasiri, Kai Zeng, Abdul Rehman, Jiheng Wang, Hojatollah Yeganeh, Qingbo Wu, Honglei Su, Qi Liu, Diqi Chen, Zhuoran Li, and Zhongling Wang. Particularly, Kede Ma helped me a lot as a senior in the lab and a close friend in life. Special thanks to Zhengfang Duanmu, who was not only an excellent collaborator on research, but also a best friend outside the lab.

I also wish to thank all my wonderful friends outside the lab: Xin Wei, Sijia He, Wenchao Xu, Nan Chen, Chaojie Ou, Mengge Chen, Weiya Ye and many others, for occasionally dragging me out of serious research and making my Ph.D journey enjoyable. I will never forget the colorful time we spent together.

Finally, I would like to thank my parents, Li Wang and Yuanchun Liu, for their endless love and support. My parents have made countless sacrifices for me, and have provided me a steady harbor to take a break and sail again. I owe both apology and gratitude to my girlfriend, Anqi He, who stuck with our love despite the distance of 13,000 km between us, and cured me so many times with her tenderness and tolerance.

Dedication

This is dedicated to Mom and Dad.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BD-PSNR Bjøntegaard-Delta peak signal-to-noise ratio 7, 85, 87, 88

BD-Rate Bjøntegaard-Delta bitrate 7, 69, 87, 88, 90

BVQA blind video quality assessment 13, 16, 17

CT Clough-Tocher 6, 8, 29, 30, 33, 35, 36, 55, 66, 68, 69, 71

DCT discrete cosine transform 12, 15

DMOS difference mean opinion score 18

EGRD eigen generalized rate-distortion 6, 8, 28, 52, 53, 55, 57, 71, 72, 76, 79–81, 91

FR full-reference 13–15, 17

GMSD gradient magnitude similarity deviation 14

GoP group of pictures 11

GRD generalized rate-distortion iv, v, 2–10, 13, 17, 19, 21, 22, 24–29, 44–49, 51–57, 59, 61–63, 65, 66, 68, 69, 71, 72, 76, 79, 81, 82, 84, 91, 92

HVS human visual system 1, 2, 14, 15, 27, 29, 63

IQA image quality assessment 14, 15

ITS information-theoretic sampling 8, 56, 60, 65, 66, 68, 69, 72, 91

MOS mean opinion score 17, 18, 53, 59

MOVIE motion-based video integrity evaluation 15

MSE mean squared error 2, 14

MV motion vector 12

OSQP operator splitting quadratic program 44, 66

PCA principal component analysis 51, 53, 71, 72, 76

PCHIP piece-wise cubic hermite interpolating polynomial 63, 68

PGRD polynomial generalized rate-distortion 71, 72, 76, 79

POCS projection onto convex sets 6, 8, 44, 91

PSNR peak signal-to-noise Ratio 14, 15, 87

QoE Quality of Experience 3, 83, 91

RAMCT robust axial-monotonic Clough-Tocher iv, 5–8, 28–30, 33, 35, 42, 44, 55, 57, 65, 66, 68, 69, 81, 82, 84, 88, 91

RD rate-distortion 1, 2, 7, 10, 19, 22, 69, 76, 92

RMSE root mean squared error 52, 53, 66, 72, 79

RR reduced-reference 13, 16

RS random sampling 68

SSIM structural similarity 14, 15, 63

STRRED spatio-temporal reduced reference entropic differencing 16

TGRD trigonometric generalized rate-distortion 71, 72, 76, 79

VIF visual information fidelity 14, 15

VQA video quality assessment 8, 10, 13–15, 17, 18, 24, 29, 59, 63, 65, 68, 81

Waterloo GRD Waterloo generalized rate-distortion iv, 6, 8, 52, 62, 64, 65, 68, 69, 71, 72, 76, 82, 84, 90, 91

Chapter 1

Introduction

1.1 Motivation

The rate-distortion (RD) theory lays a theoretical foundation for lossy data compression and is widely employed in image and video compression schemes [6]. One of the most profound outcomes from the theory is the so-called RD curve, which is defined as the function of signal distortion (or loss of quality from the original source) against data rate. The RD curve reveals the minimum bitrate required to encode a source when a fixed amount of distortion is allowed, or equivalently the highest achievable quality given limited bitrate resources [63][1]. The RD theory also proves that the RD curve varies with both the distribution of source signal and the distortion measure, making the curve a perfect description of the complexity of video content and the mechanism of the human visual system (HVS). However, the original rate-distortion theory somewhat simplistically and ideally assumed an omniscient encoder, which not only has precise knowledge of the source distribution, but also is able to fully exploit the knowledge for the coding design. If the

[1]Since a rate-distortion function can be trivially converted to a rate-quality function by subtracting the former from the maximum quality value, we use the two terms, "rate-distortion function" and "rate-quality function", interchangeably, unless otherwise specified. Without loss of generality, we further assume that the response of the rate-distortion function is the perceptual quality instead of the distortion, so the rate-distortion function is generally increasing with regards to the bitrate.

assumption ever held, any bitrate would correspond to a quality value, just as the beautiful exponential relationship between bitrate and mean squared error (MSE) for a memoryless Gaussian source [6]. Unfortunately, this assumption is utterly violated in the case of video compression. Due to limited understanding of the video signal distribution and restricted computational resources, it is currently impossible for a practical video encoder to seek out an optimal encoding profile given a fixed bitrate. To combat this dilemma, state-of-the-art video encoders often make themselves extremely customizable, leaving tens of tunable arguments to the users. Each argument influences the resulting quality in a different way and adds a new dimension to the rate-distortion relationship. This fact motivates us to propose the notion of multivariate generalized rate-distortion (GRD) functions, which becomes the main topic of the thesis. To better explain the idea of GRD functions, we illustrate the process of compressing a video with a practical video encoder in Fig. 1.1. Given a source video and a practical encoder, different encoding arguments, or media attributes may lead to different perceptual qualities of compressed video *representations*. The proposed GRD function describes the mapping from the input media attributes to the output quality as shown in Fig. 1.1. Another direction to generalize the RD function is considering various criteria of the distortion measure. The perceptual quality of a video may change substantially when viewed on different devices. In order to thoroughly describe the interaction between video content, video codec, and the HVS, we define the general class of GRD functions as follows

$$f\colon \mathbb{R}^K \to \mathbb{R}^J, \tag{1.1}$$

where the input of the function is K encoder arguments (such as the target bitrate, the spatial resolution, the framerate, the bit depth etc.), and the output of the function is a vector of perceptual distortion/quality at J viewing conditions (including variations of viewing devices).

The study of the GRD function has drawn increasing attention from the industry. Video services have become an integral part of the modern life, and the global consumer demand of such services continues to grow at an accelerating speed. According to Cisco's forecast reports [17], video traffic proportion in total mobile data traffic is expected to exceed 82% by the end of 2022, skyrocketing from modest 59% in 2017. The explosion of video data volume will quickly drain the storage, bandwidth, and computational resources

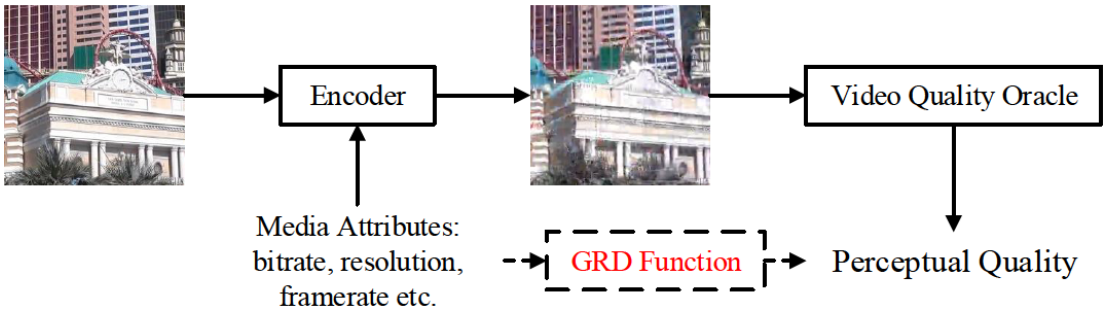


Figure 1.1: A GRD function arises from practical video encoding processes.

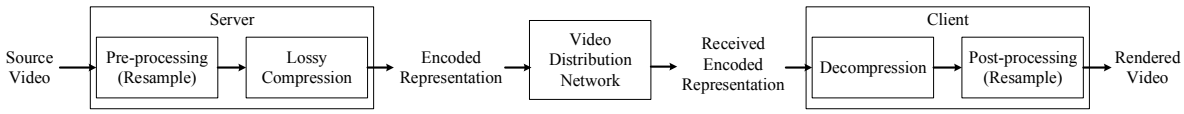


Figure 1.2: Flow diagram of video delivery chain

in the next decade. Concurrent with the surging demand of video services is the increasing expectation of Quality of Experience (QoE) from consumers. It has become a major task of video service providers to deliver high quality videos to enormous and diverse consumers with limited capacity of the video delivery systems.

A typical video delivery chain is illustrated in Fig. 1.2. At the server side, video service providers usually preprocess and encode a high quality source video into several *representations* with different settings of media attributes. At the client side, the representation of possible QoE is selected among those provided from the server side according to the consumer's viewing environment. In order to address the growing heterogeneity of consumers' viewing conditions, video service providers have to determine which representations to encode so that both optimal QoE performance and efficient bitrate usage can be achieved simultaneously. This difficult task requires precise understanding of the relationship between the perceptual quality of a video and its media attributes, which can be fully depicted by the proposed GRD function.

In summary, the GRD function can benefit a variety of existing and potential video applications in three ways.

1. The GRD function offers an approach to estimating the perceptual quality of a video representation without actually encoding it, which saves all the time-consuming computations involved in sophisticated video compression and quality assessment procedures.

2. The GRD function provides an explicit functional form of the mapping from media attributes to the perceptual quality, making it possible to directly optimize media attributes towards desired objectives of many video applications.

3. The GRD function illustrates a comprehensive description of a video content with regards to a specific video encoder, enabling thorough comparison of video codecs and deeper understanding of video complexity.

Despite the central role of the GRD function in many video applications, estimating a GRD function is difficult, expensive and time-consuming. The major difficulty arises from the lack of theoretically-grounded GRD model and the scarcity of samples in the GRD space restricted by computational resources. The function forms of existing GRD models are often heuristically designed without theoretical justification or empirical validation. The performance of these models is further impaired when only sparse attribute-quality pairs are available. This scenario often occurs in practice because probing the quality of a single sample in the GRD space involves sophisticated video encoding and quality assessment, both of which may demand excessive computational resources. For example, the recently announced highly competitive AV1 [3] video encoder and video quality assessment model VMAF [44] could be over 1000 times and 10 times slower than real-time for full high-definition (1920×1080) video contents. Given the massive volume of multimedia data on the Internet, the real challenge is to produce an accurate estimate of the GRD function with a minimal number of quality queries.

1.2 Objectives

The objectives of this thesis are to determine the theoretical space of the GRD function, and to develop novel GRD function estimation frameworks and methodologies. In particular,

we will focus on modeling the case where the video bitrate and spatial resolution are the main quality influential factors. The desirable properties of our GRD model are as follows:

- **Prediction accuracy**: The model produces asymptotically unbiased estimation of the GRD function, regardless of the source video complexity and the encoder mechanism.
- **Convergence speed**: The model requires a minimal number of samples to reconstruct the GRD surface with considerable accuracy.
- **Mathematical soundness**: The model has to be mathematically well-behaved, making it readily applicable to a variety of computational multimedia applications.

1.3 Contributions

In order to address the increasing diversity of video contents and heterogeneity of viewing devices, we proposed the notion of multivariate GRD functions as Eq. (1.1) for the first time in the literature. In the rest of the thesis, we focus on theoretically analyzing the properties of GRD functions, and developing effective algorithms to estimate a GRD function from limited queries. Specifically, we explore the behavior of GRD functions in a two-dimensional variable space of bitrate and resolution. Our analysis on real-world data shows that all GRD functions share similar regularities, but meanwhile exhibit considerable variations across different combinations of video content and encoder. Combining the practical observations with domain knowledge, we are able to define the theoretical space which a valid GRD function should live in. Determining the GRD function space not only constructs the groundwork of the form such a model should take, but also determines the constraints these functions must satisfy.

We propose two computational models for recovering the GRD surface. In the first model, we assume that the quality measurements are precise, and thus approximate the GRD function with a novel interpolation model. Specifically, an robust axial-monotonic Clough-Tocher (RAMCT) interpolation method is developed by imposing the axial-monotonicity

constraints on the classic Clough-Tocher (CT) model. With a novel re-parametrization strategy and a new continuity condition, we find that the RAMCT model can be formulated as a quadratic programming problem, whose solution provides an optimal interpolation model lying in the theoretical GRD function space. Moreover, the resulting GRD function is differentiable everywhere on the domain of interest, and possesses a certain minimum pseudo-norm property among all valid CT interpolants.

In the second model, we argue that all valid GRD functions must lie within a convex set that results from the intersection of a hyper-plane and a positive cone in a Hilbert space, and model the GRD estimation problem as a projection onto convex sets (POCS). The minimal number of required samples is thus determined by the dimensionality of the convex set of the GRD function. We further reveal that GRD functions arising in practice lie in a rather low-dimensional space according to an empirical study on a database consisting of $1,000$ real-world GRD functions. By combining the knowledge from both the theoretical reasoning and the empirical data, we formulate a low-parameter function model, namely the eigen generalized rate-distortion (EGRD) model, which is able to recover the GRD function with the access to only a small fraction of samples. The proposed EGRD model outperforms the existing methods both asymptotically and absolutely, and precludes the need for the complicated computations in the first model.

To further reduce the computational cost, we present a novel sampling scheme based on a probabilistic model and an information measure. The proposed sampling scheme picks the most informative sample every time, and updates the informativeness of the remaining candidates according to their correlations with the selected sample. Finally, a sequence of queries is generated by the proposed method, with which the overall uncertainty of the GRD function is minimized.

To evaluate the proposed algorithms, we establish the large-scale Waterloo generalized rate-distortion (Waterloo GRD) database, which is comprised of $4,750$ real-world GRD functions generated from more than $1,000$ natural videos spanning a variety of real-world contents. Extensive experiments are conducted on the database to compare the proposed algorithms with existing GRD models and sampling schemes in the literature. Experimental results show that the proposed GRD models consistently achieve higher reconstruction accuracy with fewer quality queries. Moreover, by visualizing the approximate GRD sur-

faces, we find that the two proposed models always generate well-behaved and valid GRD functions even with insufficient training samples. The power of the proposed sampling scheme is also evident in the experiments, as it always reduces the number of quality queries needed for various GRD models to achieve similar prediction accuracy.

Furthermore, we show the broad application scope of the GRD functions by three use cases that emerge in industrial practice using the proposed RAMCT model:

- **Rate-Distortion Curve Prediction at Novel Resolutions**: Given a set of RD curves at multiple resolutions, it is desirable to predict the rate-distortion performance at novel resolutions, especially when there exists a mismatch between the supported resolution of downstream content delivery network and the recommended encoding profiles. Compared to an intuitive linear interpolation method, the proposed GRD models can effectively exploit the underlying dependency among video signals at different resolutions, and accurately recover the shape of RD curve at a completely new resolution. The estimated RD curve preserves important properties of a GRD surface, thanks to the theoretical constraints imposed on the GRD model.

- **Per-Title Encoding Profile Generation**: To overcome the heterogeneity in users' network conditions and display devices, video service providers often need to encode a set of representations of various qualities. We introduce a quality-driven per-title optimization framework to automatically select the best encoding configurations where the proposed GRD model serves as the key component. The encoding profile generated by the proposed framework claims significant Bjøntegaard-Delta bitrate (BD-Rate) saving over the recommended profiles from Netflix [1], Apple [5], and Microsoft [50].

- **Codec Comparison:** In the past decade, there has been a tremendous growth in video compression algorithms, thanks to the fast development of computational multimedia. With many video encoders at hand, it becomes pivotal to compare their performance. We extend the widely used BD-Rate and Bjøntegaard-Delta peak signal-to-noise ratio (BD-PSNR) with the proposed GRD model to provide more comprehensive and more robust comparison between two video codecs.

In summary, the main contributions of the thesis are six-fold. First, we systematically summarize the properties of the GRD function, based on which the theoretical space of GRD function is well defined. Second, we introduce the RAMCT model inherited from the CT interpolation method for GRD approximation. Third, we formulate the GRD estimation problem as POCS in a Hilbert space. By approximating the function space with the most representative basis, we present a low-dimensional EGRD model for reconstructing the GRD function with minimal quality measurements. Fourth, we propose an information-theoretic sampling (ITS) scheme to further reduce the computational cost of GRD approximation. Fifth, we establish a large-scale GRD function database, the Waterloo GRD database, which is not only the first of its kind, but also the largest video quality database in the video quality assessment (VQA) community. Finally, we use three novel applications to show the potential of the proposed GRD models.

1.4 Thesis Outline

The layout of this thesis is as follows.

Chapter 2 first introduces three topics that are closely-related to the study of the GRD function: the video encoder mechanism, the objective VQA methods, and the subject-rated VQA databases. Then a review of existing models of GRD functions is presented. We find that these models are either systematically biased or computationally prohibitive.

In Chapter 3, we first comprehensively analyze the properties of the GRD function, based on which its theoretical space is determined. Then we develop a novel interpolation method, *i.e* RAMCT, to reconstruct the GRD surface from a moderate number of training samples. Finally, we show that the GRD function space can be approximated by the span of only a few principal components, based on which we propose the EGRD model.

Chapter 4 introduces a novel sampling scheme based on a probabilistic model and an information measure. The proposed sampling method generates a sequence of queries with which the uncertainty of GRD function is minimized. Then we construct a large-scale GRD function database, based on which extensive experiments are carried out to validate the superiority of the proposed methods.

Chapter 5 demonstrates the usefulness of the GRD models using three industry applications. We show that the GRD function serves as the key component in all the three applications, and the proposed GRD models significantly improve the current practice.

Chapter 6 summarizes the thesis, reemphasizing the importance of the proposed GRD function and discussing promising directions for future studies.

Chapter 2

Background

The generalized rate-distortion (GRD) function arises from the interaction of a video encoder and a video quality measurement. Therefore, it is not possible to study the GRD function without a deep understanding of video encoders, the recent advancement of objective video quality assessment, and the existing subjective experiments on real-world video data. In this chapter, we review the general workflow of a video encoder, the development of objective video quality assessment (VQA) models, and the existing subjective VQA databases in the literature. We also show how the GRD/rate-distortion (RD) function was estimated in previous researches, and conclude that existing studies on modeling the GRD function are ad-hoc and achieve only limited success.

2.1 Video Encoder

Although several advanced video encoders, such as HEVC [29], VP9 [30], AV1 [3], and AVS2 [53], have been proposed and deployed in the past decade, currently H.264 [80] is by far the most widely used video encoding standard in practice. Therefore we use the H.264 codec to introduce how a video encoder can efficiently compress a video. The other well-known encoders work in a quite similar way, even though they may differ in many technical details, such as the signal decomposition transforms and the shapes and sizes of prediction and encoding units.

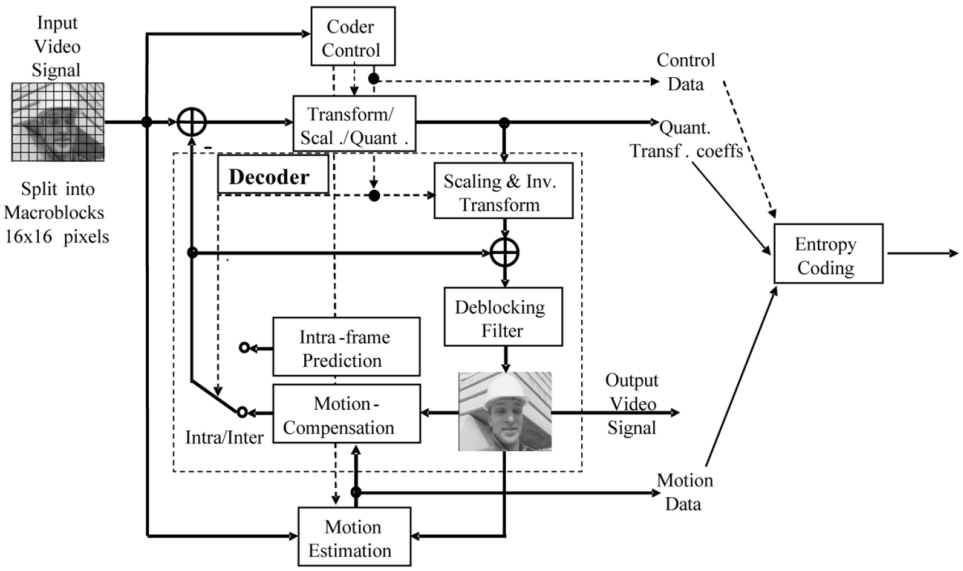


Figure 2.1: Basic coding structure of H.264 encoder for a macroblock. The dashed box illustrates a video decoder. Image by courtesy of Thomas Wiegand etc. [80].

The main idea of a video encoder is to remove the dependency between pixel values within or across frames so that the video signal can be encoded with minimal bitrates. This is done by predicting subsequent pixel values from already-encoded pixels, and only encoding the residuals with transform coding. Specifically, the H.264 encoder divides a picture into 16×16 disjoint *macroblocks*, and sequentially encodes one macroblock at a time. Fig. 2.1 illustrates the coding diagram of the H.264 encoder for a macroblock. The whole encoding process of a macroblock roughly consists of four steps. First, the pixel values of the macroblock currently under encoding are predicted by previously encoded macroblocks. Second, the prediction residual block is transformed into a frequency space, where the interdependency between residuals is further reduced. Third, the transformed values are quantized and encoded into bitstreams. Finally, the encoded macroblock is again decoded and stored in a buffer for the purpose of predicting subsequent macroblocks. In order to prevent encoding errors from spreading and to enable random access to the video stream, such prediction coding process is restricted inside a *slice*, which is formed by a sequence of macroblocks, or a group of pictures (GoP) in the case of inter-frame prediction.

We then take a look at the details of the H.264 encoder. To initialize the prediction coding process, the first macroblock has to be encoded without prediction. The H.264 encoder provides two major prediction strategies: intra- and inter-frame predictions, where the latter can be performed with only one frame (P-mode) or two frames (B-mode). The intra mode prediction only makes use of pixel values that are within the same frame, and predicts 4×4 blocks with Intra_4×4 mode in the luminance channel, or 16×16 blocks with Intra_16×16 mode in both the luminance or chroma channels. Moreover, a special mode called I_PCM allows an intra-mode macroblock to be directly encoded with its original pixel intensities. For the inter-mode macroblocks, their pixel values are predicted from other reference frames, some of which may even appear after the frame under encoding in the B-mode case as long as the reference frame is encoded earlier. This feature means that H.264-encoded videos are not decoded in the original time order. Since objects are moving across frames, motion vector (MV)s of an inter-mode macroblock relative to its reference in other frames should be estimated to achieve the best possible prediction and to minimize the residual. With the little cost of transmitting a MV, the encoding efficiency of the macroblock can be largely enhanced. To further save the bit-rate, a common technique for encoding an MV efficiently is to estimate it from the MVs of neighboring macroblocks. Then the MV residual is encoded and transmitted instead of the MV itself. It is worth noting that H.264 introduces two *skip* modes for the P- and B-mode macroblocks, respectively. For the skipped macroblocks, neither the MVs nor pixel residuals are encoded or transmitted. As a result, the optional skipped mode allows the encoder to encode a large area with constant motions with very few bits, but still preserve considerable perceptual qualities.

The pixel and motion vector residuals are then encoded by lossy transform coding. The basic unit for transformation is a 4×4 block (for both luma and chroma channels). A separable integer transformation, which behaves similarly to the traditional 4×4 discrete cosine transform (DCT), is used to achieve efficiency and effectiveness simultaneously. To further reduce the computational complexity, the quantization step on the transformed coefficients is integrated in the process of transformation. Additionally, a 4×4 or 2×2 Hadamard transform followed by a separate quantization process is applied to the DC coefficients in a macroblock for luma or chroma channels, respectively. Afterwards, these

quantized coefficients are encoded by a context-adaptive method (CAVLC or CABAC), and transmitted to the receiver.

In summary, the H.264 codec provides a hybrid block-based transform coding scheme for lossy video compression. In practice, a source video is often provided to the encoder with a target bitrate and other necessary encoding configurations, such as the spatial resolutions, the prediction modes etc. Then the encoder will then produce a corresponding video representation with visual distortions. In the case of the H.264 encoder, typical distortions include nonuniform blurring and blocking artifacts, and impairments of temporal smoothness during the video playback. These visual distortions lead to degraded perceptual quality. Such a pair of media attributes and quality is one point on the GRD surface. Therefore, the whole GRD surface can be obtained by changing the bitrates and other encoding configurations, though this approach is prohibitively time-consuming and expensive. It is worth noting that different video encoders often induce different visual distortions due to their specific designs. As a result, the GRD function may behave differently with different video encoders.

2.2 Video Quality Assessment

In order to assess the perceptual quality of countless videos, a reliable objective VQA model is thus highly desired. Existing VQA methods can be classified into full-reference (FR) VQA, reduced-reference (RR) VQA and blind video quality assessment (BVQA) based on the accessibility of the pristine reference when estimating a video's quality [62]. The FR VQA models are currently more accurate than the other two classes in predicting subjective quality scores. Since the reference video is often considered available in the context of GRD function estimation, we put our emphasis on the FR VQA methods, while a brief introduction of the RR VQA and BVQA models is also provided.

2.2.1 FR-VQA

The study of VQA started with investigating the problem of image quality assessment (IQA). mean squared error (MSE) and peak signal-to-noise Ratio (PSNR) were two earliest IQA methods that were widely used due to their simple computation, clear physical significance, and appealing mathematical properties such as differentiability, convexity, and positive definiteness. They still serve as the default quality models for rate-distortion optimization in modern video encoders [28, 29, 30, 3, 53]. While these methods could measure signal fidelity between distorted images and the references, their estimated quality scores often deviate far away from subjective ratings [73, 75, 74]. This is mainly because the classical signal fidelity measures fail to take the mechanism of the human visual system (HVS) into account.

In order to address the problem, many perceptually-meaningful FR IQA models were proposed in the past several decades. For example, the well-known structural similarity (SSIM) index [75] assesses image quality by measuring the changes in local luminance, contrast, as well as structural information. Natural scenes consist of many structures. The SSIM approach assumes that the HVS is evolved to be sensitive to the structures that recurrently occur in our everyday life. The SSIM index then quickly prevailed over a wide range of disciplines in image processing due to its high correlation with human opinions and computational simplicity, and now also becomes an option in the HEVC codec [29]. Various versions of the SSIM, such as the MS-SSIM [78] and the CW-SSIM [57], proposed later further improved the effectiveness and robustness of the SSIM index.

Another highly cited IQA model is the visual information fidelity (VIF) [64]. The VIF regards a pristine picture as a realization from a virtual "natural random source" and simulates any level of distortion annoyance by signal attenuation combined with additive white noise. The ratio of information conveyed by the HVS from a pristine image to that from a distorted image is then computed to measure visual quality of the distorted image. More recently, a very efficient and effective FR IQA algorithm was proposed in [84]. By calculating the gradient magnitude similarity deviation (GMSD) between the reference image and its distorted version, the algorithm can predict subjective opinions on image quality with state-of-the-art performance at a low computational complexity.

Directly applying the IQA methods to videos usually yields inferior performances because temporal distortions are ignored by such methods. Objective VQA models were thus designed to deal with the temporal issues. One of the earliest HVS-based VQA models was proposed in [48], where the HVS spatial-temporal filtering properties and masking effect were considered. In [79], Watson *et al.* exploited both visual filtering and masking models to estimate local just-noticeable differences of videos in the DCT domain. Then the locally adjusted errors were pooled to form a single quality score for the test video. Although following a similar philosophy, the perceptual video quality metric proposed in [36] estimated perceivable spatial distortion in the pixel domain rather than the frequency domain, and compensated it with temporal variability. In addition, color errors were also calculated and combined with the previous two factors to predict the quality of video. More recently, authors in [60] modeled the middle temporal visual area in the human brain by separable Gabor filter banks, and proposed a FR VQA method named motion-based video integrity evaluation (MOVIE). Besides the differences between Gabor filter responses, MOVIE also used the Gaussian smoothed DC errors to capture low frequency information and the optical flow fields to estimate temporal distortions. All these quality-related indices were then pooled and combined to form the MOVIE index. The performance of MOVIE was reported relatively high, but its computational burden is not bearable.

In order to exploit the success of FR IQA, some state-of-the-art VQA models base themselves upon excellent IQA indexes. For example, VMAF [44] enriched the frame-level VIF features, and trained the VQA model on a relatively large database. As another high-performance IQA model, SSIM has also been extended to evaluate the quality of videos in several works. Early trials along this direction include [47][77][76]. In [77], SSIM index for each locality is first calculated for the Y, Cb, and Cr channels independently, and then averaged with different weights assigned to each channel. Both the luminance and the motion masking effects are considered to respectively adjust the local weights in each frame and the frame weights to compute the final score. In [76], a new visual speed model was employed for weighted pooling local quality map, and improved performance was reported on SSIM and PSNR with the new pooling method. Most recently, based on SSIM and modern vision models, a novel VQA software, SSIMplus [54], has emerged, featuring not only state-of-the-art performance across different devices, resolutions and

contents, but also real-time computational speed.

2.2.2 RR-VQA and BVQA

Due to the limited bandwidth in real-world applications, full access to the reference is not always available. In such cases, RR methods may be used as a compromise. In order to represent the quality with only a few bits, early research tried to extract the significant features of perceivable video distortions [81], but only achieved limited success. In 2013, a spatio-temporal reduced reference entropic differencing (STRRED) index was proposed to evaluate the video quality by combining statistical models and perceptual principles [66]. The proposed model produced robust performance in various cases, although its relatively high computational complexity impedes its potential application.

The RR model gradually gives way to the BVQA method as the latter does not require an auxiliary channel to transmit quality features from the source video. Since a video compression codec degrades a video in a particular way, some BVQA models predict video quality by codec analysis. In [65], Søgaard *et al.* proposed to first identify whether a test video is encoded by H.264 [80] or MPEG-2 [70], and then extract respective quality features for each codec. Later, the authors proposed another set of quality features [37] for the HEVC-encoded videos [68]. Though knowledge of a specific codec helps such methods achieve decent performance, it is very hard to incorporate them into a single general-purposed model or to extend them to new codecs.

By considering a video as a stack of pictures, V-CORNIA [83] takes advantage of the successful BIQA features, CORNIA [85], to characterize frame-level perceptual qualities, and adaptively pool them into a video quality score along the temporal dimension. However, such a framework fails to take into account the following influencing factors in video perceptual quality: 1) motion-induced blindness [11, 55] to spatial distortions; 2) possible temporal artifacts or incoherence [86, 56]; 3) codec-specific distortion [86]; and 4) interaction between spatial and temporal artifacts [38].

Natural video statistics (NVS) features are employed to jointly consider spatiotemporal distortions as a whole. Normally, NVS features are first extracted [56, 82, 42, 43], and

then a regression function is trained on MOS to map extracted features to quality scores. However, due to the complex nature of the BVQA problem and our limited understanding on natural video statistics, these models have only achieved limited level of success.

Despite the specific limitations the three kinds of existing VQA models may have, they are faced with the same problem that the models are often tuned on a very limited subject-rated database, which makes their generalizability questionable in the real world. Most recently, a deep-learning-based BVQA model was trained to predict the scores from a FR model, SSIMplus [54], on a large-scale database [46]. Experimental results show that the model exhibits higher performance and better generalizability than other competing models. However, only limited distortion types were covered by the premature model.

2.3 VQA Databases

The VQA community have also made a lot of efforts on building up various databases with subjective ratings to facilitate the research. A good VQA database should contain as diverse contents and distortion types as possible so that it can not only serve as a proper validation for the proposed algorithm, but also provide representative training samples for researchers investigating new models. However, collecting subjective ratings is rather laborious, expensive and time-consuming, since at least 15 observers are required to obtain a mean opinion score (MOS). A typical subjective-rated VQA database only consists of less than a dozen of source videos and only a few hundreds of distorted videos. Moreover, it is typically the case that each video content is only encoded at 3-4 bitrates and 2-3 spatial resolutions. Such a sparse distribution of representation samples makes all the subject-rated databases inappropriate for benchmarking the GRD function estimation methods, motivating us to propose our own database of GRD functions. For the completeness of the review, we briefly list some well-known public-domain databases as follows:

1. **LIVE Video Quality Assessment Database [61]:** 150 distorted videos were generated from 10 pristine ones by undergoing four kinds of distortion, namely MPEG-2 compression, H.264 compression, simulated error-prone IP and wireless network transmission. For the compression distortions, four levels were included in this

database, while for the transmission distortions only three levels were used. The quality of each distorted video is given as difference mean opinion score (DMOS) which was computed from ratings by 29 subjectives.

2. **CSIQ Video Database [72]:** This database consists 216 distorted videos, which were generated from 12 high-quality reference videos with 6 different types of distortion. These distortions include 4 popular compression standards, i.e. H.264, HEVC, MPEG, and a wavelet based codec, SNOW, 1 simulated transmission error in wireless environment, and 1 spatial-domain additive white Gaussian noise. 35 subjects were recruited to rate all the videos, and DMOS for each test video is provided within the database.

3. **MCL-V Database [45]:** This VQA database contains 12 reference videos with the high-definition resolution (1920×1080), each of which is compressed by the H.264 encoder at two different spatial resolutions at four different levels. As a result, 96 video representations are generated, whose MOS were collected and published along with the database.

4. **IVP Subjective Quality Video Database [88]:** In this database, 10 reference videos of high definition at 1920×1088 were subjected to 4 kinds of distortions to generate a total of 128 distorted videos. 42 paid viewers, including 25 non-experts and 17 experts, were recruited to watch all the videos and rated their perceptual qualities. Subjective scores are reported in the form of DMOS plus standard deviation.

5. **MCML 4K UHD video quality database [15]:** This database covers two of the highest spatial resolutions in practical use, *i.e.* 3840×2160 and 1920×1080. Three state-of-the-art encoders, H.264, HEVC and VP9, were employed to compress ten source videos at four different quality levels. As a result, 240 distorted videos were generated in total.

2.4 GRD Function Estimation Methods

Although many recent works have noticed that various media attributes may influence the perceptual quality of encoded representations [89, 69, 22, 12, 52], this is the first time to explicitly define a multivariate GRD function as in Eq. (1.1) to describe the relationship between these attributes and the resulting quality. Therefore, existing works are only focused on modeling 1D RD functions. Although we may extend these 1D methods to multi-dimensional cases by estimating multiple RD curves at different resolutions, failing to exploit the correlations between these RD curves may lead to inferior performances in terms of both the prediction accuracy and the convergence rate. Moreover, these methods lack the capability to predict the RD behaviors at novel resolutions, since they are restricted only to the bitrate dimension and several discrete resolutions.

We then briefly review existing methods for estimating RD functions. These methods roughly fall into two categories based on their assumptions about the shape of a RD function. The first model class only makes weak assumptions about the properties of RD functions. For example, [22] assumed the continuity of RD functions and apply linear interpolation to estimate the response function after densely sampling the encoding bitrates. However, the exhaustive search process is computationally expensive, not to mention the number of samples required increases exponentially with respect to the dimension of input space.

By contrast, the second class of models makes strong *a priori* assumptions about the form of the RD function to alleviate the need of excessive training samples. For example, Toni *et al.* [69, 41] derived a reciprocal function to model the RD curve as

$$q = c - \frac{a}{|x+d|^b + 0.01},$$

where x and q respectively denote the bitrate and the perceptual quality of the video representation, while a, b, c and d are free parameters to fit. Similarly, [12] modeled the rate-quality curve at each spatial resolution with a logarithmic function, *i.e.*

$$q = a\log(x+1) + bx + c.$$

A significant limitation of these models is that the employed assumptions are often heuristic, leading to biased estimation results.

In addition to the specific limitations the two kinds of models may respectively have, they also suffer from the same problem that the training samples in the GRD space are either manually picked or randomly selected, neglecting the difference in the informativeness of samples. While many recent works acknowledge the importance of GRD function [89, 69, 22, 12], a careful analysis and modeling of the response has yet to be done. To address this void, it is highly desirable to seek for a good compromise between 1) global and rigid models depending on minimal training samples and 2) local and indefinite models requiring exhaustive search in the video representation space.

Chapter 3

Modeling Generalized Rate-Distortion Function

In this chapter, we first summarize the mathematical properties that all generalized rate-distortion (GRD) functions share through both theoretical analysis and practical observations. Such mathematical properties actually define the theoretical space of the GRD function, which not only lays the groundwork of the mathematical form a GRD model should take, but also determines the constraints a valid GRD function should conform to. Then, we develop two empirical GRD models by combining the theoretical function space with interpolation or dimension reduction techniques, respectively. Both models can accurately recover the per-title GRD function with a small number of queries, and possess desired mathematical features that will benefit subsequent applications.

3.1 Theoretical Space of GRD Functions

We have defined the GRD function as the mapping from media attributes, such as bitrate and spatial resolution, to perceptual qualities of encoded video representations with these attributes. In this section, we show how the theoretical space of GRD functions can be derived from domain knowledge and real-world data.

3.1.1 Observations from Practical GRD Surfaces

We construct a large-scale database of GRD functions, where more than $4,000$ realistic GRD functions are collected. Details about the database will be elaborated later in Section 4.2. From the practical data, we summarize several key observations of GRD function as follows.

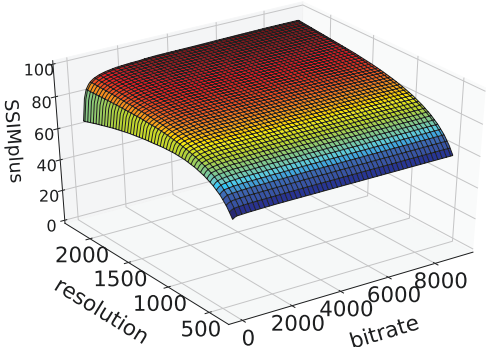
Observation 1: The family of GRD function embraces a significant amount of variability across different video contents and encoding schemes. We illustrate in Fig. 3.1 three sample GRD surfaces of different contents encoded with the same video codec, from which we can see how the surface varies with video complexity. Moreover, the capability of video encoder may also significantly influence the shape of GRD function as shown in Fig. 3.2. Such diversity turns out to be the major challenge for precisely approximating a per-title GRD function.

Observation 2: The GRD surface plateaus when encoding bitrates are high. This fact is made evident by projecting the 3D GRD surface to the bitrate-quality plane, as displayed in Fig. 3.3. Each rate-distortion (RD) curve eventually saturates when the encoding bitrate goes beyond a certain threshold. Another interesting finding from Fig. 3.3 is that the saturation threshold for a higher resolution is always greater than that for a lower one. This phenomenon may be explained by the fact that downsampling a video reduces the total information that needs to be encoded.

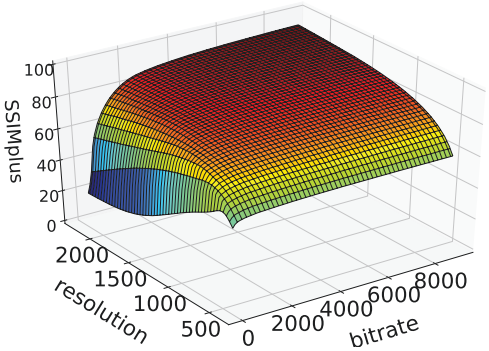
Observation 3: GRD functions are smooth in the bitrate-resolution space. In theory, the Shannon lower bound, the infimum of the required bitrate to achieve a certain quality, is guaranteed to be smooth with respect to the target distortion [6]. On the other hand, successive change in the spatial resolution would gradually deviate the frequency component and entropy of the source video, resulting in smooth transition in the perceived quality. In practice, the smoothness of GRD functions have been empirically attested in many subjective experiments [52, 87] as well as the sample GRD surfaces in Fig. 3.1 and Fig. 3.2.

Observation 4: The GRD function is axially monotonic along the bitrate dimension at any given resolution. With the resolution fixed, the GRD function degenerates to a normal rate-distortion (rate-quality) curve, which is guaranteed to be monotonic with

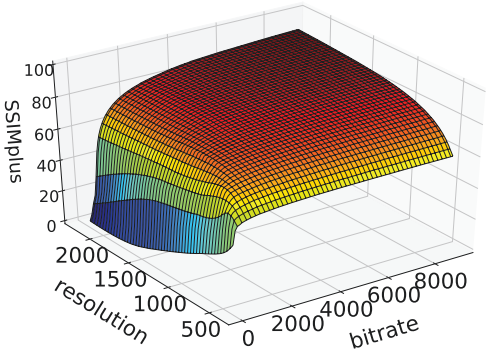
(a) Moto



(c) Soccer

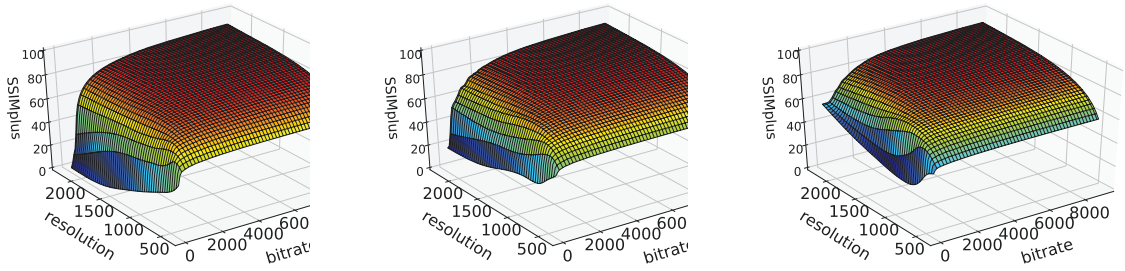


(e) Garden



(f) GRD surface of "Garden"

Figure 3.1: GRD functions of different contents compressed by the H.264 codec [28].



(a) H.264 [28] (b) HEVC [29] (c) VP9 [30]

Figure 3.2: GRD surfaces of "Garden" generated by different video encoders.

respect to the amount of consumed bitrate resources by the rate-distortion theory [6]. The axial monotonicity is also empirically validated by the increasing curves in Fig. 3.3 as well as multiple subjective tests carried out on other video quality assessment (VQA) databases [45, 15].

Observation 5: The GRD function is generally NOT monotonic along the resolution axis. We draw the resolution-quality curves in Fig. 3.4 by slicing the GRD surface at various bitrates. As we can see, high resolutions do not necessarily mean high perceptual quality when the video is encoded with short bitrate supply. To be specific, encoding at high resolution with insufficient bitrates would produce artifacts such as blocking, ringing, and contouring, whereas encoding at low resolution with upsampling and interpolation would introduce blurring [22]. Besides, the actual behavior of resolution-quality curve heavily depends on the characteristics of video content, adding extra complexity for modeling the GRD function.

Observation 6: Given adequate bitrate resources, a representation with higher resolution size always exhibits greater perceptual quality as indicated by Fig. 3.4. When a pristine video is encoded with the highest bitrate, we deem that no compression artifacts will be introduced by the encoding process. Therefore, the quality degradation only comes from the loss of high frequency components due to the lowpass filtering and the downsample-upsample process. Since the degree of frequency loss is a monotonic function of the scaling factor, the perceptual quality of the resulting representation degrades as the encoding resolution shrinks. Some advanced viewing devices, such as a smart 4K TV, may

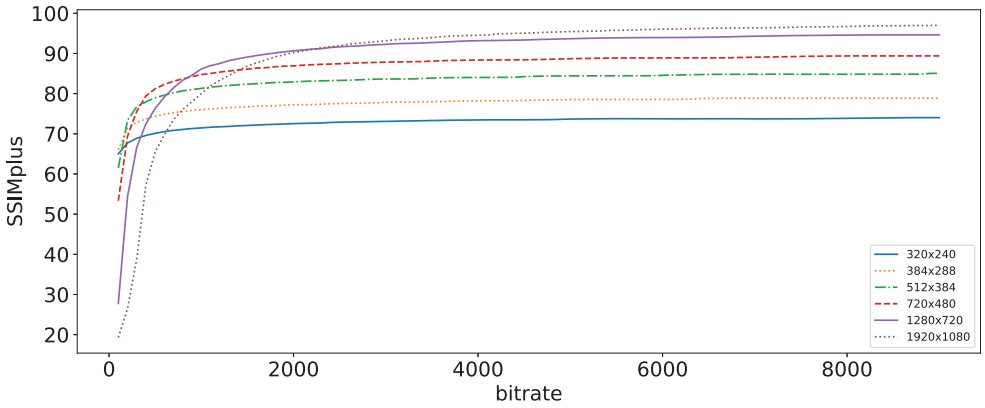


Figure 3.3: RD curves of "Soccer" at different resolutions.

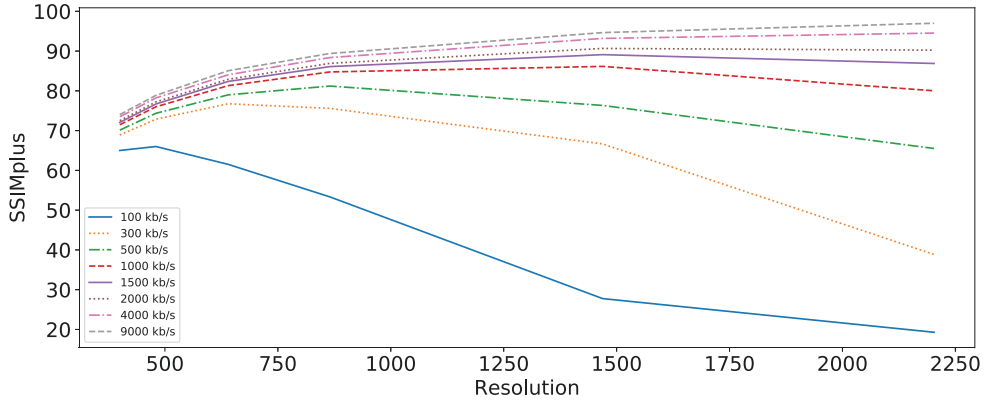


Figure 3.4: Resolution-quality curves of "Soccer" at different bitrates.

mitigate the loss of high frequency components by performing advanced super-resolution techniques, but this does not change the axial-monotonicity property. This observation also implies that the highest quality representation is achieved by encoding the source video with abundant bitrates at the original resolution.

3.1.2 Mathematical Properties of GRD Functions

The observations revealed from the real-world data provide us with a good foundation to characterize the GRD function. We are now ready to abstract and formulate the common

properties of GRD function.

Denote the domain of GRD function f by Ω. The first property is that Ω is a compact set in $\mathbb{R}^2$. In the thesis, we are focused on only two media attributes, *i.e.* the bitrate and the spatial resolution of a representation, so we define the two variables of the GRD function as the bitrate x in the unit of kilo-bits per second (kb/s), and the diagonal length y of spatial resolution in the unit of number of pixels. It is reasonable to represent a resolution by its diagonal length, because size and diagonal length are 1-on-1 mapped and well correlated among commonly used video resolutions [1, 5, 22]. Furthermore, we show that x and y arguably belong to two bounded and closed intervals $[x_{MIN}, x_{MAX}]$, and $[y_{MIN}, y_{MAX}]$, respectively, and the exact values of x_{MIN}, x_{MAX}, y_{MIN} and y_{MAX} are easily determined under reasonable assumptions. First, bitrate can never be negative, so we may set x_{MIN} to 0, suggesting that all pixel intensities are severely degraded to a single fixed value such that no bit is used to encode the video. Then, x_{MAX} may be determined by taking the maximum saturation bitrate of the highest resolution among a large number of pristine videos of diverse complexity. According to **Observation 2** in Section 3.1.1, encoding a video at bitrate higher than the saturation point will not achieve any perceptible quality improvement, so it is meaningless to model the GRD function beyond that point. For the other dimension y, we consider $y_{MIN} = 1$, when only one pixel per frame is encoded. At last, the value of y_{MAX} can be obtained from the original resolution, or commonly used encoding configuration recommendations [1, 5, 50]. Thus, we conclude that the domain Ω of GRD function should be a compact subset of a rectangular region defined by

$$\Omega_{MAX} \coloneqq \{(x, y) | x \in [x_{MIN}, x_{MAX}], y \in [y_{MIN}, y_{MAX}]\}. \tag{3.1}$$

It is worth noting that Ω_{MAX} itself is also a valid domain for the GRD function, but in practice we are often interested in the function's behavior only on a true subset of Ω_{MAX}. For example, we may restrict the minimum values of x and y to be much greater than x_{MIN} and y_{MIN}, respectively, since practically no encoder operates at 0 bitrate or 1×1 resolution. For another example, the number of resolutions used in real-world applications is actually finite, so the value of y is selected from a finite set rather than a closed interval.

Inspired by **Observation 3**, we consider smoothness as the second property of GRD function. Mathematically, "smoothness" of a function is measured by its order of continu-

ity. First, we assume that ideal GRD functions are continuous, *i.e.* $f \in C^0$, when the two independent variables x and y are continuous real variables. Furthermore, it is beneficial to impose first-order continuity on GRD functions in practice. For instance, the GRD surface is desired to be differentiable in many multimedia applications [69, 13]. Higher orders of continuity are also desirable, but may be difficult to justify in practice.

Arising from the first two properties, the third property says that the GRD function has a bounded range. Given a pair of bitrate and spatial resolution values (x, y), the encoder will compress the source video to the specified representation, whose quality is the GRD function value of (x, y). First, the quality of a compressed video is usually capped by that of its source video, if we consider the source video of perfect quality. This implies that the upper bound of the GRD function is the quality of the source video, denoted by z_{MAX}. Moreover, the upper bound is only achieved by encoding the source video at the highest profile, *i.e.* $f(x_{MAX}, y_{MAX}) = z_{MAX}$. On the contrary, the worst quality z_{MIN} is delivered by the lowest coding profile, either when $x = x_{MIN}$ or $y = y_{MIN}$. In both cases, the video content is utterly destroyed by excessive compression, and almost no information can be transmitted into the human visual system (HVS). Since the unit of perceptual quality is arbitrary, we normalize the range of GRD function such that $z_{MIN} = 0$ and $z_{MAX} = 100$ to keep consistent with the employed quality measurement tool SSIMplus [54]. Finally, we formulate this property as

$$f(x, y) \in [0, 100], \forall (x, y) \in \Omega \subseteq \Omega_{MAX}. \tag{3.2}$$

We may also assume that the following three equations

$$\begin{aligned} &f(x_{MIN}, y) = 0, \forall y \in [y_{MIN}, y_{MAX}] \\ &f(x, y_{MIN}) = 0, \forall x \in [x_{MIN}, x_{MAX}] \\ &f(x_{MAX}, y_{MAX}) = 100 \end{aligned} \tag{3.3}$$

hold for all GRD functions.

The fourth property of the GRD function is its axial monotonicity as described in **Observation 4** and **6**. Formally, we summarize the two types of axial monotonicity by

$$f(x_1, y) \leq f(x_2, y), \forall\ x_1, x_2, y,\ \text{s.t.}\ x_{MIN} \leq x_1 \leq x_2 \leq x_{MAX}, y_{MIN} \leq y \leq y_{MAX} \tag{3.4}$$

$$f(x_{MAX}, y_1) \leq f(x_{MAX}, y_2), \forall\ y_1, y_2,\ \text{s.t.}\ y_{MIN} \leq y_1 \leq y_2 \leq y_{MAX}. \tag{3.5}$$

3.1.3 Construction of the GRD Function Space

Existing GRD models [52, 69, 41, 12] typically presume a parametric form of the GRD function that violates at least one of the aforementioned properties, and are thus deemed systematically biased. Instead of constructing a parametric function form conforming to all the mathematical properties, we define the theoretical GRD function space to account for both the regularities and the significant variability of the GRD function as demonstrated in the previous two subsections.

Denote by W_{GRD} the theoretical space of GRD function. The actual definition of W_{GRD} may depend on specific problem settings. Here we only give two sample definitions of W_{GRD}, based on which we develop two empirical GRD models, *i.e.* robust axial-monotonic Clough-Tocher (RAMCT) and eigen generalized rate-distortion (EGRD), in next two sections, respectively.

In the first case, we are only interested in a smaller domain

$$\Omega_1 = \{(x,y)|x \in [x_{min}, x_{max}], y \in [y_{min}, y_{max}], \\ 0 < x_{min} < x_{max} < x_{MAX}, 1 < y_{min} < y_{max} < y_{MAX}\}.$$

Note that in this case, properties indicated by Eq. (3.3) and (3.5) are not applicable. To achieve better mathematical properties, we impose C^1 continuity on f as suggested, and thus replace Eq. (3.4) with a derivative condition. The function space is defined by

$$W_1 \coloneqq \left\{ f : \mathbb{R}^2 \to \mathbb{R} | f \in C^1(\Omega_1), f(x,y) \in [0,100], \frac{\partial f}{\partial x}(x,y) \geq 0, \forall (x,y) \in \Omega_1 \right\}. \tag{3.6}$$

In the second case, we consider a larger domain

$$\Omega_2 = \{(x,y)|x \in [0, x_{MAX}], y \in [y_{min}, y_{MAX}], 1 < y_{min} < y_{MAX}\}, \tag{3.7}$$

where we are able to determine x_{MAX}, y_{min} and y_{MAX} with a large-scale database. As a result, we can define a function space with more constraints as

$$W_2 \coloneqq \big\{ f : \mathbb{R}^2 \to \mathbb{R} | f \in C^0(\Omega_2), f(0,y) = 0, f(x_{MAX}, y_{MAX}) = 100, \\ f(x_a, y) \leq f(x_b, y), f(x_{MAX}, y_a) \leq f(x_{MAX}, y_b), \forall x_{MIN} \leq x_a < x_b \leq x_{MAX}, \\ y_{MIN} \leq y \leq y_{MAX}, y_{MIN} \leq y_a < y_b \leq y_{MAX}\} \tag{3.8}$$

3.2 Modeling GRD Function by Robust Axial-Monotonic Clough-Tocher Interpolation

In this section, we develop a Robust Axial-Monotonic Clough-Tocher (RAMCT) interpolation method to accurately estimate the GRD function with a moderate number of queries. The proposed model imposes the mathematical constraints on a localized interpolation model, and thus strikes the balance between flexibility and regularity. By re-parametrizing the piece-wise cubic Bézier function, we derive the desired affine-invariant C^1 continuity and the axial monotonicity constraints as well as the objective function in the Clough-Tocher (CT) framework, and model the interpolation problem as a quadratic programming problem.

3.2.1 Problem Formulation

We further assume that the quality measurement is precise for the following reasons. Because the HVS is the ultimate receiver in most applications, subjective evaluation is a straightforward and reliable approach to evaluate the quality of digital videos. Traditional subjective experiment protocol models a subject's perceived quality as a random variable, assuming the quality labeling process to be stochastic. Because subjective experiment is expensive and time consuming, it is hardly used in the GRD function approximation process. In practice, objective VQA methods that produce precise quality predictions are often employed to generate ground truth samples in the GRD function. Therefore, a GRD function should pass through the quality scores of objective VQA evaluated on the encoded video representations.

With this additional assumption, we formulate the GRD function approximation problem as

$$\begin{aligned} & \arg\min_{f} \mathcal{L}[f] \\ \text{s.t. } & f \in W_1 \\ & f(x_n, y_n) = z_n, n = 1, \cdots, N \end{aligned} \tag{3.9}$$

where W_1 is defined in Eq. (3.6), and where $\mathcal{L}$, N, x_n, y_n, and z_n represent a functional of f, the total number of training samples, bitrate, spatial resolution, and quality of the n-th training sample, respectively. Usually, the functional $\mathcal{L}$ measures desired mathematical properties, such as curvature, of f.

We deal with Eq. (3.9) as a multivariate interpolation problem, and develop a novel RAMCT interpolation algorithm to solve it in the subsequent subsections. Section 3.2.2 reviews the traditional CT method, from which the proposed model is inherited. The proposed C^1 continuity condition, axial monotonicity condition, objective function, and robust axial-monotonic CT algorithm are novel contributions that are detailed in Section 3.2.3, 3.2.4, 3.2.5, and 3.2.6, respectively.

3.2.2 Review of Clough-Tocher Method

Basic Settings and Notations of CT method

Since first introduced in 1960's [18], the CT method has been the most widely used multi-dimensional scattered data interpolation method, thanks to its C^1 continuity and low computational complexity [2, 4, 10]. Consider the scattered points (x_n, y_n) located in the x, y plane and their values z_n over the plane. The CT method basically looks for a piece-wise bivariate cubic function f that passes through all the points (x_n, y_n, z_n). The CT method first performs the Delaunay triangulation [23] in the x, y plane, dividing the whole plane into multiple non-overlapped triangles with the scatter points being the triangle vertices. Fig. 3.5 shows an example of triangulating a 2-dimensional input space with given scatter points. Then a piece-wise cubic function is employed as the interpolant for each triangle. Specifically, each triangle is further divided from its center point into three equivalent subtriangles, where a cubic function in the form of Bézier surface is estimated. Hereafter, we refer to the overall triangle as the macrotriangle and its subtriangles as microtriangles. For clarity and brevity, we also denote the macrotriangle edge that is opposite to the vertex $V_i, i = 0, 1, 2$ by E_i, and the internal microtriangle edge that connects V_i and S by $\hat{E}_i$. Let $\{i, j, k\}$ be a cyclic permutation of $\{0, 1, 2\}$ for the rest of this section.

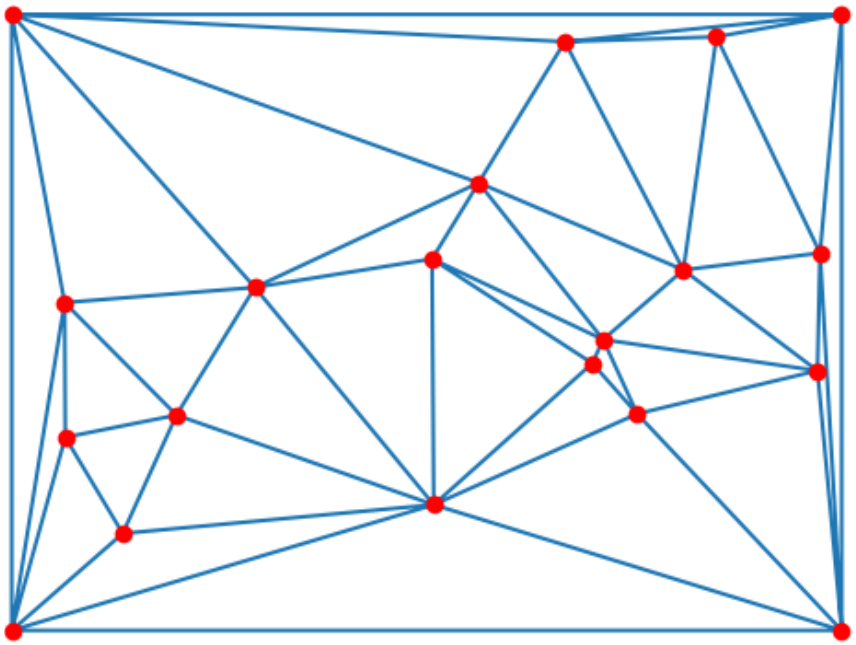


Figure 3.5: Result of Delaunay triangulation in a 2-dimensional input space. The red points are the given scatter samples for interpolation.

Introduction to Bézier Function

A cubic function in the Bézier form is completely determined by a so-called *control net* as demonstrated in Fig. 3.6 and 3.7. Take a microtriangle ΔV_0V_1S in Fig. 3.6 as an example. The control net over ΔV_0V_1S contains 10 *control points*, including 3 vertices V_0, V_1, S, 6 trisection points on 3 edges T_{01}, T_{10}, I_{01}, I_{02}, I_{11}, I_{12}, and the center C_2 of the microtriangle. The 10 control points form 6 *control net patches*, namely $\Delta_{V_0T_{01}I_{01}}$, $\Delta_{T_{01}T_{10}C_2}$, $\Delta_{T_{10}V_1I_{11}}$, $\Delta_{I_{01}C_2I_{02}}$, $\Delta_{C_2I_{11}I_{12}}$ and $\Delta_{I_{02}I_{12}S}$, whose positions and orientations are closely related to the properties of the resulting interpolant as will be seen soon. Associated with each control point V is a real number c_V, which is often termed as the *ordinate* of the control point in this context. The 10 ordinates also serve as the coefficients of the cubic Bézier function.

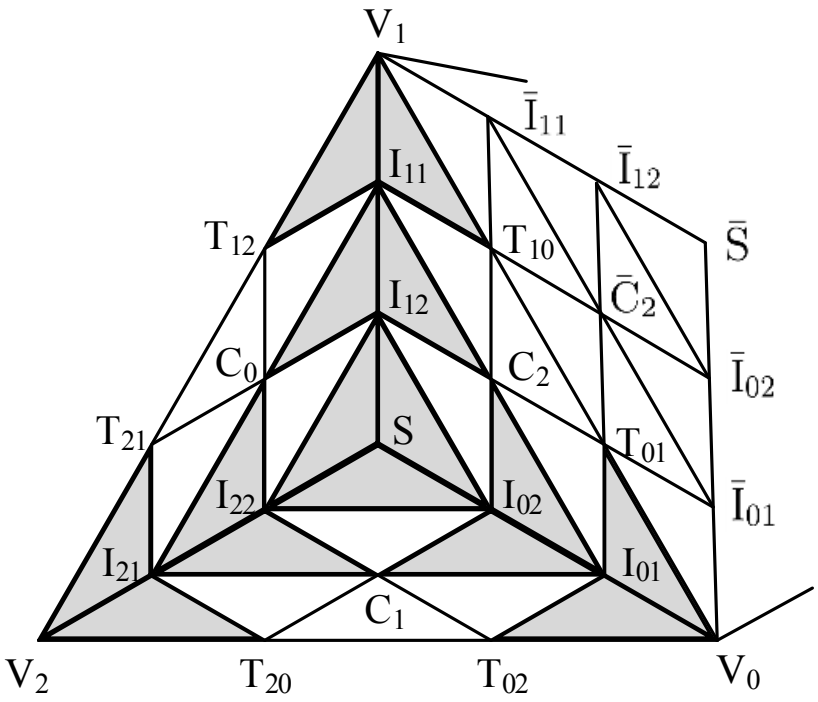


Figure 3.6: Top view of the CT split in one triangle of the triangulation, showing three divided microtriangles and its 19 Bézier control points.

Mathematically, we formulate the cubic Bézier surface in the microtriangle $\Delta_{V_iV_jS}$ as

$$\begin{aligned} f(\alpha,\beta,\gamma) =& c_{V_i}\alpha^3 + 3c_{T_{ij}}\alpha^2\beta + 3c_{I_{i1}}\alpha^2\gamma + c_{V_j}\beta^3 + \\ & 3c_{T_{ji}}\alpha\beta^2 + 3c_{I_{j1}}\beta^2\gamma + c_S\gamma^3 + 3c_{I_{i2}}\alpha\gamma^2 + \\ & 3c_{I_{j2}}\beta\gamma^2 + 6c_{C_k}\alpha\beta\gamma, \end{aligned} \tag{3.10}$$

where (α,β,γ) is the barycentric coordinates with regard to the three vertices of the microtriangle. A 3-dimensional visualization of the control net and its corresponding Bézier surface can be found in Fig. 3.7. The barycentric coordinates of a point $P=(x,y)$ with regard to $\Delta_{V_iV_jS}$ can be defined as

$$\alpha = \frac{A_{PV_jS}}{A_{V_iV_jS}}, \beta = \frac{A_{PSV_i}}{A_{V_jSV_i}}, \gamma = \frac{A_{PV_iV_j}}{A_{SV_iV_j}}, \tag{3.11}$$

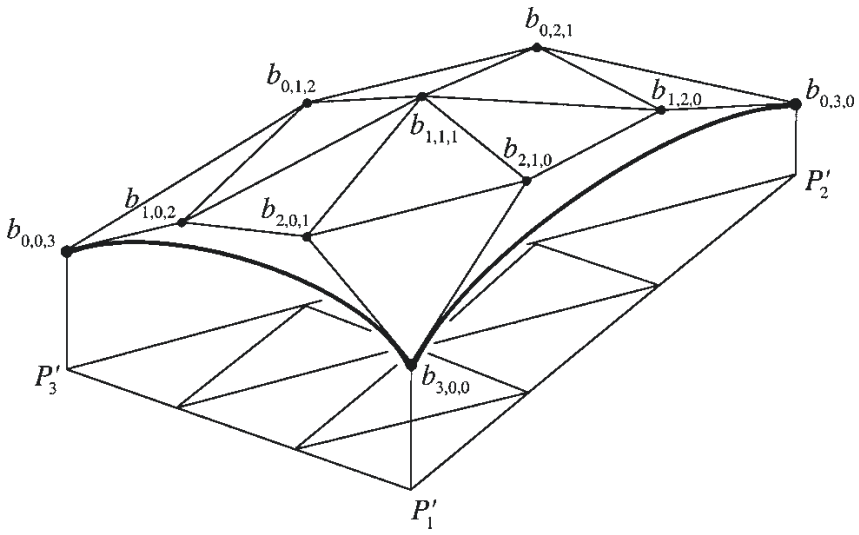


Figure 3.7: A bivariate cubic Bézier function over a triangle $P_1'P_2'P_3'$. The curved surface lies beneath its triangulated control net defined by the ordinates $b_{u,v,w}$, and is tangent with the control net at three vertices. Image by courtesy of Isaac Amidror [4].

where A_{UVW} means the directional area of the triangle formed by points U, V, W and is positive when U, V, W is counter-clockwise. Immediately, we have $\alpha + \beta + \gamma = 1$. Besides, the conversion from Cartesian coordinate to barycentric coordinate is obtained by representing the numerators in Eq. (3.11) with (x, y) [2], *i.e.*

$$\alpha = \frac{(x_{V_j} - x)(y_S - y) - (x_S - x)(y_{V_j} - y)}{A_{V_iV_jS}}, \tag{3.12a}$$

$$\beta = \frac{(x_S - x)(y_{V_i} - y) - (x_{V_i} - x)(y_S - y)}{A_{V_jSV_i}}, \tag{3.12b}$$

$$\gamma = \frac{(x_{V_i} - x)(y_{V_j} - y) - (x_{V_j} - x)(y_{V_i} - y)}{A_{SV_iV_j}}. \tag{3.12c}$$

The Bézier surface features a series of mathematical properties. Some of them are essential to both the CT interpolation framework, and the proposed RAMCT algorithm. We list these useful properties of the bicubic Bézier function, but omit their lengthy proofs here. Their usefulness will be elaborated later. Interested readers may refer to [26, 27, 35, 51].

Theorem 3.1. *Restricting a bivariate cubic Bézier function f, as defined in Eq. (3.11), to one of its triangle edge results in a univariate cubic Bézier polynomial, whose Bézier ordinates are the same as those of f associated with that edge.*

Corollary 3.1.1. *Assuming that two triangles have a shared edge, the Bézier functions defined on the two triangles are C^0 continuous across the edge if and only if they share their Bézier ordinates associated with this edge.*

Theorem 3.2. *The Bézier surface in Eq. (3.10) defined on a triangle is tangent with its control net at three triangle vertices V_i, V_j and S.*

Corollary 3.2.1. *Assuming that two triangles have the same vertex at V, the Bézier functions defined on the two triangles are C^1 continuous at V if and only if their respective control net patches associated with V are coplanar.*

Theorem 3.3. *Consider the Bézier function defined on $\Delta_{V_0V_1S}$. Let $\mathbf{l} = (l_0, l_1, l_2)^T$ denote any direction not parallel to the edge E_2. Here, l_0, l_1, and l_2 denote the barycentric representation of the direction. Then the directional derivative of the Bézier function is a univariate quadratic Bézier polynomial with Bézier ordinates*

$$3(l_0c_{V_0} + l_1c_{T_{01}} + l_2c_{I_{01}}),\ 3(l_0c_{T_{01}} + l_1c_{T_{10}} + l_2c_{C_2}),\ 3(l_0c_{T_{10}} + l_1c_{V_1} + l_2c_{I_{11}}).$$

Corollary 3.3.1. *Consider the two Bézier functions defined on $\Delta_{V_0V_1S}$ and $\Delta_{V_0V_1\bar{S}}$ sharing an edge V_0V_1 as shown in Fig. 3.6. They are C^1 continuous across V_0V_1 if and only if they are C^1 continuous at V_0 and V_1, and the two control net patches, $\Delta C_2T_{01}T_{10}$ and $\Delta\bar{C}_2T_{01}T_{10}$, are coplanar.*

Theorem 3.4. *A Bézier function is axial-monotonic when its corresponding control net is axial-monotonic.*

Classic CT Method

We note that 10 Bézier ordinates are required to define a Bézier surface on each micro-triangle. At the first glance, we need to determine 30 parameters for the interpolant in one macrotriangle with only 3 equality constraints given by $c_{V_0} = z_{V_0}$, $c_{V_1} = z_{V_1}$, and

$c_{V_2} = z_{V_2}$. Fortunately, the degree of freedom can be dramatically reduced with certain smoothness constraints. Under the C^0 assumption within macrotriangles, each two Bézier surfaces share the same Bézier ordinates at the their common boundaries V_0S, V_1S, and V_2S (Corollary 3.1.1), leaving 19 free parameters in the macrotriangle $\Delta_{V_0V_1V_2}$ as shown in Fig. (3.6). The inner-triangle C^1 continuity removes 7 additional degree of freedoms by enforcing the shaded neighboring control net patches in Fig. 3.6 to be coplanar (Corollary 3.2.1 and 3.3.1) [26]. To ensure inter-triangle C^1 continuity, a standard approach is to pick a directional derivative $d^e_{E_i}$ not parallel to the shared triangle edge, and assume the derivative to be linear on the edge, which further reduces the degree of freedom to 9 (Theorem 3.3). Taking into account the three known values at V_0, V_1, and V_2, we eventually need 6 variables to parametrize each macrotriangle. The classic CT method typically considers the 6 partial derivatives of f at the 3 triangle vertices as the variables. Although the gradients are not always available in practice, in most cases they can be estimated by considering the known values not only in the vertices of the triangle in question, but also in its neighbors. The most commonly used method is to estimate the gradients by minimizing the second-order derivatives along all Bézier curves [51]. Readers who are interested in the details of the CT method may refer to [51, 27, 2, 4].

The original CT method suffers from at least three limitations in approximating GRD functions. First, it picks the normal of the edges as the direction of cross-boundary derivative $d^e_{E_i}$. However, this choice gives an interpolant that is not invariant under affine transforms. This has some undesirable consequences: for a very narrow triangle, the spline can develop huge oscillations [27]. Second, the interpolant composite of piece-wise Bézier polynomials is not axial-monotonic, even when the given points are axial monotonic. Third, the CT algorithm achieves the external smoothness by estimating the gradients at three vertices $V_i, i = 0, 1, 2$, and by assuming the normal derivative at the triangle boundary E_i to be linear. The linear assumption is somewhat arbitrary and may violate monotonicity we want to achieve. We will address the three limitations in the proposed RAMCT model as described in the rest of this section.

3.2.3 Parametrization and Affine-Invariant C1 Continuity

In this subsection, we propose a novel parametrization for the piece-wise bicubic Bézier function and an associated affine-invariant C^1 constraint. In the traditional CT method [51], the C^1 continuity across macrotriangle boundaries is guaranteed by a linear assumption on the normal derivatives. We improve this arbitrary assumption from two aspects. First, instead of the normal derivative of the interpolant f at the triangle boundary E_i, we consider the directional derivative $d^e_{E_i}$ of the control net patch $\Delta C_i T_{jk} T_{kj}$ along the direction of the vector $\overrightarrow{\bar{C}_i C_i}$. Since this derivative transforms similarly as the gradient under affine transforms, the resulting interpolant is affine-invariant [27]. Second, we lift the unwanted linear constraints on the cross-boundary derivatives, elevating the number of parameters in a macrotriangle back to 9. The increased freedom allows the interpolant to be monotonic if monotone data are given.

Now we can re-parametrize the interpolant f on $\Delta V_0 V_1 V_2$ with the 3 additional variables $d^e_{E_i}, i = 1, 2, 3$ and 6 partial derivatives of f at 3 macrotriangle vertices. Denote the x- and y-partial derivatives at V_i by $d^x_{V_i}$ and $d^y_{V_i}$, respectively. According to Corollary 3.2.1, the Bézier ordinates adjacent to the three macrotriangle vertices V_0, V_1, and V_2 are immediately available by

$$c_{T_{ij}} = (x_{T_{ij}} - x_{V_i}) d^x_{V_i} + (y_{T_{ij}} - y_{V_i}) d^y_{V_i} + z_{V_i} \tag{3.13a}$$

$$c_{T_{ik}} = (x_{T_{ik}} - x_{V_i}) d^x_{V_i} + (y_{T_{ik}} - y_{V_i}) d^y_{V_i} + z_{V_i} \tag{3.13b}$$

$$c_{I_{i1}} = (x_{I_{i1}} - x_{V_i}) d^x_{V_i} + (y_{I_{i1}} - y_{V_i}) d^y_{V_i} + z_{V_i} \tag{3.13c}$$

where $i = 0, 1, 2$. Then we can represent c_{C_i} in terms of $c_{T_{jk}}$, $c_{T_{kj}}$ and $d^e_{E_i}$ as

$$\begin{aligned} c_{C_i} =& \theta_{kj} c_{T_{jk}} + \theta_{jk} c_{T_{kj}} + \eta_i d^e_{E_i}, \\ =& (x_{T_{kj}} - x^*_i) d^x_{V_j} + (y_{T_{kj}} - y^*_i) d^y_{V_j} + (x_{T_{jk}} - x^*_i) d^x_{V_k} \\ &+ (y_{T_{jk}} - y^*_i) d^y_{V_k} + \eta_i d^e_{E_i} + \theta_{kj} z_{V_j} + \theta_{jk} z_{V_k}, \end{aligned} \tag{3.14}$$

where

$$
\begin{aligned}
x_i^* =& \frac{(x_{\bar{C}_i} - x_{C_i})(x_{V_j}y_{V_k} - x_{V_k}y_{V_j}) - (x_{V_k} - x_{V_j})(x_{C_i}y_{\bar{C}_i} - x_{\bar{C}_i}y_{C_i})}{(x_{\bar{C}_i} - x_{C_i})(y_{V_k} - y_{V_j}) - (y_{\bar{C}_i} - y_{C_i})(x_{V_k} - x_{V_j})},\\
y_i^* =& \frac{(y_{\bar{C}_i} - y_{C_i})(x_{V_j}y_{V_k} - x_{V_k}y_{V_j}) - (y_{V_k} - y_{V_j})(x_{C_i}y_{\bar{C}_i} - x_{\bar{C}_i}y_{C_i})}{(x_{\bar{C}_i} - x_{C_i})(y_{V_k} - y_{V_j}) - (y_{\bar{C}_i} - y_{C_i})(x_{V_k} - x_{V_j})},\\
\theta_{kj} =& \frac{x_{T_{kj}} - x_i^*}{x_{T_{kj}} - x_{T_{jk}}} = \frac{y_{T_{kj}} - y_i^*}{y_{T_{kj}} - y_{T_{jk}}},\\
\theta_{jk} =& \frac{x_{T_{jk}} - x_i^*}{x_{T_{jk}} - x_{T_{kj}}} = \frac{y_{T_{jk}} - y_i^*}{y_{T_{jk}} - y_{T_{kj}}},\\
\eta_i =& \sqrt{(x_{C_i} - x_i^*)^2 + (y_{C_i} - y_i^*)^2}.
\end{aligned}
$$

Note that (x_i^*, y_i^*) is actually the intersection of the boundary E_i and the segment $C_i\bar{C}_i$. Finally, the last 4 Bézier ordinates, c_S and $c_{I_{i2}}, i = 0, 1, 2$ are expressed by

$$
\begin{aligned}
c_{I_{i2}} =& \frac{1}{3}(c_{I_{i1}} + c_{C_j} + c_{C_k})\\
=& \frac{1}{3}[(x_{I_{i1}} - x_{V_i}) + (x_{T_{ki}} - x_j^*) + (x_{T_{ji}} - x_k^*)]d_{V_i}^x\\
&+ \frac{1}{3}[(y_{I_{i1}} - y_{V_i}) + (y_{T_{ki}} - y_j^*) + (y_{T_{ji}} - y_k^*)]d_{V_i}^y\\
&+ \frac{1}{3}(x_{T_{ij}} - x_k^*)d_{V_j}^x + \frac{1}{3}(y_{T_{ij}} - y_k^*)d_{V_j}^y + \frac{1}{3}\eta_k d_{E_k}^e\\
&+ \frac{1}{3}(x_{T_{ik}} - x_j^*)d_{V_k}^x + \frac{1}{3}(y_{T_{ik}} - y_j^*)d_{V_k}^y + \frac{1}{3}\eta_j d_{E_j}^e\\
&+ \frac{1}{3}[z_{V_i} + (\theta_{ki}z_{V_i} + \theta_{ik}z_{V_k}) + (\theta_{ij}z_{V_j} + \theta_{ji}z_{V_i})]
\end{aligned} \tag{3.15}
$$

$$
\begin{aligned}
c_S =& \frac{1}{3}\sum_{i=0}^{2} c_{I_{i2}}\\
=& \frac{1}{9}\sum_{i=0}^{2}[(x_{I_{i1}} - x_{V_i}) + 2(x_{T_{ki}} - x_j^*) + 2(x_{T_{ji}} - x_k^*)]d_{V_i}^x\\
&+ \frac{1}{9}\sum_{i=0}^{2}[(y_{I_{i1}} - y_{V_i}) + 2(y_{T_{ki}} - y_j^*) + 2(y_{T_{ji}} - y_k^*)]d_{V_i}^y\\
&+ \frac{2}{9}\sum_{i=0}^{2}\eta_i d_{E_i}^e + \frac{1}{9}\sum_{i=0}^{2}[(1 + 2\theta_{ji} + 2\theta_{ki})z_{V_i}],
\end{aligned} \tag{3.16}
$$

thanks to the C^1 continuity of f and Corollary 3.3.1.

In summary, we re-parametrize the interpolant f over $\Delta V_0V_1V_2$ by Eq. (3.13), (3.14), (3.15), and (3.16). For simplicity, these equations can be factorized into the matrix form

$$\mathbf{c} = \mathbf{R}\mathbf{d} + \mathbf{f}, \tag{3.17}$$

where $\mathbf{c} \in \mathbb{R}^{16\times 1}$, $\mathbf{R} \in \mathbb{R}^{16\times 9}$, $\mathbf{d} \in \mathbb{R}^{9\times 1}$, $\mathbf{f} \in \mathbb{R}^{16\times 1}$, $\mathbf{c}$ and $\mathbf{d}$ represent the ordinates of control net and unknown derivatives, respectively. Therefore, finding the interpolant of the macrotriangle corresponds to determining the 9 unknown parameters in $\mathbf{d}$. Details of these matrices can be found in Appendix A.1.

Besides the inner macrotriangle constraints, we also want to keep $d^e_{E_i}$ consistent across the triangle boundary to ensure external C^1 smoothness. As a result, the following equality constraints need to be added for each edge with adjacent triangles

$$d^e_{E_i} + d^e_{\bar{E}_i} = 0, \tag{3.18}$$

where $d^e_{\bar{E}_i}$ denotes the directional derivative of the control net patch $\Delta\bar{C}_iT_{jk}T_{kj}$, which belongs to the neighboring macrotriangle, and is along the opposite direction of $d^e_{E_i}$.

Combining (3.17) and (3.18), we guarantee that the resulting interpolant is C^1 continuous and affine-invariant.

3.2.4 Axial Monotonicity

This subsection aims to derive the sufficient constraints on $\mathbf{d}$ for the Bézier surface in the macrotriangle $\Delta_{V_0V_1V_2}$ to be axial-monotonic. In general, the interpolant composite of piece-wise Bézier polynomials is not monotonic even though the sampled points are monotonic. Several works have been done to derive sufficient conditions for a univariate or bivariate Bézier function [31, 35]. We adopt the sufficient condition proposed in [35], which is also summarized in Theorem 3.4. By combining the sufficient conditions in all three microtriangles and the inner triangle continuity condition as indicated in Corollary 3.3.1, we claim the following corollary regarding the sufficient condition for f to be axial monotonic.

Corollary 3.4.1. *f is a piece-wise cubic Bézier function defined on $\Delta V_0V_1V_2$, which is split to three microtriangles by its center S as shown in Fig. 3.6. The definition of f on each microtriangle is given in Eq. (3.10). One sufficient condition for f to be axial-monotonic is that all the coplanar patches of the control net of f, i.e. $\Delta V_iT_{ij}T_{ik}, \Delta I_{i1}C_kC_j, \Delta T_{ij}T_{ji}C_k$, $\Delta I_{02}I_{12}I_{22}, i = 0, 1, 2$, are axial-monotonic.*

Mathematically, Corollary 3.4.1 can be formulated by the following inequalities:

$$(y_{V_i} - y_{V_k})c_{T_{ij}} + (y_{V_j} - y_{V_i})c_{T_{ik}} \leq (y_{V_j} - y_{V_k})z_{V_i} \tag{3.19a}$$

$$(y_{V_k} - y_{V_j})c_{I_{i1}} + (y_{V_i} - y_{V_k})c_{C_k} + (y_{V_j} - y_{V_i})c_{C_j} \leq 0 \tag{3.19b}$$

$$(y_{V_2} - y_{V_1})c_{I_{02}} + (y_{V_0} - y_{V_2})c_{I_{12}} + (y_{V_1} - y_{V_0})c_{I_{22}} \leq 0 \tag{3.19c}$$

$$(y_S - y_{V_j})c_{T_{ij}} + (y_{V_i} - y_S)c_{T_{ji}} + (y_{V_j} - y_{V_i})c_{C_k} \leq 0. \tag{3.19d}$$

We can summarize the monotonicity constraint in matrix form

$$\mathbf{Gc} \leq \mathbf{h}, \tag{3.20}$$

where $\mathbf{G} \in \mathbb{R}^{10\times16}$ and $\mathbf{f} \in \mathbb{R}^{10\times1}$. Further substituting (3.17) into (3.20), we obtain the monotonicity constraint in terms of $\mathbf{d}$

$$\mathbf{GRd} \leq \mathbf{h} - \mathbf{Gf}. \tag{3.21}$$

More details on how we construct $\mathbf{G}$ and $\mathbf{h}$ are given in Appendix A.2.

3.2.5 Objective Function

To determine the unknown derivatives, we propose to minimize the total curvature of the interpolated surface under the smoothness assumption. Directly computing the total curvature is computationally intractable. Alternatively, we minimize the curvature of Bézier curves at the edges of each microtriangle as an approximation. Specifically, in $\Delta_{V_0V_1V_2}$, the objective function is written as

$$\mathcal{L}[f]_{V_0V_1V_2} = \frac{1}{2}\sum_{i=0}^{2}\int_{E_i}\left[\frac{\partial^2 f}{\partial E_i^2}\right]^2 ds_{E_i} + \sum_{i=0}^{2}\int_{\hat{E}_i}\left[\frac{\partial^2 f}{\partial \hat{E}_i^2}\right]^2 ds_{\hat{E}_i}, \tag{3.22}$$

where the weight $\frac{1}{2}$ is introduced to cancel the double counting of the external edges, and ds_{E_i} and $ds_{\hat{E}_i}$ represent the element of arc length on the restrictions of f to E_i and $\hat{E}_i$, respectively.

Consider an external boundary E_i, whose Bézier control net coefficients are z_{V_j}, $c_{T_{jk}}$, $c_{T_{kj}}$, and z_{V_k}. The integral of the second order derivative of the Bézier curve on E_i can be represented in terms of the four coefficients as

$$
\begin{aligned}
&\int_{E_i}\left[\frac{\partial^2 f}{\partial E_i^2}\right]^2 ds_{E_i} = \frac{1}{\|E_i\|^3}\int_0^1 \left[f''_{E_i}(t)\right]^2 dt \\
=&\frac{18}{\|E_i\|^3}(2c_{T_jk}^2 + 2c_{T_kj}^2 - 2c_{T_jk}c_{T_kj}) + \\
&\frac{-36}{\|E_i\|^3}(z_{V_j}c_{T_jk} + z_{V_k}c_{T_kj}) + \frac{12}{\|E_i\|^3}(z_{V_j}^2 + z_{V_k}^2 + z_{V_j}z_{V_k}) \\
=&\begin{bmatrix} c_{T_{jk}} & c_{T_{kj}} \end{bmatrix}\begin{bmatrix} \frac{36}{\|E_i\|^3} & \frac{-18}{\|E_i\|^3} \\ \frac{-18}{\|E_i\|^3} & \frac{36}{\|E_i\|^3} \end{bmatrix}\begin{bmatrix} c_{T_{jk}} \\ c_{T_{kj}} \end{bmatrix} + \\
&\begin{bmatrix} \frac{-36z_{V_j}}{\|E_i\|^3} & \frac{-36z_{V_k}}{\|E_i\|^3} \end{bmatrix}\begin{bmatrix} c_{T_{jk}} \\ c_{T_{kj}} \end{bmatrix} + \\
&\frac{12}{\|E_i\|^3}(z_{V_j}^2 + z_{V_k}^2 + z_{V_j}z_{V_k}),
\end{aligned} \tag{3.23}
$$

where

$$\|E_i\| = \sqrt{(x_{V_j} - x_{V_k})^2 + (y_{V_j} - y_{V_k})^2}$$

is the length of E_i.

Similarly, we get the other part of the objective function from an internal boundary $\hat{E}_i$, whose coefficients are $z_{V_i}, c_{I_{i1}}, c_{I_{i2}}$, and c_S.

$$
\begin{aligned}
&\int_{\hat{E}_i}\left[\frac{\partial^2 f}{\partial \hat{E}_i^2}\right]^2 ds_{\hat{E}_i} = \frac{1}{\|\hat{E}_i\|^3}\int_0^1 \left[f''_{\hat{E}_i}(t)\right]^2 dt \\
=&\frac{6}{\|\hat{E}_i\|^3}(6c_{I_{i1}}^2 + 6c_{I_{i2}}^2 + 2c_S^2 - 6c_{I_{i1}}c_{I_{i2}} - 6c_{I_{i2}}c_S) + \\
&\frac{12z_{V_i}}{\|\hat{E}_i\|^3}(-3c_{I_{i1}} + c_S) + \frac{12}{\|\hat{E}_i\|^3}z_{V_i}^2 \\
=&\begin{bmatrix} c_{I_{i1}} & c_{I_{i2}} & c_S \end{bmatrix}\begin{bmatrix} \frac{36}{\|\hat{E}_i\|^3} & \frac{-18}{\|\hat{E}_i\|^3} & 0 \\ \frac{-18}{\|\hat{E}_i\|^3} & \frac{36}{\|\hat{E}_i\|^3} & \frac{-18}{\|\hat{E}_i\|^3} \\ 0 & \frac{-18}{\|\hat{E}_i\|^3} & \frac{12}{\|\hat{E}_i\|^3} \end{bmatrix}\begin{bmatrix} c_{I_{i1}} \\ c_{I_{i2}} \\ c_S \end{bmatrix} \\
&+\begin{bmatrix} \frac{-36z_{V_i}}{\|\hat{E}_i\|^3} & 0 & \frac{12z_{V_i}}{\|\hat{E}_i\|^3} \end{bmatrix}\begin{bmatrix} c_{I_{i1}} \\ c_{I_{i2}} \\ c_S \end{bmatrix} + \frac{12z_{V_i}^2}{\left\|\hat{E}_i\right\|^3},
\end{aligned} \tag{3.24}
$$

where

$$
\|\hat{E}_i\| = \sqrt{(x_S - x_{V_i})^2 + (y_S - y_{V_i})^2}
$$

is the length of $\hat{E}_i$.

Substituting (3.23) and (3.24) into (3.22), we obtain the loss function for $\Delta_{V_0V_1V_2}$ in matrix form

$$
\mathcal{L}[f]_{V_0V_1V_2} = \mathbf{c}^T\mathbf{U}_{V_0V_1V_2}\mathbf{c} + \mathbf{w}_{V_0V_1V_2}^T\mathbf{c} + const, \tag{3.25}
$$

where $\mathbf{U}_{V_0V_1V_2} \in \mathbb{R}^{16\times 16}$ and $\mathbf{w}_{V_0V_1V_2} \in \mathbb{R}^{16\times 1}$. How to compute each entry of $\mathbf{U}_{V_0V_1V_2}$ and $\mathbf{w}_{V_0V_1V_2}$ is detailed in Appendix A.3.

Further substituting $\mathbf{c} = \mathbf{R}\mathbf{d} + \mathbf{f}$ into (3.25), we represent the local loss in terms of $\mathbf{d}$

$$
\begin{aligned}
\mathcal{L}[f]_{V_0V_1V_2} =&(\mathbf{R}\mathbf{d} + \mathbf{f})^T\mathbf{U}_{V_0V_1V_2}(\mathbf{R}\mathbf{d} + \mathbf{f}) + \mathbf{w}_{V_0V_1V_2}^T(\mathbf{R}\mathbf{d} + \mathbf{f}) + const \\
=&\mathbf{d}^T(\mathbf{R}^T\mathbf{U}_{V_0V_1V_2}\mathbf{R})\mathbf{d} + (\mathbf{f}^T\mathbf{U}_{V_0V_1V_2} + \mathbf{w}_{V_0V_1V_2}^T)\mathbf{R}\mathbf{d} + const.
\end{aligned} \tag{3.26}
$$

3.2.6 Robust Axial-Monotonic Clough-Tocher Method

Here we propose our Robust Axial-Monotonic Clough-Tocher (RAMCT) method. The inequality constraints in (3.20) are sufficient conditions for x-axial monotonicity. However, the sufficient conditions cannot be satisfied in some extreme cases, making the primary optimization problem infeasible. To relax these constraints, we introduce hinge loss to some of these inequalities, motivated by the success of the support vector machine employing hinge loss to deal with the inseparable case [19]. Specifically, the modified inequality constraints are formulated as

$$(y_{V_i} - y_{V_k})c_{T_{ij}} + (y_{V_j} - y_{V_i})c_{T_{ik}} \leq (y_{V_j} - y_{V_k})z_{V_i} \tag{3.27a}$$

$$(y_{V_k} - y_{V_j})c_{I_{i1}} + (y_{V_i} - y_{V_k})c_{C_k} + (y_{V_j} - y_{V_i})c_{C_j} + \xi_{i1} \leq 0 \tag{3.27b}$$

$$(y_{V_2} - y_{V_1})c_{I_{02}} + (y_{V_0} - y_{V_2})c_{I_{12}} + (y_{V_1} - y_{V_0})c_{I_{22}} \leq 0 \tag{3.27c}$$

$$(y_S - y_{V_j})c_{T_{ij}} + (y_{V_i} - y_S)c_{T_{ji}} + (y_{V_j} - y_{V_i})c_{C_k} + \xi_{C_k} \leq 0, \tag{3.27d}$$

where $\boldsymbol{\xi} \in \mathbb{R}^{6\times 1}$ and $\boldsymbol{\xi} \leq \mathbf{0}$. Note that (3.27a),(3.27c) are identical to (3.19a),(3.19c) because they are also necessary conditions of axial monotonicity (See Appendix A.4 for proof). By rewriting these constraints in the matrix form, we obtain

$$\begin{bmatrix} \mathbf{G} & \mathbf{J}_1 \\ \mathbf{O} & \mathbf{J_2} \end{bmatrix} \begin{bmatrix} \mathbf{c} \\ \boldsymbol{\xi} \end{bmatrix} \leq \begin{bmatrix} \mathbf{h} \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{G}$ and $\mathbf{h}$ are the same as in (3.20),(3.21). $\mathbf{J}_2$ is a 6×6 identity matrix, while $\mathbf{J}_1 \in \mathbb{R}^{10\times 6}$ is obtained by padding $\mathbf{J}_2$ with 3 rows of zeros to its top and inserting a row of zeros between the 3rd and 4th rows of $\mathbf{J}_2$.

By substituting (3.17) into the inequality above, we finally obtain the inequality constraints in terms of the unknowns $\mathbf{d}$ and the auxiliary variables $\boldsymbol{\xi}$ as

$$\begin{bmatrix} \mathbf{GR} & \mathbf{J}_1 \\ \mathbf{O} & \mathbf{J}_2 \end{bmatrix} \tilde{\mathbf{d}} \leq \begin{bmatrix} \mathbf{h} - \mathbf{Gf} \\ \mathbf{0} \end{bmatrix}, \tag{3.28}$$

where

$$\tilde{\mathbf{d}} \coloneqq \begin{bmatrix} \mathbf{d} \\ \boldsymbol{\xi} \end{bmatrix}.$$

The objective function is then modified accordingly,

$$\mathcal{L}[f]_{V_0V_1V_2} = \mathbf{c}^T\mathbf{U}_{V_0V_1V_2}\mathbf{c} + \mathbf{w}_{V_0V_1V_2}^T\mathbf{c} - \boldsymbol{\lambda}^T\boldsymbol{\xi} + const, \tag{3.29}$$

where $\boldsymbol{\lambda} = [\lambda, \lambda, \cdots, \lambda]^T$ is the weighting parameter. Substituting $\mathbf{c} = \mathbf{R}\mathbf{d} + \mathbf{f}$ into (3.29), we get

$$\begin{aligned}
\mathcal{L}[f]_{V_0V_1V_2} =& \mathbf{d}^T(\mathbf{R}^T\mathbf{U}_{V_0V_1V_2}\mathbf{R})\mathbf{d} + (\mathbf{f}^T\mathbf{U}_{V_0V_1V_2} + \mathbf{w}_{V_0V_1V_2}^T)\mathbf{R}\mathbf{d} \\
& - \boldsymbol{\lambda}^T\boldsymbol{\xi} + const \\
=& \begin{bmatrix} \mathbf{d}^T & \boldsymbol{\xi}^T \end{bmatrix} \begin{bmatrix} \mathbf{R^T}\mathbf{U}_{V_0V_1V_2}\mathbf{R} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \boldsymbol{\xi} \end{bmatrix} + \\
& \begin{bmatrix} (\mathbf{f}^T\mathbf{U}_{V_0V_1V_2} + \mathbf{w}_{V_0V_1V_2}^T)\mathbf{R} & -\boldsymbol{\lambda}^T \end{bmatrix} \begin{bmatrix} \mathbf{d} \\ \boldsymbol{\xi} \end{bmatrix} + const \\
=& \tilde{\mathbf{d}} \begin{bmatrix} \mathbf{R^T}\mathbf{U}_{V_0V_1V_2}\mathbf{R} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \tilde{\mathbf{d}} + \\
& \begin{bmatrix} (\mathbf{f}^T\mathbf{U}_{V_0V_1V_2} + \mathbf{w}_{V_0V_1V_2}^T)\mathbf{R} & -\boldsymbol{\lambda}^T \end{bmatrix} \tilde{\mathbf{d}} + const,
\end{aligned}$$

which becomes a positive semidefinite quadratic function of the augmented variable vector $\tilde{\mathbf{d}}$.

To achieve a globally optimal solution, we need to bring together the local constraints and the objective function defined by Eq. (3.18), (3.28) and (3.26). Denote by $\boldsymbol{\Delta}$ and $\boldsymbol{\mathcal{E}}$ the set of all macrotriangles and the set of shared macrotriangle boundaries obtained after performing triangulation [23] on the given N scattered samples, respectively. In one macrotriangle, we have in total 15 variables, including 9 derivatives and 6 auxiliary variables. Among the 9 derivatives, 6 partial derivatives are associated with 3 macrotriangle vertices, *i.e.* the given samples, while the other 3 directional derivatives are associated with the 3 triangle edges. In total, there are $2N+3|\boldsymbol{\Delta}|$ derivatives plus $6|\boldsymbol{\Delta}|$ auxiliary variables, amounting to $2N + 9|\boldsymbol{\Delta}|$ unknowns to solve. Denote the vector of all unknowns by $\hat{\mathbf{d}}$. Finding the axial-monotonic interpolant corresponds to solving the following optimization

problem

$$\begin{aligned} \arg\min_{\hat{\mathbf{d}}} \quad & \sum_{\Delta\in\boldsymbol{\Delta}} \mathcal{L}[f]_\Delta \\ \text{subject to} \quad & \begin{bmatrix} \mathbf{G}_\Delta\mathbf{R} & \mathbf{J}_1 \\ \mathbf{O} & \mathbf{J}_2 \end{bmatrix} \hat{\mathbf{d}}_\Delta \le \begin{bmatrix} \mathbf{h}_\Delta - \mathbf{G}_\Delta\mathbf{f} \\ \mathbf{0} \end{bmatrix}, \quad \forall\Delta\in\boldsymbol{\Delta} \\ & \begin{bmatrix} 1 & 1 \end{bmatrix} \hat{\mathbf{d}}_E = 0, \quad \forall E\in\boldsymbol{\mathcal{E}}, \end{aligned} \tag{3.30}$$

where $\hat{\mathbf{d}}_\Delta$ and $\hat{\mathbf{d}}_E$ indicate the unknowns (including two types of derivative and the auxiliary variables) associated with a given macrotriangle Δ and the two opposite directional derivatives of a given edge E, respectively. The exact entry values of $\mathbf{G}_\Delta$ and $\mathbf{h}_\Delta$ depend on the shape and orientation of Δ. Note that the constraints are linear with respect to $\hat{\mathbf{d}}$, and the summation of $\mathcal{L}[f]$ is still a positive semidefinite quadratic function. Thus, finding $\hat{\mathbf{d}}$ turns into a standard quadratic programming problem, which can be solved by existing convex programming softwares. In this work, we adopt the operator splitting quadratic program (OSQP) package [67] as the solver due to its accuracy and efficiency. Specifically, the OSQP software can solve a quadratic programming method of $1,000$ variables, which corresponds to the RAMCT problem Eq. (3.30) with more than 50 samples, in less than 0.1 second. With all the derivatives solved, we insert them back to Eq. (3.17) to get the coefficients of Bézier function, and thus achieve the optimal interpolant f satisfying Eq. (3.9).

3.3 Modeling GRD Function by Dimension Reduction

In this section, we are focused on the theoretical space W_2 of GRD function, and reveal that the GRD function estimation problem can be modeled as a projection onto convex sets (POCS). Further analysis on the real-world data from a large-scale GRD function database allows us to approximate the theoretical space by a low-dimensional one. By reducing the dimension of the GRD function space, we propose an empirical GRD model which can reconstruct the GRD surface with few attribute-quality pairs.

3.3.1 Problem Formulation

In this subsection, we propose novel formulations for the GRD function estimation problem, following which an empirical GRD model arises naturally. We find that the parameter number of the proposed model depends on the dimension of the GRD function space, and thus a corresponding dimension reduction problem is proposed.

GRD Function in a Hyperplane and a Cone

We begin by analyzing the structure of the theoretical space of GRD function. The theoretical function space W_2 as defined in Eq. (3.8) actually arises from the intersection of two common function spaces.

The equality constraints in W_2 jointly form a hyperplane

$$H_1 \coloneqq \left\{ f : \mathbb{R}^2 \to \mathbb{R} | f \in C^0(\Omega_2), f(0, y) = 0, f(x_{MAX}, y_{MAX}) = 100 \right\},$$

which can be described as the Minkowski sum of a linear function subspace

$$H_0 \coloneqq \left\{ f : \mathbb{R}^2 \to \mathbb{R} | f \in C^0(\Omega_2), f(0, y) = 0, f(x_{MAX}, y_{MAX}) = 0 \right\}$$

and any function $f_0 \in H_1$. Formally,

$$H_1 = f_0 + H_0, \quad \forall f_0 \in H_1. \tag{3.31}$$

The inequality constraints form a closed convex cone

$$\begin{aligned} V \coloneqq \{ f : \mathbb{R}^2 \to \mathbb{R} | f(x_a, y) \leq f(x_b, y), f(x_{MAX}, y_a) \leq f(x_{MAX}, y_b), \\ \forall 0 \leq x_a < x_b \leq x_{MAX}, y_{min} \leq y \leq y_{MAX}, y_{min} \leq y_a < y_b \leq y_{MAX} \} \end{aligned} \tag{3.32}$$

as it is readily shown that $\forall \alpha, \beta > 0$ and $v_0, v_1 \in V$, $\alpha v_0 + \beta v_1 \in V$.

Finally, we conclude that the theoretical space W_2 can be described as the intersection of the hyperplane H_1 and the convex cone V as shown in Fig. 3.8:

$$W_2 = H_1 \cap V. \tag{3.33}$$

Thus W_2 is convex thanks to the convexity of H_1 and V.

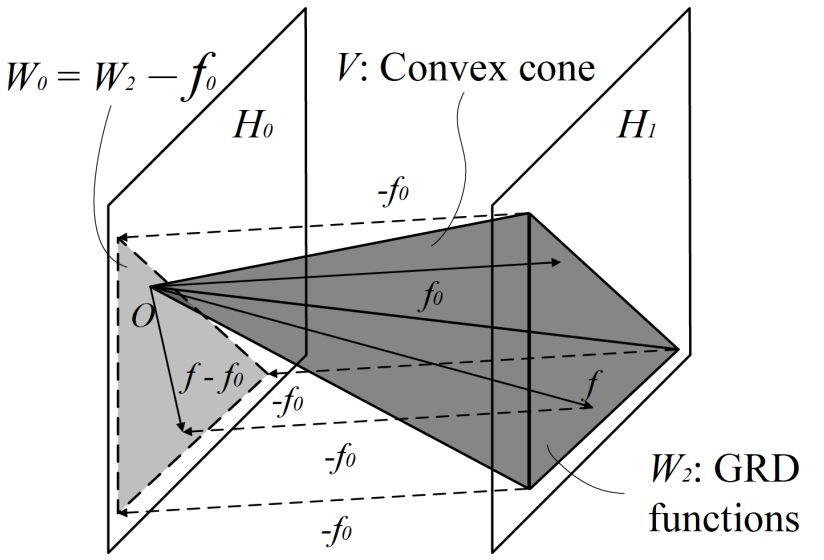


Figure 3.8: Visualization of theoretical GRD space. Monotonic functions form the darkly shaded polygonal cone V, while the theoretical GRD space $W_2 = H_1 \cap V$ is represented by the lightly shaded triangle in the hyperplane H_1.

GRD Function in a Hilbert Space

Compared to the original GRD function space W_2, it is easier to study another set

$$W_0 \coloneqq W_2 - f_0, \tag{3.34}$$

where f_0 can be any fixed function residing in W_2. According to Eq. (3.31), (3.33) and (3.34), W_0 is a convex subset of H_0 as illustrated in Fig. 3.8.

As we have defined in Eq. (3.7), Ω_2 is a compact subset in $\mathbb{R}^2$, so $\forall h \in H_0$, the image $h(\Omega_2)$ is a compact subset in $\mathbb{R}$. Without loss of generality, assuming that $h(\Omega_2)$ is absolutely bounded by M, or formally, $|h(x,y)| \leq M(h) \in \mathbb{R}, \forall (x,y) \in \Omega_2$, we show that $h \in H_0$ is square-integrable:

$$\iint_{\Omega_2} |h(x,y)|^2 \, dxdy \leq M(h)^2 |\Omega_2| < \infty,$$

where $|\Omega_2|$ means the area of Ω_2. Therefore, it is possible to equip the space H_0 with an

inner product

$$\langle h, g\rangle \coloneqq \iint_{\Omega_2} h(x,y)g(x,y)dxdy.$$

and denote by d^2 the induced metric. As a result, H_0 is a subset of a Hilbert space

$$L^2(\Omega_2) \coloneqq \left\{ h \middle| \iint_{\Omega_2} |h(x,y)|^2 \, dxdy < \infty \right\}.$$

According to the theorems of generalized Fourier series, there exists a *maximal* orthonormal basis $\{h_n \in L^2(\Omega_2), n = 1, 2, 3, \cdots\}$ that spans the Hilbert space $L^2(\Omega_2)$. Here, *maximal* means that $\langle h, h_n\rangle = 0, \forall n$ implies $h = 0$ in the L^2 sense. Formally,

$$h \sim \sum_{n=1}^{\infty} c_n h_n, \forall h \in L^2(\Omega_2), \tag{3.35}$$

where $\sim$ denotes the equality in the L^2 sense, and $c_n = \langle h, h_n\rangle \in \mathbb{R}$. Without loss of generality, we may restrict all the basis $\{h_n\}$ belonging to H_0, *i.e.*

$$h_n(0,y) = h_n(x_{MAX}, y_{MAX}) = 0, \forall n, y_{min} \leq y \leq y_{MAX},$$

since functions that differ only at finite points are considered equivalent in the L^2 space. Note that the selection of basis $\{h_n\}$ is, of course, not unique. We thus express W_0 in the context of the Hilbert space $L^2(\Omega_2)$ and its basis $\{h_n\}$

$$W_0 = \left\{ h \middle| h = \textstyle\sum_{n=1}^{\infty} c_n h_n, h(0,y) = h(x_{MAX}, y_{MAX}) = 0, \forall y_{min} \leq y \leq y_{MAX}, f_0 + h \in V \right\}. \tag{3.36}$$

GRD Function Estimation as a Projection Problem

Eq. (3.34) also defines a bijection between the elements in W_0 and the valid GRD functions in W_2. Estimating a GRD function thus corresponds to searching the optimal element in W_0. However, W_0 is of infinite dimensionality, implying that an infinite number of observations are required to determine a point in W_0. To resolve the dilemma, we assume that practical GRD functions only arise around a low-dimensional space $\tilde{H}_0^N \subset L^2(\Omega_2)$, which is spanned by N function basis $\{h_n \in H_0, n = 1, 2, 3, \cdots, N\}$. We will validate the assumption in next section.

Let $\tilde{W}_0^N = W_0 \cap \tilde{H}_0^N$, so we have

$$\tilde{W}_0^N = \left\{ h \middle| h = \sum_{n=1}^{N} c_n h_n, f_0 + h \in V \right\}. \tag{3.37}$$

Given a set of observations $\{f(x_i, y_i) = z_i, i \in \mathcal{I}\}$, we formally express the GRD function estimation problem as

$$\begin{aligned} \underset{\{c_n, n=1,2,\cdots,N\}}{\arg\min} \quad & \sum_{i \in \mathcal{I}} |z_i - f_0(x_i, y_i) - \sum_{n=1}^{N} c_n h_n(x_i, y_i)|^2 \\ \text{s.t.} \quad & f_0 + \sum_{n=1}^{N} c_n h_n \in V. \end{aligned} \tag{3.38}$$

Eq. (3.38) formulates a projection problem onto a closed convex set $\tilde{W}_0^N$ in the Hilbert space $L^2(\Omega_2)$. If the projection problem is well defined, it will be nicely concluded that there exists a unique solution in $\tilde{W}_0^N$ that best fits the observed data. Therefore, the number of samples must be at least equal to the dimension N of $\tilde{W}_0^N$ for Eq. (3.38) to be a valid projection problem.

Optimal Approximation of GRD Function Space

Recall that we aim for accurately estimating the GRD function with minimal number of samples. Lower dimension of the approximated space $\tilde{W}_0^N$ means fewer samples to probe, while introducing more estimation errors. Given a fixed number N of basis, we want to determine a set of orthonormal basis $\{h_n\}$ that spans the optimal approximation of real-world GRD function space. Consider a real-world GRD function $f \in W_2$. Its best approximation in $f_0 + \tilde{W}_0^N$ is given by

$$\tilde{f} \coloneqq f_0 + \sum_{n=1}^{N} \langle f - f_0, h_n \rangle h_n \tag{3.39}$$

with an approximation error

$$\mathcal{R}[f] \coloneqq |f - \tilde{f}|_2^2.$$

The optimal basis is thus obtained by minimizing the expected approximation error, *i.e.*

$$\begin{aligned} \arg\min_{\{h_n\}} \quad & \mathbb{E}_{p(f)}\mathcal{R}[f], \\ \text{s.t.} \quad & |h_n|_2^2 = 1, \quad n = 1, \cdots, N \end{aligned} \tag{3.40}$$

where $\mathbb{E}_{p(f)}$ means expectation over the distribution of real-world GRD function.

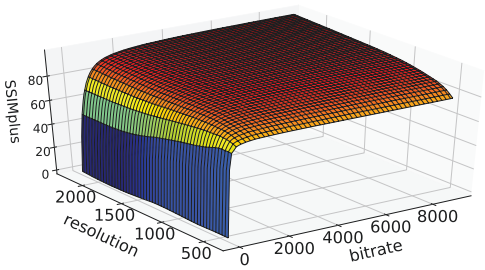
3.3.2 Real-World GRD Function Space

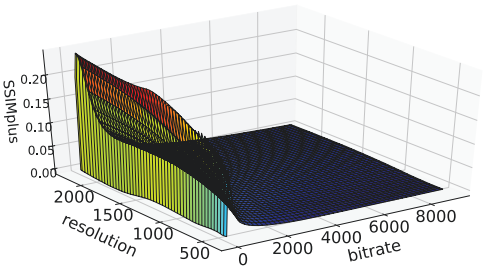
Although f is a continuous function in theory, in practice we often work with discrete samples of the GRD function on a dense and uniform rectangular sampling grid. Therefore, we collect all the samples into a vector and treat a GRD function f interchangeably as a vector and a continuous function. A real-world GRD function may be thought of as such a vector of perceptual qualities of representations encoded with a finite set of encoding profiles obtained on a fixed grid in the domain. This approximation can be done due to the following two facts. First, a limited number of processes is involved in video encoding, suggesting that only a finite number of samples on the GRD surface are practically achievable. Second, we have assumed the GRD function is smooth enough so that we can recover it from its dense samples by interpolation. In particular, when the GRD function is band-limited, it can be fully recovered when the sampling density is larger than the Nyquist rate. In this work, we densely sample 540 points on each GRD surface, and consider the resulting 540-dimensional vector as the ground-truth of a GRD function.

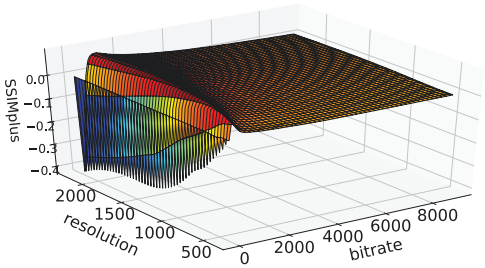
Given M empirical GRD functions in the vector version, Eq. (3.40) may be rewritten as

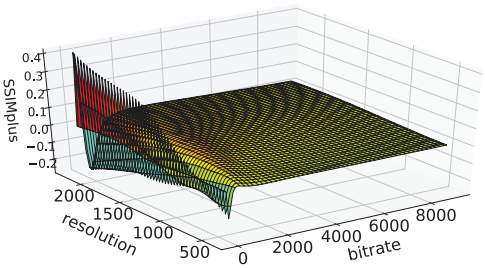
$$\begin{aligned} \arg\min_{\{h_n\}} \quad & \frac{1}{M}\sum_{m=1}^{M}\left| f_m - f_0 - \sum_{n=1}^{N}[(f_m - f_0)^T h_n]h_n \right|_2^2, \\ \text{s.t.} \quad & |h_n|_2^2 = 1, \quad n = 1, \cdots, N \end{aligned} \tag{3.41}$$

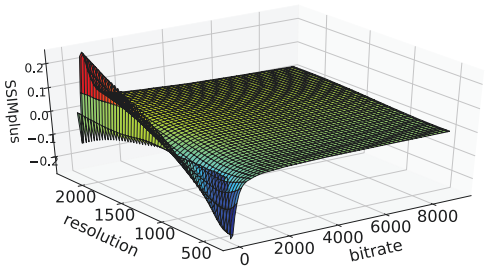
where the inner product of functions is replaced with the normal vector inner product, and $|\bullet|_2$ indicates the l_2 norm in a vector space. f_0 can be any valid GRD function in W_2.

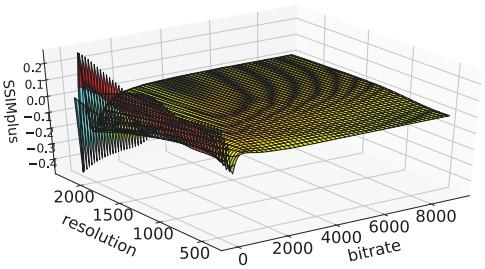


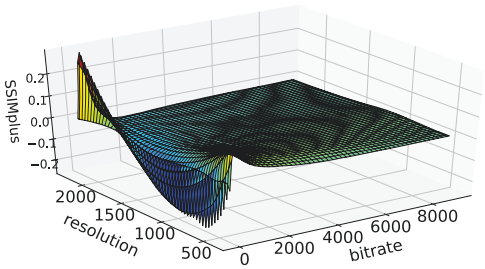


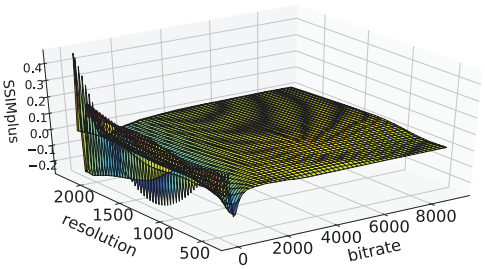












(g) Sixth principal component

(h) Seventh principal component

Figure 3.9: (a) The mean and (b)-(h) the first seven principal components of 800 real-world GRD functions.

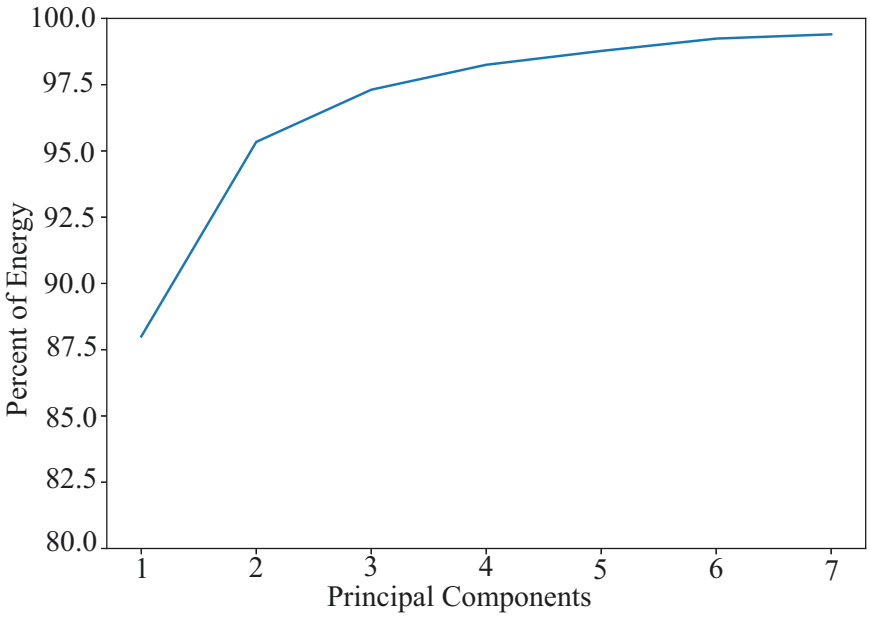


Figure 3.10: A plot showing the percentages of the energies captured by $\tilde{H}_0^N$, the span of the first N principal components. The subspace corresponding to the seven largest eigenvalues encapsulates 99.5 percent of total energy.

Noticing that W_2 is convex, we set

$$f_0 = \bar{f} \coloneqq \frac{1}{M} \sum_{m=1}^{M} f_m$$

as it is readily shown that $\bar{f} \in W_2$. As a result, Eq. (3.41) can be solved by the classic principal component analysis (PCA) algorithm. It is worth noting that we need to perform the PCA algorithm only once, as the solved principal components can be used to model any GRD function. $\{h_n, n = 1, 2, \cdots, N\}$ are thus the eigenvectors associated with the largest N eigenvalues of the empirical covariance matrix of $f - f_0$, and the optimal N-dimensional approximation $\tilde{H}_0^N$ to H_0 is the span of the corresponding eigenspaces.

In this work, we train the eigenvectors $\{h_n\}$ on 800 real-world GRD functions, each represented by a 540-dimensional vector. The training GRD functions are actually part of

Table 3.1: Performance of EGRD on the training set. The second and fourth columns correspond to the average performance over all training GRD functions, while the third and last columns show performance on the worst fit curve.

# of basis	Average RMSE	Worst RMSE	Average l^∞	Worst l^∞
0	4.22	19.32	29.96	67.19
1	2.02	14.18	18.72	56.32
2	1.21	8.93	11.53	41.85
3	0.90	4.43	7.88	33.11
4	0.79	4.37	7.05	32.92
5	0.65	4.01	5.73	33.47
6	0.47	2.80	4.01	26.14
7	0.42	2.78	3.43	15.86

the newly-collected Waterloo generalized rate-distortion (Waterloo GRD) database, whose construction processes will be elaborated in Section 4.2. We draw the the mean GRD surface f_0 and the first seven empirical eigen GRD surfaces in Fig. 3.9. The cumulative energies explained by the first few principal components increase rapidly, as seen in Fig. 3.10. In fact, seven eigenvalues explain more than 99.5 percent of the energy. This verifies our assumption that practical GRD function space can be well approximated by a low-dimensional one, and also suggests that even a seven-parameter model should work reasonably well for most GRD functions found in practice.

We also empirically show how well the training data are represented by the learned basis $\{h_n\}$. Specifically, the EGRD model, which will be expounded in next section, is used to reconstruct the GRD functions with the first N principal components. Table 3.1 shows the reconstruction accuracy in terms of the root mean squared error (RMSE) and l^∞ distance between the reconstructed and original surfaces for $N = 0, 1, \cdots, 7$, where $N = 0$ means that all the GRD functions are approximated by their mean $\bar{f}$. As seen in the table, practical GRD functions can be described by only 7 principal components at a surprising precision. Specifically, the average RMSE and l^∞ distance between the fit and the reference GRD function are reduced to 0.42 and 3.370 with 7 basis, respectively. According

to previous studies [49, 32], such small differences of quality are often regarded as imperceptible to human eyes in a common subjective test [40]. Despite the overall performance, robustness of an algorithm is revealed by the worst performed scenario. As we can see, the worst RMSE is only 2.78, meaning that 7 principal components can explain more than 97% of the shape variation of the most eccentric GRD function in the database. The last column lists the greatest quality score estimation error ever arising over all video samples in the training set. Considering the full score range $[0, 100]$, the greatest relative error on a single video sample only amounts to 16%. As a comparison, the largest error between an individual's score and the mean opinion score (MOS) can easily reach 50 percent of the full score range [15]. Another interesting finding is that even 3 principal components can achieve an average RMSE less than 1, further validating the low-dimensional assumption.

The results from PCA not only imply the possibility of a GRD model with very few parameters, but also give us more insights into the nature of the regularity of GRD function. For example, the shape of principal components as shown in Fig. 3.9 suggests that the variance of GRD function is primarily localized in the low bitrate region, and is slightly larger at high resolution compared to low resolution. This implies that a probe on the GRD function may convey different amount of information in different regions. We will present an information-theoretic sampling method in Chapter 4.

3.3.3 Eigen Generalized Rate-Distortion Model

In the previous subsection, we have obtained f_0 as the average and the principal components $\{h_n\}$ from a training set of real-world GRD functions. The optimal N-dimensional approximation of $\bar{W}_0$ is thus determined by $\tilde{W}_0^N = W_0 \cap \tilde{H}_0^N$, the eigen-space spanned by the first N principal components $\{h_n, 1, \cdots, N\}$. Based on the eigen-space, an N-parametric model for GRD function, namely EGRD, is proposed as

$$\tilde{f} = f_0 + \sum_{n=1}^{N} c_n h_n, \tag{3.42}$$

where the coefficients $\{c_n\}$ are determined by solving the projection problem in Eq. (3.38).

The remaining difficulty is how to formulate the convex cone V. Thanks to the convexity of V, we may approximate the cone by a set of linear inequalities. For brevity, we summarize Eq. (3.42) in the matrix form

$$\tilde{f} = f_0 + H^N \mathbf{c}, \tag{3.43}$$

where $H^N \coloneqq [h_1, h_2, \cdots, h_N]$ and $\mathbf{c} = [c_1, c_2, \cdots, c_N]^T$. Denote by D_x the first order difference matrix along the x-axis, and by D_y the matrix that derives the first order difference along the y-axis only when $x = x_{MAX}$. The discrete form of V can be expressed by

$$\begin{bmatrix} D_x \\ D_y \end{bmatrix} \tilde{f} \geq 0. \tag{3.44}$$

By substituting Eq. (3.43) into Eq. (3.44), we have

$$-\begin{bmatrix} D_x \\ D_y \end{bmatrix} H^N \mathbf{c} \leq \begin{bmatrix} D_x \\ D_y \end{bmatrix} f_0, \tag{3.45}$$

which imposes constraints on the parameters $\mathbf{c}$, and thus regularizes shape of the resultant GRD function.

Now we are able to rewrite Eq. (3.38) in the matrix form with $\mathbf{c}$ being the optimization arguments:

$$\begin{aligned} \arg\min_{\mathbf{c}} \quad & \mathbf{c}^T H^N[\mathcal{I}]^T H^N[\mathcal{I}]\mathbf{c} + 2\left(f_0[\mathcal{I}] - \mathbf{z}\right)^T H^N[\mathcal{I}]\mathbf{c} \\ \text{s.t.} \quad & -\begin{bmatrix} D_x \\ D_y \end{bmatrix} H^N \mathbf{c} \leq \begin{bmatrix} D_x \\ D_y \end{bmatrix} f_0, \end{aligned} \tag{3.46}$$

where $\mathbf{z} = [z_i]^T$, $f_0[\mathcal{I}] = [f_0(x_i, y_i)]^T, i \in \mathcal{I}$, and $H^N[\mathcal{I}] = [h_1[\mathcal{I}], \cdots, h_N[\mathcal{I}]]$. It is not difficult to show that $H^N[\mathcal{I}]^T H^N[\mathcal{I}]$ is positive definite, and all the inequalities are linear constraints, so Eq. (3.46) becomes a standard quadratic programming problem. By solving it with the convex optimization tool [67], we can get the optimal projection $\mathbf{c}^*$ in $\tilde{W}_0^N$, and reconstruct the GRD function by Eq. (3.43).

3.4 Summary

In this chapter, we defined the theoretical space of the GRD function, based on which two computational GRD function models were developed. Both models were able to reconstruct a GRD surface by probing several samples on the function. The proposed models fundamentally distinguish themselves from the other GRD estimation methods. The traditional methodology used by the existing methods can be typically summarized by two steps: 1) presuming a fixed function form with several parameters, and 2) fitting the model to a few queries of the target GRD function. However, it has been shown that these presumed models are systematically biased as they in their own analytical forms violate at least one of the mathematical properties that a valid GRD function must conform to. On the contrary, two proposed methods operate in the theoretical space of GRD functions and thus ensure that the obtained estimate is always a valid GRD function. Furthermore, the two methods may also be extended to accommodate with the case where three or more media attributes come into play. For example, the RAMCT method may estimate a GRD function with 3 independent variables by adopting a tri-variate CT scheme, and the EGRD model can handle the case with even more than 3 dimensions since a multi-dimensional GRD function is reorganized as a long vector in this method. However, the GRD function estimation task is not fully resolved without a proper sampling strategy, as both GRD function models would require a number of queries to fit. In the next chapter, an information-theoretic sampling algorithm will be developed, after which the performances of the proposed GRD models will be evaluated and compared to the state-of-the-art given a sequence of queries selected by the sampling strategy.

Chapter 4

Reconstruction from Sampling of Generalized Rate-Distortion Function

In this chapter, we present an information-theoretic sampling (information-theoretic sampling (ITS)) strategy that generates an optimal order of queries on the generalized rate-distortion (GRD) function. In order to do this, we propose an informativeness measure of each query on the GRD function via a probabilistic model. Equipped with the effective sampling method, we experimentally compare the performances of the two proposed GRD estimation models with state-of-the-art algorithms on a large scale database consisting of thousands of practical GRD functions. The effectiveness of the proposed sampling method is also validated by comparing to a naïve random sampling strategy.

4.1 Information-Theoretic Sampling

In this section, we expound the information-theoretic sampling algorithm in detail. We begin by defining and measuring the informativeness of each query, based on which the sampling algorithm is designed to minimize the uncertainty of the remaining part of the function. Finally, we take a closer look at the resulting query sequence, and show that the proposed sampling algorithm pays more attention to the regions with greater variability

and sparser queries.

4.1.1 Informativeness of Samples

Existing GRD estimation algorithms [8, 9, 52, 69, 22, 12, 41] as well as the robust axial-monotonic Clough-Tocher (RAMCT) and the eigen generalized rate-distortion (EGRD) models proposed in Section 3.2 and 3.3 would require at least a few queries of the target GRD function at different locations in the representation space. Each query, which involves the sophisticated video encoding and quality assessment processes, takes up a significant part of the total computational time to approximate a GRD surface. It is thus desirable to develop an optimal sampling scheme to minimize the number of required quality queries.

Intuitively, different representations do not provide the same information for estimating the shape of GRD surface. For example, encoding a video at very high bitrate always yields saturated quality at any given resolution (Fig. 3.3), providing very limited information for determining the GRD function. In order to measure the informativeness of each sample, we discretize the GRD function by uniformly sampling the bitrate-resolution space as in Section 3.3.2. Let $\mathbf{x} = (x_1, ..., x_N)$ denote the vector of discrete samples on a GRD function, where N is the total number of sample points on the sampling grid. Given that the GRD function is smooth, and that the sampling grid is dense, these discrete samples provide an approximate description of the continuous GRD function. Assuming $\mathbf{x}$ is created from GRD functions of real-world video content, we model $\mathbf{x}$ as an N-dimensional random variable. Since it has been proven that practical GRD functions can be efficiently approximated by linear representations of few principal components (Section 3.3.2), it is reasonable to assume $\mathbf{x}$ follows a multivariate Normal distribution, i.e. its probability density function $p_{\mathbf{x}}(\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The total uncertainty of $\mathbf{x}$ is characterized by its joint entropy given by

$$H_{\mathbf{x}}(\mathbf{x}) = \frac{1}{2} \log |\boldsymbol{\Sigma}| + const, \tag{4.1}$$

where $|\cdot|$ is the determinant operator. If the full vector $\mathbf{x}$ is further divided into two parts such that $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, and the $\mathbf{x}_2$ portion has been resolved by

$\mathbf{x}_2 = \mathbf{a}$, then the remaining uncertainty is given by the conditional entropy

$$H_{\mathbf{x}_1|\mathbf{x}_2}(\mathbf{x}_1|\mathbf{x}_2 = \mathbf{a}) = \frac{1}{2}\log|\bar{\mathbf{\Sigma}}| + const, \tag{4.2}$$

where

$$\bar{\mathbf{\Sigma}} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \tag{4.3}$$

is the covariance matrix of the conditional distribution $p_{\mathbf{x}_1|\mathbf{x}_2}(\mathbf{x}_1|\mathbf{x}_2 = \mathbf{a})$. Note that the exact value of $\mathbf{x}_2$ does not affect the remaining uncertainty of $\mathbf{x}_1$. Thus, the informativeness of sampling $\mathbf{x}_2$ can be measured by the mutual information between $\mathbf{x}$ and $\mathbf{x}_2$, or formally

$$\begin{aligned} I(\mathbf{x}; \mathbf{x}_2) =& H_{\mathbf{x}}(\mathbf{x}) - H_{\mathbf{x}_1|\mathbf{x}_2}(\mathbf{x}_1|\mathbf{x}_2) \\ =& \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\log|\bar{\mathbf{\Sigma}}| + const. \end{aligned} \tag{4.4}$$

4.1.2 Information-Theoretic Sampling Algorithm

Since finding a subset $\mathbf{x}_2$ that maximizes Eq. (4.4) is an NP-hard problem [39], we adopt a greedy strategy to find one sample at a time that most efficiently reduces the uncertainty of GRD estimation. It is known that the greedy approximation is within a constant multiplicative factor from optimal [14, 90]. Similar to Eq. (4.4), given $\mathbf{x}_2 = \mathbf{a}$, the informativeness of the i-th sample in $\mathbf{x}_1$ is calculated by

$$I(\mathbf{x}_1; x_i|\mathbf{x}_2 = \mathbf{a}) = \frac{1}{2}\log|\bar{\mathbf{\Sigma}}| - \log|\bar{\mathbf{\Sigma}}_{ii} - \frac{\bar{\boldsymbol{\sigma}}_i^T\bar{\boldsymbol{\sigma}}_i}{\bar{\sigma}_{ii}}| + const, \tag{4.5}$$

where $\bar{\sigma}_{ii}, \bar{\boldsymbol{\sigma}}_i$, and $\bar{\mathbf{\Sigma}}_{ii}$ denote the i-th diagonal entry, the i-th column without $\bar{\sigma}_{ii}$, and the remaining submatrix by excluding the i-th row and column from $\bar{\mathbf{\Sigma}}$. Thus, with $\mathbf{x}_2$ being already sampled, to find the next most informative sample is equivalent to minimize the log determinant of the conditional covariance matrix of remaining sample points [7]

$$\arg\max_i \quad I(\mathbf{x}_1; x_i|\mathbf{x}_2 = \mathbf{a}) = \arg\min_i \quad \log|\bar{\mathbf{\Sigma}}_{ii} - \frac{\bar{\boldsymbol{\sigma}}_i^T\bar{\boldsymbol{\sigma}}_i}{\bar{\sigma}_{ii}}|. \tag{4.6}$$

Algorithm 1: Information Theoretic Sampling

Result: A list of samples S

Initialize S as an empty list; $\bar{\mathbf{\Sigma}}^{(1)} = \mathbf{\Sigma}$;

for $k := 1$ *to* N **do**

  $i^{(k)} = \arg\min_i \ \mathrm{tr}(\bar{\mathbf{\Sigma}}_{ii}^{(k)} - \frac{\bar{\boldsymbol{\sigma}}_i^{(k)^T} \bar{\boldsymbol{\sigma}}_i^{(k)}}{\bar{\sigma}_{ii}^{(k)}})$;

  $x^{(k)} = \mathrm{VQA}(\mathrm{Encode}(\mathbf{r}_i^{(k)}))$;

  Set $S = S \cup x^{(k)}$;

  $\bar{\mathbf{\Sigma}}^{(k+1)} = \bar{\mathbf{\Sigma}}_{ii}^{(k)} - \frac{\bar{\boldsymbol{\sigma}}_i^{(k)^T} \bar{\boldsymbol{\sigma}}_i^{(k)}}{\bar{\sigma}_{ii}^{(k)}}$;

  if *Stopping criterion is met* **then**

    Break ;

  end

end

Minimizing (4.6) directly is computationally expensive, especially when the dimensionality is high. Alternatively, we minimize the upper bound of the conditional entropy

$$\arg\min_i \quad \mathrm{tr}(\bar{\mathbf{\Sigma}}_{ii} - \frac{\bar{\boldsymbol{\sigma}}_i^T \bar{\boldsymbol{\sigma}}_i}{\bar{\sigma}_{ii}}), \tag{4.7}$$

where

$$\log|\bar{\mathbf{\Sigma}}_{ii} - \frac{\bar{\boldsymbol{\sigma}}_i^T \bar{\boldsymbol{\sigma}}_i}{\bar{\sigma}_{ii}}| \leq \mathrm{tr}(\bar{\mathbf{\Sigma}}_{ii} - \frac{\bar{\boldsymbol{\sigma}}_i^T \bar{\boldsymbol{\sigma}}_i}{\bar{\sigma}_{ii}} - \mathbf{I})$$

and $\mathbf{I}$ denotes identity matrix. The sample with the minimum average loss in (4.7) over all viewing devices is most informative. Once the optimal sample index is obtained, we encode the video at the i-th representation, evaluate its quality with objective video quality assessment (VQA) algorithms, and update the conditional covariance matrix in (4.3). The process is applied iteratively until a stopping criterion is satisfied. A possible stopping criterion is that the overall uncertainty of the GRD function is reduced below a certain threshold, *e.g.* 540 (the total number of representation samples in the discretized GRD function space) $\times$ 10 (the standard deviation of mean opinion score (MOS) in the LIVE VQA database [61]), since when $tr(\Sigma)$ is below the threshold, we may ascribe the uncertainty in the system to the disagreement between subjects. With this stopping criterion, we find that 38 samples are required to achieve this level of uncertainty on average. We

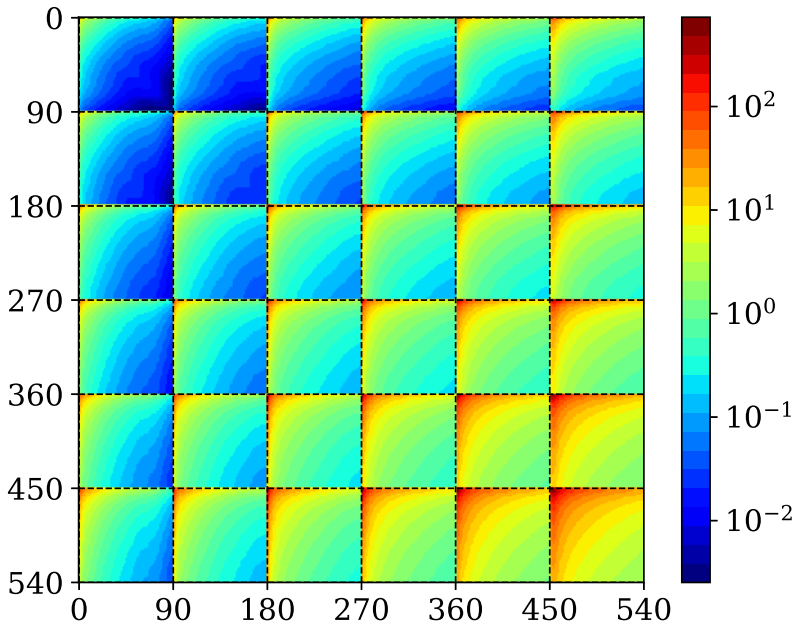


Figure 4.1: Empirical covariance matrix of the GRD functions. Warmer color represents higher values, meaning higher variance of the quality of a sample on the diagonal, or higher correlation between those of two samples off the diagonal. Samples in the same 90-sample segment have the same spatial resolution, which elevates every 90 samples. Within each 90-sample segment, bitrates are presented in an ascending order.

summarize the proposed uncertainty sampling method in Algorithm 1, where $\mathbf{r}_i$ represents the bitrate and spatial resolution of the i-th representation.

4.1.3 Discussion

To get a sense of what type of samples will be chosen by the proposed algorithm, we analyze several influencing factors in the objective function (4.7):

- By the basic properties of trace, the objective function in the ITS algorithm can be

factorized as

$$\begin{aligned}
&\mathrm{tr}(\bar{\boldsymbol{\Sigma}}_{ii} - \frac{\bar{\boldsymbol{\sigma}}_i^T \bar{\boldsymbol{\sigma}}_i}{\bar{\sigma}_{ii}}) \\
=&\mathrm{tr}(\bar{\boldsymbol{\Sigma}}_{ii}) - \frac{\mathrm{tr}(\bar{\boldsymbol{\sigma}}_i^T \bar{\boldsymbol{\sigma}}_i)}{\bar{\sigma}_{ii}} \\
=&\mathrm{tr}(\bar{\boldsymbol{\Sigma}}) - (\bar{\sigma}_{ii} + \frac{1}{\bar{\sigma}_{ii}} \sum_{j \neq i} \bar{\sigma}_{ij}^2),
\end{aligned}$$

where $\bar{\sigma}_{ij}$ denotes the entry of $\bar{\boldsymbol{\Sigma}}$ at i-th row and j-th column. Thus, $\mathrm{tr}(\bar{\boldsymbol{\Sigma}}_{ii} - \frac{\bar{\boldsymbol{\sigma}}_i^T \bar{\boldsymbol{\sigma}}_i}{\bar{\sigma}_{ii}})$ is a decreasing function with respect to $\bar{\sigma}_{ii}$ when $\bar{\sigma}_{ii} > \sqrt{\sum_{j \neq i} \bar{\sigma}_{ij}^2}$. This indicates that samples with large uncertainty are more likely to be selected than those with small uncertainty.

- According to (4.3), $\forall j \neq i$,

$$\bar{\sigma}_{jj}^{(k+1)} = \bar{\sigma}_{jj}^{(k)} - \frac{\bar{\sigma}_{ij}^{(k)^2}}{\bar{\sigma}_{ii}^{(k)}},$$

suggesting the rate of reduction in the uncertainty of sample j is proportional to its squared correlation with the selected sample i in the k-th iteration. Fig. 4.1 shows an empirical covariance matrix $\boldsymbol{\Sigma}$ estimated from our video database that will be detailed in the next chapter, from which we observe that the GRD functions typically exhibit high correlation in a local region. Combining the first observation above, we conclude that the next optimal choice of sample should be selected from the region where labeled samples are sparse.

- Video contents with different complexities are encoded to various quality levels at a low to medium bitrate, while to a similar high quality given sufficient encoding resources. As a result, more uncertainty of the GRD function is observed in the low bitrate region than the high bitrate region at the same spatial resolution. On the other hand, the uncertainty of GRD function also shifts with respect to the encoding spatial resolution. Specifically, video downsampling significantly narrows the video content space, indirectly reducing the variability of the GRD function. Moreover, the

visual quality is often dominated by the blurring artifacts due to downsampling and upsampling operations when a video is encoded at low spatial resolutions, leading to small quality variance. Both trends are reflected in Fig. 4.1, as warmer colors are found in regions of lower bitrates and higher spatial resolutions. We may conclude from the trends that most samples will be drawn from the low bitrate and high resolution region.

- Note that knowing that $\mathbf{x}_2 = \mathbf{a}$ alters the variance, though the new variance does not depend on the specific value of $\mathbf{a}$. The independence has two important consequences. First, the proposed sampling scheme is general enough to accommodate all kinds of GRD estimator. More importantly, the algorithm results in a unique sampling sequence for all GRD functions. In other words, we can generate a lookup table of optimal querying order, making the sampling process fully parallelizable.

4.2 Waterloo Generalized Rate-Distortion Database

4.2.1 Obtaining Ground-truth GRD Functions

In order to evaluate the performance of a GRD model, one needs to know the ground truth of a GRD function. Given a high quality source video and an encoder, we obtain the corresponding GRD ground-truth by densely sampling the bitrate-resolution domain. Specifically, the video is processed by the following steps sequentially:

- Spatial downsample: In the Waterloo generalized rate-distortion (Waterloo GRD) database, all source videos are in 1080P, so maximum encoding resolution cannot exceed the original video 1920×1080, *i.e.* $y_{MAX} = 2203$. We downsample the source video using bi-cubic filter to six spatial resolutions (1920×1080, 1280×720, 720×480, 512×384, 384×288, 320×240) according to the list of Netflix certified devices [22]. Consequently, the actual lower bound of spatial resolution in the database is $y_{min} = 400$.

- Compression: We encode the downsampled sequences using specified video encoders (which may include H.264 [28], HEVC [29] and/or VP9 [30]) with target bitrate ranging from 100 kbps to 9 Mbps with a step size of 100 kbps.

In total, we get $6 \times 90 = 540$ representations for each combination of video content and encoder. Then the perceptual quality of a representation is assessed as the response of GRD function. Ideally, the GRD function response, or the representation quality, should be measured by subjective evaluation, because the human visual system (HVS) is the ultimate receiver in most applications. However, subjective experiment is expensive and time consuming. Therefore, we employ objective VQA measurements to generate ground-truth samples of the GRD function in practice. We evaluate the quality of each representation at a given spatial resolution and bitrate using SSIMplus [54] for the following reasons. First, SSIMplus is shown to outperform other state-of-the-art quality measures in terms of accuracy and speed [54, 25]. Second, it is currently the only objective VQA model that offers meaningful cross-resolution and cross-device scoring, an essential property of GRD. Third, its precedent models structural similarity (SSIM) [75] index and MS-SSIM [78] index have been demonstrated to perform well in estimating the GRD functions [12] and been widely used in industry practice. The resulting dense samples of SSIMplus are regarded as the ground truth of GRD functions (The range of SSIMplus is from 0 to 100 with 100 indicating perfect quality). It is worth noting that our GRD modeling approaches do not constrain themselves on any specific VQA methods. When other ways of generating dense ground-truth samples are available, the same GRD modeling approach may also be applied.

We post-process the raw data to obtain GRD functions on a regular grid. First, the lossless encoding bitrate may be lower than 9,000 kbps when the complexity of source video is low. In such case, we pad the highest achievable quality at each resolution to the end of GRD function along the bitrate dimension. Second, the rate-control of video encoder is inaccurate, suggesting that the actual encoding rate may be different from the target bitrate. Therefore, We resampled the rate-distortion curves at each resolution uniformly with a step-size of 100 kbps using piece-wise cubic hermite interpolating polynomial (PCHIP) to preserve monotonicity.

Figure 4.2: Sample frames of source videos in the WGRD set. All images are cropped for better visibility.

4.2.2 Constructing the WGRD Database

The Waterloo GRD database consists of two parts. In the Waterloo GRD Phase I database, we construct a new video database which contains 250 pristine natural videos that span a great diversity of video content. An important consideration in selecting the videos is that they need to be representative of the videos we see in the daily life. Therefore, we resort to the Internet and elaborately select 200 keywords to search for creative common licensed videos. The keywords are broadly classified into 8 categories: human, animal, plant, landscape, cityscape, still-life, transportation, and computer-synthesized videos. We initially obtain more than 700 4K videos. Many of these videos contain significant distortions, including heavy compression artifacts, noise, blur, and other distortions due to improper operations during video acquisition and sharing. To make sure that the videos are of pristine quality, we carefully inspect each of the videos multiple times by zooming in and remove those videos with visible distortions. We further reduce artifacts and other unwanted con-

tamination by downsampling the videos to a size of 1920×1080 pixels, from which we extract 10 second semantically coherent video clips. Eventually we end up with 250 high quality source videos. Then we compress each source videos into the aforementioned 540 representations with three commonly used encoders, namely H.264, HEVC and VP9, with two-pass encoding settings [34, 22, 41]. The perceptual qualities of each representation at five commonly used display devices including cellphone, tablet, laptop, desktop, and TV are computed by SSIMplus [54]. Finally, we obtain $250 \times 3 \times 5 = 3750$ GRD functions for the Waterloo GRD Phase I database.

For the Waterloo GRD Phase II database, we further collect more than $3,000$ video contents from the Internet using the same keywords. Similar quality screening and post-processing procedures are applied to the videos, resulting in $1,000$ high quality 10-second video clips. This time we compress the $1,000$ videos only using the H.264 encoder, and evaluate the encoded representations with only one device, *i.e.* desktop. Consequently, the Waterloo GRD Phase II database contains $1,000$ GRD functions of greater video complexity diversity compared to the Phase I database.

In summary, the Waterloo GRD database consists of $4,750$ real-world GRD functions, which are formed by $4,750 \times 540 = 2,565,000$ objective scores from $945,000$ different video representations featuring $1,250$ different contents. The Waterloo GRD database is not only the first of its kind, but also the largest and most diverse video database in the VQA community. Some representative video contents and GRD functions from the database can be found in Fig. 4.2 and Fig. 3.1, respectively.

4.3 Performance Evaluation of RAMCT and ITS

In this section, we evaluate the performance of RAMCT and ITS proposed in Section 3.2 and Section 4.1, respectively, on the Waterloo GRD Phase I database.

4.3.1 Experimental Setups

Implementation Details: We initialize the scattered network with Delaunay triangulation [23], inherited from Clough-Tocher (CT) method [18]. The balance weight λ in (3.29) is set to 10^{-4}. In our current experiments, the performance of the proposed RAMCT is fairly insensitive to variations of the value. We employed operator splitting quadratic program (OSQP) [67] to solve the quadratic programming problem, where the maximum number of iterations is set to 10^6. In this experiment, we perform the ITS until the optimal selection order of all 540 samples is generated, and increase the number of samples following this order when testing the performance of a GRD model. Specifically, we report the estimation performances when 20, 30, 50, 75, 100, and 540 samples are available to represent the capability of competing and proposed GRD models. Since a triangulation only covers the convex hull of the scattered point set, extrapolation beyond the convex hull is not possible. In order to make a fair comparison, we initialize the training set S as the representations with the maximum and minimum bitrates at all spatial resolutions. To construct the covariance matrix described in Section 4.1 as well as test the proposed algorithm, we randomly segregated the database into a training set which contains the GRD functions from 80 percent of video contents, and a testing set with remaining GRD functions. The random split is repeated 50 times and the median performance is reported.

Evaluation Criteria: We test the performance of the GRD estimators in terms of both accuracy and rate of convergence. Specifically, we used two metrics to evaluate the accuracy. The root mean squared error (RMSE) and l^∞ norm of the error values are computed between the estimated function and the actual function for each source content. The median results are then computed over all testing functions. All GRD estimation models achieve greater accuracy as the sample number increases, and especially the increasingly complex GRD functions can be well-fitted at the cost of using many parameters. What distinguishes these models from each other is the rate and manner with which the quality of the approximation varies with the number of training samples.

Table 4.1: RMSE performance of the competing GRD function models with different number of labeled samples selected by random sampling (RS) and the proposed information theoretic sampling (ITS).

sample #	Reciprocal [41]		Logarithmic [12]		PCHIP [31]		CT [51]		RAMCT	
	RS	ITS	RS	ITS	RS	ITS	RS	ITS	RS	ITS
20	N.A.	N.A.	4.80	**3.65**	8.29	5.15	9.41	7.49	11.63	4.01
30	7.89	9.13	3.71	3.25	5.56	**1.44**	6.16	4.77	3.31	1.81
50	6.17	8.60	3.07	2.60	2.94	0.26	3.43	3.49	2.17	**0.24**
75	5.50	6.99	2.27	2.22	1.75	**0**	2.20	1.81	1.00	**0**
100	5.24	6.20	2.14	2.04	1.33	**0**	1.66	1.12	0.36	**0**
540	4.95	4.95	1.66	1.66	**0**	**0**	**0**	**0**	**0**	**0**

Table 4.2: l^{∞} performance of the competing GRD function models with different number of labeled samples selected by random sampling (RS) and the proposed information theoretic sampling (ITS).

sample #	Reciprocal [41]		Logarithmic [12]		PCHIP [31]		CT [51]		RAMCT	
	RS	ITS	RS	ITS	RS	ITS	RS	ITS	RS	ITS
20	N.A.	N.A.	19.40	**16.56**	38.87	28.11	36.50	29.51	45.15	21.88
30	48.32	45.36	17.85	12.28	33.04	11.07	29.84	18.70	27.07	**6.13**
50	52.48	45.48	15.75	12.37	24.33	**2.10**	21.82	14.30	23.99	2.13
75	54.49	49.08	14.59	13.53	18.22	0.47	17.89	7.76	16.51	**0.11**
100	55.54	51.26	14.22	14.44	16.00	0.26	15.59	5.84	14.23	**0**
540	58.04	58.04	18.33	18.14	**0**	**0**	**0**	**0**	**0**	**0**

4.3.2 Experimental Results

We test five GRD function models including reciprocal regression [69], logarithmic regression [12], PCHIP interpolation [31], CT interpolation [51], and the proposed RAMCT on the Waterloo GRD Phase I database. To evaluate the performance of the information theoretic sampling (ITS) algorithm, we apply it on the five GRD models above and compare its performance with random sampling (RS) scheme as the baseline. For RS, the initial set of the training sample S is set as the representations with the maximum and minimum bitrates at all spatial resolutions for fair comparison. The random sampling process was repeated 50 times and the median performance is reported.

Tables 4.1 and 4.2 show the prediction accuracy on the database, from which the key observations are summarized as follows. First, the models that assume a certain analytic functional form are consistently biased, failing to accurately fit GRD functions even with all samples probed. On the other hand, the existing interpolation models usually take more than 100 random samples to converge, although they are asymptotically unbiased. By contrast, the proposed RAMCT model converges with only a moderate number of samples. Second, we analyze the core contributors of RAMCT with deliberate selection of competing models. Specifically, the 1D monotonic interpolant PCHIP significantly outperforms the 2-dimensional generic interpolant CT, suggesting the importance of axial monotonicity. Furthermore, RAMCT achieves even better performance by exploiting the 2D structure and jointly modeling the GRD functions. Third, we observe strong generalizability of the proposed information theoretic sampling strategy evident by the significant improvement over random sampling on all models. The performance improvement is most salient on the proposed model. In general, RAMCT is able to accurately model GRD functions with only 30 labeled samples, based on which the reciprocal model merely have sufficient known variables to initialize fitting. To gain a concrete impression, we also recorded the execution time of the entire GRD estimation pipeline including video encoding, objective VQA, and GRD function approximation with the competing algorithms on a computer with 3.6GHz CPU and 16G RAM. RAMCT with uncertainty sampling takes around 10 minutes to reduce l^{∞} below 5, which is more than 100 times faster than the tradition regression models with random sampling.

4.3.3 Visual Illustration

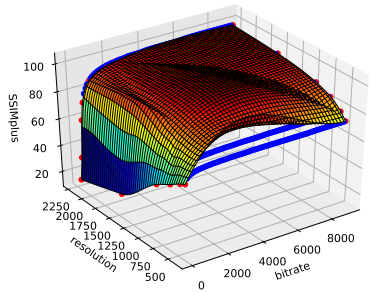
We select a sample GRD function from the Waterloo GRD Phase I database, and visualize the corresponding GRD surfaces reconstructed by the classic CT and the proposed RAMCT interpolation methods in Fig. 4.3. Red and blue points represent the training and testing ground-truth quality scores of the actual GRD function, respectively. By taking a close look at these examples, we emphasize several important advantages of the proposed RAMCT and ITS methods.

- The RAMCT method is good at reconstructing the complex shape of the GRD function. Comparing Fig. 4.3e and 4.3f, the RAMCT-reconstructed surface preserves the subtle structure and the flat region of the ground truth in the low and high bitrate regions, respectively, while the CT method fails in both regions. It is worth noting that, even though only one point other than the minimum and maximum bitrates was sampled at the lowest two resolutions, the RAMCT method still managed to precisely reconstruct the curving of GRD function by taking into account information from samples at other resolutions.

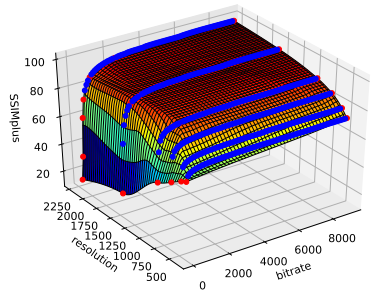
- The RAMCT method always reproduces valid GRD surfaces even when the training samples are scarce, while the reconstructed function by the classic CT method may severely violate the constraints. The validity, such as the axial monotonicity with regards to the bitrate, of GRD function is the key factor for many subsequent video applications to be feasible and solvable. For example, reasonable Bjøntegaard-Delta bitrate (BD-Rate) can only be computed between two monotonic rate-distortion (RD) curves [9].

- The sampling order of the ITS method is visually shown in Fig. 4.3b, 4.3d, and 4.3f. As we can see, the selected samples are localized in the low bitrate range, and lean towards high resolutions, where real-world GRD functions exhibit great diversity. Therefore, such a sampling strategy can efficiently eliminate the uncertainty of a GRD function.

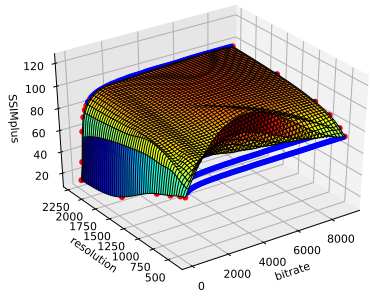
In summary, we conclude that, besides the prediction accuracy and the convergence rate, the proposed methods also provide better shaped and more reasonable GRD approxima-



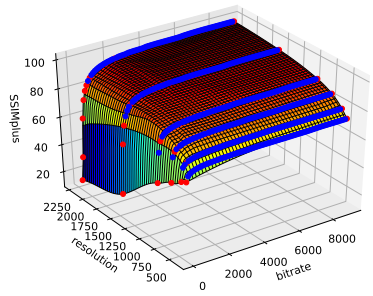
(a) CT with 20 training samples



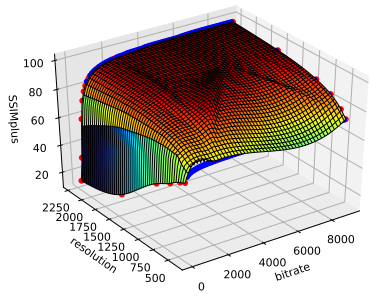
(b) RAMCT with 20 training samples



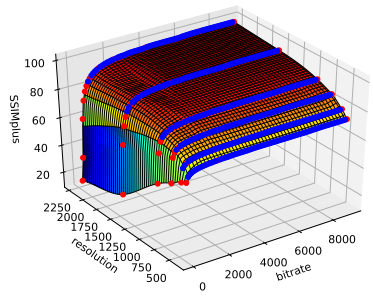
(c) CT with 30 training samples



(d) RAMCT with 30 training samples



(e) CT with 40 training samples



(f) RAMCT with 40 training samples

Figure 4.3: CT and RAMCT performance given 20, 30, 40 training samples selected by the proposed ITS method. Red and blue points represent the training and testing ground-truth quality scores, respectively.

tions compared to the classic CT method, validating our theoretical analysis on the GRD function.

4.4 Performance Evaluation of EGRD

In this section, we compare the performance of principal components to alternative basis in terms of approximating the practical GRD function space, and EGRD to other regression models on the Waterloo GRD Phase II database.

4.4.1 Experimental Setups

We compare the learned principal components with two general orthonormal basis, *i.e.* 2D polynomial and 2D trigonometric basis, based on which we may get two competing GRD models, namely the polynomial generalized rate-distortion (PGRD) and the trigonometric generalized rate-distortion (TGRD) models, following the steps below. First, the basis vectors are sorted in the descending order according to the variance of the training set explained by each dimension. This sorting is similar to sorting the principal components from the greatest to the smallest eigen values, and thus guarantees that the optimal N-dimensional approximated space is spanned by the first N basis vectors. Second, we replace the principal component analysis (PCA) basis with one of the alternative options in Eq. (3.43) and (3.46), resulting a similar quadratic programming problem. Finally, the PGRD and the TGRD models are obtained by solving their respective problems.

In the first experiment, we compare the capability of the three basis to represent the practical GRD function space. All the three basis can approximate the space better at the cost of involving more basis vectors. Therefore, we fix the number of employed components from 0 to 20 in each basis and evaluate the estimation error of using these basis to approximate the actual GRD functions with different dimensions, where the 0-dimension case means that all the test GRD functions are approximated by the mean function of the training set. The approximation capability of basis is thus judged by the rate and manner

with which the estimation error varies with the number of basis vectors. It is worth noting that in this experiment all samples of a test GRD function are available to the models.

Then we conduct another experiment to test the reconstruction performance of GRD models when only sparse samples are available. In this experiment, we fix the number of basis vectors to 7 in the EGRD model since the first seven principal components are sufficient to explain 99.5% variance of the training set. As for a fair comparison, we also choose the first seven components of the polynomial and trigonometric basis for the PGRD and the TGRD models, respectively. The other two competing GRD models are 1D reciprocal regression [69], and 1D logarithmic regression [12] models. Since it has been shown in the previous experiments that the ITS is far better than random sampling, we employ the ITS algorithm to select the training samples for the GRD models. For the two 1D models, we particularly modify the ITS algorithm so that the sampling process can be performed in turn at each resolution. The sample number ranges from 7 to 50 to show the reconstruction performance and the convergence rate of GRD models with limited samples.

Similar to the previous experiments in Section 4.3, the approximation error is also evaluated by the RMSE and l^{∞} distance between the reconstructed and the reference GRD surfaces. Besides the average performance, we also report the worst result in the test set to show the robustness of each model.

Since the proposed EGRD model needs training, we randomly segregate the Waterloo GRD Phase II database into a training set of 800 GRD functions and a test set with remaining 200 GRD functions. All the GRD models are evaluated on the test set, and the average and the worst-case performances are recorded. We repeat the process for 50 times, and report the median results in subsequent sections.

4.4.2 Approximation Capability of Basis

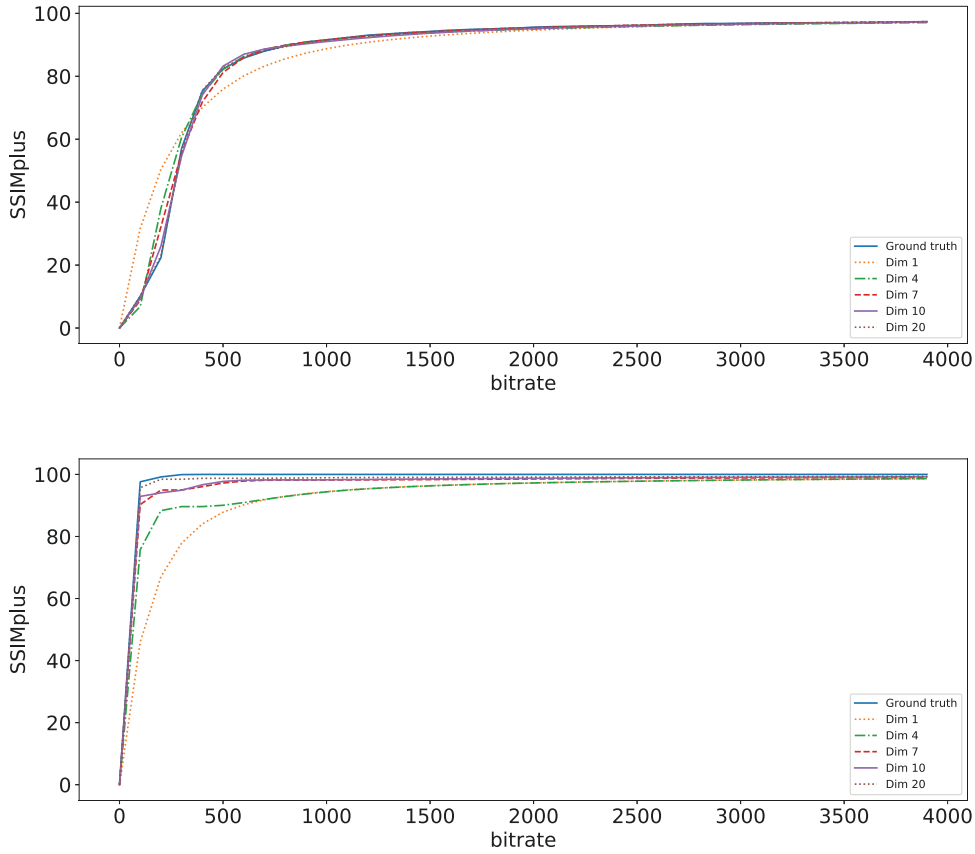
Tables 4.3 and 4.4 summarize the approximation capability of the three basis with 0 to 20 components being employed. Although all the basis can provide more accurate approximations by using more elements, the proposed PCA basis significantly outperform the other two general basis. In fact, the performance of 20 polynomial or trigonometric basis vectors

Table 4.3: RMSE performance of the competing GRD function basis with different number of basis vectors. Best average and worst-case performances are highlighted by italics and boldface, respectively

Vector #	Polynomial		Trigonometric		PCA	
	Average	Worst	Average	Worst	Average	Worst
0	*4.22*	**14.86**	*4.22*	**14.86**	*4.22*	**14.86**
1	4.22	14.86	4.20	14.80	*2.00*	**9.02**
2	4.05	12.77	4.08	13.76	*1.20*	**3.69**
3	4.01	12.30	4.00	13.12	*0.88*	**3.05**
4	3.97	12.28	3.98	12.45	*0.79*	**2.68**
5	3.63	8.57	3.86	12.11	*0.65*	**2.45**
6	3.41	7.22	3.86	11.97	*0.47*	**2.04**
7	3.22	6.69	3.85	11.64	*0.42*	**2.00**
8	3.00	6.48	3.70	11.11	*0.40*	**1.90**
9	2.81	6.31	3.65	10.67	*0.37*	**1.86**
10	2.65	6.12	3.63	10.41	*0.34*	**1.45**
11	2.49	5.85	3.57	10.16	*0.31*	**1.27**
12	2.36	5.75	3.53	10.05	*0.31*	**1.20**
13	2.27	5.57	3.45	9.58	*0.31*	**1.11**
14	2.18	5.52	3.41	9.35	*0.29*	**0.97**
15	2.18	5.47	3.37	9.10	*0.28*	**0.89**
16	2.16	5.41	3.34	9.02	*0.26*	**0.89**
17	2.13	5.31	3.27	8.48	*0.25*	**0.80**
18	2.10	5.24	3.26	8.46	*0.24*	**0.70**
19	2.04	5.15	3.23	8.43	*0.23*	**0.69**
20	1.98	5.02	3.19	8.05	*0.22*	**0.69**

Table 4.4: l^∞ performance of the competing GRD function basis with different number of basis vectors. Best average and worst-case performances are highlighted by italics and boldface, respectively

Vector #	Polynomial		Trigonometric		PCA	
	Average	Worst	Average	Worst	Average	Worst
0	*29.93*	**65.04**	*29.93*	**65.04**	*29.93*	**65.04**
1	29.93	65.04	29.91	65.04	*18.49*	**54.42**
2	29.58	65.04	29.80	65.04	*11.41*	**37.74**
3	29.38	65.04	29.58	65.04	*7.84*	**29.41**
4	29.16	64.61	29.51	65.04	*6.99*	**28.08**
5	27.83	64.76	29.13	65.01	*5.72*	**27.52**
6	27.02	64.21	29.12	64.99	*4.04*	**20.77**
7	26.28	63.32	29.09	64.99	*3.43*	**14.10**
8	25.38	62.66	28.48	64.85	*3.04*	**12.10**
9	24.54	61.63	28.37	64.82	*2.71*	**11.10**
10	23.52	60.15	28.16	64.82	*2.38*	**10.20**
11	22.42	57.91	27.85	64.71	*2.21*	**9.52**
12	21.42	56.75	27.78	64.64	*2.16*	**8.99**
13	20.65	54.88	27.48	64.53	*2.08*	**8.58**
14	19.46	51.68	27.39	64.53	*1.95*	**8.83**
15	19.41	51.61	27.14	64.43	*1.82*	**8.17**
16	19.31	51.49	27.11	64.38	*1.68*	**8.12**
17	18.96	50.84	26.84	64.29	*1.60*	**7.61**
18	18.56	50.29	26.82	64.27	*1.49*	**6.99**
19	18.01	49.10	26.78	64.23	*1.39*	**6.64**
20	17.57	47.51	26.54	64.05	*1.32*	**6.42**



(b) The worst case when 10 principal components are used.

Figure 4.4: Ground truth and approximate RD curves at 1920×1080 with 1, 4, 7, 10 and 20 PCA basis vectors.

is easily beaten by that of only 2 principal components. The advantage of the PCA basis is made more clear when considering the l^∞ performance in Table 4.4. The average l^∞ error of PCA basis diminishes as the dimensionality increases, and drops below 10 with only 3 basis vectors. By contrast, the other two basis fail to achieve a l^∞ error less than 17 even with 20 parameters. The PCA basis also approximate the most difficult curve well in the test set as shown by the worst-case performances in Table 4.3 and 4.4. With the default dimensionality of 7, the worst-case performance of PCA basis is much better than the average performance of the two competing basis with 20 vectors. Moreover, the worst-case performance of PCA basis improves significantly as the number of basis vectors increases, while the other two basis improve at a much lower rate, if there is any.

In order to qualitatively illustrate how the representation power of PCA basis improves with dimensionality, we show in Fig. 4.4 two of the most difficult GRD functions (associated with the worst l^∞ error) in the test set from one of the 50 random splits. For better visualization, we draw the approximation curves with 1, 4, 7, 10, and 20 dimensions at 1920×1080 only. From the two cases, we observe that the approximation with 7 basis grossly captures the main trend of RD curves. With the dimensionality of 20, there is little difference between the approximate and actual curves. The worst-case curves also show qualitatively how fitting improves with the number of basis vectors involved.

4.4.3 Performance of GRD models

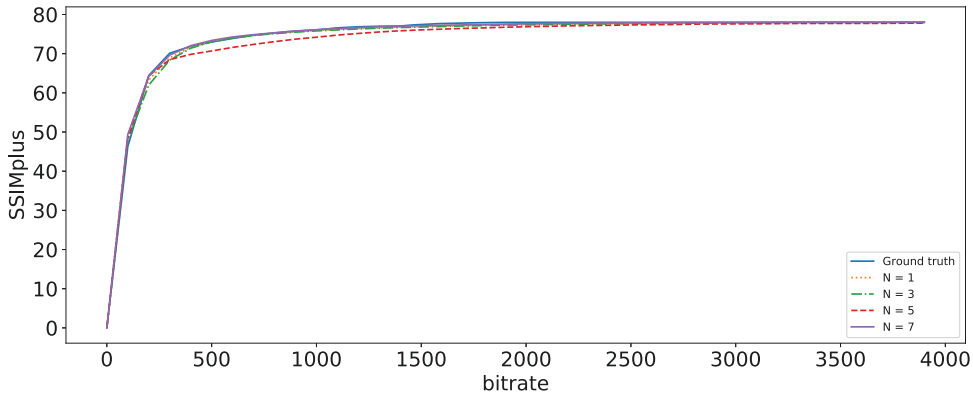
In practice, probing the quality of a single sample in the GRD space is expensive and time-consuming due to sophisticated video encoding and quality assessment processes. It is desirable to accurately reconstruct the rate-distortion surface with minimal number of samples. In this experiment, we compare the performance of the EGRD model to four competing GRD models, reciprocal regression [69], logarithmic regression [12], PGRD, and TGRD on the Waterloo GRD Phase II database. Tables 4.5 and 4.6 list the reconstruction performance on the database, from which we have several key observations as follows. First, the three basis-projection-based methods in general outperform the other models, which assume a heuristic analytic functional form in advance. Second, among the three basis-projection-based methods, the EGRD model delivers by far the best performance

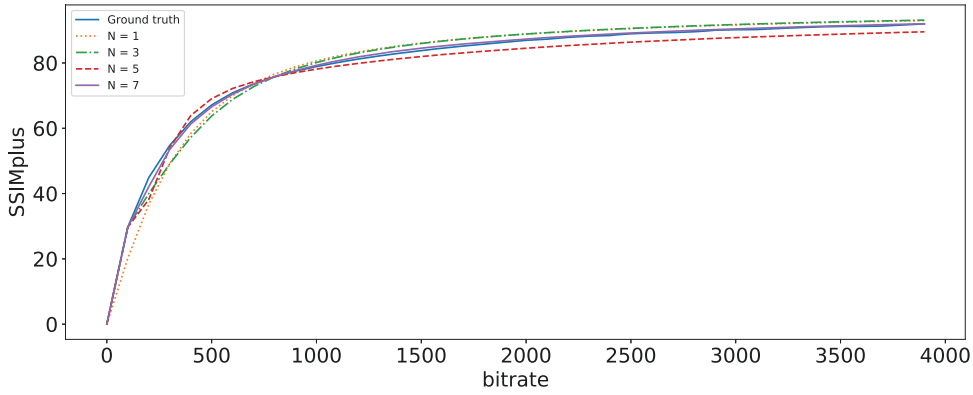
Table 4.5: RMSE performance of the competing GRD models with different available samples. Best average and worst-case performances are highlighted by italics and boldface, respectively

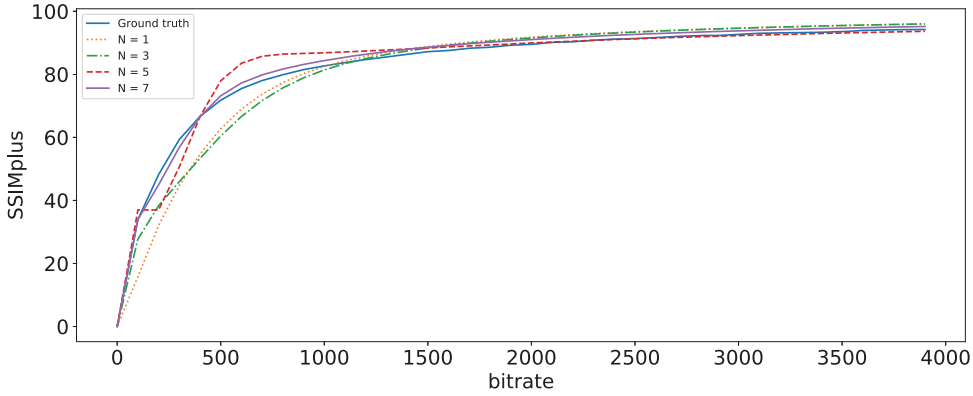
Sample #	Reciprocal [69]		Logarithmic [12]		PGRD		TGRD		EGRD	
	Average	Worst	Average	Worst	Average	Worst	Average	Worst	Average	Worst
7	N.A.	N.A.	N.A.	N.A.	3.83	8.93	4.90	11.69	*0.67*	**2.52**
10	N.A.	N.A.	N.A.	N.A.	3.59	7.83	4.58	11.74	*0.58*	**2.33**
15	N.A.	N.A.	N.A.	N.A.	3.54	7.37	4.27	11.77	*0.58*	**2.32**
20	N.A.	N.A.	25.14	32.99	3.42	6.95	4.14	11.79	*0.54*	**2.25**
25	N.A.	N.A.	21.30	29.91	3.44	6.92	4.12	11.80	*0.52*	**2.20**
30	16.44	35.82	18.53	27.04	3.38	6.78	4.05	11.81	*0.51*	**2.18**
35	16.31	35.14	16.64	24.76	3.36	6.76	4.04	11.76	*0.51*	**2.20**
40	15.34	36.90	14.74	23.08	3.35	6.72	4.01	11.76	*0.51*	**2.17**
45	13.98	33.47	12.91	20.09	3.34	6.71	3.97	11.75	*0.50*	**2.17**
50	12.18	33.02	12.05	18.95	3.32	6.70	3.94	11.77	*0.49*	**2.17**

Table 4.6: l^{∞} performance of the competing GRD models with different available samples. Best average and worst-case performances are highlighted by italics and boldface, respectively

Sample #	Reciprocal [69]		Logarithmic [12]		PGRD		TGRD		EGRD	
	Average	Worst	Average	Worst	Average	Worst	Average	Worst	Average	Worst
7	N.A.	N.A.	N.A.	N.A.	25.23	63.21	28.57	64.99	*4.80*	**15.88**
10	N.A.	N.A.	N.A.	N.A.	25.24	63.21	28.58	64.99	*3.42*	**13.04**
15	N.A.	N.A.	N.A.	N.A.	25.19	63.21	28.68	64.99	*3.02*	**12.14**
20	N.A.	N.A.	49.53	65.73	25.34	63.25	28.72	64.99	*2.99*	**11.86**
25	N.A.	N.A.	43.74	57.90	25.34	63.28	28.71	64.99	*2.96*	**11.60**
30	39.48	59.37	39.00	54.03	25.50	63.28	28.74	64.99	*2.92*	**11.78**
35	39.37	59.37	38.92	54.03	25.47	63.28	28.75	64.99	*2.93*	**11.68**
40	41.29	63.36	34.10	51.59	25.54	63.28	28.77	64.99	*2.94*	**11.58**
45	43.87	66.37	30.74	45.19	25.55	63.28	28.79	64.99	*2.96*	**11.51**
50	36.73	62.63	30.74	47.47	25.60	63.28	28.82	64.99	*2.98*	**11.84**







(c) The RD curve at 1920×1080.

Figure 4.5: Ground truth and approximate GRD functions with the same number N of basis vectors and samples.

Table 4.7: Performance of EGRD with different number N of basis and samples on the testing set. The second and fourth columns correspond to the average performance over all training GRD functions, while the third and last columns show performance on the worst fit curve.

N	Average RMSE	Worst RMSE	Average l^∞	Worst l^∞
1	2.16	9.03	19.41	55.86
2	1.36	3.62	14.70	40.02
3	1.11	3.36	9.25	31.31
4	1.14	3.46	8.22	30.98
5	0.97	3.39	6.99	27.99
6	0.65	2.38	6.12	26.81
7	0.67	2.52	4.80	15.88

in all circumstances, while the PGRD model performs slightly better than TGRD. This is consistent with the approximation capability of their employed basis as validated in Section 4.4.2. Third, the performances of the three basis-projection-based methods do not improve much with regards to the increase of sample numbers, which implies that the performance of this kind of GRD model is primarily determined by how well the employed basis vectors can represent the actual GRD function space. Fourth, the EGRD model can precisely recover the whole GRD surface with the minimal number of samples, based on which the reciprocal and the logarithmic regression model cannot even initialize the fitting process. In general, the default EGRD model is able to reconstruct the GRD function with much higher accuracy using only 7 quality probes than other models using 50 samples.

To see how the EGRD model performs with even less samples, we gradually reduce the number N of employed basis vectors and samples at the same time until $N = 1$. As we can see from Table 4.7, the EGRD model generally improves its accuracy with more basis vectors and samples used, and starts outperforming the reciprocal and the logarithmic models from $N = 2$. In fact, the EGRD model can reduce the average RMSE to less than 1 with only 5 samples. Fig. 4.5 illustrates a sample GRD function approximated by the EGRD model with $N = 1, 3, 5, 7$. For brevity, we only show the ground-truth

and approximate curves at three resolutions of 320×240, 1280×720, and 1920×1080, from which we can see that $N = 7$ provides the best estimation of the actual curves at all resolutions, even though other options of N already fit the reference very well except at the highest resolution. Besides, it is worth taking a closer look at the $N = 5$ curve at 1920×1080. The two abrupt turning points on the approximate curve are not only the evidence of occasional regression failures of the EGRD model, but also validating the effectiveness of the employed monotonicity constraints.

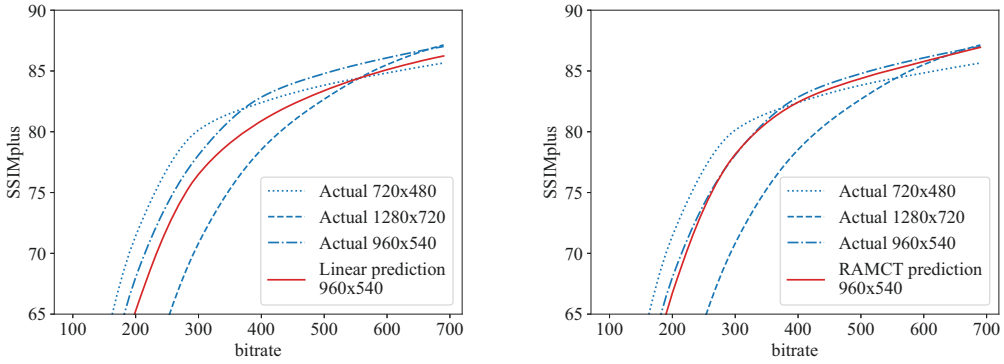
Chapter 5

Applications of GRD Functions

The application scope of generalized rate-distortion (GRD) model is much broader than video quality assessment (VQA). In this chapter, we demonstrate three practical use cases of GRD model. Although we adopt the robust axial-monotonic Clough-Tocher (RAMCT) model for demonstration, the eigen generalized rate-distortion (EGRD) model can be applied similarly with minimal modifications.

5.1 Rate-Distortion Curve at Novel Resolutions

Given a set of rate-distortion curves at multiple resolutions, it is desirable to predict the rate-distortion performance at novel resolutions, especially when there exists a mismatch between the supported viewing device of downstream content delivery network and the recommended encoding profiles. Traditional methods linearly interpolate the rate-distortion curve at novel resolutions [22], neglecting the characteristics of GRD functions. Fig. 5.1 compares the linearly interpolated and RAMCT-interpolated rate-distortion curves at 960×540 with the ground truth SSIMplus curve, from which we have several observations. First, the linearly interpolated curve shares the same intersection with the neighboring curves at 740×480 and 1280×720, inducing consistent bias to the prediction. The proposed RAMCT model is able to accurately predict the quality at the intersection of the



(a) (b)

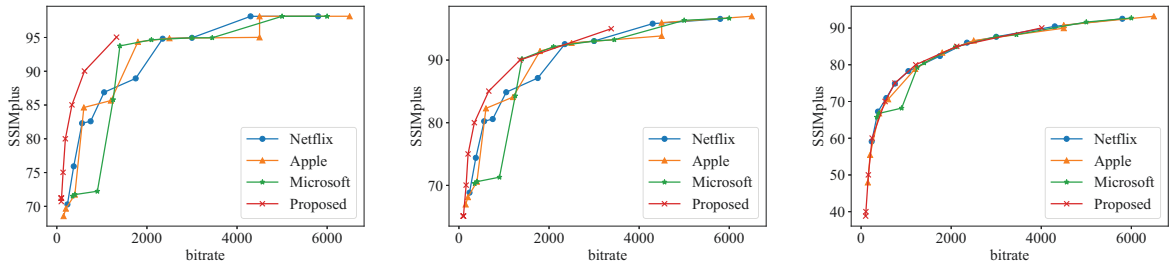
Figure 5.1: Prediction of RD curve of novel resolution from known RD curves of other resolutions. (a) Prediction of 960×540 RD curve from 720×480 and 1280×720 curves using linear interpolation; (b) Prediction of 960×540 RD curve using the proposed method.

neighboring curves by taking all known rate-distortion curves into consideration. Second, the linearly interpolated rate-distortion curve always lies between its neighboring curves, suggesting that the predicted quality at any bitrate is lower than the quality on one of its neighboring curves. This behavior contradicts the fact that each resolution may have a bitrate region in which it outperforms other resolutions [22]. By contrast, RAMCT better preserves the general trend of the resolution-quality curve at different bitrate, thanks to the regularization imposed by the C^1 condition at given nodes. Third, RAMCT outperforms the linear interpolation model in predicting the ground truth rate-distortion curve across all bitrates. The experimental results also justify the effectiveness of the C^1 and smoothness prior used in RAMCT.

To further validate the performance of the proposed GRD model at novel spatial resolutions, we predict the rate-distortion curves of 20 randomly selected source videos from the Waterloo generalized rate-distortion (Waterloo GRD) database at three novel resolutions (640×360, 960×540, and 1600×900). The evaluated bitrate ranges from 100 kbps to 9 Mbps with a step size of 100 kbps. The results are listed in Table 5.1. We can observe that RAMCT outperforms the linear model [22] with a clear margin at novel resolutions.

Table 5.1: Performance of linear interpolation and RAMCT on predicting RD curve at novel resolution.

Resolution	l_∞		MSE	
	Linear	RAMCT	Linear	RAMCT
640×360	7.68	3.12	7.89	1.56
960×540	7.66	4.83	6.61	3.14
1600×900	8.77	7.87	5.18	4.98
Average	8.04	5.27	6.56	3.23



(a) Title with low complexity (b) Title with moderate complexity (c) Title with high complexity

Figure 5.2: Bitrate ladders generated by the recommendations and the proposed algorithm for three contents.

5.2 Per-Title Encoding Profile Generation

To overcome the heterogeneity in users' network conditions and display devices, video service providers often encode videos at multiple bitrates and spatial resolutions. However, the selection of the encoding profiles are either hard-coded, resulting in sub-optimal Quality of Experience (QoE) due to the negligence of the difference in source video complexities, or selected based on interactive objective measurement and subjective judgment that are inconsistent and time-consuming. To deliver the best quality video to consumers, each title should receive a unique bitrate ladder, tailored to its specific complexity characteristics.

We introduce a quality-driven per-title optimization framework to automatically select the best encoding configurations, where the proposed GRD model serves as the key component.

Content delivery networks often aim to deliver videos at certain quality levels to satisfy different viewers. It is beneficial to minimize the bitrate usage in the encoding profile when achieving the objective. Mathematically, the quality-driven bitrate ladder selection problem can be formulated as a constrained optimization problem. Specifically, for the i-th representation,

$$\begin{aligned} \underset{\{x,y\}}{\arg\min} \quad & x \\ \text{subject to} \quad & f(x,y) \geq C_i,\ i = 1, \ldots, m, \end{aligned} \tag{5.1}$$

where x, y, $f(\cdot,\cdot)$, C_i and m represent the bitrate, the spatial resolution, the GRD function, the target quality level of video representation i, and the total number of video representations, respectively. Solving the optimization problem requires precise knowledge of the GRD function. Thanks to the effectiveness and differentiability of RAMCT, the proposed model can be incorporated with gradient-based optimization tools [33] to solve the per-title optimization problem. Specifically, the optimality condition of the constrained optimization problem in Eq. (5.1) is achieved when the following conditions hold. First, if the quality at the lowest bitrate is higher than the target quality C_i, the i-th optimal representation is (x_L, y^*), where x_L is the lowest bitrate and $y^* = \arg\max_y f(x_L, y)$. Second, when the target quality cannot be obtained by the lowest bitrate, the optimality condition is fulfilled when the inequality constraints are active (*i.e.*, the equation condition is achieved). Consequently, the optimization problem can be transformed into finding the bitrate x_i at each resolution such that $C_i = f(x_i, y)$, $\forall y$. We apply the Dichotomous-based search method [89] to compute the optimal bitrate at 320×240, 384×288, 512×384, 640×360, 720×480, 960×540, 1280×720, 1600×900, 1920×1080 [59]. The final bitrate is then obtained by taking the minimum value across all resolutions.

To validate the proposed per-title encoding profile selection algorithm, we apply the algorithm to generate bitrate ladders using H.264 [80] for 50 randomly selected videos in the Waterloo GRD database. We set the target quality levels $\{C_i\}_{i=1}^{10}$ as {30, 40, 50, 60, 70, 75, 80, 85, 90, 95} to cover diverse quality range and to match the total number of representations in standard recommendations [69]. For simplicity, we optimize

Table 5.2: Average bitrate saving of encoding profiles. Negative values indicate actual bitrate reduction.

	Microsoft [50]	Apple [5]	Netflix [1]	Proposed
Microsoft	0	-	-	-
Apple	-25.3%	0	-	-
Netflix	-29.3%	-5.6%	0	-
Proposed	-62.0%	-48.9%	-46.8%	0

the representation sets for only one viewing device (cellphone), while the procedure can be readily extended to multiple devices to generate a more comprehensive representation set. In Fig. 5.2, we compare the rate-quality curve of representation sets generated by the proposed algorithm, recommendations by Netflix [1], Apple [5], and Microsoft [50] for three videos with different complexities, from which the key observations are as follows. First, contrasting the hand-crafted bitrate ladders, the encoding profile generated by the proposed algorithm is content adaptive. Specifically, the encoding bitrate increases with respect to the complexity of the source video as illustrated in Fig. 5.2. Second, the proposed method achieves the highest quality at all bitrate levels. The performance improvement is mainly introduced by the encoding strategy at the convex hull encompassing the individual per-resolution rate-distortion curves [22]. Table 5.2 provides a full summary of the Bjøntegaard-Delta bitrate (BD-Rate) [8], indicating the required overhead in bitrate to achieve the same SSIMplus values. We observe that the proposed framework outperforms the existing hard-coded representation sets by at least 47%.

5.3 Codec Comparison

In the past decade, there has been a tremendous growth in video compression algorithms, thanks to the fast development of computational multimedia. With many video encoders at hand, it becomes pivotal to compare their performance, so as to find the best algorithm as well as directions for further advancement. Bjøntegaard-Delta model [8, 9] has become the most commonly used objective coding efficiency measurement. Bjøntegaard-Delta peak

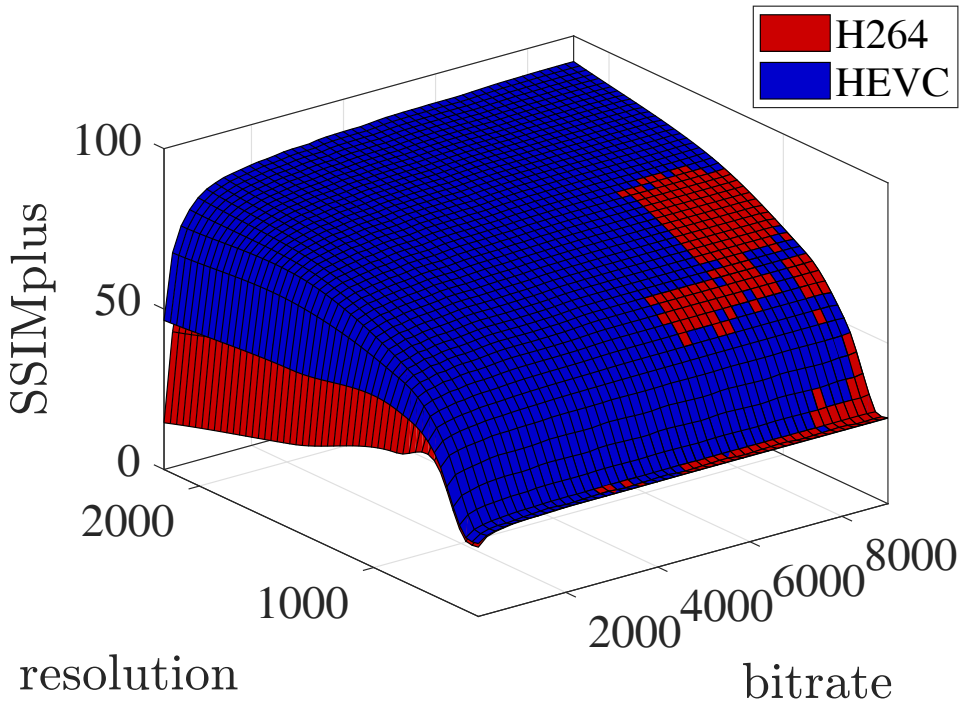


Figure 5.3: Generalized rate-distortion surfaces of H.264 and HEVC encoders for a sample source video.

signal-to-noise ratio (BD-PSNR) and Bjøntegaard-Delta bitrate (BD-Rate) are typically computed as the difference in bitrate and quality (measured in peak signal-to-noise Ratio (PSNR)) based on the interpolated rate-distortion curves

$$Q_{BD} = \frac{\int_{x_L}^{x_H} [z_B(x) - z_A(x)]dx}{\int_{x_L}^{x_H} dx}, \tag{5.2a}$$

$$R_{BD} \approx 10^{\frac{\int_{z_L}^{z_H} [x_B(z) - x_A(z)]dz}{\int_{z_L}^{z_H} dz}} - 1, \tag{5.2b}$$

where z_A and z_B are the interpolated rate-quality, x_A and x_B are the quality-rate functions of encoder A and B, respectively. $[x_L, x_H]$ and $[z_L, z_H]$ are the effective domain and range of the rate-distortion functions.

However, there are at least two major drawbacks of the widely used measures, which severely undermine their efficiency and reliability as codec comparison criteria. First, BD-PSNR and BD-Rate do not take spatial resolution and viewing condition into consideration. Fig. 5.3 shows two GRD surfaces generated by H.264 [80] and HEVC encoders for a source video. Although H.264 performs on par with HEVC at low resolutions, it requires higher bitrate to achieve the same target quality at high resolutions. Therefore, applying BD-PSNR and BD-Rate on a single resolution is not sufficient to fairly compare the overall performance between encoders. Second, the rate-distortion function $z(x)$ and the distortion-rate function $x(z)$ are independently interpolated and thus may not be the inverse to each other as the theory suggests. Such mismatch sometimes even leads to opposite conclusions when two codecs are compared with the two measures.

To resolve the two limitations, we propose the generalized quality gain (Q_{gain}) and rate gain (R_{gain}) models as

$$Q_{gain} = \frac{\int_U \int_{y_L}^{y_H} \int_{x_L}^{x_H} p(u)[z_B(x, y, u) - z_A(x, y, u)]dxdydu}{\int_{y_L}^{y_H} \int_{x_L}^{x_H} dxdy}, \tag{5.3a}$$

$$R_{gain} \approx 10^{\frac{\int_U \int_{y_L}^{y_H} \int_{z_L}^{z_H} p(u)[x_B(z,y,u) - x_A(z,y,u)]dzdydu}{\int_U \int_{y_L}^{y_H} \int_{z_L}^{z_H} p(u)dzdydu}} - 1, \tag{5.3b}$$

where $p(u)$, U, and $[y_L, y_H]$ represent the probability density of viewing devices, the set of all device of interests, and the domain of video spatial resolution, respectively. The generalized

Q_{gain} and R_{gain} models represent the expected quality gain and the expected bitrate gain (saving when R_{gain} negative) across all spatial resolutions and viewing devices, leading to a more comprehensive evaluation of codecs. Moreover, the effect of any individual influencing factor can be obtained by taking the marginal expectation in the corresponding dimension, which is more robust than BD-PSNR and BD-Rate at a single resolution. It should be noted that $z_A(x, y, u)$ is essentially the GRD function of codec A, which can be efficiently approximated by the RAMCT model. The "inverse" function $x_A(z, y, u)$ can also be estimated numerically from the interpolated surface thanks to its smoothness, axial monotonicity and differentiability. The implementation details of Q_{gain} and R_{gain} are discussed below.

Computing Q_{gain}: We apply the *trapz* function in Numpy [24] to compute the numerical integration in Eq. (5.3a) at each resolution from 500 kbps to 4000 kbps with 100 steps. The reason to select the bitrate range is that theoretically Bjøntegaard-Delta models should be computed based on the actual bitrate instead of the target bitrate. However, some encoders are not able to precisely control the encoding bitrate at very low or high bitrate ranges. We find the actual bitrate variability of encoders do not exceed 10% of target bitrate within the this range, suggesting the error introduced by bitrate control is insignificant [5]. The bitrate integration is computed at 320×240, 384×288, 512×384, 640×360, 720×480, 960×540, 1280×720, 1600×900, 1920×1080 [59], and integrated over spatial resolution to obtain the final score.

Computing R_{gain}: The computation procedure of generalized R_{gain} at one resolution is illustrated in Fig. 5.4. Specifically, the shaded area S_{target} represents the overall bitrate saving at one resolution, and can be computed as

$$S_{target} = AUC_{CDE} + S_{EHLK} - AUC_{FGH} - S_{CFJI},$$

where S and AUC stand for the area of rectangle and area under curve, respectively. The integration requires precise knowledge regarding to the bitrates x_{L_A}, x_{L_B}, x_{H_A}, x_{H_A}, and the effective quality interval $[z_L, z_H]$ at each resolution. At any resolution y, the integration interval is defined as

$$z_L = max(z_A(x_L, y), z_B(x_L, y)),$$
$$z_H = min(z_A(x_H, y), z_B(x_H, y)),$$

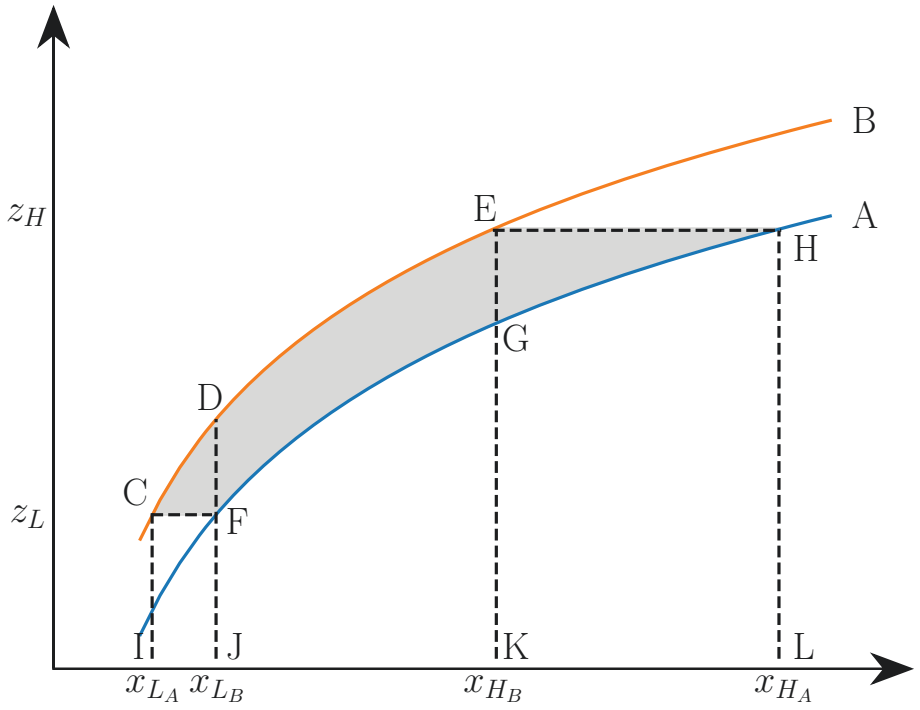


Figure 5.4: Computation procedure of generalized Bjøntegaard-Delta Rate on a 1D RD Curve. The shaded area represents the overall bitrate saving.

Table 5.3: Performance of VP9, and HEVC in terms of the generalized Q_{gain} and R_{gain} models with H.264 as the baseline codec

Codec	Q_{gain}	R_{gain}
VP9	1.9	-27.5%
HEVC	0.2	-15.9%

in the original BD-Rate model. The bitrates (x_{L_A}, x_{L_B}, x_{H_A}, x_{H_A}) where the two rate-distortion functions achieve z_L and z_H are obtained with the Dichotomous-based search method [89]. We apply the *trapz* function in Numpy [24] to compute the numerical area under curves with step number of 100. In the end, we integrate the overall bitrate saving over all spatial resolutions and devices to obtain the generalized R_{gain}.

Using the proposed measures on the Waterloo GRD Phase I database, we evaluate the performance of three video encoders, namely H.264, VP9, and HEVC. In order to simplify the expression, we set $p(u)$ to be uniform distribution for five display devices including cellphone, tablet, laptop, desktop, and TV here. Table 5.3 shows the performance of VP9 and HEVC in terms of the proposed measures with H.264 as the reference codec, from which we can observe that VP9 outperforms HEVC by an average of 1.7 and 11% in Q_{gain} and R_{gain}, respectively. The results of our objective codec evaluation are in general consistent with the recent video quality assessment results [71, 16].

Chapter 6

Conclusion

The rate-distortion theory has uncovered a secret trade between the bitrate resource and the resultant Quality of Experience (QoE), based on which many video-related businesses are made possible. In a free market, consumers pay the video service provider for their desired QoE, while the latter invests money on bitrate resources, which are in turn traded for perceptual quality of digital videos. Therefore, it is of vital importance and great interest for the video industry to understand their supply curve of QoE, which is comprehensively described by the proposed generalized rate-distortion (GRD) function.

In the thesis, we mainly focus on how to precisely estimate a GRD function with a minimal number of quality queries. This may be regarded as either an interpolation or a projection onto convex sets (POCS) problem. We actually fulfilled both ideas by proposing the robust axial-monotonic Clough-Tocher (RAMCT) and the eigen generalized rate-distortion (EGRD) models, respectively. To further reduce the number of required quality queries, we proposed the information-theoretic sampling (ITS) scheme, which preferentially select the most informative samples to minimize the uncertainty of GRD function. Extensive experiments on the Waterloo generalized rate-distortion (Waterloo GRD) database show that the proposed GRD estimation models outperform competing algorithms by a large margin. Despite the joint modeling of the multi-dimensional GRD function and the delicately designed sampling strategy, we think that the effectiveness of the proposed models stems from the appropriate use of domain knowledge of the GRD function. Such prior knowledge

imposes a strong regularity on the learned model, making it more predictable and interpretable, and less possible to be overfitting. The same methodology may be extended to a broad range of applications, such as machine learning [7] and data visualization [58].

We also demonstrate the power of an accurate GRD model by three practical video applications at the end of the thesis. Precise understanding of the GRD function provides the possibility of optimizing many video processing procedures content by content, resulting in efficient bitrate usage and thus cost savings. Moreover, the concept of GRD functions may be extended to many other signal types, such as digital images, remote sensing signals etc.

While not discussed deeply in the thesis, the GRD function also plays a key role in understanding video, the most common visual signal in our daily life. Just as the 1D rate-distortion (RD) curve, the GRD function provides more thorough description of video complexity. Our research reveals an interesting fact that a GRD function can be fully characterized by only a few parameters, though its shape varies a lot with content complexity, video encoder, and viewing device. This fact implies the possibility of describing the video complexity with even less than ten features. Once they really existed, the perceptual quality of an encoded video representation can be precisely predicted only by analyzing the source content, saving a lot of computational time wasted in repeatedly encoding the video.

Our results show that, the study of GRD functions in the thesis is a promising start for scientifically investigating many longstanding and emerging problems in various video applications. There is still a lot of room for further improvement and optimization for current video services after several decades of fast development. With the challenge of meeting the growing consumer demand using limited resources, I hope the study in the thesis can shed light on the long-neglected problem, and provide a new view point to the research community and the industry.

References

[1] A. Aaron, Z. Li, M. Manohara, J. De Cock, and D. Ronca. Per-Title encode optimization. https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2., 2015. Accessed on Jul. 20, 2018.

[2] P. Alfeld. A trivariate Clough-Tocher scheme for tetrahedral data. *Computer Aided Geometric Design*, 1(2):169–181, Jun. 1984.

[3] Alliance for Open Media. AV1 bitstream and decoding process specification. https://aomedia.org/av1-bitstream-and-decoding-process-specification/., 2018. Accessed on Apr. 1, 2018.

[4] I. Amidror. Scattered data interpolation methods for electronic imaging systems: A survey. *Journal of Electronic Imaging*, 11(2):157–177, Apr. 2002.

[5] Apple. Best practices for creating and deploying HTTP live streaming media for iPhone and iPad. http://is.gd/LBOdpz., 2016. Accessed on Jul. 20, 2018.

[6] T. Berger. Rate distortion theory and data compression. In *Advances in Source Coding*, pages 1–39. Springer, 1975.

[7] C. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006.

[8] G. Bjøntegaard. Calculation of average PSNR differences between rd-curves. Technical Report VCEG-M33, ITU-T SG 16/Q6, 13th VCEG Meeting, Video Coding Experts Group (VCEG), Austin, TX, USA, Apr. 2001.

[9] G. Bjøntegaard. Improvements of the BD-PSNR model. Technical Report VCEG-AI11, ITU-T SG 16/Q6, 35th VCEG Meeting, Video Coding Experts Group (VCEG), Berlin, Germany, Jul. 2008.

[10] W. Böhm, G. Farin, and J. Kahmann. A survey of curve and surface methods in CAGD. *Computer Aided Geometric Design*, 1(1):1–60, Jul. 1984.

[11] Y. S. Bonneh, A. Cooperman, and D. Sagi. Motion-induced blindness in normal observers. *Nature*, 411(6839):798, Jun. 2001.

[12] C. Chen, S. Inguva, A. Rankin, and A. Kokaram. A subjective study for the design of multi-resolution ABR video streams with the VP9 codec. In *Electronic Imaging*, pages 1–5, 2016.

[13] C. Chen, Y. Lin, A. Kokaram, and S. Benting. Encoding bitrate optimization using playback statistics for HTTP-based adaptive video streaming. *arXiv preprint arXiv:1709.08763*, Sep. 2017.

[14] Y. Chen, S. Hamed Hassani, A. Karbasi, and A. Krause. Sequential information maximization: When is greedy near-optimal? In *Proc. Conf. Learning Theory*, pages 338–363, 2015.

[15] M. Cheon and J. Lee. Subjective and objective quality assessment of compressed 4K UHD videos for immersive experience. *IEEE Trans. Circuits and Systems for Video Tech.*, 28(7):1467–1480, Jul. 2018.

[16] F. Christian. Multi-codec DASH dataset: An evaluatation of AV1, AVC, HEVC and VP9. https://bitmovin.com/av1-multi-codec-dash-dataset/, 2018. Accessed on Jul. 20, 2018.

[17] Cisco Mobile VNI. Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021 white paper. https://www.cisco.com/c/en/us/

solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, 2019. Accessed on Mar. 5, 2019.

[18] R. Clough and J. Tocher. Finite element stiffness matrices for analysis of plates in bending. In *Proceedings of Conf. on Matrix Methods in Structural Analysis*, 1965.

[19] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, Sep. 1995.

[20] P. Costantini and C. Manni. A local shape-preserving interpolation scheme for scattered data. *Computer Aided Geometric Design*, 16(5):385–405, Jun. 1999.

[21] G. Cote, B. Erol, M. Gallant, and F. Kossentini. H.263+: video coding at low bit rates. *IEEE Trans. Circuits and Systems for Video Tech.*, 8(7):849–866, Nov. 1998.

[22] J. De Cock, Z. Li, M. Manohara, and A. Aaron. Complexity-based consistent-quality encoding in the cloud. In *Proc. IEEE Int. Conf. Image Proc.*, pages 1484–1488, 2016.

[23] B. Delaunay. Sur la sphere vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk*, 7(793-800):1–2, Oct. 1934.

[24] Numpy developers. Numpy. http://www.numpy.org/, 2005. Accessed on Jul. 20, 2018.

[25] Z. Duanmu, K. Ma, and Z. Wang. Quality-of-Experience of adaptive video streaming: Exploring the space of adaptations. In *Proc. ACM Int. Conf. Multimedia*, MM '17, pages 1752–1760, 2017.

[26] G. Farin. Bézier polynomials over triangles and the construction of piecewise C^R polynomials. *Brunel University Mathematics Technical Papers collection*, Jan. 1980.

[27] G. Farin. A modified Clough-Tocher interpolant. *Computer Aided Geometric Design*, 2(1-3):19–27, Sep. 1985.

[28] FFmpeg team. FFmpeg v.2.8.15. https://trac.ffmpeg.org/wiki/Encode/H.264, 2018. Accessed on Jan. 20, 2018.

[29] FFmpeg team. FFmpeg v.2.8.15. https://trac.ffmpeg.org/wiki/Encode/H.265, 2018. Accessed on Jan. 20, 2018.

[30] FFmpeg team. FFmpeg v.2.8.15. https://trac.ffmpeg.org/wiki/Encode/VP9, 2018. Accessed on Jan. 20, 2018.

[31] N. Fritsch and R. Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, Apr. 1980.

[32] F. Gao, D. Tao, X. Gao, and X. Li. Learning to rank for blind image quality assessment. *IEEE Trans. Neural Networks and Learning Systems*, 26(10):2275–2290, Oct. 2015.

[33] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, 2014. Accessed on Jul. 20, 2018.

[34] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar. Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders. In *Picture Coding Symposium*, pages 394–397, 2013.

[35] L. Han and L. Schumaker. Fitting monotone surfaces to scattered data using c1 piecewise cubics. *SIAM Journal on Numerical Analysis*, 34(2):569–585, Mar. 1997.

[36] A. P. Hekstra, J. G. Beerends, D. Ledermann, F. E. de Caluwe, S. Kohler, R. H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss. PVQM − A perceptual video quality measure. *Signal Processing: Image Communication*, 17(10):781–798, Nov. 2002.

[37] X. Huang, J. Søgaard, and S. Forchhammer. No-reference pixel based video quality assessment for HEVC decoded video. *Journal of Visual Communication and Image Representation*, 43:173–184, Feb. 2017.

[38] Q. Huynh-Thu and M. Ghanbari. Temporal aspect of perceived quality in mobile video broadcasting. *IEEE Trans. Broadcasting*, 54(3):641–651, Sep. 2008.

[39] S. Ma N. Odintsova A. Beygelzimer G. Grabarnik I. Rish, M. Brodie and K. Hernandez. Adaptive diagnosis in distributed systems. *IEEE Trans. Neural Networks*, 16(5):1088–1109, Sep. 2005.

[40] ITU-R. Methodology for the subjective assessment of the quality of television pictures. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf, 2012. Accessed on Feb. 17, 2019.

[41] C. Kreuzberger, B. Rainer, H. Hellwagner, L. Toni, and P. Frossard. A comparative study of DASH representation sets using real user characteristics. In *Proc. Int. Workshop on Network and OS Support for Digital Audio and Video*, pages 1–4, 2016.

[42] X. Li, Q. Guo, and X. Lu. Spatiotemporal statistics for video quality assessment. *IEEE Trans. Image Processing*, 25(7):3329–3342, Jul. 2016.

[43] Y. Li, L. Po, C. Cheung, X. Xu, L. Feng, F. Yuan, and K.-W. Cheung. No-reference video quality assessment with 3d shearlet transform and convolutional neural networks. *IEEE Trans. Circuits and Systems for Video Tech.*, 26(6):1044–1057, Jun. 2016.

[44] Z. Li, A. Aaron, L. Katsavounidis, A. Moorthy, and M. Manohara. Toward a practical perceptual video quality metric. http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html., 2016. Accessed on Jul. 20, 2018.

[45] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo. MCL-V: A streaming video quality assessment database. *Journal of Visual Communication and Image Representation*, 30:1–9, Jul. 2015.

[46] W. Liu, Z. Duanmu, and Z. Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *Proc. ACM Int. Conf. Multimedia*, MM '18, pages 546–554, 2018.

[47] L. Lu, Z. Wang, A. C. Bovik, and J. Kouloheris. Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video, Aug. 2002.

[48] F. Lukas and Z. L. Budrikis. Picture quality prediction based on a visual model. *IEEE Trans. Communications*, 30(7):1679–1692, Jul. 1982.

[49] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Trans. Image Processing*, 26(8):3951–3964, Aug. 2017.

[50] G. Michael, T. Christian, H. Hermann, C. Wael, N. Daniel, and B. Stefano. Combined bitrate suggestions for multi-rate streaming of industry solutions. http://alicante.itec.aau.at/am1.html., 2013. Accessed on Jul. 20, 2018.

[51] G. M. Nielson. A method for interpolating scattered data based upon a minimum norm network. *Mathematics of Computation*, 40(161):253–271, Jan. 1983.

[52] Y. Ou, Y. Xue, and Y. Wang. Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions. *IEEE Trans. Image Processing*, 23(6):2473–2486, Jun. 2014.

[53] PKU-VCL. AVS2 codec source code repository. https://github.com/pkuvcl/xavs2, 2018. Accessed on Jan. 27, 2018.

[54] A. Rehman, K. Zeng, and Z. Wang. Display device-adapted video Quality-of-Experience assessment. In *Proc. SPIE*, pages 939406.1–939406.11, 2015.

[55] J. G. Robson. Spatial and temporal contrast-sensitivity functions of the visual system. *Journal of Optical Society of America*, 56(8):1141–1142, Aug. 1966.

[56] M. A Saad, A. C. Bovik, and C. Charrier. Blind prediction of natural video quality. *IEEE Trans. Image Processing*, 23(3):1352–1365, Mar. 2014.

[57] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE Trans. Image Processing*, 18(11):2385–2401, Nov. 2009.

[58] M. Sarfraz and M. Hussain. Data visualization using rational spline interpolation. *Journal of Computational and Applied Mathematics*, 189(1):513–525, May 2006.

[59] Screensiz. Viewport sizes and pixel densities for popular devices. http://screensiz.es/, 2018. Accessed on Apr. 20, 2019.

[60] K. Seshadrinathan and A. C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Trans. Image Processing*, 19(2):335–350, Feb. 2010.

[61] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack. Study of subjective and objective quality assessment of video. *IEEE Trans. Image Processing*, 19(6):1427–1441, Jun. 2010.

[62] 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*, 40(1):1–32, Aug. 2014.

[63] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 7:142–163, 1959.

[64] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Trans. Image Processing*, 15(2):430–444, Feb. 2006.

[65] J. Søgaard, S. Forchhammer, and J. Korhonen. No-reference video quality assessment using codec analysis. *IEEE Trans. Circuits and Systems for Video Tech.*, 25(10):1637–1650, Feb. 2015.

[66] R. Soundararajan and A. C. Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Trans. Circuits and Systems for Video Tech.*, 23(4):684–694, Apr. 2013.

[67] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: An operator splitting solver for quadratic programs. *ArXiv preprint arXiv:1711.08013*, Nov. 2017.

[68] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans. Circuits and Systems for Video Tech.*, 22(12):1649–1668, Dec. 2012.

[69] L. Toni, R. Aparicio-Pardo, K. Pires, G. Simon, A. Blanc, and P. Frossard. Optimal selection of adaptive streaming representations. *ACM Trans. Multimedia Computing, Communications, and Applications*, 11(2):1–43, Feb. 2015.

[70] P.N. Tudor. Mpeg-2 video compression. *Electronics & Communication Engineering Journal*, 7(6):257–264, Dec. 1995.

[71] D. Vatolin, D. Kulikov, E. Mikhail, D. Stanislav, and Z. Sergey. MSU codec comparison 2017 part V: High quality encoders. http://www.compression.ru/video/codec_comparison/hevc_2017/MSU_HEVC_comparison_2017_P5_HQ_encoders.pdf. Accessed on Jul. 20, 2018.

[72] P. V. Vu and D. M. Chandler. ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1):1–24, Feb. 2014.

[73] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, Mar. 2002.

[74] Z. Wang and A.C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, Jan. 2009.

[75] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, Apr. 2004.

[76] Z. Wang and Q. Li. Video quality assessment using a statistical model of human visual speed perception. *Journal of Optical Society of America A*, 24(12):61–69, Dec. 2007.

[77] Z. Wang, L. Lu, and A. C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, 19(2):121–132, Feb. 2004.

[78] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers*, volume 2, pages 1398–1402, Nov. 2003.

[79] A. B. Watson, J. Hu, and J. F. McGowan III. Digital video quality metric based on human vision. *Journal of Electronic Imaging*, 10(1):20–29, Jan. 2001.

[80] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the H. 264/AVC video coding standard. *IEEE Trans. Circuits and Systems for Video Tech.*, 13(7):560–576, Jul. 2003.

[81] S. Wolf and M. H. Pinson. Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system. In *Proc. SPIE*, volume 3845, pages 266–277, Sep. 1999.

[82] X. Xia, Z. Lu, L. Wang, M. Wan, and X. Wen. Blind video quality assessment using natural video spatio-temporal statistics. In *Proc. IEEE Int. Conf. Multimedia and Expo*, pages 1–6, 2014.

[83] J. Xu, P. Ye, Y. Liu, and D. Doermann. No-reference video quality assessment via feature learning. In *Proc. IEEE Int. Conf. Image Proc.*, pages 491–495, 2014.

[84] W. Xue, L. Zhang, X. Mou, and A. C. Bovik. Gradient Magnitude Similarity Deviation: A highly efficient perceptual image quality index. *IEEE Trans. Image Processing*, 23(2):684–695, Feb. 2014.

[85] P. Ye, J. Kumar, L. Kang, and D. Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, pages 1098–1105, 2012.

[86] K. Zeng, T. Zhao, A. Rehman, and Z. Wang. Characterizing perceptual artifacts in compressed video streams. In *Proc. SPIE 9014, Human Vision and Electronic Imaging XIX*, page 90140Q, 2014.

[87] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh. Cross-dimensional perceptual quality assessment for low bit-rate videos. *IEEE Trans. Multimedia*, 10(7):1316–1324, Nov. 2008.

[88] F. Zhang, S. Li, L. Ma, Y. C. Wong, and K. N. Ngan. IVP subjective quality video database. http://ivp.ee.cuhk.edu.hk/research/database/subjective/, 2011. Accessed on Apr. 03, 2019.

[89] W. Zhang, Y. Wen, Z. Chen, and A. Khisti. QoE-driven cache management for HTTP adaptive bit rate streaming over wireless networks. *IEEE Trans. Multimedia*, 15(6):1431–1445, Oct. 2013.

[90] A. X. Zheng, I. Rish, and A. Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. In *Proc. Conf. Uncertainty in Artificial Intelligence*, pages 675–682, 2005.

APPENDICES

Appendix A

Implementation and Mathematical Details about RAMCT

A.1 Details of Re-parametrization for Bézier Ordinates

We show details of the matrices in Eq. (3.17) as follows.

$$\mathbf{c} = \begin{bmatrix} c_{T_{01}} \\ c_{T_{02}} \\ c_{I_{01}} \\ c_{T_{12}} \\ c_{T_{10}} \\ c_{I_{11}} \\ c_{T_{20}} \\ c_{T_{21}} \\ c_{I_{21}} \\ c_{C_0} \\ c_{C_1} \\ c_{C_2} \\ c_{I_{02}} \\ c_{I_{12}} \\ c_{I_{22}} \\ c_S \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix}
\frac{x_{V_1}-x_{V_0}}{3} & \frac{y_{V_1}-y_{V_0}}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{x_{V_2}-x_{V_0}}{3} & \frac{y_{V_2}-y_{V_0}}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{x_{V_1}+x_{V_2}-2x_{V_0}}{9} & \frac{y_{V_1}+y_{V_2}-2y_{V_0}}{9} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{x_{V_2}-x_{V_1}}{3} & \frac{y_{V_2}-y_{V_1}}{3} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{x_{V_0}-x_{V_1}}{3} & \frac{y_{V_0}-y_{V_1}}{3} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{x_{V_0}+x_{V_2}-2x_{V_1}}{9} & \frac{y_{V_0}+y_{V_2}-2y_{V_1}}{9} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{x_{V_0}-x_{V_2}}{3} & \frac{y_{V_0}-y_{V_2}}{3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{x_{V_1}-x_{V_2}}{3} & \frac{y_{V_1}-y_{V_2}}{3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{x_{V_0}+x_{V_1}-2x_{V_2}}{9} & \frac{y_{V_0}+y_{V_1}-2y_{V_2}}{9} & 0 & 0 & 0 \\
0 & 0 & \frac{\theta_{21}(x_{V_2}-x_{V_1})}{3} & \frac{\theta_{21}(y_{V_2}-y_{V_1})}{3} & \frac{\theta_{12}(x_{V_1}-x_{V_2})}{3} & \frac{\theta_{12}(y_{V_1}-y_{V_2})}{3} & \eta_0 & 0 & 0 \\
\frac{\theta_{20}(x_{V_2}-x_{V_0})}{3} & \frac{\theta_{20}(y_{V_2}-y_{V_0})}{3} & 0 & 0 & \frac{\theta_{02}(x_{V_2}-x_{V_0})}{3} & \frac{\theta_{02}(y_{V_2}-y_{V_0})}{3} & 0 & \eta_1 & 0 \\
\frac{\theta_{01}(x_{V_1}-x_{V_0})}{3} & \frac{\theta_{01}(y_{V_1}-y_{V_0})}{3} & \frac{\theta_{10}(x_{V_0}-x_{V_1})}{3} & \frac{\theta_{10}(y_{V_0}-y_{V_1})}{3} & 0 & 0 & 0 & 0 & \eta_2 \\
\frac{(3\theta_{01}+1)x_{V_1}+(3\theta_{20}+1)x_{V_2}-(3\theta_{20}+3\theta_{01}+2)x_{V_0}}{27} & \frac{(3\theta_{01}+1)y_{V_1}+(3\theta_{20}+1)y_{V_2}-(3\theta_{20}+3\theta_{01}+2)y_{V_0}}{27} & \frac{(3\theta_{10}+1)x_{V_0}+(3\theta_{21}+1)x_{V_2}-(3\theta_{10}+3\theta_{21}+2)x_{V_1}}{27} & \frac{(3\theta_{10}+1)y_{V_0}+(3\theta_{21}+1)y_{V_2}-(3\theta_{10}+3\theta_{21}+2)y_{V_1}}{27} & \frac{\theta_{02}(x_{V_0}-x_{V_2})}{9} & \frac{\theta_{02}(y_{V_0}-y_{V_2})}{9} & 0 & \frac{\eta_0}{3} & \frac{\eta_2}{3} \\
\frac{\theta_{01}(x_{V_1}-x_{V_0})}{9} & \frac{\theta_{01}(y_{V_1}-y_{V_0})}{9} & \frac{\theta_{10}(x_{V_0}-x_{V_1})}{9} & \frac{\theta_{10}(y_{V_0}-y_{V_1})}{9} & \frac{\theta_{12}(x_{V_1}-x_{V_2})}{9} & \frac{\theta_{12}(y_{V_1}-y_{V_2})}{9} & \frac{\eta_0}{3} & 0 & \frac{\eta_2}{3} \\
\frac{\theta_{20}(x_{V_2}-x_{V_0})}{9} & \frac{\theta_{20}(y_{V_2}-y_{V_0})}{9} & \frac{\theta_{21}(x_{V_2}-x_{V_1})}{9} & \frac{\theta_{21}(y_{V_2}-y_{V_1})}{9} & \frac{(3\theta_{12}+1)x_{V_0}+(3\theta_{12}+1)x_{V_1}-(3\theta_{02}+3\theta_{12}+2)x_{V_2}}{27} & \frac{(3\theta_{12}+1)y_{V_0}+(3\theta_{12}+1)y_{V_1}-(3\theta_{02}+3\theta_{12}+2)y_{V_2}}{27} & \frac{\eta_1}{3} & \frac{\eta_1}{3} & 0 \\
\frac{(6\theta_{02}+1)x_{V_0}+(6\theta_{12}+1)x_{V_1}-(6\theta_{02}+6\theta_{12}+2)x_{V_2}}{27} & \frac{(6\theta_{02}+1)y_{V_0}+(6\theta_{12}+1)y_{V_1}-(6\theta_{02}+6\theta_{12}+2)y_{V_2}}{27} & \frac{(6\theta_{12}+1)x_{V_0}+(6\theta_{21}+1)x_{V_2}-(6\theta_{10}+6\theta_{21}+2)x_{V_1}}{27} & \frac{(6\theta_{10}+1)y_{V_0}+(6\theta_{21}+1)y_{V_2}-(6\theta_{10}+6\theta_{21}+2)y_{V_1}}{27} & \frac{(6\theta_{02}+1)x_{V_0}+(6\theta_{12}+1)x_{V_1}-(6\theta_{02}+6\theta_{12}+2)x_{V_2}}{27} & \frac{(6\theta_{02}+1)y_{V_0}+(6\theta_{12}+1)y_{V_1}-(6\theta_{02}+6\theta_{12}+2)y_{V_2}}{27} & \frac{2\eta_0}{3} & \frac{2\eta_1}{3} & \frac{2\eta_2}{3}
\end{bmatrix} \tag{A.1}$$

$$\mathbf{d} = \begin{bmatrix}
d^x_{V_0} \\ d^y_{V_0} \\ d^x_{V_1} \\ d^y_{V_1} \\ d^x_{V_2} \\ d^y_{V_2} \\ d^e_{E_0} \\ d^e_{E_1} \\ d^e_{E_2}
\end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} z_{V_0} \\ z_{V_0} \\ z_{V_0} \\ z_{V_1} \\ z_{V_1} \\ z_{V_1} \\ z_{V_2} \\ z_{V_2} \\ z_{V_2} \\ \theta_{21} z_{V_1} + \theta_{12} z_{V_2} \\ \theta_{02} z_{V_2} + \theta_{20} z_{V_0} \\ \theta_{10} z_{V_0} + \theta_{01} z_{V_1} \\ \frac{(\theta_{20}+\theta_{10}+1) z_{V_0} + \theta_{21} z_{V_1} + \theta_{02} z_{V_2}}{3} \\ \frac{\theta_{10} z_{V_0} + (\theta_{21}+\theta_{01}+1) z_{V_1} + \theta_{12} z_{V_2}}{3} \\ \frac{\theta_{20} z_{V_0} + \theta_{01} z_{V_1} + (\theta_{12}+\theta_{02}+1) z_{V_2}}{3} \\ \frac{(2\theta_{20}+2\theta_{10}+1) z_{V_0} + (2\theta_{21}+2\theta_{01}+1) z_{V_1} + (2\theta_{12}+2\theta_{02}+1) z_{V_2}}{3} \end{bmatrix}$$

A.2 Details of Inequality Constraint

Summarizing Eq. (3.20) as matrix form, we obtain

$$\mathbf{G}\mathbf{c} \leq \mathbf{h},$$

where

$$\mathbf{G} = \begin{bmatrix} y_{V_0}-y_{V_2} & y_{V_1}-y_{V_0} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & y_{V_1}-y_{V_0} & y_{V_2}-y_{V_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & y_{V_2}-y_{V_1} & y_{V_0}-y_{V_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{V_2}-y_{V_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{V_1}-y_{V_0} & y_{V_0}-y_{V_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & y_{V_0}-y_{V_2} & 0 & 0 & 0 & y_{V_1}-y_{V_0} & 0 & y_{V_2}-y_{V_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{V_1}-y_{V_0} & y_{V_0}-y_{V_2} & y_{V_2}-y_{V_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{V_2}-y_{V_1} & y_{V_0}-y_{V_2} & y_{V_1}-y_{V_0} & 0 \\ y_S-y_{V_1} & 0 & 0 & 0 & y_{V_0}-y_S & 0 & 0 & 0 & 0 & 0 & 0 & y_{V_1}-y_{V_0} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & y_S-y_{V_2} & 0 & 0 & 0 & y_{V_1}-y_S & 0 & y_{V_2}-y_{V_1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & y_{V_2}-y_S & 0 & 0 & 0 & 0 & y_S-y_{V_0} & 0 & 0 & 0 & y_{V_0}-y_{V_2} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{A.2}$$

and

$$\mathbf{h} = \begin{bmatrix} (y_{V_1} - y_{V_2})z_{V_0} \\ (y_{V_2} - y_{V_0})z_{V_1} \\ (y_{V_0} - y_{V_1})z_{V_2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

A.3 Details of Loss Function

Expanding Eq. (3.29), we obtain

$$\mathbf{U}_{V_0V_1V_2} = \begin{bmatrix}
\frac{18}{||E_2||^3} & 0 & 0 & 0 & \frac{-9}{||E_2||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{18}{||E_1||^3} & 0 & 0 & 0 & 0 & \frac{-9}{||E_1||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{36}{||E_0||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-18}{||E_0||^3} & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{18}{||E_0||^3} & 0 & 0 & 0 & \frac{-9}{||E_0||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{-9}{||E_2||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{36}{||\hat{E}_1||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-18}{||\hat{E}_1||^3} & 0 & 0 \\
0 & \frac{-9}{||E_1||^3} & 0 & 0 & 0 & 0 & \frac{18}{||E_1||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{-9}{||E_0||^3} & 0 & 0 & 0 & \frac{18}{||E_0||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{36}{||\hat{E}_2||^3} & 0 & 0 & 0 & 0 & 0 & \frac{-18}{||\hat{E}_2||^3} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{-18}{||\hat{E}_0||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{36}{||\hat{E}_0||^3} & 0 & 0 & \frac{-18}{||\hat{E}_0||^3} \\
0 & 0 & 0 & 0 & 0 & \frac{-18}{||\hat{E}_1||^3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{36}{||\hat{E}_1||^3} & 0 & \frac{-18}{||\hat{E}_1||^3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-18}{||\hat{E}_2||^3} & 0 & 0 & 0 & 0 & 0 & \frac{36}{||\hat{E}_2||^3} & \frac{-18}{||\hat{E}_2||^3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-18}{||\hat{E}_0||^3} & \frac{-18}{||\hat{E}_1||^3} & \frac{-18}{||\hat{E}_2||^3} & \frac{12}{||\hat{E}_0||^3} + \frac{12}{||\hat{E}_1||^3} + \frac{12}{||\hat{E}_2||^3}
\end{bmatrix},$$

and

$$\mathbf{w}_{V_0V_1V_2} = \begin{bmatrix} \frac{-18z_{V_0}}{||E_2||^3} \\ \frac{-18z_{V_0}}{||E_1||^3} \\ \frac{-36z_{V_0}}{||\hat{E}_0||^3} \\ \frac{-18z_{V_1}}{||E_0||^3} \\ \frac{-18z_{V_1}}{||E_2||^3} \\ \frac{-36z_{V_1}}{||\hat{E}_1||^3} \\ \frac{-18z_{V_2}}{||E_1||^3} \\ \frac{-18z_{V_2}}{||E_0||^3} \\ \frac{-36z_{V_2}}{||\hat{E}_2||^3} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{12z_{V_0}}{||\hat{E}_0||^3} + \frac{12z_{V_1}}{||\hat{E}_1||^3} + \frac{12z_{V_2}}{||\hat{E}_2||^3} \end{bmatrix}.$$

A.4 Proof of Necessary Conditions of Axial Monotonicity

Since the interpolated surface is a continuous piece-wise cubic function, it being monotonic everywhere is equivalent to every cubic function in its own triangle being monotonic. Denote (i, j, k) a cyclic permutation of $(0, 1, 2)$. Consider a microtriangle $\Delta_{V_iV_jS}$. The Bézier surface $z(\alpha, \beta, \gamma)$ over the microtriangle can be formulated as Eq. (3.10), where the barycentric coordinates (α, β, γ) are definded as in Eq. (3.11). The chain rule of derivatives shows that

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial \alpha}\frac{\partial \alpha}{\partial x} + \frac{\partial f}{\partial \beta}\frac{\partial \beta}{\partial x} + \frac{\partial f}{\partial \gamma}\frac{\partial \gamma}{\partial x}. \tag{A.1}$$

From Eq. (3.10), we have

$$\frac{\partial f}{\partial \alpha} = 3c_{V_i}\alpha^2 + 6c_{T_{ij}}\alpha\beta + 6c_{I_{i1}}\alpha\gamma + 3c_{T_{ji}}\beta^2 + 3c_{I_{i2}}\gamma^2 + 6c_{C_k}\beta\gamma \tag{A.2a}$$

$$\frac{\partial f}{\partial \beta} = 3c_{T_{ij}}\alpha^2 + 6c_{T_{ji}}\alpha\beta + 6c_{C_k}\alpha\gamma + 3c_{V_j}\beta^2 + 3c_{I_{j2}}\gamma^2 + 6c_{I_{j1}}\beta\gamma \tag{A.2b}$$

$$\frac{\partial f}{\partial \gamma} = 3c_{I_{i1}}\alpha^2 + 6c_{C_k}\alpha\beta + 6c_{I_{i2}}\alpha\gamma + 3c_{I_{j1}}\beta^2 + 3c_S\gamma^2 + 6c_{I_{j2}}\beta\gamma, \tag{A.2c}$$

and from Eq. (3.11),

$$\frac{\partial \alpha}{\partial x} = \frac{y_{V_j} - y_S}{2A}, \frac{\partial \beta}{\partial x} = \frac{y_S - y_{V_i}}{2A}, \frac{\partial \gamma}{\partial x} = \frac{y_{V_i} - y_{V_j}}{2A} \tag{A.3}$$

where A denotes the area of $\Delta_{V_iV_jS}$. Substitute (A.2) and (A.3) into (A.1). After some rearrangements, we reach

$$\begin{aligned}
\frac{\partial f}{\partial x} =& \frac{3}{2A}[c_{V_0}(y_{V_j} - y_S) + c_{T_{ij}}(y_S - y_{V_i}) + c_{I_{i1}}(y_{V_i} - y_{V_j})]\alpha^2 \\
&+ \frac{3}{2A}[c_{T_{ji}}(y_{V_j} - y_S) + c_{V_j}(y_S - y_{V_i}) + c_{I_{j1}}(y_{V_i} - y_{V_j})]\beta^2 \\
&+ \frac{3}{2A}[c_{I_{i2}}(y_{V_j} - y_S) + c_{I_{j2}}(y_S - y_{V_i}) + c_S(y_{V_i} - y_{V_j})]\gamma^2 \\
&+ \frac{6}{2A}[c_{T_{ij}}(y_{V_j} - y_S) + c_{T_{ji}}(y_S - y_{V_i}) + c_{C_k}(y_{V_i} - y_{V_j})]\alpha\beta \\
&+ \frac{6}{2A}[c_{I_{i1}}(y_{V_j} - y_S) + c_{C_k}(y_S - y_{V_i}) + c_{I_{i2}}(y_{V_i} - y_{V_j})]\alpha\gamma \\
&+ \frac{6}{2A}[c_{C_k}(y_{V_j} - y_S) + c_{I_{j1}}(y_S - y_{V_i}) + c_{I_{j2}}(y_{V_i} - y_{V_j})]\beta\gamma.
\end{aligned} \tag{A.4}$$

f is x-axial monotonic if and only if (A.4) is nonnegative within ΔV_iV_jS. Note that α, β, γ are all nonnegative inside or on the edge of ΔV_iV_jS. One sufficient condition for (A.4) to be nonnegative is its all coefficients are nonnegative [35], as indicated in Eq. (3.19). To show that Eq. (3.19a) and (3.19c) are necessary conditions for $\frac{\partial z}{\partial x}$ to be nonnegative, we simply check the value of $\frac{\partial z}{\partial x}$ at three vertices of $\Delta_{V_iV_jS}$, where (α, β, γ) equals $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, respectively. Substituting the three barycentric coordinates back into

(A.4), we obtain three inequalities as the necessary condition in $\Delta V_i V_j S$

$$c_{V_i}(y_{V_j} - y_S) + c_{T_{ij}}(y_S - y_{V_i}) + c_{I_{i1}}(y_{V_i} - y_{V_j}) \geq 0 \tag{A.5a}$$

$$c_{T_{ji}}(y_{V_j} - y_S) + c_{V_j}(y_S - y_{V_i}) + c_{I_{j1}}(y_{V_i} - y_{V_j}) \geq 0 \tag{A.5b}$$

$$c_{I_{i2}}(y_{V_j} - y_S) + c_{I_{j2}}(y_S - y_{V_i}) + c_S(y_{V_i} - y_{V_j}) \geq 0. \tag{A.5c}$$

The inner-triangle C^1 continuity implies that points $V_i, T_{ij}, I_{i1}, T_{ik}$ are coplanar, *i.e.* $c_{I_{i1}} = (c_{V_i} + c_{T_{ij}} + c_{T_i k})/3$. Substitute this equation and $y_S = (y_{V_i} + y_{V_j} + y_{V_k})/3$ into (A.5a). Some rearrangements will yield

$$c_{V_i}(y_{V_k} - y_{V_j}) + c_{T_{ij}}(y_{V_i} - y_{V_k}) + c_{T_{ik}}(y_{V_j} - y_{V_i}) \leq 0, \tag{A.6}$$

which is exactly Eq. (3.19a). Further, the summation of (A.5c) over all possible (i, j, k) gives Eq. (3.27c). Now we have proved Eq. (3.19a) and Eq. (3.19c) in Section 3.2.6 are necessary for the interpolant to be axial monotonic in $\Delta V_0 V_1 V_2$.